Maejo Int. J. Sci. Technol. **2026**, *20*(01), 108-123

Maejo International Journal of Science and Technology

e-ISSN 2697-4746
Available online at www.mijst.mju.ac.th

Full Paper

Minimum covariance determinant-based estimators for estimating population mean using two auxiliary variables

Tolga Zaman [1,*], Ebrucan Islamoglu[2] and Hasan Bulut [3]

[1]Department of Mathematical Engineering, Gumushane University, Gumushane 29100, Turkey
[2]Department of Finance and Banking, Nevsehir Hacı Bektas Veli University, Nevsehir 50300, Turkey
[3]Department of Statistics, Ondokuz Mayıs University, Samsun 55139, Turkey
*Corresponding author, e-mail: tolga.zaman@gumushane.edu.tr

Received: 30 May 2025 / Accepted: 14 April 2026 / Published: 27 April 2026

Abstract: This study addresses the challenge of outliers in sample surveys, a common issue that can distort results and lead to inaccurate conclusions. To tackle this problem, researchers have developed various robust regression techniques including least trimmed squares, least median of squares, least absolute deviations, Huber's, Hample's and Tukey's maximum likelihood estimators, and Huber's maximum likelihood estimator for location and scale estimators. We propose a new approach: minimum covariance determinant-based ratio estimators for estimating the population mean in sample surveys. To assess the performance of the proposed minimum covariance determinant-based estimators, we derive their mean square error expressions and determine the conditions under which they surpass existing estimators in efficiency. To validate our theoretical findings, we conduct a numerical illustration using real-world data and perform simulations with R-Programme software. The results demonstrate that our method effectively handles outliers and improves the overall accuracy of survey-based estimations, making it a valuable tool for researchers working with sample survey data.

Keywords: ratio estimators, robust regression, minimum covariance determinant estimate, mean sqaure error

INTRODUCTION

Many researchers assume that the error term in regression analysis follows a Gaussian (normal) distribution and use the least squares estimation (LSE) method to estimate regression parameters. However, in practice even if the normality assumption is accepted, residuals often do

not follow a normal distribution. In particular, outlier observations or observations suspected of being outliers tend to violate the normality assumption, leading to biased (incorrect) parameter estimates when using the LSE method. To address such issues, researchers increasingly rely on robust methods, which have been widely used in recent years. Maximum likelihood estimators are called M-estimators in the literature and these M-estimators are commonly used robust methods. The M-estimator is a generalised version of the maximum likelihood estimation, and the LSE method itself is also known as an M-estimator. The M-estimator method iteratively obtains parameter estimates by minimising an appropriate objective function for the given data set. Specifically, when there are outlier observations in the dependent variable Y, the Huber-M and Tukey-M estimation methods often provide better results than the LSE method [1]. Outliers and data contamination can substantially affect the accuracy of classical estimation methods, making robust statistical techniques essential in practical applications. Robust regression techniques have been extensively studied in the literature to address the influence of outliers. Notably, Mitra et al. [2] provided a comprehensive treatment of key methods and their theoretical foundations. Over the years, various robust regression methods including M-estimators, scale (S)-estimator and minimum covariance determinant (MCD)-based approaches have been developed to improve efficiency and reliability under contamination. These studies highlight the importance of robust methods in mitigating the impact of atypical observations while maintaining desirable statistical properties. Building on this foundation, the present study proposes multivariate MCD-based estimators using two auxiliary variables under simple random sampling, providing a novel extension of existing robust estimation frameworks. In multiple linear regression analysis based on the LSE method, a high correlation between independent or predictor variables can lead to the issue of multi-collinearity [3]. It has been reported that this multicolinearity problem reduces the reliability of estimates by increasing the standard errors of the regression coefficients [3, 4]. As a result, although LSE estimates remain unbiased in a model affected by multicolinearity, interpreting the effects of predictor variables becomes challenging.

Robust regression techniques are designed to minimise the influence of outliers when estimating model parameters. By reducing the effect of extreme data points, these methods help the model better identify underlying patterns, leading to less bias and more trustworthy parameter estimates. This makes them especially useful in research where data might be affected by measurement errors or inconsistencies.

In sampling theory outliers negatively affect the estimators. Recently robust regression methods and robust covariance estimates have been frequently used to reduce the negative effect caused by outliers. For example, Kadılar and Çıngı [5] proposed ratio estimators using Huber-M estimates. Zaman and Bulut [6] proposed ratio estimators using various robust regression methods. Shahzad et al. [7] provided a class of ratio estimators in the case of missing data using robust regression and covariance estimates. Bulut and Zaman [8] extended the estimators provided by Zaman and Bulut [6] by utilising MCD robust covariance estimation. Yadav and Prasad [9] presented exponential estimators using robust regression methods in the presence of outliers.

Zaman et al. [10] proposed the estimators by considering least trimmed squares (LTS), scale (S), least median of squares (LMS), and Huber-M estimators in two auxiliary variables. Shahzad et al. [11] provided an MCD estimate with quantile regression to develop quantile-regression-type mean estimators and demonstrated that these estimators outperform existing alternatives in both simulation studies and real-data applications. Durrani et al. [12] introduced robust exponential ratio-type variance estimators using MCD and minimum volume ellipsoid techniques, consistently

showing lower mean squared errors across theoretical, simulation-based and empirical evaluations. Kalina [13] presented a minimum-weighted covariance determinant estimator and proposed stronger multivariate adaptations. Subzar et al. [14] proposed robust ratio estimators based on Huber's M-estimation for simple random sampling. Rather et al. [15] introduced a ratio-type estimator that employs the undesceding kernel redescending M-estimator, which demonstrates strong performance even in the presence of extreme contamination in the data. Ahmed et al. [16] proposed robust ratio estimators designed to improve efficiency under skewed or contaminated distributions. Shahzad et al. [17] presented robust ratio estimators for mean estimation under missing data conditions. Subzar et al. [18] suggested that robust regression approaches such as least absolute deviations (LAD) and M-estimators can substantially improve the efficiency of ratio estimators in contaminated data sets. Zaman [19] presented ratio estimators based on robust regression, outperforming earlier robust designs. Raza et al. [20] provided regression-based ratio estimators using redescending M-estimators, demonstrating that these robust approaches yield more reliable population parameter estimates in the presence of outliers.

Zaman and Bulut [21] proposed a family of ratio estimators to estimate population means using robust regression methods such as Huber's maximum likelihood estimator for location, LTS and LMS in double sampling. Zaman and Bulut [22] presented efficient robust-type estimators for population variance under simple and stratified random sampling. Grover and Kaur [23] introduced an estimator, which is enhanced in this paper by applying Searls' technique to develop a more robust population mean estimator in simple random sampling without replacement. Gulzar et al. [24] proposed two new classes of exponential-cum-ratio estimators for estimating finite population mean using dual auxiliary variables, demonstrating their superior efficiency over existing estimators through mean squared error (MSE) comparisons, simulation, robustness studies and empirical analysis with real data sets. Anas et al. [25] presented a modified class of ratio-type regression estimators for estimating the population mean under simple random sampling in the presence of missing data, incorporating quantile regression to address extreme observations and improve the efficiency of estimators.

