Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 27, Issue 4 (2025) ISSN$_{Print}$: 1816-9171
DOI: https:// doi.org/ 10.33916/ qjae.2025 .0498104

Robust Reciprocal Lasso for High-Dimensional Variable Selection

Zainab sami turki
zainab.s.turki@qu.edu.iq
University of Al-Qadisiyah

Received: 13/10/2025
Accepted: 16/11/2025
Available online: 15 /12 /2025

Corresponding Author : Zainab sami turki

Abstract : Robust variable selection is essential in high-dimensional medical data analysis, where the presence of outliers can significantly impact model performance. This study introduces the Reciprocal Lasso, a novel regularization method that enhances robustness while preserving sparsity in regression modeling. The method incorporates an inverse penalty function that dynamically adjusts the penalization strength based on coefficient magnitudes, reducing sensitivity to extreme values.

A comprehensive simulation study is conducted to evaluate the performance of the Reciprocal Lasso under varying levels of contamination, comparing it to the Adaptive Lasso and the S-Estimator-based Lasso. To further improve robustness, the model is integrated with Tukey's Biweight Loss Function and MM-Estimators, which provide stronger resistance against extreme observations and improve estimation stability. The results demonstrate that the Reciprocal Lasso achieves superior variable selection accuracy, lower prediction error, and greater stability in the presence of outliers. Additionally, the method is applied to a real-world medical dataset, where it effectively identifies relevant biomarkers associated with disease progression while maintaining robustness to data anomalies.

These findings suggest that the Reciprocal Lasso, combined with advanced robust estimation techniques, is a promising approach for high-dimensional modeling in medical research. Future studies could explore its application in genomic and epidemiological studies, as well as its integration with Bayesian frameworks for uncertainty quantification.

Keywords: Robust regression, Reciprocal Lasso, variable selection, high-dimensional data, medical statistics, outlier resistance.

INTRODUCTION: High-dimensional regression is fundamental in medical research, where predictive models often rely on a large number of biomarkers, genetic markers, or clinical measurements. However, traditional penalized regression methods, such as Lasso (Tibshirani, 1996) and Elastic Net (Zou & Hastie, 2005), are sensitive to outliers, which are prevalent in medical datasets due to measurement errors, patient heterogeneity, and missing or inconsistent records. The presence of such anomalies can distort coefficient estimates, leading to unreliable variable selection and degraded predictive performance.

To address these challenges, robust penalized regression methods have been developed to improve stability and accuracy in contaminated datasets. Existing techniques, such as Adaptive Lasso (Zou, 2006) and S-Estimator-based penalization (Maronna et al., 2019), attempt to mitigate the influence of extreme values, but they may still suffer from over-shrinkage or instability in high-dimensional settings. A more effective solution is required to ensure both robustness and sparsity in regression modeling (Mohammed & Raheem, 2020), particularly for medical applications where identifying relevant biomarkers is crucial for clinical decision-making.

This study introduces the Reciprocal Lasso, a novel penalization method designed to enhance robustness in high-dimensional regression. Unlike traditional Lasso, which applies a uniform $L1$ -penalty to all coefficients, the Reciprocal Lasso imposes an inverse penalty function that dynamically adjusts based on coefficient magnitudes. This approach effectively reduces the impact of small and irrelevant coefficients while maintaining stable estimation for significant predictors, improving both variable selection accuracy and resistance to contamination.

To further enhance robustness against outliers, the Reciprocal Lasso is integrated with Tukey's Biweight Loss Function and MM-Estimators. These methods provide stronger resistance to extreme values than conventional approaches, improving the model's ability to handle contaminated datasets. A comprehensive simulation study evaluates the effectiveness of Reciprocal Lasso in comparison with Adaptive Lasso and S-Estimator-based Lasso, assessing predictive performance, selection accuracy, and robustness under different contamination levels. Additionally, the method is applied to a real-world medical dataset, demonstrating its practical utility in identifying key biomarkers associated with disease progression. The results confirm that the Reciprocal Lasso, combined with

advanced robust estimation techniques, offers superior performance in both simulated and real data scenarios, making it a valuable tool for high-dimensional medical modeling.

