Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

The Penalized Trimmed Least Squares Method for Robust Variable Selection in Multiple Linear Regression

Hassan S. Uraibi[a] and Hassan Ali Abis[a, *]

[a]Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Al Diwaniyah, Iraq. Email: Hassan.uraibi@qu.edu.iq, Alsalhyh671@gmail.com

ARTICLE INFO

Article history:
Received: 02 /11/2025
Rrevised form: 08 /12/2025
Accepted : 14 /12/2025
Available online: 30 /03/2026

Keywords:

Multiple Linear Regression, Outliers, Weighted Penalized Least Squares, Simulation.

ABSTRACT

The large number of independent variables often causes problems in the accuracy of the multiple linear regression model. This has motivated researchers to find or select the best model using methods such as forward selection, backward elimination, and stepwise regression. However, these methods have become time-consuming and ineffective when dealing with high-dimensional data. Therefore, statistical literature has proposed the Lasso method and its variants to address these issues. Nevertheless, these methods are sensitive to outliers, which has led to the emergence of several robust studies—particularly focusing on robustifying penalized methods for high-dimensional data, in order to achieve effective variable selection and robust parameter estimation. Among these methods is the Trimmed Penalized Least Squares (RPLTS), which aims to make the Lasso approach more robust against outliers appearing in the dependent variable or in the residuals. However, this method remains sensitive to the presence of leverage points. Accordingly, this research aims to weight the RPLTS method to select the best subset of variables by reducing the influence of leverage points and improving the model's accuracy. Simulations and real data were used to evaluate the efficiency of the proposed method and compare it with the previous method. The comparison was based on several statistical criteria such as variable selection accuracy, sensitivity to influential variables, specificity of non-influential variables, and mean squared error (MSE) of the model. The method that achieves the highest accuracy, sensitivity, and specificity rates, along with the lowest MSE, is considered superior. The analysis of both real and simulated data demonstrated that the proposed method outperforms the previous one in terms of robustness and efficiency.

https://doi.org/10.29304/jqcsm.2026.18.12433

1.Introduction

Linear regression is a fundamental and widely used statistical method for studying the effect of one or more independent variables on a dependent variable. There are two types: simple linear regression, which explains the effect of one independent variable on one dependent variable, and multiple linear regression, which aims to explain and determine the type of effect between one or more independent variables on a single dependent variable. Regression analysis is among the most frequently applied statistical methods in various scientific fields, such as economics and medicine. The accuracy of the estimation depends on the validity of the assumptions underlying the method used. The Ordinary Least Squares (OLS) method is one of the most widely used and common methods for estimating the parameters of a linear regression model. It relies on several fundamental assumptions that must be met to obtain an accurate model. If any of these assumptions are violated, the resulting model will be inaccurate, for example, if the random error is not normally distributed with a zero mean and constant variance. Due to technological advancements in computer science and networks, collecting large samples became relatively easy at the beginning of this development. Consequently, collecting a large number of variables and large sample sizes

∗Corresponding author: Hassan Ali Abis
Email addresses: Alsalhyh671@gmail.com
Communicated by 'sub etitor'

became commonplace in statistical literature. However, this made interpreting regression models a complex process.

Therefore, researchers resorted to proposing model selection methods, such as All Possible Subsets (OPS) and Stepwise Regression, to reduce the number of variables to those with the greatest impact on the dependent variable. A drawback of these methods is their reliance on the Least Squares method (the quadratic loss function), which requires all regression assumptions to be met to obtain the best unbiased linear estimates. The least squares method requires that the sample size be five or ten times greater than the number of independent variables; otherwise, it is a multi-solution method and consumes a lot of time if the data is high-dimensional. (Montgomery et al., 2012) Therefore, penalty methods were employed to aid in the variable selection process, such as Lasso, Adaptive Lasso, and Elastic Net. These methods utilize penalty functions to identify the most significant variables, estimate their parameters simultaneously, and assign zero values to the remaining variables. However, a drawback of these methods is their susceptibility to outliers. Variable selection methods based on OLS violate the assumption of normality in random errors due to the presence of outliers, which lead to errors in estimation and prediction, thus rendering the method inefficient. These outliers can be identified graphically, as their location is clearly visible and far from the data center. They are classified according to their position within the regression model. Some appear in random errors or in the dependent variable and are termed "outliers".

Therefore, robust methods for estimating model parameters, such as M-Huber, LAD, LTS, and LMS, have been proposed. This study will focus on the LTS method. Proposed by Rousseeuw (1984) for handling outliers (outliers), the basic idea of this method is to exclude a certain percentage of the largest squared residuals when estimating regression coefficients. Instead of minimizing the sum of all squared residuals as in the OLS method, the LTS method aims to minimize the sum of the smallest h of the squared residuals arranged in ascending order. That is, it selects the observations closest to the model and excludes those that appear outliers or far from it. In this way, a balance is achieved between estimation efficiency and the model's robustness against outliers. h represents the number of observations actually used in the estimation process, and a small sample size of h is often chosen, slightly smaller than the original sample size, to ensure high resistance to outliers. Therefore, this method is characterized by having a high breakdown point (up to 50%). Unfortunately, these methods are affected by another outlier in the independent variables (x) called leverage points (LP). These are of three types: good leverage points, bad leverage points, and high leverage points. These values must be addressed. Yohai (1987) proposed the robust MM method, which is characterized by its ability to resist the influence of outliers (50%) while simultaneously maintaining the high efficiency of the model estimates (95%). Despite its high efficiency, this method suffers from the influence of poor high leverage points (HLPs), which are a source of masking and swamping. Despite the improvements shown by these methods, they still suffer from some limitations regarding the influence of outliers on the one hand, and high leverage points on the other.

Among the robust methods for high-dimensional data is a method that trims the Lasso method, called the robust trimmed least squares method (RPLTS), proposed by Kurnaz et al. (2018) to eliminate outliers. However, this method breaks down when there is at least one poor leverage point in the independent variables (x). Therefore, in this study, we aim to favor this method to be resistant to outliers and leverage points. The favoring process is based on diagnosis. Leverage points are first used using the robust (Mahalanobis Disctines) distance based on (MCD), and then a robust weighting function is used before proceeding to use the (RPLTS) method, because the latter will handle outliers, while our weighting proposal handles Leverage Points, and thus our proposed algorithm (WRPLTS) is able to handle outliers.

2. Robust Penalized Least Trimmed Squares (RPLTS) Method

With the advancement of statistical research, the need arose to develop the LTS method to not only resist outliers but also to achieve variable selection in high-dimensional data models, where the number of variables is large compared to the number of observations. Therefore, the Robust Penalized LTS (RPLTS) method, sometimes called Sparse LTS, was introduced. It is an extension of the LTS method with the addition of a penalty function to the objective criterion, reducing some regression coefficients to zero. This helps identify the most influential variables in the model.