Zaman et al. [26] proposed ratio estimators which incorporate robust techniques to enhance the use of auxiliary variables for estimating the population mean in simple random sampling. Audu et al. [27] provided a robust modified class of estimators using non-conventional dispersion measures, which are less sensitive to outliers. Pandey et al. [28] suggested robust estimators for population mean estimation in ranked set sampling scenarios. Yadav and Prasad [29] presented a novel exponential estimator for estimating the finite population mean, utilising robust regression techniques, specifically Huber M-estimation, to effectively handle outliers, with theoretical MSE comparisons showing superior performance over ordinary least squares in the presence of outliers. Abid et al. [30] introduced a robust dual auxiliary variables-based exponential-cum-ratio class of estimators to enhance the efficiency of ratio-type estimators for estimating the population mean, particularly in the presence of extreme observations, by integrating both conventional and non-conventional measures under simple random sampling without replacement.

Zaman and Bulut [31] proposed a class of robust estimators based on the MCD and the minimum volume ellipsoid robust covariance estimates for estimating population variance in the presence of outliers in simple random sampling. Hashem et al. [32] provided a family of robust estimators based on the orthogonalised Gnanadesikan-Kettenring covariance matrix, offering a computationally feasible and reliable alternative to the MCD estimator in survey sampling, especially in the presence of outliers. The efficiency of finite population mean estimation has been

substantially improved through modified ratio-type estimators that use auxiliary information [33, 34]. To address the adverse effects of outliers and non-normality, robust covariance-based approaches, particularly those relying on the MCD, have gained prominence in survey sampling [35]. Recent contributions have extended MCD-based calibration and mean estimation procedures to systematic sampling, showing improvements in bias and efficiency compared to classical estimators [36]. While the MCD methodology has previously been applied to robust mean estimation under systematic sampling with a single auxiliary variable [36], the present study extends this approach in several important ways. Specifically, we develop a new class of MCD-based estimators under simple random sampling that simultaneously incorporates two auxiliary variables. This extension requires multivariate covariance estimation, introduces additional cross-covariance terms, and results in distinct bias and MSE structures. The theoretical derivations, optimality conditions and efficiency analyses for the dual-auxiliary case are therefore fundamentally different. By addressing the multivariate auxiliary setting under simple random sampling, our methodology provides new insights into robustness and efficiency gains, highlighting the methodological contribution and practical relevance of using multiple auxiliary variables in population mean estimation. Another line of research has applied MCD-based methods to median ranked set sampling, demonstrating notable efficiency gains over traditional approaches [37].

Moreover, the use of dual auxiliary information has been shown to significantly enhance estimator performance in both the ranked set sampling and simple random sampling contexts [38, 39]. Parallel to these developments, neutrosophic frameworks have recently been introduced to accommodate indeterminacy and uncertainty in auxiliary information, offering a flexible extension to classical estimation theory [40].

Collectively, these studies highlight the continued relevance of robust, auxiliary-based and hybrid estimation strategies for efficient population mean estimation. The MCD-based ratio estimators are less sensitive to extreme values and offer superior efficiency in scenarios where data contain a high proportion of outliers. We highlight the primary advantages of MCD-based estimators, which are their higher degree of robustness and their ability to efficiently handle the data sets with a high proportion of outliers. Unlike methods such as LTS and Huber-M, which may still be influenced by extreme values in certain scenarios, MCD estimators are less sensitive to outliers, making them more reliable in cases where data are contaminated by extreme observations.

REVIEW OF EXISTING ESTIMATORS

Zaman et al. [10] proposed the following estimators by considering LTS, S, LMS and Huber-M estimations in two auxiliary variables:

$$\bar{y}_{z1} = \bar{y}\left(\frac{\bar{X}_1}{\bar{x}_1}\right)^{\alpha_1}\left(\frac{\bar{X}_2}{\bar{x}_2}\right)^{\alpha_2} + b_{1(LTS)}(\bar{X}_1 - \bar{x}_1) + b_{2(LTS)}(\bar{X}_2 - \bar{x}_2), \quad (1)$$

$$\bar{y}_{z2} = \bar{y}\left(\frac{\bar{X}_1}{\bar{x}_1}\right)^{\alpha_1}\left(\frac{\bar{X}_2}{\bar{x}_2}\right)^{\alpha_2} + b_{1(S)}(\bar{X}_1 - \bar{x}_1) + b_{2(S)}(\bar{X}_2 - \bar{x}_2), \quad (2)$$

$$\bar{y}_{z3} = \bar{y}\left(\frac{\bar{X}_1}{\bar{x}_1}\right)^{\alpha_1}\left(\frac{\bar{X}_2}{\bar{x}_2}\right)^{\alpha_2} + b_{1(LMS)}(\bar{X}_1 - \bar{x}_1) + b_{2(LMS)}(\bar{X}_2 - \bar{x}_2), \quad (3)$$

$$\bar{y}_{z4} = \bar{y}\left(\frac{\bar{X}_1}{\bar{x}_1}\right)^{\alpha_1}\left(\frac{\bar{X}_2}{\bar{x}_2}\right)^{\alpha_2} + b_{1(HubM)}(\bar{X}_1 - \bar{x}_1) + b_{2(HubM)}(\bar{X}_2 - \bar{x}_2). \quad (4)$$

where $b_{1(LTS)}$, $b_{1(S)}$, $b_{1(LMS)}$, $b_{1(HubM)}$ are coefficients of slope obtained from LTS, S, LMS and Huber-M methods respectively.

The MSE expressions of the estimators proposed by Zaman et al. [10] are given as

$$MSE(\bar{y}_{zi}) \cong \frac{1-f}{n}\Big[S_y^2 + \left(\alpha_1^* R_1 + B_{1rob(k)}\right)^2 S_{x_1}^2 + \left(\alpha_2^* R_2 + B_{2rob(k)}\right)^2 S_{x_2}^2 - 2\left(\alpha_1^* R_1 + B_{1rob(k)}\right)S_{yx_1} - 2\left(\alpha_2^* R_2 + B_{2rob(k)}\right)S_{yx_2} + 2\left(\alpha_1^* R_1 + B_{1rob(k)}\right)\left(\alpha_2^* R_2 + B_{2rob(k)}\right)S_{x_1x_2}\Big], \quad (5)$$

where $f = (n/N)$ denotes the sampling fraction. Here $B_{1rob(k)}$ and $B_{2rob(k)}$ are obtained from robust regression methods, namely LTS, S, LMS and Huber-M estimators (with $k = LTS, S, LMS, Huber-M$). The quantities $\mathrm{R_1} = (\overline{\mathrm{Y}}/\overline{\mathrm{X}}_1)$ and $\mathrm{R_2} = (\overline{\mathrm{Y}}/\overline{\mathrm{X}}_2)$ are ratio parameters while the optimal coefficients α_1^* and α_2^* are defined as

$$\alpha_1^* = \frac{\left(\mathrm{S_y}\rho_{\mathrm{x_1x_2}}\right)\left(\rho_{\mathrm{yx_1}}\rho_{\mathrm{x_1x_2}} - \rho_{\mathrm{yx_2}}\right)}{\left(\mathrm{R_1 S_{x_1}}\left(1 - \rho^2_{\mathrm{x_1x_2}}\right)\right)}, \alpha_2^* = \frac{\left(\mathrm{S_y}\rho_{\mathrm{x_1x_2}}\right)\left(\rho_{\mathrm{yx_2}}\rho_{\mathrm{x_1x_2}} - \rho_{\mathrm{yx_1}}\right)}{\left(\mathrm{R_2 S_{x_2}}\left(1 - \rho^2_{\mathrm{x_1x_2}}\right)\right)}.$$