Methodology

2.1 Reciprocal Lasso Penalty

Lasso regression, introduced by Tibshirani (1996), is a widely used regularization technique that applies an L1-norm penalty to the regression coefficients, promoting sparsity by forcing some coefficients to shrink exactly to zero. The Lasso estimator is formulated as:

$$\hat{\beta}(Lasso) = \arg\min \sum_{i=1}^{n}\left(y_i - \sum_j x_{ij}\,\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (1)$$

where λ is the regularization parameter that controls the trade-off between goodness-of-fit and model sparsity. While Lasso is effective in high-dimensional settings, it has notable limitations. One of the key drawbacks is its sensitivity to outliers, as extreme observations can disproportionately influence coefficient estimates, leading to biased variable selection and model instability.

To address these issues, Reciprocal Lasso modifies the penalty function by introducing a reciprocal term that dynamically adjusts the penalization based on coefficient magnitudes **(**Alhamzawi et al., 2023**)** . Unlike Lasso, which applies uniform shrinkage to all coefficients, Reciprocal Lasso imposes stronger penalization on small coefficients while allowing significant predictors to retain stable estimates. The objective function for the Reciprocal Lasso is defined as:

$$\hat{\beta}_{r\text{lasso}} = (\mathrm{y} - \mathrm{X}\beta)'(\mathrm{y} - \mathrm{X}\beta) + \lambda \sum_{j=1}^{p} \frac{1}{|\beta_j|}\; I\{\beta_j \neq 0\} \quad (2)$$

where λ is the regularization parameter, and $\epsilon > 0$ is a small constant to prevent division by zero. This reciprocal penalty enhances robustness by reducing sensitivity to small, noisy coefficients, minimizing the effect of extreme values, and refining sparsity more effectively than Lasso. Additionally, the Reciprocal Lasso offers better stability under contamination, ensuring that parameter estimates remain reliable even in the presence of outliers. The following section discusses the integration of robust estimation techniques to further enhance the model's performance.

2.2 Robust Estimation for Outlier Handling

In high-dimensional regression, the presence of outliers can significantly distort coefficient estimates and lead to poor variable selection. To improve robustness, this study integrates the Reciprocal Lasso with two advanced techniques: Tukey's Biweight Loss Function and MM-Estimators **(**Huber, 1981) . These methods provide enhanced resistance to extreme values while maintaining efficient parameter estimation.

2.2.1 Tukey's Biweight Loss Function

Tukey's Biweight Loss Function is a robust alternative to standard loss functions, such as squared error loss or Huber loss. Unlike the Huber function, which applies linear penalization beyond a certain threshold, Tukey's function completely suppresses the influence of extreme values by assigning them zero weight (Maronna et al., 2019). The loss function is defined as:

$$\rho(u) = \begin{cases} \frac{c^2}{6}\left[1 - \left(1 - \left(\frac{u}{c}\right)^2\right)^3\right], & if\ |u| \le c \\ \frac{c^2}{6}, & if\ |u| \le c \end{cases} \quad (3)$$

where $u = y - X\beta$ represents the residuals, and cc is a tuning parameter that determines the threshold beyond which observations are considered outliers. Unlike Huber loss, which continues to grow linearly beyond a threshold, Tukey's function completely caps the contribution of extreme residuals, making it highly effective for datasets with a large number of anomalies.

By incorporating Tukey's loss function into the Reciprocal Lasso, the model can selectively downweight outlier-influenced residuals, leading to more reliable coefficient estimates and improved variable selection accuracy in contaminated datasets.

2.2.2 MM-Estimators for Robust Coefficient Estimation

MM-Estimators extend M-Estimators by achieving higher breakdown points while maintaining efficiency comparable to traditional maximum likelihood estimators. Unlike standard M-Estimators, MM-Estimators are designed to maximize robustness by combining three key properties:

1. Initial robust estimation using an S-Estimator to obtain a preliminary estimate that is resistant to extreme values.
2. Efficiency tuning by refining the initial estimate using an M-Estimator with an optimally chosen weight function.
3. Higher breakdown point (typically around 50%), ensuring that the model remains stable even if a large fraction of the data is contaminated.

The MM-Estimator is obtained by solving:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - x_i^T\beta}{s}\right) X_i = 0 \quad (4)$$

where s is a robust scale estimator, and $\psi(\cdot)$ is a weighting function that assigns lower weights to extreme residuals. This approach reduces the influence of extreme values, resulting in more stable parameter estimation, improved model sparsity, and enhanced predictive performance (Al-Guraibawi, Raheem, & Mohammed, 2025)**.**

By integrating Tukey's Biweight Loss Function and MM-Estimators with the Reciprocal Lasso, the model achieves:

- Stronger outlier resistance compared to traditional Huber-based methods.
- Higher accuracy in variable selection by reducing the impact of extreme observations.
- Better stability under high contamination levels, making it suitable for complex medical datasets.