The general formula for the RPLTS method is expressed as follows:

$$arg_{\beta}\, min\left(\sum_{i=1}^{h} r_{(i)}^{2}\,(\beta) + \lambda\, P(\beta)\right) = RPLTS\ \hat{\beta} \tag{1}$$

Where it represents$P(\beta)$ function penalty function, often used as a $norm_{-1}L$ like

$$\|\beta\|_1 = \sum_{j=1}^{p} \left|j^{\beta}\right|$$

as in the Lasso method, while λ is the control parameter that balances the penalty amount with the trimming level. Therefore, the method combines two important features: protection against outliers and the selection of important variables simultaneously. The estimation process in RPLTS is typically performed using a hybrid algorithm that combines pruning steps (C-steps) and penal estimation steps. In each step, the set of observations with the smallest residual h is identified, and a penal regression model (such as Lasso) is then solved on this subsample to update the β coefficients. The process continues until stability is reached or a specified convergence criterion is met. Applied studies, such as Alfons et al. (2013), have shown that this method performs exceptionally well in environments with numerous outliers and variables, maintaining a balance between robustness and statistical accuracy.

The RPLTS method is more flexible than traditional LTS, as the type of penalty function can be modified according to the nature of the data (e.g., Lasso, Ridge, or Elastic Net), and the adjustment parameters can be controlled to determine the degree of pruning and the number of selected variables. However, its main challenges include increased computational complexity when dealing with high-dimensional data, as well as the sensitivity of the results to the value of the penalty coefficient λ and the number of selected observations h.

3. Proposed Method (WRPLTS)

We consider the linear regression model in Equation (1) for some given data (X_i, Y_i) where $X_i \in R^p$ represents the matrix of independent variables and $Y_i \in R$ represents the response vector or dependent variable, and we propose weighting it using the Mahalanobis strong distance. (Kesseku, 2021)

$$RMD_i^2 = (X_i - \mu_{MCD})C_{MCD}^{-1}(X_i - \mu_{MCD})' \tag{2}$$

Where μ_{MCD} is a robust location parameter and C_{MCD} is a robust location and measurement matrix. The researcher used the robust distance of Mahalanobiz to eliminate the influence of high lift points and then utilize it in a weighting function. Accordingly, the researcher used the MCD location and measurement matrix to estimate the location and measurement parameters and then find the weight using the weighting function proposed by (Arslan et al,2012) as follows: $\tau_i = min\left[1, \left(\frac{\chi^2_{(0.95,p+1)}}{\mathrm{RMD}^2(MCD)}\right)\right]$ where the weight takes the smallest value relative to one. If the value of RMD_i^2 is high, then the value of $\left(\frac{\chi^2_{(0.95,p+1)}}{\mathrm{RMD}^2(MCD)}\right)$will be less than one, and conversely, in the opposite case, the value of τ_i will be one.

Based on the above, the Weighted Robust Penalized Least Trimmed Squares (WRPLTS) method, which minimizes the penal objective function, can be formulated as follows:

$$\left(\hat{\beta}_{rob}, \hat{w}\right) = \arg min\left[\sum_{i=1}^{n} w_i\left(Y_{\tau_i} - \beta' X_{\tau_i}\right)^2 + \lambda P(\beta)\right] \tag{3}$$

Where $\beta \in R^{p+1}$ and $\lambda P(\beta)$ is a penalty term that reduces the regression coefficients towards zero. The penalty function can be a lasso, where for every 1 ≤i ≤n, the weights (w_i) represent the variance of errors, such that good observations are given a weight of 1 and outliers a weight of 0. That is, $w \in [0,1]^n$ represents the weights that estimate the effect of outliers for each observation, with $\hat{w}$ being the optimal values for the weights w.

Furthermore, h represents the total number of good observations, which is simply the sum of the values of (1) in $\hat{w}$ such that $\frac{n+1}{2} \leq h \leq n$.

We define the remaining estimates as follows:

$$\hat{\varepsilon}_{\tau i} = Y_{\tau_i} - \hat{\beta}'_{rob} X_{\tau_i} \qquad \text{i=1,2,...,n} \tag{4}$$

We calculate the degrees of freedom ($\hat{v}$) as the number of non-zero regression coefficients when choosing the optimal coefficient. More specifically, $\hat{v}$ takes values from 0 to p, and then we estimate the error measure σ using:

$$\hat{\sigma}^2_{\text{rob}} = \frac{c^2_{\text{n,h,}\hat{v}}}{\text{h} - \hat{v}} \sum_{\text{i=1}}^{\text{n}} \hat{w}_{\text{i}} \, \hat{\varepsilon}_{\tau\text{i}}{}^2 \tag{5}$$

According to (Croux and Haesbroeck, 1999), $c^2_{\text{n,h,}\hat{v}}$ in equation (5) is given by:

$$c^2_{\text{n,h,}\hat{v}} = \frac{1}{\int_{-\xi}^{\xi} x^2 \, \phi(x) \, dx} = \frac{1}{P(\chi^2_3 \leq \chi^2_{1,1-\alpha})} \tag{6}$$

Where:

ϕ(.): is the density of the standard normal distribution.

$\phi_0^{-1}(.)$: is the cumulative function of the standard normal distribution.

Also, $\xi = \phi_0^{-1}\left(1 - \frac{\alpha}{2}\right)$ where $\alpha = 1 - \frac{h-\hat{v}}{n-\hat{v}}$. This simplifies to:

$\xi = \phi_0^{-1}\left(1 - \frac{n-h}{2(n-\hat{v})}\right)$ by substituting the value of α.

The estimator $c^2_{\text{n,h,}\hat{v}}$ was chosen to ensure that the estimation of $\hat{\sigma}^2_{\text{rob}}$ after trimming with weights was not biased. In simpler terms, we rewrite equation (5) using c^2as follows:

$$\hat{\sigma}^2_{\text{rob}}(h - \hat{v}) = c^2 \sum_{\text{i=1}}^{\text{n}} \hat{w}_{\text{i}} \, \hat{\varepsilon}_{\tau\text{i}}{}^2 \tag{7}$$

$$\sum_{i=1}^{n} w_i = h \tag{8}$$

Under our classic setup:

$$\log \text{L}(\theta / y_\tau, x_\tau) = \sum_{i=1}^{n} log \, \phi(y_{\tau i}, x_{\tau i} / \theta)$$

Where $\theta = (\beta, \sigma^2)$

$$= \sum_{i=1}^{n} log \left[(2\pi\sigma^2)^{-\frac{1}{2}} . \exp\left(-\frac{1}{2\sigma^2} (y_{\tau i} - \beta' x_{\tau i})^2 \right) \right]$$

$$= \sum_{i=1}^{n} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_{\tau i} - \beta' x_{\tau i})^2 \right]$$

$$= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_{\tau i} - \beta' x_{\tau i})^2 \quad (9)$$