In addition, $\mathrm{B_1} = \left(\mathrm{S_{yx_1}}/\mathrm{S^2_{x_1}}\right)$ and $\mathrm{B_2} = \left(\mathrm{S_{yx_2}}/\mathrm{S^2_{x_2}}\right)$ are the classical least square estimators. The quantities $\mathrm{S^2_{x_1}}$ and $S_{x_2}^2$ denote the population variances of $\mathrm{x_{1i}}$ and $\mathrm{x_{2i}}$ respectively, while S_{yx_1} and S_{yx_2} represent the population covariances between y_i, x_{1i} and between y_i, x_{2i} respectively. $\mathrm{S^2_{x_1}}$ and $S_{x_2}^2$ are the variances of population of $\mathrm{x_{1i}}$ and $\mathrm{x_{2i}}$ respectively, and S_{yx_1} and S_{yx_2} are the covariance of population between y_i, x_{1i} and between y_i, x_{2i} respectively. The robust regression methods used by Zaman et al. [10] can be explained as follows. The LTS estimation method works by first sorting the squared error terms in ascending order. Instead of considering all error terms, LTS focuses only on the smallest φ values, summing them up to create a more robust estimation process. The objective is to minimise this sum, ensuring that the estimation is less influenced by extreme outliers. Mathematically, the function to be minimised is expressed as

$$\min\sum_{i=1}^{\varphi}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i)^2 \quad (6)$$

where $\varphi = \frac{\tau}{2} + 1$ and τ is the number of observations.

Rousseeuw and Yohai [41] introduced S estimation, a robust regression method that builds on M-estimation by incorporating M-scales to improve resistance to outliers. Unlike traditional methods, which may be overly sensitive to extreme values, S estimation relies on the residual scale of M estimation to achieve better robustness. The key idea behind S estimation is to use the residual standard deviation instead of the median to address some of the median's limitations in robust regression. This approach ensures more stable estimates while maintaining robustness against outliers. The function to be minimised in S estimation is expressed as

$$\min\sum_{i=1}^{n}\rho\left(\frac{y_i - \sum_{i=1}^{n} x_{ij}\beta}{\widehat{\sigma_S}}\right) \quad (7)$$

where $\sqrt{(1/nL)\sum_{i=1}^{n} w_i e_i^2}$, L=0.199 and $w_i = (\rho(u_i)/{u_i}^2)$ [42].

Rousseeuw and Leroy [43] improved the LMS estimation, a robust regression method designed to handle outliers effectively. Unlike traditional regression approaches, LMS focuses on minimising the median of the squared errors rather than the sum. By using the median instead of the

Maejo Int. J. Sci. Technol. **2026**, *20*(01), 108-123

mean, LMS provides higher resistance to extreme values, ensuring that some outliers do not disproportionately influence the estimation. The objective of LMS is to minimise the following:

$$\min(\text{median}(\varepsilon_i^2)) \tag{8}$$

The LMS estimator is highly robust to unusual observations that deviate significantly in both the x and y directions. One of its key strengths is its high breakdown point of 0.5, meaning it can tolerate up to 50% of the data being outliers without compromising the accuracy of the estimation.

Huber [44] presented Huber-M estimation, a robust regression method that minimises the influence of outliers by using a different function instead of directly minimising the sum of squared errors. Unlike traditional least squares estimation, which treats all errors equally, Huber-M estimation applies a weighted approach treating small errors similarly to least squares while reducing the impact of larger errors (outliers). The objective function of the M-estimation is presented as

$$min \sum_{i=1}^{n} \rho(e_i) \tag{9}$$

and is a symmetric function of outliers.

Huber's function ρ, is designed as

$$\rho(e) = \begin{cases} \frac{e^2}{2}, & |e| \le m, \\ m|e| - \frac{m^2}{2}, & |e| > m \end{cases} \tag{10}$$

The derivative of the function given in equation (10) is expressed as

$$\varphi(y) = \begin{cases} e, & |e| \le m, \\ msgn(e), & |e| > m \end{cases} \tag{11}$$

In the Huber-M estimation method the sign function (denoted as sgn(.)) plays a key role in determining how residuals are handled. The function is defined as

$$Sgn(x) \begin{cases} -1 & if\ x < m \\ 0 & if\ |x| \le m \\ 1 & if\ x > m \end{cases} \tag{12}$$

Here, m is a threshold value and for Huber's estimate**,** it is set to 1.345 as a constant. This threshold determines how residuals are treated. If the residual is smaller than m, the error is treated normally (with a linear weight), but if it is larger, the influence of that residual is reduced to prevent it from having an excessive impact on the estimation process. This makes Huber's estimate robust to outliers while still being efficient for the majority of the data points [45]. For a more detailed discussion on robust regression techniques**,** the work of Zaman and Bulut [6] and Zaman et al. [10] provide valuable insights and can be consulted for further exploration.

In this study we generalise the estimators presented by Zaman et al. [10] by considering MCD robust covariance estimation in addition to LTS, S, LMS and Huber-M regression coefficients in simple random sampling.

PROPOSED CLASS OF ESTIMATOR

We provide the proposed MCD-based estimators as given below:

Maejo Int. J. Sci. Technol. **2026**, *20*(01), 108-123

$$\bar{y}_{pr1} = \bar{y}_{MCD}\left(\frac{\bar{X}_{1MCD}}{\bar{x}_{1MCD}}\right)^{\alpha_1}\left(\frac{\bar{X}_{2MCD}}{\bar{x}_{2MCD}}\right)^{\alpha_2} + b_{1(LTS)}(\bar{X}_{1MCD} - \bar{x}_{1MCD}) + b_{2(LTS)}(\bar{X}_{2MCD} - \bar{x}_{2MCD}) \quad (13)$$

$$\bar{y}_{pr2} = \bar{y}_{MCD}\left(\frac{\bar{X}_{1MCD}}{\bar{x}_{1MCD}}\right)^{\alpha_1}\left(\frac{\bar{X}_{2MCD}}{\bar{x}_{2MCD}}\right)^{\alpha_2} + b_{1(S)}(\bar{X}_{1MCD} - \bar{x}_{1MCD}) + b_{2(S)}(\bar{X}_{2MCD} - \bar{x}_{2MCD}) \quad (14)$$

$$\bar{y}_{pr3} = \bar{y}_{MCD}\left(\frac{\bar{X}_{1MCD}}{\bar{x}_{1MCD}}\right)^{\alpha_1}\left(\frac{\bar{X}_{2MCD}}{\bar{x}_{2MCD}}\right)^{\alpha_2} + b_{1(LMS)}(\bar{X}_{1MCD} - \bar{x}_{1MCD}) + b_{2(LMS)}(\bar{X}_{2MCD} - \bar{x}_{2MCD}) \quad (15)$$

$$\bar{y}_{pr4} = \bar{y}_{MCD}\left(\frac{\bar{X}_{1MCD}}{\bar{x}_{1MCD}}\right)^{\alpha_1}\left(\frac{\bar{X}_{2MCD}}{\bar{x}_{2MCD}}\right)^{\alpha_2} + b_{1(HubM)}(\bar{X}_{1MCD} - \bar{x}_{1MCD}) + b_{2(HubM)}(\bar{X}_{2MCD} - \bar{x}_{2MCD}) \quad (16)$$