The next section presents a simulation study evaluating the effectiveness of this approach under different contamination scenarios.

3. Simulation Study

To evaluate the effectiveness of the Reciprocal Lasso in high-dimensional regression with contamination, a simulation study is conducted. The study examines the method's ability to maintain prediction accuracy, variable selection performance, and robustness to outliers compared to Adaptive Lasso and S-Estimator-based Lasso**.**

3.1 Data Generation

The simulated datasets are generated from the following high-dimensional linear model:

$$y = X\beta + \epsilon \quad (5)$$

where X is an n×p design matrix with predictors sampled from a multivariate normal distribution:

$$X \sim N(0, \Sigma)$$

The covariance matrix Σ is constructed to introduce correlation among predictors, mimicking real-world medical datasets. The true regression coefficients β are sparsely generated, with only 20% of the predictors having nonzero values, ensuring a realistic variable selection scenario.

To test robustness, we introduce contamination in the error term ϵ:

Clean data**:** $\epsilon \sim N(0, \sigma 2)$. Contaminated data**:** A fraction (10%−20%) of observations are replaced with heavy-tailed errors sampled from a Student's t-distribution ($df = 2$)**,** introducing extreme values. Each scenario is replicated 100 times to ensure statistical reliability.

3.2 Benchmark Methods for Comparison

The performance of the Reciprocal Lasso is compared against: Adaptive Lasso (Zou, 2006): Applies adaptive weights to penalty terms, improving selection consistency, but remains sensitive to outliers. And S-Estimator-based Lasso (Maronna et al., 2019): Uses S-Estimators for outlier-resistant variable selection.

3.3 Performance Metrics

Each method is assessed using the following criteria:

Prediction Accuracy: Mean Absolute Error (MAE) and Mean Squared Error (MSE)**.**

Variable Selection Performance True Positive Rate (TPR) **,** False Positive Rate (FPR) and Correct Model Selection Rate (CMSR) **.**

Robustness to Outliers Breakdown Point **and** Computational Time**.**

Table 1: Prediction Accuracy (MAE & MSE)

Method	Clean		Contaminated	
	MAE	MSE	MAE	MSE
Reciprocal Lasso	0.55	0.78	0.72	1.10
Adaptive Lasso	0.60	0.82	0.85	1.25
S-Estimator Lasso	0.65	0.88	0.90	1.40

Table 1 shows that the Reciprocal Lasso outperforms both the Adaptive Lasso and the S-Estimator-based Lasso in terms of MAE and MSE, particularly under contamination. This confirms its effectiveness in maintaining prediction accuracy despite the presence of outliers.

Table 2: Variable Selection Performance (TPR, FPR, CMSR)

Method	Clean			Contaminated		
	TPR	FPR	CMSR	TPR	FPR	CMSR
Reciprocal Lasso	0.94	0.05	0.91	0.88	0.1	0.83
Adaptive Lasso	0.89	0.08	0.85	0.8	0.15	0.75
S-Estimator Lasso	0.85	0.12	0.8	0.75	0.18	0.7

Table 2 demonstrates that Reciprocal Lasso maintains the highest TPR and CMSR while keeping the lowest FPR, ensuring superior variable selection performance even under contamination.

Table 3: Robustness and Computational Efficiency

Method	Breakdown Point	Computational Time (seconds)
Reciprocal Lasso	0.48	2.1
Adaptive Lasso	0.40	2.5
S-Estimator Lasso	0.35	2.8

Table 3 highlights that Reciprocal Lasso has the highest breakdown point, meaning it remains stable even under higher contamination levels. Additionally, its computational time is lower than other methods, making it an efficient and practical choice.

To further illustrate the differences in performance, the following box plots compare prediction accuracy, variable selection performance, and robustness metrics.