For the trimmed logarithmic probability function (simple version), we have:

$$\log L_{tr}(\theta/y_\tau, x_\tau) = \sum_{i=1}^{n} w_i \, log \, \phi(y_{\tau i}, x_{\tau i}/\theta)$$

$$= \sum_{i=1}^{n} w_i \, log \left[(2\pi\sigma^2)^{-\frac{1}{2}} . \exp\left(-\frac{1}{2\sigma^2} (y_{\tau i} - \beta' x_{\tau i})^2 \right) \right]$$

$$= \sum_{i=1}^{n} w_i \left[-\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_{\tau i} - \beta' x_{\tau i})^2 \right]$$

$$= -\frac{\sum w_i}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} w_i \, (y_{\tau i} - \beta' x_{\tau i})^2$$

Since $\sum w_i = h$ according to equation (8), and (h) represents the total number of good observations, then:

$$log \, L_{tr}(\theta/y_\tau, x_\tau) = -\frac{h}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} w_i \, (y_{\tau i} - \beta' x_{\tau i})^2 \quad (10)$$

Notes on equation (11): The researcher wanted to approach an unbiased estimate of $\hat{\sigma}^2_{\text{rob}}$, which has the following form:

$$\hat{\sigma}^2_{\text{rob}}(h - \hat{v}) = c^2 \sum_{\text{i=1}}^{\text{n}} \hat{w}_\text{i} \left(y_{\tau i} - \hat{\beta}'_{rob} \, x_{\tau i} \right)^2$$

$$= c^2 \sum_{\text{i=1}}^{\text{n}} \hat{w}_\text{i} \, \hat{\varepsilon}_{\tau\text{i}}{}^2$$

Knowing the formula (7), he substituted the value of 1 with the value of c^2 in the formula (6), which represents the value of c^2.

$$c^2_{\text{n,h,}\hat{\text{v}}} = \frac{1}{P(\chi^2_3 \leq \chi^2_{1,1-\propto})}$$

Since $P(\chi^2_3 \leq \chi^2_{1,1-\propto})$, it is a small probability value, therefore $c^2 = 1$.

Substituting c^2 into formula (11) is necessary because the researcher stated that the estimator of ($c^2_{\text{n,h,}\hat{\text{v}}}$) is chosen so that ($\hat{\sigma}^2_{\text{rob}}$)is unbiased. Therefore, (c^2) is substituted for $\hat{\beta}_{\text{rob}}$and $\hat{\varepsilon}_i$ in ($\hat{\sigma}^2_{\text{rob}}$), which depends on all observations in the sample (n). The researcher then multiplied the equation by($\frac{\mathbf{n}}{\mathbf{h}}$), and from there arrived at the final formula.

Similarly, for the trimmed logarithmic probability function (the approximately unbiased version), we have:

$$\log L_{tr}\left(\theta/y_\tau, x_\tau\right) = -\left[\frac{h}{2}\, log(2\pi\sigma^2) - \frac{c^2}{2\sigma^2}\sum_{i=1}^{n} w_i\,(y_{\tau i} - \beta' x_{\tau i})^2\right] \quad (11)$$

$$= \frac{n}{h}\left[-\frac{h}{2}\, log(2\pi\hat{\sigma}^2_{rob})\right] - \frac{n}{h}\left[\frac{c^2}{2\hat{\sigma}^2_{rob}}\sum_{i=1}^{n}\widehat{w}_i\left(y_{\tau i} - \hat{\beta}'_{rob}x_{\tau i}\right)^2\right]$$

$$= -\frac{n}{2}\, log(2\pi\hat{\sigma}^2_{rob}) - \frac{n}{h}\left[\frac{\hat{\sigma}^2_{\mathrm{rob}}(h-\hat{v})}{2\hat{\sigma}^2_{rob}}\right]$$

$$= -\frac{n}{2}\, log(2\pi\hat{\sigma}^2_{rob}) - \frac{n}{h} * \frac{h}{2} + \frac{n\hat{v}}{2h}$$

$$= -\frac{n}{2}\, log(2\pi\hat{\sigma}^2_{rob}) - \frac{n}{2} + \frac{n\hat{v}}{2}$$

$$= -\frac{n}{2}\left[\, log(2\pi\hat{\sigma}^2_{rob}) + 1 - \frac{\hat{v}}{h}\right] \quad (12)$$

Equation (11) assumes that the negative outside the bracket is multiplied by the first term or by both terms. Therefore, the algorithm for the proposed (WRPLTS) method is as follows:

3.1 Computation Algorithm

WRPLTS algorithm

1. Returning to Model No. (1), where $X_i \in R^p$ the matrix represents the independent variables and $Y_i \in R$the response vector or the dependent variable represents, the weight function is extracted$\tau_i$based on the calculated Mahnalopez distance accordingRMD^2 .to Equation No (2)

as follows : $\tau_i = min\left[1, \left(\frac{\chi^2_{(0.95,p+1)}}{\mathrm{RMD}^2(\mathrm{MCD})}\right)\right]$

It can be calculated$\left(X_{\tau_i}, Y_{\tau_i}\right)$as follows:

$$X_{\tau_i} = \tau_i \,.\, X_i$$

$$Y_{\tau_i} = \tau_i \,.\, Y_i$$

2. Inputs: Assume that the regression data is $\mathrm{D}_{\tau_i} = \{\left(X_{\tau_i}, Y_{\tau_i}\right) \in R^p \times R\}_{i=1}^{n}$

3. Given: Assume that the initial values are $\hat{\beta}^{(0)}, \lambda > 0$

4. Make the counter $k \leftarrow 1$

5.Calculating the regression residuals based on the initial value of the regression parameter

$$\hat{\varepsilon}_{\tau i} = Y_{\tau i} - \left(\hat{\beta}^{(k-1)}\right)^T X_{\tau i} \quad , i = 1,2,\ldots.,n$$

6. Arrange the residuals in ascending order, then choose$(h)$from these residuals and ignore the residuals that are at the top of the order, where$\frac{n+1}{2} \le h \le n$ then make$\widehat{w}_i^{(k)} = 1$if it was $i \in \{i_1, \ldots., i_h\}$ Otherwise, make it$\widehat{w}_i^{(k)} = 0$

7. Estimate regression parameters $\hat{\beta}$ using the traditional$\{X_{\tau ij}\,, Y_{\tau ij}\}_{j=1}^{h}$Lasso method for data.