where $\bar{y}_{MCD}$, $\bar{X}_{1MCD}$, $\bar{X}_{2MCD}$, $\bar{x}_{1MCD}$, $\bar{x}_{2MCD}$ are computed from the MCD estimation. The MSE of the estimators given in equations (13-16) consists of variance and covariance components based on robust MCD measures, and is given as

$$\begin{aligned} MSE(\bar{y}_{pri}) \cong \frac{1-f}{n}\Big[& S^2_{y(MCD)} + \left(\alpha^*_{1(MCD)}R_{1(MCD)} + B_{1rob(k)}\right)^2 S^2_{x_1(MCD)} \\ & + \left(\alpha^*_{2(MCD)}R_{2(MCD)} + B_{2rob(k)}\right)^2 S^2_{x_2(MCD)} \\ & - 2\left(\alpha^*_{1(MCD)}R_{1(MCD)} + B_{1rob(k)}\right)S_{yx_1(MCD)} - 2\left(\alpha^*_2 R_{2(MCD)} + B_{2rob(k)}\right)S_{yx_2(MCD)} \\ & + 2\left(\alpha^*_{1(MCD)}R_{1(MCD)} + B_{1rob(k)}\right)\left(\alpha^*_{2(MCD)}R_{2(MCD)} + B_{2rob(k)}\right)S_{x_1x_2(MCD)}\Big] \end{aligned} \quad (17)$$

where $\mathrm{R_{1(MCD)}} = (\overline{\mathrm{Y}}_{(\mathrm{MCD})}/\overline{\mathrm{X}}_{1(\mathrm{MCD})})$, $\mathrm{R_{2(MCD)}} = (\overline{\mathrm{Y}}_{(\mathrm{MCD})}/\overline{\mathrm{X}}_{2(\mathrm{MCD})})$,

$\alpha^*_{1(\mathrm{MCD})} = (\mathrm{S_{y(MCD)}}\rho_{x_1x_2(MCD)})(\rho_{yx_1(MCD)}\rho_{x_1x_2(MCD)} - \rho_{yx_2(MCD)})/\mathrm{R_{1(MCD)}}\mathrm{S_{x_1(MCD)}}(1-\rho^2_{x_1x_2(MCD)})$,

$\alpha^*_{2(\mathrm{MCD})} = (\mathrm{S_{y(MCD)}}\rho_{x_1x_2(MCD)})(\rho_{yx_2(MCD)}\rho_{x_1x_2(MCD)} - \rho_{yx_1(MCD)})/\mathrm{R_{2(MCD)}}\mathrm{S_{x_2(MCD)}}(1-\rho^2_{x_1x_2(MCD)})$.

We remark that the MSE equations of the proposed MCD-based estimators are in the same form as the MSE equations given in equation (5), but it is apparent that $R_1, R_2, S^2_{x_1}, S^2_{x_2}, S_{x_1x_2}$ S_{yx_1}, S_{yx_2}, S^2_y, α^*_1 and α^*_2 in equation (5) should be replaced by $R_{1(MCD)}$, $R_{2(MCD)}$, $S^2_{x_1(MCD)}$, $S^2_{x_2(MCD)}$, $S_{x_1x_2(MCD)}$, $S_{yx_1(MCD)}$, $S_{yx_2(MCD)}$, $S^2_{y(MCD)}$, $\alpha^*_{1(MCD)}$ and $\alpha^*_{2(MCD)}$ whose values are computed MCD covariance estmation.

The MCD estimation for the location parameter of multivariate data was introduced by Rousseeuw [46]. It is defined as the mean of the h points from the data set X, where the determinant of the covariance matrix is minimised. The trimming ratio, denoted by a, determines the number of points considered for the estimation, with h being equal to $n(1 - a)$, where n is the total number of data points. The covariance matrix of this subset is then used as the MCD estimator for the scatter parameter, where $T(X)$ represents the mean vector of the subset of size h from the data set X whose covariance matrix has the smallest determinant.

One of the key features of the MCD estimation is its breakdown point, which is equal to the trimming ratio a. When h is set to 0.5 n, the estimator reaches its maximum breakdown point of

Maejo Int. J. Sci. Technol. **2026**, *20*(01), 108-123

50%. However, to achieve a balance between robustness and efficiency, it is typically set to h = 0.75 n, which results in a breakdown point of 25% [47, 48, 49]. In simpler terms, the MCD estimation focuses on identifying the most representative subset of data by trimming away outliers, ensuring that the estimate is robust (resistant to outliers) and efficient (providing accurate results with minimal data loss).

The proposed estimator $\bar{\boldsymbol{y}}_{\boldsymbol{pri}}$ is said to be more efficient than the estimator $\bar{\boldsymbol{y}}_{\boldsymbol{zi}}$ by Zaman et al. [10] if its MSE is smaller. Accordingly, the condition for superiority is obtained as follows:

$$MSE(\bar{y}_{pri}) < MSE(\bar{y}_{zi})$$

$$\begin{aligned} S^2_{y(MCD)} - S^2_y + \left(\alpha^*_{1(MCD)}R_{1(MCD)} + B_{1rob(k)}\right)^2 S^2_{x_1(MCD)} - \left(\alpha^*_1 R_1 + B_{1rob(k)}\right)^2 S^2_{x_1} \\ + \left(\alpha^*_{2(MCD)}R_{2(MCD)} + B_{2rob(k)}\right)^2 S^2_{x_2(MCD)} - \left(\alpha^*_2 R_2 + B_{2rob(k)}\right)^2 S^2_{x_2} \\ - 2\left(\alpha^*_{1(MCD)}R_{1(MCD)} + B_{1rob(k)}\right)S_{yx_1(MCD)} + 2\left(\alpha^*_1 R_1 + B_{1rob(k)}\right)S_{yx_1} \\ - 2\left(\alpha^*_2 R_{2(MCD)} + B_{2rob(k)}\right)S_{yx_2(MCD)} + 2\left(\alpha^*_2 R_2 + B_{2rob(k)}\right)S_{yx_2} \\ + 2\left(\alpha^*_{1(MCD)}R_{1(MCD)} + B_{1rob(k)}\right)\left(\alpha^*_{2(MCD)}R_{2(MCD)} + B_{2rob(k)}\right)S_{x_1x_2(MCD)} \\ - 2\left(\alpha^*_1 R_1 + B_{1rob(k)}\right)\left(\alpha^*_2 R_2 + B_{2rob(k)}\right)S_{x_1x_2} < 0 \end{aligned} \quad (18)$$

When condition (18) is satisfied, the proposed MCD-based estimators given in (13)-(16) are more efficient than the estimators given in (1)-(4) [10].