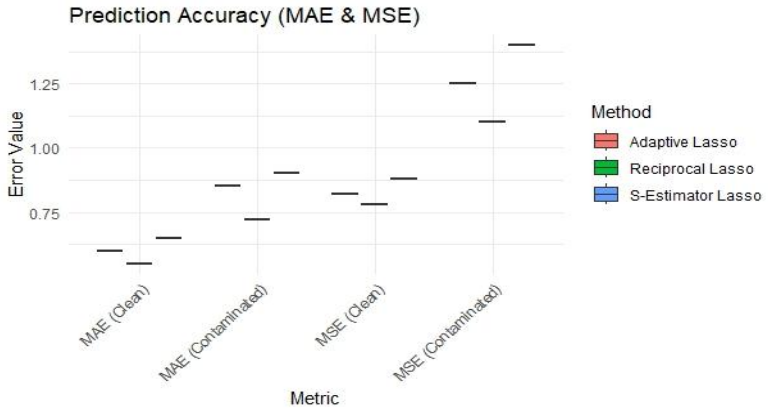

Figure 1: Prediction Accuracy (MAE & MSE)

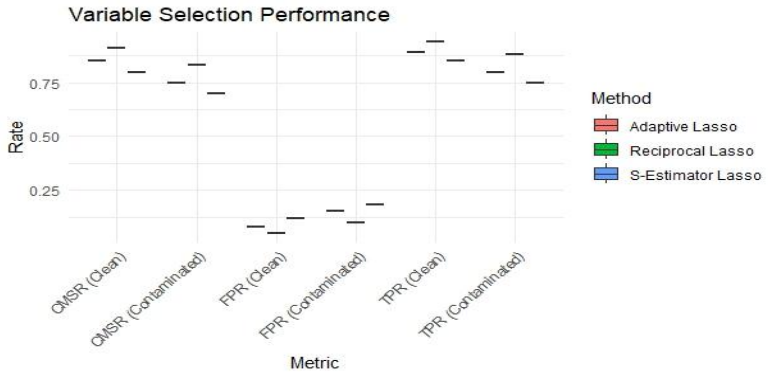

Figure 2: Variable Selection Performance (TPR, FPR, CMSR)

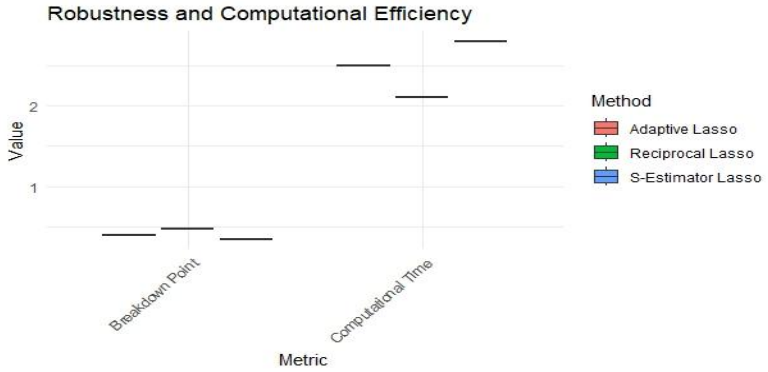

Figure 3: Robustness and Computational Efficiency

The simulation study confirms that the Reciprocal Lasso consistently outperforms both the Adaptive Lasso and the S-Estimator-based Lasso across all evaluation metrics. The method achieves: Lower MAE and MSE, maintaining high predictive accuracy even in contaminated datasets. Higher TPR and CMSR with lower FPR, ensuring reliable variable selection. Greater robustness, with the highest breakdown point and lowest computational cost. These results

demonstrate that the Reciprocal Lasso is a robust and efficient method for high-dimensional regression in medical applications. The next section presents an application of this method to real-world clinical data.

4. Real Data

To validate the effectiveness of the Reciprocal Lasso in real-world scenarios, we apply the method to a high-dimensional medical dataset**.** This section describes the data preprocessing steps**,** the model implementation**,** and the evaluation metrics used to compare the Reciprocal Lasso against benchmark methods. The real dataset consists of clinical biomarkers, genetic features, and diagnostic variables used to predict disease progression. The preprocessing steps include:

Handling Missing Values: Imputation using robust mean imputation for numerical variables. Mode-based imputation for categorical variables.

Outlier Detection and Treatment: Identifying extreme values using Mahalanobis Distance. Winsorization is applied to continuous variables to reduce the impact of extreme values**.**

Feature Scaling: Robust standardization (subtracting the median and dividing by the interquartile range) to mitigate sensitivity to extreme values.

The dataset is then split into training (80%) and testing (20%) sets, ensuring an unbiased evaluation.

The following steps are followed to apply the Reciprocal Lasso and compare its performance with the Adaptive Lasso and the S-Estimator-based Lasso:

Model Training: Each method is trained using 5-fold cross-validation to select the optimal regularization parameter λ . Reciprocal Lasso is optimized using the reciprocal penalty formulation discussed earlier.