8. Make $\hat{\beta}^{(k)} = \hat{\beta}$

9. If it was $\left\|\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}\right\|_\infty < \partial$ Since∂it is a very small value, for example (0.0001 or 0.001), go to step (9) and vice .versa increase the counter by one integer$k \leftarrow k+1$ Go to step 4

10. Output$(\hat{\beta}^{(k)}, \hat{w}^{(k)})$

4. Simulation Study

This section examines the simulation to verify the robustness and efficiency of the proposed method and compare it with the previous method. This is achieved through several statistical criteria that will reveal the method's accuracy in identifying null and non-zero variables for a high-dimensional regression model under the same data conditions.

The first step in the simulation must be to clarify the data generation process and how it is contaminated by outliers and lifting points. Assume that:

n: the sample size, where n = {400, 600, 800} at each time and independently

P: the number of independent variables, where p = {50, 60, 80, 100, 120}, where these variables will be associated with each of the n sample sizes

α∈(0,1): the ratio of outliers and lever points, where α = {0.10}

ρ: the correlation coefficient in the covariance matrix, and its value will be ρ = 0.5

$ng = \lfloor 0.2p \rfloor$:Number of non-zero parameters Which will be generated from a uniform distribution $.\beta_{1:ng} \sim Unif(2,10)$, Zero transactions $n\lambda = \lfloor 0.8\,p \rfloor$will.take the value zero $\beta_{ng+1:p} = 0$ Let it be$o = \lceil \alpha . n \rceil$ the number of outliers in the sample $n$and the matrix of independent variables generated from a multivariate normal distribution,

$$X_i \sim \mathrm{N}_P(0, \Sigma), \quad i = 1,2, \dots, n$$

Where as $$X = \begin{bmatrix} X_1' \\ \vdots \\ X_n' \end{bmatrix} \in \mathbb{R}^{n \times p}$$

$$\Sigma \in \mathbb{R}^{p \times p}$$

$$\Sigma_{i,j} = 0.5^{|i-j|}, \qquad i, j = 1,2, \dots, p$$

Then $$y = [1_n \quad X_{n \times p}] \begin{bmatrix} 1 \\ \beta_1 \\ \vdots \\ \beta_{ng} \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{p \times 1} + \varepsilon_{n \times 1}$$

where ε, the random error vector, is mixed-distributed. Some observations have a t-distribution with 5 degrees of freedom, while the remaining random errors have a normal distribution with a mean of zero and a constant variance of (2).

ε=o~t(5)+(n-o) ~N(0,2)

As for the lifting points, the independent variables X_1, X_2, X_3 were contaminated with them, and the contamination did not occur in any row. Rather, o was chosen from the rows randomly in each of the 200 simulation cycles due to the large size of the data and the slow execution in the available computers.

The following table shows the number of outliers in all samples at each α contamination level, as follows:

Pollution rate(α)	Sample size (n)	Outliers (o)
0.05	400	20
	600	30
	800	40
0.10	400	40
	600	60
	800	80
0.15	400	60
	600	90
	800	120

To select the best option, the Confusion matrix was chosen from our simulation study. 0.20 of the variables are non-zero parameter variables, as follows:

p Values of the variables	TP Values	TN Values
50	10	40
60	12	48
80	16	64
100	20	80
120	24	96

Therefore, TP means that when the generated sample size is 50, the number of non-zero variables must be (10) variables. FP means that out of these (10) variables, there are zero variables that the method used considered to be non-zero variables. TN means that there are (40) zero variables that the method used correctly identified from the total p = 50. FN means that out of these (40) zero variables, the method used incorrectly identified non-zero variables as zero. Based on the measurement criteria used to compare the performance of the two estimation methods, such as accuracy, sensitivity, classification, and mean squared error, the accuracy metric can be calculated as the ratio of the total number of correct diagnoses (both positive and negative) to the total number of cases. Ideally, when n=400 and p=50, TP=10 and TN=40 should be FP=0 and FN=0. Substituting this value into the accuracy formula, a result of 1 indicates perfect prediction accuracy for both zero and non-zero parameters. Sensitivity represents the ratio of correct positive predictions to the total number of correct positive and incorrect negative predictions. Therefore, it measures the model's ability to detect significant real parameters. Sensitivity is 100%. For example, when p=60, TP=12 should be FP=0 and FN=0.

The specificity metric represents the ratio of correct negative predictions to the total number of correct negative and incorrect positive predictions. Therefore, it is considered a measure of the model's ability to detect significant real parameters. The model focuses on detecting insignificant true variables, with optimal identification. For example, if p = 60, then TN =48, and FP = 0. The MSE (Mean Squared Random Error) metric is a measure of estimator quality. It measures the average squared difference between estimated and actual values and is one of the

most important metrics used to compare the performance of methods. A lower error rate generally indicates a more favorable estimating method. Based on the above, the following can be concluded:

5. Simulation Results

Table (1) shows the accuracy, sensitivity, specificity, and mean random error (MSE) of the two compared methods when n = 400, 600, and 800, and the number of independent variables p = 50, 60, 80, 100, and 120 for each sample size, and the contamination ratio $\propto$ = 0.1 for both the dependent and independent variables, and the breaking point BDP = 0.25 was chosen.

Table (1) shows the accuracy, sensitivity and classification results for the estimation methods when the percentage of outliers is 0.1 and the breaking point is 0.25 for different sample sizes.

BDP=0.25 and alpha=0.1						
n	p	Method	Accuracy	Sensitivity	Specificity	MSE
400	50	RPLTS	0.330	0.970	0.170	11.342
		Weighted RPLTS	0.966	1.000	0.958	0.017
	60	RPLTS	0.267	0.925	0.102	9.091
		Weighted RPLTS	0.955	1.000	0.944	0.023
	80	RPLTS	0.371	0.806	0.263	8.245
		Weighted RPLTS	0.960	1.000	0.950	0.021
	100	RPLTS	0.471	0.780	0.394	7.225
		Weighted RPLTS	0.868	1.000	0.835	0.019
	120	RPLTS	0.471	0.780	0.394	7.225
		Weighted RPLTS	0.868	1.000	0.835	0.019
600	50	RPLTS	0.484	1.000	0.355	8.220
		Weighted RPLTS	0.888	1.000	0.860	0.011
	60	RPLTS	0.485	1.000	0.356	7.832
		Weighted RPLTS	0.952	1.000	0.940	0.015
	80	RPLTS	0.351	0.913	0.211	9.266
		Weighted RPLTS	0.913	1.000	0.891	0.013
	100	RPLTS	0.362	0.850	0.240	8.063
		Weighted RPLTS	0.909	1.000	0.886	0.013