RESULTS AND DISCUSSION

Population 1: Practical Study

We use the epilepsy data set from Zaman et al. [10] to compare the efficiencies between the proposed estimators and those by Zaman et al. [10] based on MSE equations in simple random sampling. The data set consists of observations $n = 59$ and is publicly available in the R statistical software (package: robustbase, data: epilepsy), ensuring full reproducibility of the analysis [50]. The data originate from a clinical study on epileptic patients and include one study variable and two auxiliary variables. The study variable y represents the number of epileptic seizures recorded during the observation period while the auxiliary variables are x_1, the baseline number of seizures prior to randomisation and x_2, the age of patients. These auxiliary variables are expected to be correlated with the study variables and are therefore incorporated to improve estimation efficiency. It is worth noting that the variables in this data set are primarily count-based and may exhibit skewness and potential outliers, which makes the data particularly suitable for evaluating the robustness and efficiency of the proposed estimators. The definitions of the variables in the data set are shown in Table 1. The data set contains two auxiliary variables and one study variable. The statistics of the data for real data set are indicated in Table 2.

Table 1. Description of auxiliary and study variables for data set

Data set	Variable	Description
Epilepsy	x_1	Total number of epilepsy attacks
	x_2	Age values of patients
	y	Number of epileptic attacks recorded prior to randomisation

Maejo Int. J. Sci. Technol. **2026**, *20*(01), 108-123

Table 2. Data statistics

Statistic	Classical	MCD
$\bar{Y}$	33.051	16.182
$\bar{X}_1$	28.339	28.795
$\bar{X}_2$	31.220	19.841
S_y^2	2077.911	107.125
S_{x1}^2	-15.724	42.309
S_{x2}^2	1018.730	126.732
r_{yx1}	-0.055	0.004
r_{yx2}	0.831	0.698
r_{x1x2}	-0.189	-0.076
R_1	1.166	0.562
R_2	1.059	0.816
a_1	0.103	0.056
a_2	0.003	0.002

Table 3 introduces the robust regression coefficients obtained using the LTS, S, LMS and Huber-M. We have calculated the MSE values of our proposed MCD-based estimators and Zaman et al. [10] estimators, as defined in Sections 2 and 3.

Table 3. Robust regression coefficients

Model	B_1	B_2
LTS	-0.1479567	0.6002718
S	0.1262206	0.6653604
LMS	-0.1545064	0.695279
Huber-M	0.4919941	1.0793328

Figure 1 indicates the scatter graph of the auxiliary and study variables; it is clear that the data have outliers according to the Figure. We assume n = 10, 20, 30, 40, and outlier ratios τ = 0.10, 0.20, 0.30 by considering simple random sampling.

Table 4 denotes the MSE values of our proposed MCD-based estimators and those in Zaman et al. [10]. When Table 4 is examined, it is seen that MSE values of the estimators decrease as sample size increases.

Table 5 presents the numerical results of the efficiency conditions obtained theoretically in equation 18. Based on these outcomes, the proposed estimators can be considered highly robust and effective. These results are also illustrated in Figure 2.

The proposed MCD-based estimator based on S estimates produces the lowest MSE values for the epilepsy data set. The proposed MCD-based estimators have the highest MSE values when compared with those by Zaman et al. [10] estimators in the data sets. These results are not surprising because the condition (19) is satisfied. This situation is clearly presented in Table 4.

Maejo Int. J. Sci. Technol. **2026**, *20*(01), 108-123

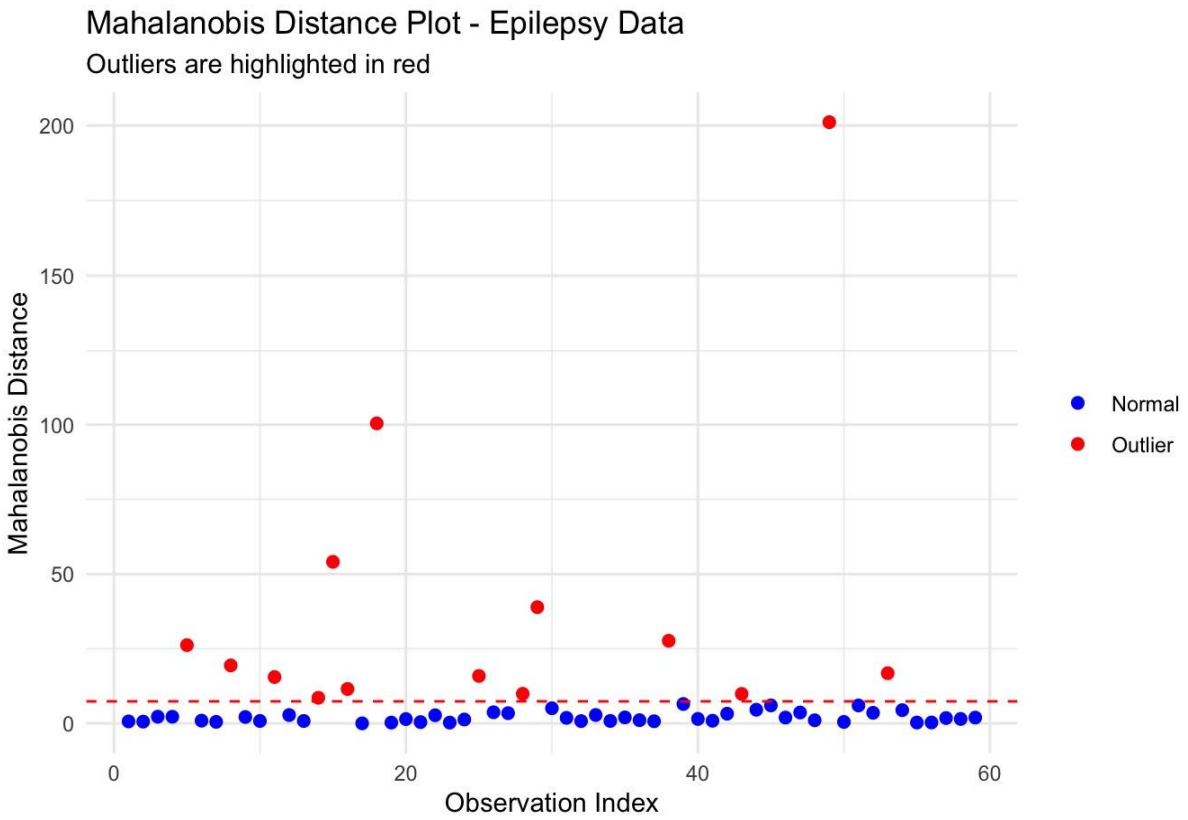

Figure 1. Scatter graph of auxiliary and study variables

Table 4. MSE values of Zaman et al. [10] and proposed MCD-based estimators

Sample size (n)	Regression methods	Zaman et al. [10] estimators (Classic)	Proposed estimators (MCD)
10	LTS	92.3497	4.6939
	S	86.2496	4.5496
	LMS	85.3814	5.0346
	Huber-M	59.0219	7.0075
20	LTS	36.7514	1.8680
	S	34.3238	1.8105
	LMS	35.8069	1.8509
	Huber-M	23.4883	2.7887
30	LTS	18.2187	0.9260
	S	17.0152	0.8975
	LMS	15.5738	0.9591
	Huber-M	11.6438	1.3824
40	LTS	8.9523	0.4550
	S	8.3609	0.4410
	LMS	8.1152	0.4584
	Huber-M	5.7215	0.6793

Maejo Int. J. Sci. Technol. **2026**, *20*(01), 108-123

Table 5. Results of the condition in Equation (18)