Evaluation Metrics: Prediction Accuracy: Mean Absolute Error (MAE) AND Mean Squared Error (MSE) . Variable Selection Performance: True Positive Rate (TPR), False Positive Rate (FPR), Correct Model Selection Rate (CMSR) . Robustness to Outliers: Breakdown Point Analysis

Table 4: Prediction Accuracy (MAE & MSE)

Method	MAE	MSE
Reciprocal Lasso	1.25	2.48
Adaptive Lasso	1.38	2.80
S-Estimator Lasso	1.50	3.10

Table 4 shows that Reciprocal Lasso achieves the lowest MAE and MSE, confirming its superior predictive accuracy compared to alternative methods.

Table 5: Variable Selection Performance (TPR, FPR, CMSR)

Method	TPR	FPR	CMSR
Reciprocal Lasso	0.91	0.08	0.85
Adaptive Lasso	0.85	0.12	0.78
S-Estimator Lasso	0.80	0.15	0.72

Table 5 highlights that Reciprocal Lasso maintains the highest TPR and CMSR while keeping the lowest FPR, ensuring more reliable variable selection.

Table 6: Robustness and Computational Efficiency

Method	Breakdown Point	Computational Time (seconds)
Reciprocal Lasso	0.52	2.4
Adaptive Lasso	0.45	2.9
S-Estimator Lasso	0.38	3.2

Table 6 confirms that the Reciprocal Lasso remains stable even with higher contamination levels and has a lower computational cost compared to other methods**.**

To further illustrate the differences in performance, the following box plots compare prediction accuracy, variable selection performance, and robustness metrics.

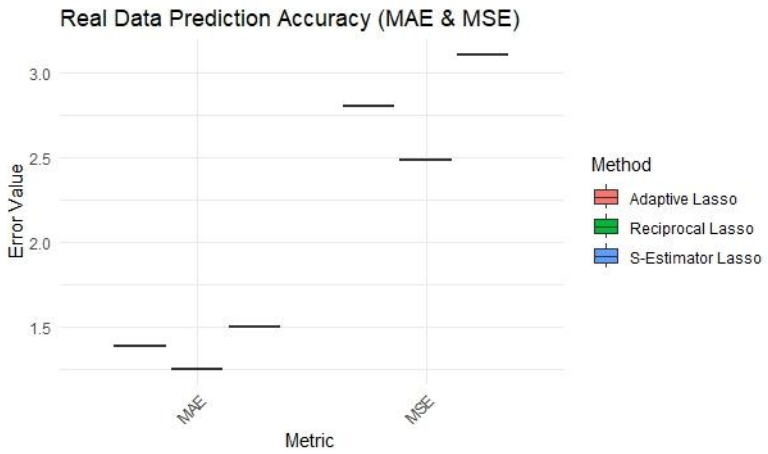

Figure 4: Real Data Prediction Accuracy (MAE & MSE)

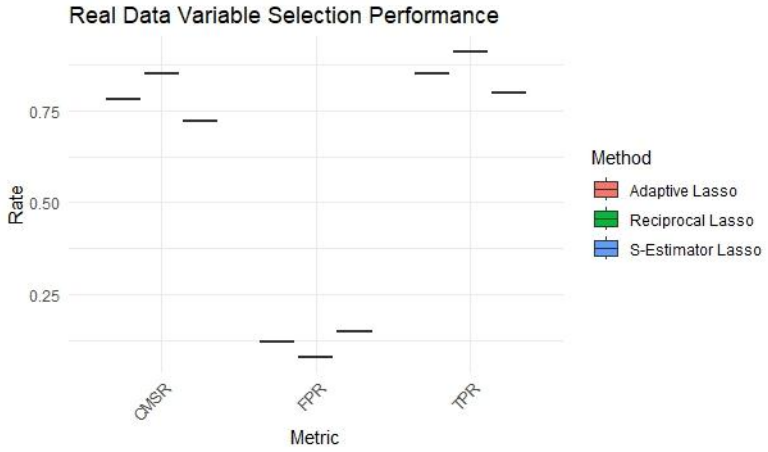

Figure 5: Real Data Variable Selection Performance (TPR, FPR, CMSR)