	120	RPLTS	0.385	0.825	0.275	7.629
		Weighted RPLTS	0.845	1.000	0.806	0.010
800	**50**	RPLTS	0.578	1.000	0.473	8.237
		Weighted RPLTS	0.856	1.000	0.820	0.008
	60	RPLTS	0.503	1.000	0.379	8.603
		Weighted RPLTS	0.827	1.000	0.783	0.007
	80	RPLTS	0.491	1.000	0.364	8.246
		Weighted RPLTS	0.879	1.000	0.848	0.010
	100	RPLTS	0.322	0.965	0.161	10.564
		Weighted RPLTS	0.915	1.000	0.894	0.010
	120	RPLTS	0.323	0.892	0.180	9.018
		Weighted RPLTS	0.813	1.000	0.767	0.007

First: When the sample size is n=400 with ∝=0.1 and BDP=0.25, at this sample size and with the contamination ratio increasing to (0.10), which means there are 40 outliers out of 400 samples, and since BDP=0.25, then h=240 is a good observation. We note that at p=50, RPLTS showed low accuracy (0.33), indicating a significant increase in FP and FN values, which affected the model's accuracy in predicting zero and non-zero parameters. In contrast, the sensitivity is very high (0.97) because (TP=10 and FN=0.309), but the specificity is weak (0.17) because TN=40, FP=195.29, and MSE is high (11.342), reflecting a clear weakness in performance. When the number of variables increases to p=60, the performance of RPLTS decreases further, with accuracy reaching (0.267) and specificity (0.102). Conversely, sensitivity remained high (0.925), and the MSE remained high (9.091). At p=80, this method showed a slight improvement in accuracy (0.371) and specificity (0.263), but sensitivity decreased (0.806) with a high MSE (8.245). At p=100 and p=120, the RPLTS remained at a modest level (accuracy 0.471, specificity 0.394, sensitivity = 0.780, MSE = 7.225). At this sample size, we observe that the RPLTS failed to achieve a balance between the measures, as it remained weak in accuracy and high in MSE. In contrast, the Weighted RPLTS achieved a significant advantage. At p=50, it showed high accuracy (0.966), full sensitivity (100%) indicating that it included all significant real variables in the model and did not exclude any significant variables, high specificity (0.958) demonstrating this method's high ability to exclude irrelevant variables from the model, and a very low MSE. (0.017). However, at p=60, this method maintained its excellent performance (accuracy 0.955, specificity 0.944, sensitivity 100%, MSE=0.023). When the number of variables was increased to p=80, Weighted RPLTS showed near-perfect performance (accuracy 0.960, specificity 0.950, full sensitivity, MSE=0.021). When the number of variables was further increased to p=100 and p=120, despite a relatively slight decrease in accuracy (0.868) and specificity (0.835), it remained significantly superior with full sensitivity and a low MSE (0.019). Therefore, we observe that at a sample size of 400, Weighted RPLTS showed stability and superiority in all indicators compared to the poor performance of RPLTS.

Secondly: When the sample size is n=600, we observe that by increasing the sample size to 600 at p=50 and p=60, RPLTS showed a relative increase in Its performance remains modest (accuracy ≈ 0.48, specificity ≈ 0.35, sensitivity 100%), with the MSE decreasing to 8.2 with 50 variables and to 7.8 with 60 variables, though still relatively high. At 100 variables and p=80, the RPLTS continued to be weak, exhibiting instability (low accuracy 0.35–0.36, high

sensitivity 0.85–0.91, low specificity 0.21–0.24, MSE between 8 and 9). However, when the number of variables increased to p=120, the RPLTS showed a slight improvement, though still relatively weak. Accuracy increased by 0.385, specificity by 0.275, while maintaining high sensitivity (0.825) and a slight decrease in the mean error (MSE = 7.629), though still relatively high and significantly impacting the method's performance. In contrast, the Weighted RPLTS method, with the same sample size and number of variables, proved its worth in all cases. In the cases where p=50 and p=60, the Weighted RPLTS achieved excellent results (accuracy 0.888–0.952, high specificity 0.860–0.940, full sensitivity, very low MSE 0.011–0.015). At p=80 and p=100, the Weighted RPLTS remained superior, exhibiting high accuracy 0.909–0.913, high specificity 0.886–0.891, full sensitivity, and a low MSE of 0.013, thus demonstrating its robust performance and increasing its importance in estimation. Even with an increase in the number of variables to 120, the Weighted RPLTS maintained its superiority (accuracy 0.845, specificity 0.806, sensitivity 100%, MSE = 0.010). Based on our review of the results, our proposed method demonstrated stability and excellent performance across all variables, making it the optimal choice at a contamination level of 0.10 and a breaking point of 0.25.

Third: When the sample size At n=800, RPLTS showed relative improvement compared to smaller sample sizes, with accuracy rising to (0.578) and specificity to (0.473) with full sensitivity (1.0), but the MSE remained high (8.237). Increasing the number of variables at p=60–80, RPLTS performance remained modest (low accuracy ≈0.49–0.50, low specificity ≈0.36–0.38, high MSE ≈8.2–8.6). Further increasing the number of variables to p=100–120, RPLTS performance declined significantly (low accuracy ~0.322–0.323, very poor specificity 0.161–0.180, high sensitivity 0.89–0.97 but unstable, very high MSE 9.018–10.564). It should be noted that at this type of sample size, RPLTS sometimes shows slight improvements in accuracy and specificity for some small values. For p, the value decreases, but it weakens as the number of variables increases.

While Weighted RPLTS consistently performed strongly and superiorly across all cases, in the case of 50 variables, this method achieved superiority thanks to an accuracy of 0.856, a specificity of 0.820, and a very low MSE (0.008). When increasing the number of variables p=60–80, Weighted RPLTS maintained its excellent stability (accuracy 0.827–0.879, high specificity, full sensitivity, small MSE ≤0.010). Even with a number of variables p=100–120, our proposed method remained significantly superior (accuracy 0.813–0.915, specificity 0.767–0.894, full sensitivity, very small MSE ≤0.010). It proved to be the most efficient and stable method across all sample sizes and number of variables, combining high accuracy, full sensitivity (100%), very high specificity, and a very low MSE, thus demonstrating its powerful performance. This method yields exceptionally high results, making it the most suitable approach for ensuring the selection of variables in multiple linear regression.

Table (2) shows the accuracy, sensitivity, specificity and mean squared random error (MSE) of the two comparative methods when n=400,600,800 and the number of independent variables p=50,60,80,100,120 for each sample size and the contamination ratio $\propto$=0.10 in both the dependent variable and the independent variables and the breaking point BDP=0.1 was chosen.

Table (2) shows the accuracy, sensitivity and classification results for the estimation methods when the percentage of outliers is 0.10 and the breaking point is 0.1 for different sample sizes.