Sample size (n)	Regression method	Classic	MCD	Efficiency comparison	
				Difference (Eq.18)	Decision
10	LTS	92.3497	4.6939	-1055.448	TRUE
	S	86.2496	4.5496	-983.735	TRUE
	LMS	85.3814	5.0346	-967.441	TRUE
	Huber-M	59.0219	7.0075	-626.296	TRUE
20	LTS	36.7514	1.8680	-1055.448	TRUE
	S	34.3238	1.8105	-983.735	TRUE
	LMS	35.8069	1.8509	-1027.388	TRUE
	Huber-M	23.4883	2.7887	-626.297	TRUE
30	LTS	18.2187	0.9260	-1055.448	TRUE
	S	17.0152	0.8975	-983.735	TRUE
	LMS	15.5738	0.9591	-892.001	TRUE
	Huber-M	11.6438	1.3824	-626.296	TRUE
40	LTS	8.9523	0.4550	-1055.448	TRUE
	S	8.3609	0.4410	-983.735	TRUE
	LMS	8.1152	0.4584	-951.045	TRUE
	Huber-M	5.7215	0.6793	-626.297	TRUE

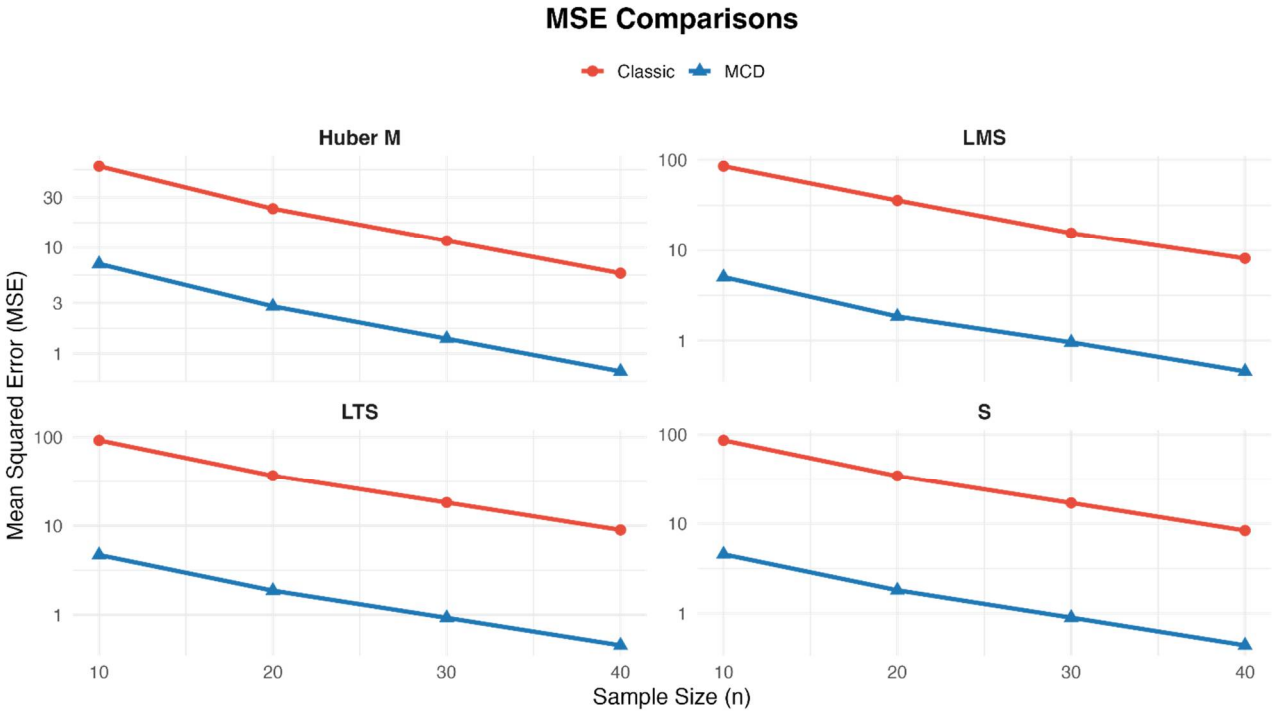

Figure 2. MSE comparisons of classical and proposed estimators for population-1

Maejo Int. J. Sci. Technol. **2026**, *20*(01), 108-123

Population 2: Simulation Study

A simulation study was conducted to evaluate the performance of the proposed MCD-based estimators. We used the same data set as that used by Zaman et al. [10] in this simulation study. The sample size was taken as n = 10, 20, 30, 40. Random samples were generated so that each time there would be an outlier rate of τ = 0.10, 0.20, 0.30 in the data. The estimated values were calculated using the sample data. Details of the data generation mechanism and outlier structure are provided by Zaman et al. [10]. This process was repeated 1000 times for each case and the MSE values were calculated according to equation (19):

$$MSE = \frac{1}{1000}\sum_{i=1}^{1000}(\widehat{\overline{Y}}_{\mathrm{i}} - \overline{Y})^{\mathbf{2}}, \tag{19}$$

where $\widehat{\overline{Y}}_{\mathrm{i}}$ shows the estimators for $i = 1,2,\ldots,999,1000$ and $\overline{Y}$ was calculated from the remaining data after the observations with outliers were removed.

Table 6 and Figure 3 show the simulation results. According to the results, the proposed MCD-based estimators outperform those given by Zaman et al. [10] in all sample sizes. The proposed MCD-based estimators have the lowest MSE values. Under all conditions, the proposed MCD-based estimators perform better than those by Zaman et al. [10]. The MSE of the proposed estimators using MCD in both real data and simulation decreases while efficiency increases.

Table 6. Simulation results of MSE values by Zaman et al. [10] and proposed MCD-based estimators

		Outlier ratio 0.10		Outlier ratio 0.20		Outlier ratio 0.30	
Sample size (n)	Regression methods	Zaman et al. [10] (Classic)	Proposed Estimators (MCD)	Zaman et al. [10] (Classic)	Proposed Estimators (MCD)	Zaman et al. [10] (Classic)	Proposed Estimators (MCD)
10.00	LTS	127583.9	322.4619	239541.6	33.63198	328693	25186.33
	S	128658.5	426.86095	235327.7	24.84621	326607.9	25673.32
	LMS	140896.5	12.58267	247014.3	0.803325	331176.4	27294.1
	Huber-M	195239.3	192.51561	227550.2	364.8209	240900.6	7362.121
20.00	LTS	103657.3	2051.107	210187	1948.91	299373.5	34.37133
	S	114897.8	2186.049	213484.5	2022.126	299854	5.479167
	LMS	98239.4	1896.501	208101.5	1726.059	301338.1	26.38234
	Huber-M	156511.8	3023.96	220910.1	5032.36	240381.1	5079.87
30.00	LTS	103816.7	2460.3	203706.9	3015.017	294844.5	541.9844
	S	117064.6	2889.34	207651.5	3331.969	293611	621.7268
	LMS	101791.6	2463.819	201975.7	2924.21	296508.5	463.2565
	Huber-M	153227.7	3680.696	224269.6	6193.818	251084.8	6062.222
40.00	LTS	105125	3503.19	200896.9	3056.733	304437.6	653.7693
	S	122315.5	4100.051	204523.4	3608.595	303058.8	709.8339
	LMS	108050.4	3621.717	200960.9	3240.945	303567.8	692.0687
	Huber-M	156093.2	4871.581	219579.7	5995.623	258219.9	6480.672