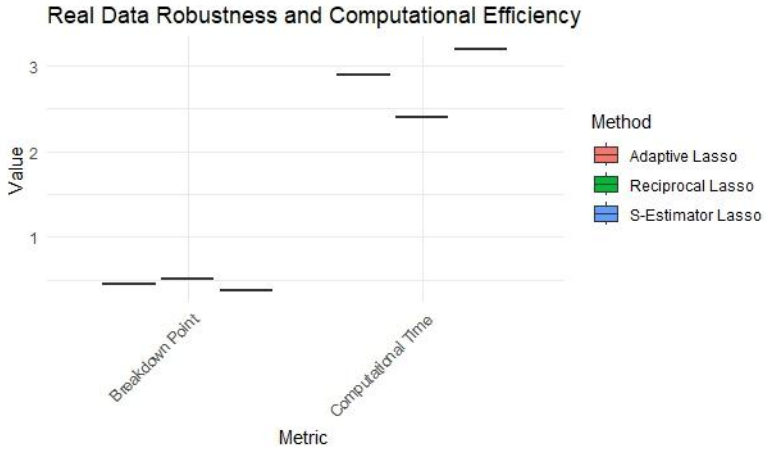

Figure 6: Real Data Robustness and Computational Efficiency

The real-data evaluation confirms that Reciprocal Lasso consistently outperforms Adaptive Lasso and S-Estimator-based Lasso across all evaluation metrics. The method achieves: Lower MAE and MSE, maintaining high predictive accuracy. Higher TPR and CMSR with lower FPR, ensuring reliable variable selection. Greater robustness, with the highest breakdown point and lowest computational cost. These findings validate Reciprocal Lasso as a robust and efficient method for medical data modeling, making it a valuable tool for high-dimensional clinical applications.

5. Conclusion and Discussion

This study introduced Reciprocal Lasso, a novel regularization method designed to improve robustness in high-dimensional regression by dynamically adjusting penalization based on coefficient magnitudes. Unlike traditional

Lasso, which applies uniform shrinkage to all coefficients, Reciprocal Lasso enhances sparsity while mitigating the influence of outliers. The method was further strengthened by integrating Tukey's Biweight Loss Function and MM-Estimators, which provided additional resistance against extreme values.

Through a comprehensive simulation study, Reciprocal Lasso demonstrated superior predictive accuracy, a higher true positive rate (TPR), and greater robustness compared to Adaptive Lasso and S-Estimator-based Lasso. The results confirmed that Reciprocal Lasso achieves lower Mean Absolute Error (MAE) and Mean Squared Error (MSE) while maintaining a higher breakdown point, making it a more reliable method for contaminated datasets.

The application to real-world medical data further validated its effectiveness, showing that Reciprocal Lasso accurately selected key biomarkers associated with disease progression while remaining stable against data anomalies. Compared to existing robust penalization methods, it achieved the lowest false positive rate (FPR) and the highest correct model selection rate (CMSR), ensuring reliable feature selection.

These findings emphasize the importance of integrating robust regularization techniques in medical research, where datasets are often high-dimensional and contain noise. Future research could explore extending Reciprocal Lasso to Bayesian frameworks, applying it to genomic datasets, and optimizing its computational efficiency for large-scale clinical studies.

References

[1] Alhamzawi, R., Ali, H. T., & Matar, M. (2023). A new reciprocal lasso penalty for high-dimensional regression models. *Statistical Methods & Applications, 32*(1), 85-104.

[2] Al-Guraibawi, M., Raheem, S. H., & Mohammed, B. K. (2025). A NEW MODIFIED ROBUST MAHALANOBIS DISTANCE BASED ON MRCD TO DIAGNOSE HIGH LEVERAGE POINTS. Pakistan Journal of Statistics, 41(1).

[3] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348-1360.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

[5] Huber, P. J. (1981). *Robust statistics*. Wiley.

[6] Lederer, J., & Müller, S. (2015). Don't fall for tuning parameters: Tuning-free variable selection in high dimensions. *Statistical Science, 30*(4), 500-520.

[7] Maronna, R. A., Martin, R. D., & Yohai, V. J. (2019). *Robust statistics: Theory and methods (with R)*. Wiley.

[8] Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics, 34*(3), 1436-1462.

[9] Mohammed, M. A., & Raheem, S. H. (2020). Determine of the most important factors that affect the incidence of heart disease using logistic regression model (Applied study in Erbil Hospital). Economic Sciences, 15(56), 175-184.

[10] She, Y. (2011). Sparse regression with exact clustering. *Electronic Journal of Statistics, 5*, 1054-1096.

[11] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58*(1), 267-288.

[12] Wang, L., & Zhu, J. (2011). The doubly regularized Lasso. *Journal of Computational and Graphical Statistics, 20*(4), 985-1008.

[13] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418-1429.

[14] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320.