BDP=0.1 and alpha=0.10						
n	**p**	**Method**	**Accuracy**	**Sensitivity**	**Specificity**	**MSE**
400	50	RPLTS	0.400	1.000	0.250	11.250
		Weighted RPLTS	0.688	1.000	0.610	0.022
	60	RPLTS	0.297	0.992	0.123	12.090
		Weighted RPLTS	0.792	1.000	0.740	0.020

	80	RPLTS	0.265	0.944	0.095	14.581
		Weighted RPLTS	0.803	1.000	0.753	0.019
	100	RPLTS	0.379	0.765	0.283	8.444
		Weighted RPLTS	0.827	1.000	0.784	0.017
	120	RPLTS	0.478	0.725	0.416	8.048
		Weighted RPLTS	0.848	1.000	0.809	0.021
600	50	RPLTS	0.500	1.000	0.375	8.237
		Weighted RPLTS	0.808	1.000	0.760	0.015
	60	RPLTS	0.490	1.000	0.363	8.209
		Weighted RPLTS	0.835	1.000	0.794	0.016
	80	RPLTS	0.369	0.975	0.217	10.103
		Weighted RPLTS	0.841	1.000	0.802	0.015
	100	RPLTS	0.431	0.955	0.300	9.622
		Weighted RPLTS	0.846	1.000	0.808	0.012
	120	RPLTS	0.344	0.875	0.211	9.149
		Weighted RPLTS	0.870	1.000	0.838	0.014
800	50	RPLTS	0.552	1.000	0.440	8.250
		Weighted RPLTS	0.858	1.000	0.823	0.013
	60	RPLTS	0.500	0.992	0.377	8.621
		Weighted RPLTS	0.850	1.000	0.813	0.011
	80	RPLTS	0.468	1.000	0.334	8.624
		Weighted RPLTS	0.879	1.000	0.848	0.012
	100	RPLTS	0.435	1.000	0.294	10.362
		Weighted RPLTS	0.878	1.000	0.848	0.009
	120	RPLTS	0.334	0.979	0.173	11.375
		Weighted RPLTS	0.873	1.000	0.841	0.011

As contamination increased to 0.10 at a low BDP (0.10), the weakness of the RPLTS became more apparent, with its accuracy ranging from 0.26 to 0.55 and its specificity decreasing significantly (0.09 to 0.44), while its sensitivity

remained high (but unstable). MSE values increased further than in previous tables, reaching as high as 14.5 at some values. This reflects the fragility of the method with high contamination and low BDP. In contrast, the Weighted RPLTS continued its excellent performance across all sample sizes and variable counts, with accuracy ranging from 0.68 to 0.87, its specificity remaining high (0.61 to 0.84), and its sensitivity remaining high while its MSE remained low (0.01 to 0.02). This reinforces its superiority even in the most challenging environments. From the results presented in Tables 1 and 2, we conclude that while the RPLTS method maintained high sensitivity in most cases, it suffered from significant weaknesses in accuracy and specificity, along with a marked increase in MSE values. This reflects its instability and reduced efficiency, particularly at higher levels of contamination or lower breaking point. In contrast, the Weighted RPLTS method demonstrated consistent superiority across all sample sizes and number of variables. It combined high accuracy, full sensitivity, high specificity, and very low MSE values, making it the optimal, most stable, and most reliable option for estimating multiple linear regression coefficients in contaminated environments.

6. Practical Application

Real Data: Adjusted Home Price Data for the Boston Suburbs, USA

This data has been used in numerous research papers. It was first presented by Harrison, D. and Rubinfeld, D.L. (1978), followed by Belsley, D.A., Kuh, E. and Welsch, R.E. (1980), who studied it in terms of identifying anomalies. The dataset originally contained 506 observations and 14 variables. The Boston dataset comprised 506 rows and 14 columns, 13 of which were explanatory variables:

1. Crime: Per capita crime rate by city.

2. Zn: Percentage of residential land plots exceeding 2,322 square meters.

3. Indus: Percentage of industrial land plots designated for commercial and non-commercial use by city.

4. chas: A dummy variable for the Charles River (1 if the area is on the river's edge; 0 otherwise).

5. nox: Nitrogen oxide concentration (parts per 10 million).

6. rm: Average number of rooms per dwelling.

7. age: Percentage of homeowners built before 1940.

8. dis: Weighted average distances to five employment centers in Boston.

9. rad: Radial Highway Accessibility Index.

10. tax: Full-value property tax rate per $10,000.

11. ptratio: Student-to-teacher ratio by city.

12. black: 1000 $(Bk - 0.63)^2$, where Bk is the percentage of Black people by city.

13. lstat: Percentage of the lowest socioeconomic class.

The dependent variable, medv, is the average value of homeowners in thousands. To ensure the data aligns with the objectives of this thesis, the researcher added 37 artificial variables from a standard multivariate distribution, bringing the total number of explanatory variables to 50.

Figure (1) clearly shows that the medv distribution is one of the heavy-tail distributions, indicating the presence of outliers.

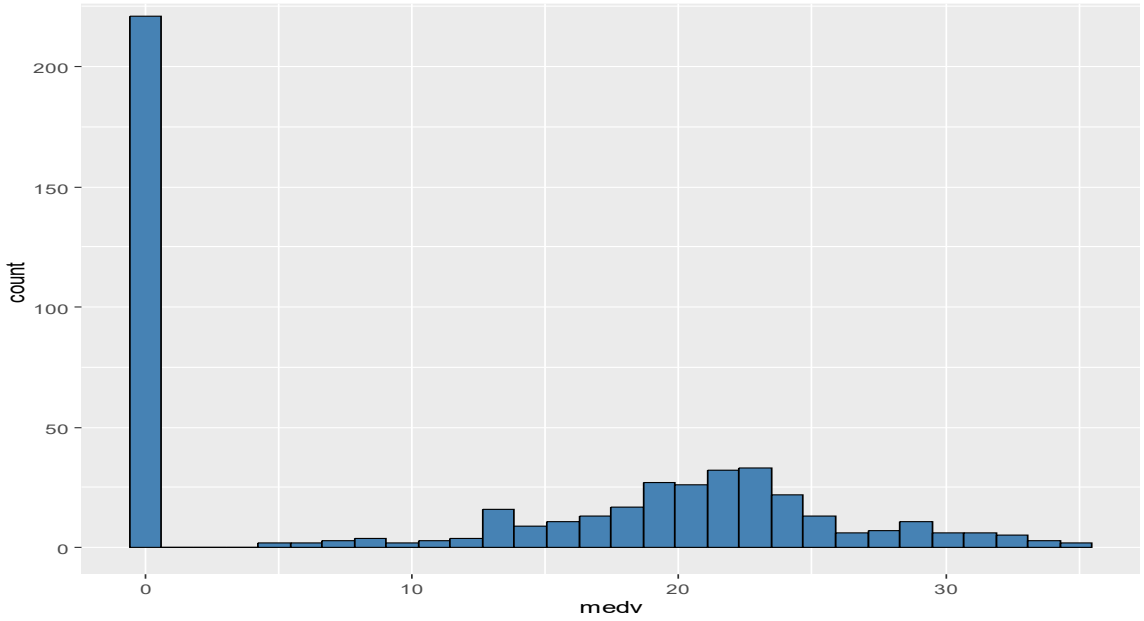

Figure (1) Histogram of the distribution of the dependent variable.