Maejo Int. J. Sci. Technol. **2026**, *20*(01), 108-123

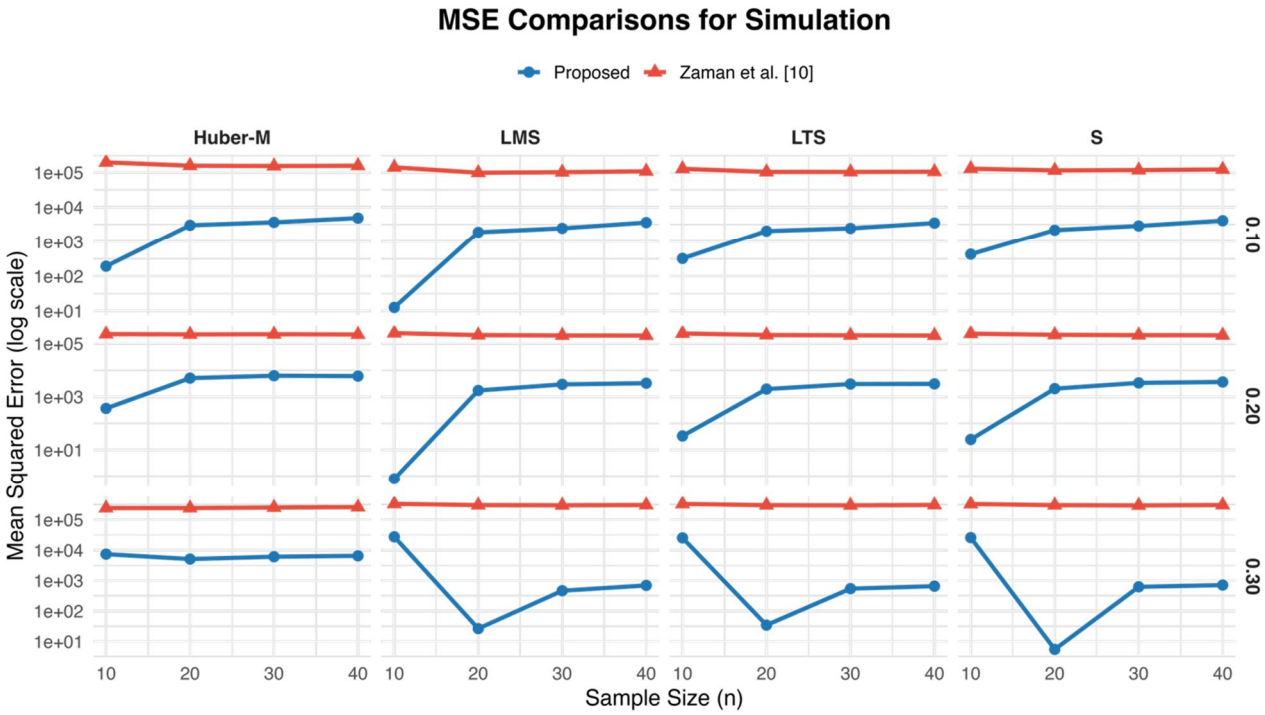

Figure 3. MSE comparisons of classical and proposed estimators for population-1 at outlier ratios of 0.10, 0.20 and 0.30

In practical applications MCD-based estimators are particularly effective in several situations. They are most useful when data sets contain a high proportion of outliers, as traditional methods like least squares can be heavily influenced by extreme values, leading to inaccurate results. Additionally, MCD estimators perform well when data distributions deviate from normality, offering more robust and reliable estimates in such cases. Furthermore, these estimators are particularly advantageous in handling complex survey data with error structures that cannot be adequately modelled using standard regression techniques. Thus, MCD-based estimators are recommended when robustness to outliers, non-normal distributions, or complex survey designs is critical for obtaining accurate population estimates.

CONCLUSIONS

We have presented MCD-based estimators using LTS, S, LMS and Huber-M robust regression techniques to handle the robustness task for the estimator given by Zaman et al. [10]. The results demonsrate that the proposed MCD-based estimators for estimating the population mean of the study variables using outlier data for simple random sampling are more efficient. The proposed MCD-based estimators provide lower MSEs than those of Zaman et al. [10]. These results have been demonstrated theoretically and supported by empirical and simulation results.

REFERENCES

1. Y. Büyükkör and A. K. Şehirlioğlu, "Robust regression: A comparative simulation study", *Eur. J. Sci. Technol.*, **2020**, *18*, 188-195.
2. K. Mitra, A. Veeraraghavan and R. Chellappa, "Analysis of sparse regularization based robust regression approaches", *IEEE Trans. Signal Process.*, **2013**, *61*, 1249-1257.

3. D. C. Montgomery, E. A. Peck and G. G. Vining, "Introduction to Linear Regression Analisis", 6th Edn., John Wiley and Sons, Hoboken, **2021**, pp.63-66.
4. A. S. Albayrak, "Alternative biased estimation techniques of least squares technique in case of multiple linear correlation and an application", *Int. J. Manage. Econ. Bus.*, **2005**, *1*, 105-126.
5. C. Kadılar, M. Candan and H. Cıngı, "Ratio estimators using robust regression", *Hacettepe J. Math. Stat.*, **2007**, *36*, 181-188.
6. T. Zaman and H. Bulut, "Modified ratio estimators using robust regression methods", *Commun. Stat. Theory Meth.*, **2019**, *48*, 2039-2048.
7. U. Shahzad, N. H. Al-Noor, M. Hanif, I. Sajjad and M. M. Anas, "Imputation based mean estimators in case of missing data utilizing robust regression and variance covariance matrices", *Commun. Stat. Simul. Comput.*, **2022**, *51*, 4276-4295.
8. H. Bulut and T. Zaman, "An improved class of robust ratio estimators by using the minimum covariance determinant estimation", *Commun. Stat. Simul. Comput.*, **2022**, *51*, 2457-2463.
9. V. K. Yadav and S. Prasad, "Some exponential estimators in sample survey using robust regression method in the presence of outliers", *Lobachevskii J. Math.*, **2024**, *45*, 1674-1690.
10. T. Zaman, E. Dünder, A. Audu, D. A. Alilah, U. Shahzad and M. Hanif, "Robust regression-ratio-type estimators of the mean utilizing two auxiliary variables: A simulation study", *Math. Prob. Eng.*, **2021**, *2021*, Art.no.6383927.
11. U. Shahzad, N. H. Al-Noor, N. Afshan, D. A. Alilah, M. Hanif and M. M. Anas, "Minimum covariance determinant–based quantile robust regression type estimators for mean parameter", *Math. Prob. Eng.*, **2021**, *2021*, Art.no.5255839.
12. Q.-Ul-A. S. Durrani, N. Ali, U. Shahzad, M. Hanif and S. Ghaffar, "Robust estimation of variance using supplementary information", *J. Asian Dev. Stud.*, **2024**, *13*, 310-335.
13. J. Kalina, "The minimum weighted covariance determinant estimator revisited", *Commun. Stat. Simul. Comput.*, **2020**, *51*, 3888-3900.
14. 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", *Rev. Investig. Oper.*, **2019**, *40*, 201-209.
15. K. Ul I. Rather, E. G. Koçyiğit, R. Onyango and C. Kadılar, "Improved regression in ratio type estimators based on robust M-estimation", *PLoS One*, **2022**, *17*, Art.no.e0278868.
16. A. Ahmed, A. Sanaullah, E. Oral and M. Hanif, "Robust ratio estimators of population mean for skewed and contaminated population", *J. Stat. Comput. Simul.*, **2022**, *93*, 800-817.
17. U. Shahzad, N. H. Al-Noor, M. Hanif, I. Sajjad and M. M. Anas, "Imputation based mean estimators in case of missing data utilizing robust regression and variance-covariance matrices", *Commun. Stat. Simul. Comput.*, **2020**, *51*, 4276-4295.
18. M. Subzar, C. N. Bouza and A. I. Al-Omari, "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", *Rev. Investig. Oper.*, **2019**, *40*, 600-609.
19. T. Zaman, "Improvement of modified ratio estimators using robust regression methods", *Appl. Math. Comput.*, **2019**, *348*, 627-631.
20. A. Raza, M. Noor-ul-Amin and M. Hanif, "Regression-in-ratio estimators in the presence of outliers based on redescending M-estimator", *J. Reliab. Stat. Stud.*, **2019**, *12*, 1-10.

21. T. Zaman and H. Bulut, "A simulation study: Robust ratio double sampling estimator of finite population mean in the presence of outliers", *Sci. Iran.*, **2024**, *31*, 1330-1341.
22. T. Zaman and H. Bulut, "An efficient family of robust type estimators for the population variance in simple and stratified random sampling", *Commun. Stat. Theory Meth.*, **2023**, *52*, 2610-2624.
23. L. K. Grover and A. Kaur, "An improved regression type estimator of population mean with two auxiliary variables and its variant using robust regression method", *J. Comput. Appl. Math.*, **2021**, *382*, Art.no.113072.
24. M. A. Gulzar, W. Latif, M. Abid, H. Z. Nazir and M. Riaz, "On enhanced exponential-cum-ratio estimators using robust measures of location", *Concurr. Comput. Pract. Exp.*, **2022**, *34*, Art.no.e6763.
25. M. M. Anas, Z. Huang, U. Shahzad, T. Zaman and S. Shahzadi, "Compromised imputation based mean estimators using robust quantile regression", *Commun. Stat.Theory Meth.*, **2022**, *53*, 1700-1715.
26. T. Zaman, H. Bulut and S. K. Yadav, "Robust ratio-type estimators for finite population mean in simple random sampling: A simulation study", *Concurr. Comput. Pract. Exp.*, **2022**, *34*, Art.no.e7273.
27. A. Audu, A. Gidado, N. S. Dauran, S. A. Abdulazeez, M. A. Yunusa and I. Abubakar, "Modified robust regression-type estimators with multi-auxiliary variables using non-conventional measures of dispersion", *Niger. J. Basic Appl. Sci.*, **2023**, *31*, 8-25.
28. M. K. Pandey, G. N. Singh and A. Bandyopadhyay, "Efficiency study of a robust regression-type estimator for population mean under different ranked set sampling methods with outlier handling", *Braz. J. Probab. Stat.*, **2024**, *38*, 232-252.
29. V. K. Yadav and S. Prasad, "Some exponential estimators in sample survey using robust regression method in the presence of outliers", *Lobachevskii J. Math.*, **2024**, *45*, 1674-1690.
30. M. Abid, M. Sun, W. Latif and T. Nawaz, "A novel robust class of estimators for estimation of finite population mean: A simulation study", *Pak. J. Stat. Oper. Res.*, **2024**, *20*, 417-444.
31. T. Zaman and H. Bulut, "A new class of robust ratio estimators for finite population variance", *Sci. Iran.*, **2022**, *32*, Art.no.5100.
32. A. F. Hashem, A. O. Alshammari, U. Shahzad and S. Iftikhar, "OGK approach for accurate mean estimation in the presence of outliers", *Math.*, **2025**, *13*, Art.no.3251.
33. J. Subramani, "Two parameter modified ratio estimators with two auxiliary variables for the estimation of finite population mean", *Biom. Biostat. Int. J.*, **2018**, *7*, 559-568.
34. N. Dansawad, "Efficient modified estimators for the population mean using auxiliary variables", *Asian Health Sci. Technol. Rep.*, **2022**, *30*, 84-93.
35. T. Zaman and H. Bulut, "Modified regression estimators using robust regression methods and covariance matrices in stratified random sampling", *Commun. Stat. Theory Meth.*, **2020**, *49*, 3407-3420.
36. A. M. Alomair, U. Shahzad, N. H. Al-Noor and M.A. Alomair, "Minimum-covariance-determinant-based mean estimators under systematic sampling", *Maejo Int. J. Sci. Technol.*, **2025**, *19*, 67-79.
37. A. M. Alomair and U. Shahzad, "Optimizing mean estimators with calibrated minimum covariance determinant in median ranked set sampling", *Symmetry*, **2023**, *15*, Art.no.1581.

38. R. Alharbi, M. S. Mustafa, A. Al Mutairi, M. Hussein, M. Yusuf, A. Elshenawy and S. G. Nassr, "Enhancing mean estimators in median ranked set sampling with dual auxiliary information", *Heliyon*, **2023**, *9*, Art.no.e21427
39. M. Tahir, B. Yude, S. Bashir, S. Hussain and T. Munir, "A new improved estimator for the population mean using twofold auxiliary information under simple random sampling", *Manag. Sci. Lett.*, **2023**, *13*, 265-276.
40. P. Singh and S. Gupta, "Combining two auxiliary variables for elevated estimation of finite population mean under neutrosophic framework", *Neutrosophic Sets Syst.*, **2025**, *76*, 275-287.
41. P. J. Rousseeuw and V. Yohai, "Robust regression by means of S-estimators in robust and nonlinear time series analysis", in "Lecture Notes in Statistics" (Ed. J. Franke, W. Härdle and R. D. Martin), Springer-Verlag, New York, pp.256-272.
42. M. Salibian-Barrera and V. J. Yohai, "A fast algorithm for S-regression estimates", *J. Comput. Graph. Stat.*, **2006**, *15*, 414-427.
43. P. J. Rousseeuw and A. M. Leroy, "Robust Regression and Outlier Detection", John Wiley and Sons, New York, **1987**, pp.29-39.
44. P. J. Huber, "Robust regression: Asymptotics, conjectures and Monte Carlo", *Ann. Stat.*, **1973**, *1*, 799-821.
45. J. Fox, "Nonparametric regression: Appendix to an R and S-Plus companion to applied regression", **2002**, https://socserv.socsci.mcmaster.ca/jfox/Books/Companion-1E/appendix.html (Accessed: May 2025).
46. P. J. Rousseeuw, "Multivariate estimation with high breakdown point", *Math. Stat. Appl.*, **1985**, *8*, 283-297.
47. P. J. Rousseeuw and K. van Driessen, "A fast algorithm for the minimum covariance determinant estimator", *Technometrics*, **1999**, *41*, 212-223.
48. H. Bulut and Y. Oner, "The evaluation of socio-economic development of development agency regions in Turkey using classical and robust principal component analyses", *J. Appl. Stat.*, **2017**, *44*, 2936-2948.
49. H. Bulut, "Use of robust statistics in multivariate statistical analysis", *Master Thesis*, **2014**, Ondokuz Mayıs University, Turkey.
50. P. F. Thall and S. C. Vail, "Some covariance models for longitudinal count data with overdispersion", *Biomet.*, **1990**, *46*, 657-671.