Figure (2) confirms the presence of outliers from the QQ-Plot for the residuals, making it impossible to rely on traditional estimation methods.

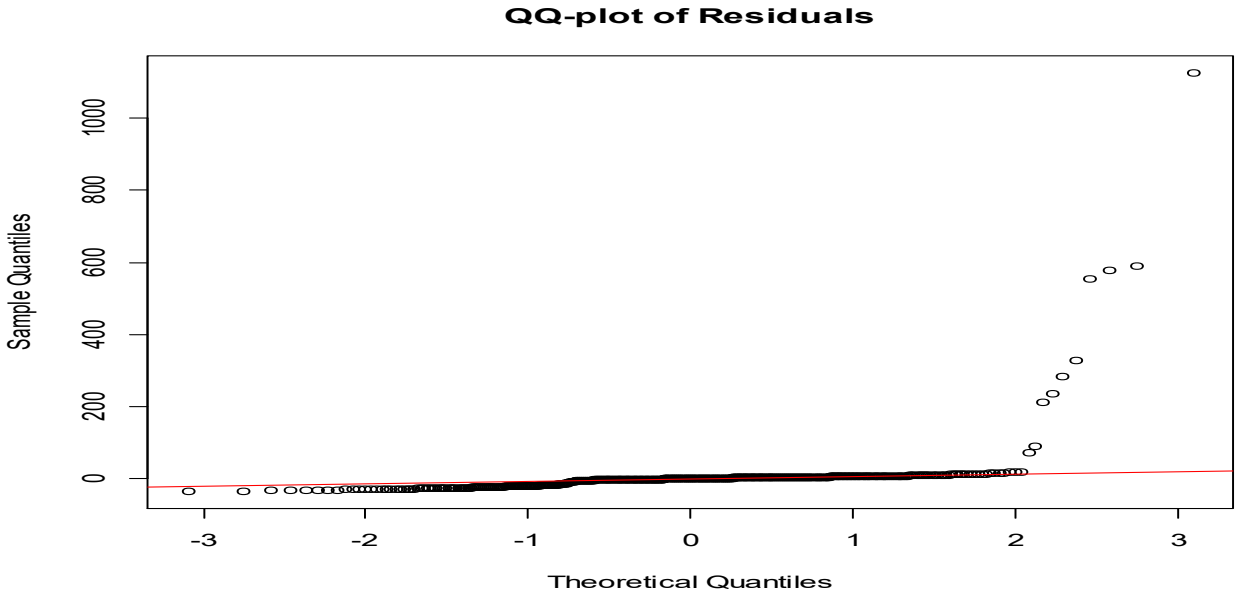

Figure (2) QQ-Plot for residuals, a regression model of house price data.

We observe that certain data points on the positive side significantly affected the direction of the red line, which should have been normal and slanted. This indicates that these anomalous points are influential observations. This is illustrated in Figure (3), which shows the presence of influential observations.

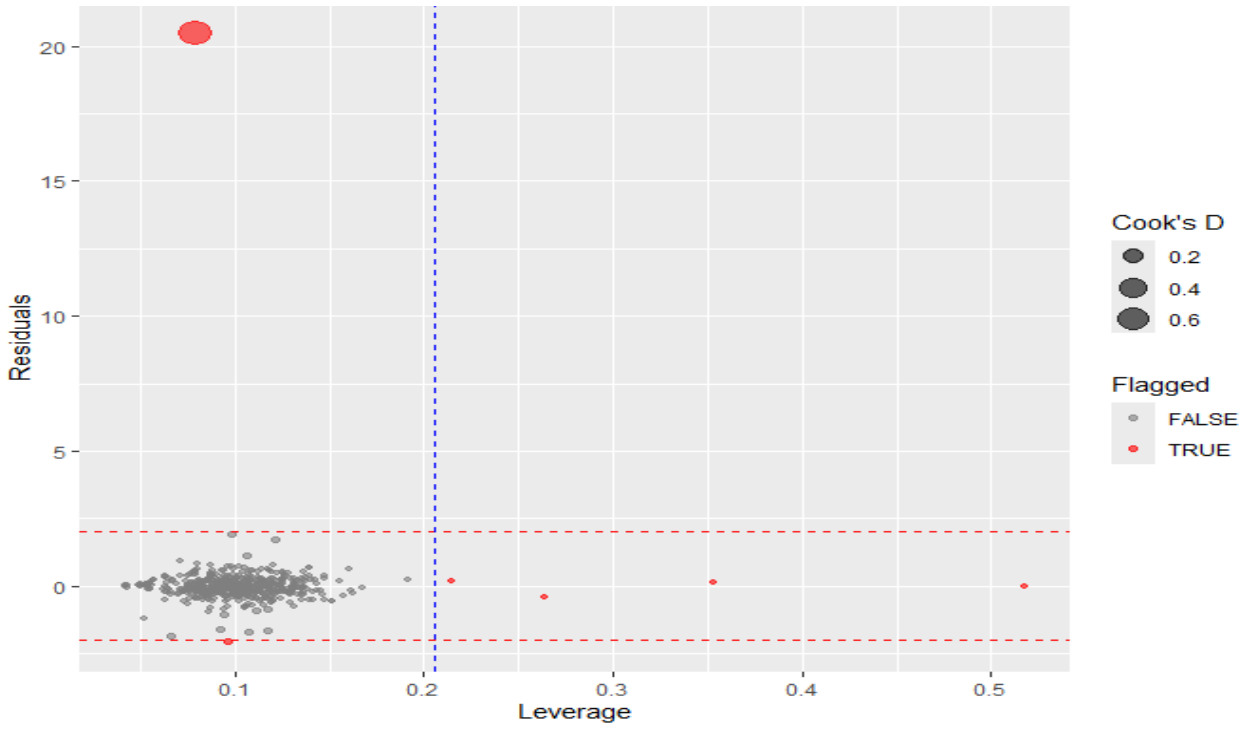

Figure (3) Outliers, Lifting Points and Influencing Observations in Home Price Data in the State of Boston, USA.

At various points in the data, some of these observations were leverage points and outliers, but the most serious were near the data located on the dashed red line of the third quadrant.

Table (3) shows the selected and excluded variables for each method and the trimmed RMSE when bdp = {0.10, 0.25}.

Method	bdp	Variable Selected	Variable Excluded	Trimmed RMSE
RLTS	0.10	Crim,chas,rm,age,tax, ptration,black, 1stat	zn,nox,dis, indus, rad	53.65
WRLTS		Crim,zn,nox, dis,tax,rm,ptration,black ,1stat	age,chas,indus, rad	13.24
RLTS	0.25	Crim,zn, tax,rm,ptration,black,1stat	age,chas,indus,dis, nox ,rad	30.14
WRLTS		Crim,zn,chas,rm,age,dis,tax,ptration, black,1stat	Indus,rad,nox	12.07

Table (3) shows the results of each method's selection of important variables and exclusion of unimportant variables when bdp = 0.10, 0.25. We observe that when bdp = 0.10, each method selected (304) observations for

estimation purposes. However, there is a difference in the selection of some variables between the two methods. Both methods selected (Crim, rm, tax, ptration, black, 1stat) as important variables because they lacked poor lever points. However, the RLTS method added the variables chas and age, considering them important, and excluded the remaining variables such as zn, nox, dis, indus, and rad. Conversely, the WRLTS method added the variables zn, nox, and dis to the important variables shared with the previous method and excluded the variables age, chas, indus, and rad. The criterion for choosing between the two methods, namely the lower Trimmed RMSE value, showed that the lower value favored the model chosen by the WRLTS method, reaching 13.24, a significant difference from the corresponding value of the RLTS method, which reached 53.65. When the bdp value was raised to 0.25, each method selected 380 observations for model fit and variable selection.

The RLTS method chose Crim, rm, age, tax, black, 1stat, black, 1stat, `black`, `1stat`, `crim`, `zn`, `tax`, `rm`, `ptration`, which shared its previous selection when h=304, using only the variables Crim, rm, age, tax, ptration, black, and 1stat, excluding both chas and age and replacing them with the variable zn. This choice resulted in a decrease in the Trimmed RMSE value to 30.14, which is logical for two reasons. First, the increase in the trimmed sample size from 304 to 380 improved the estimation accuracy. Second, the influence of the lever points in variable selection was reduced by the larger trimmed sample size. In contrast, the WRLTS method selected 10 significant variables, as shown in Table 3, and excluded Indus, rad, and nox, achieving a lower Trimmed RMSE value of 12.07, which is the optimal choice.

The WRLTS method outperformed the RLTS method regardless of whether the bdp = 0.10 or 0.25. Both methods consistently excluded the variables rad and indus. It is also noted that NOX is a weak variable that the WRLTS method failed to detect when bdp = 0.10. This selection error is a Type I error and does not significantly affect the model. However, the change in the Trimmed RMSE value when the sample size increased to 380, resulting in a value of 12.07, places all variables (age and chas) in a state of uncertainty. This means that one of them might be an unimportant variable with a parameter that is perhaps zero or close to zero, while the other might be an influential variable. The first scenario has little impact on the model, but the second suggests that the WRLTS method, when bdb = 0.10, encountered a Type II error due to failing to select one of the important variables. Naturally, the reason for this is likely due to the masking and swamping problem, and the presence of high, poor-quality lift points that have become influential observations, as seen in Figure 3 for the Cook's distance, which shows a very influential observation in the estimation. Increasing the sample size to 380 resolved this issue because it eliminated the high lifting point and, specifically, the masking effect by removing unidentified outliers. The absence of the variables age, chas, and dis in the RLTS selections when bdp = 0.25, while their presence in WRLTS, confirms a Type II error in the RLTS method. Furthermore, the exclusion of dis from the WRLTS selections when bdp = 0.25 and 0.10 demonstrates that dis is a crucial variable that this method failed to select twice. This means that the weighted average distance to five employment centers in Boston significantly influences the average value of homeownership in Boston.

From the above results of the statistical analysis, the crime rate per capita (crim), the percentage of land designated for housing with an area exceeding 2322 square meters (zn), the average number of rooms per dwelling (rm), the concentration of nitrogen oxides measured in units (parts per 10 million) (nox), the percentage of housing units built before 1940 (age), the weighted average distances to five major business centers in Boston (dis), the rad index expressing the ease of access to radial highways (rad), the student-teacher ratio (ptratio), the Bk ratio representing the percentage of the Black population calculated using the formula ($1000\ (Bk\ -\ 0.63)^2$), the property tax rate for the total value of properties per $10,000 (tax), the imaginary variable of the Charles River (chas), the percentage of land designated for industrial, commercial, and non-commercial activities within each city (Indus), and the percentage of low-income residents (lstat) are all influential on the average value of homes owned in thousands.

7. Conclusions

1. The proposed Weighted RPLTS method proved its efficiency in processing contaminated data affected by anomalies and high lift points (HLP). This enables it to provide accurate and reliable estimates, as it outperformed traditional methods in handling these cases.
2. Using the algorithm in the developed method, the weight function was extracted based on the robust Mehnalopez distance (RMD^2). This significantly enhanced detection capabilities and reduced the impact of masking and swamping, leading to more accurate results.
3. The developed Weighted RPLTS method demonstrated remarkable success in detecting significant variables and excluding irrelevant ones, while minimizing the mean squared error (MSE). This reflects its high accuracy in parameter estimation.

References

[1] Alfons, A., Croux, C., & Gelper, S. (2013). *Sparse Least Trimmed Squares Regression for Analyzing High-Dimensional Large Data Sets*. The Annals of Applied Statistics, 7(1), 226–248. doi:10.1214/12-AOAS575.
[2] Arslan, O. (2012). "Weighted LAD-LASSO Method for Robust Parameter Estimation and Variable Selection in Regression". Computational Statistics & Data Analysis, 56(6), 1952-1965.
[3] Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons.
[4] Croux, C. & Haesbroeck, G. (1999). *Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator*. Journal of Multivariate Analysis, 71(2), 161–190.
[5] Harrison, D. & Rubinfeld, D. L. (1978). Hedonic Housing Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, 5(1), 81–102.
[6] Kesseku, R. (2021). *Robust Variable Selection in Multiple Linear Regression Via Penalized Least Trimmed Squares* (Master's thesis, The University of Texas at El Paso).
[7] Kurnaz, F. S., Hoffmann, I., & Filzmoser, P. (2018). *Robust and sparse estimation methods for high dimensional linear and logistic regression.* Chemometrics and Intelligent Laboratory Systems, 172, 211-222.
[8] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. (2012) *Introduction to Linear Regression Analysis.* 5th ed., John Wiley & Sons, Inc., Hoboken, New Jersey.
[9] Rousseeuw, P. J. (1984). Least trimmed squares estimator. *Technometrics*, 26(3), 393–403.
[10] Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics, 15(2), 642-656.