Tikrit Journal of Administrative and Economic Sciences, Vol. 21, No. 71, Part (2): 560-571
Doi: www.doi.org/10.25130/tjaes.21.71.2.28

Tikrit Journal of Administrative and Economics Sciences
مجلة تكريت للعلوم الإدارية والاقتصادية
EISSN: 3006-9149
PISSN: 1813-1719

Estimating the Optimal Tune Parameter for the Binary Weighted Function in a Robust Regression Model

Saman Hussein Mahmood*, Dlshad Mahmood Saleh,
Alan Ghafur Rahim, Taha Hussein Ali
College of Administration and Economics/Salahaddin University- Erbil

Keywords:
Robust Regression Model, Outliers, Binary Weighted Function, and Tune Parameter.

A R T I C L E I N F O

Article history:

Received	12 Feb. 2025
Received in revised form	04 Mar. 2025
Accepted	26 Mar. 2025
Available online	30 Sep. 2025

©2023 THIS IS AN OPEN ACCESS ARTICLE UNDER THE CC BY LICENSE
http://creativecommons.org/licenses/by/4.0/

*Corresponding author:
Saman Hussein Mahmood
College of Administration and Economics/Salahaddin University- Erbil

Abstract: Outliers affect the accuracy of estimating multiple linear regression model parameters and lead to unacceptably large residual values; therefore, robust estimator methods such as the binary square function must be used to obtain more accurate parameters and robustness versus outliers. The proposed method involves choosing the optimal tune parameter value that produces the minimum mean square error of the multiple linear regression parameters and treating outliers for every data set. Simulation and real data were used to compare the efficiency of the models estimated based on the classical robust method for binary square function and the algorithm proposed through a MATLAB program dedicated to this purpose. The research results showed that the proposed algorithm is effective in processing extreme values, estimating the most optimal tuning parameters and precisely estimating the multiple linear regression model coefficients.

Tikrit Journal of Administrative and Economic Sciences, Vol. 21, No. 71, Part (2): 560-571
Doi: www.doi.org/10.25130/tjaes.21.71.2.28

تقدير معلمة الضبط الأمثل للدالة الثنائية الموزونة في نموذج الانحدار الحصين

سامان حسين محمود دلشاد محمود صالح آلان غفور رحيم طه حسين علي
كلية الإدارة والاقتصاد / جامعة صلاح الدين-اربيل

المستخلص

تؤثر القيم المتطرفة على دقة تقدير معلمات نموذج الانحدار الخطي المتعدد وتؤدي إلى قيم متبقية كبيرة بشكل غير مقبول، لذلك يجب استخدام طرق تقدير حصينة مثل الدالة التربيعية الثنائية للحصول على معلمات أكثر دقة ومتانة مقابل القيم المتطرفة. تتضمن الطريقة المقترحة اختيار قيمة معلمة الضبط المثلى التي تنتج الحد الأدنى لخطأ المربع المتوسط لمعلمات الانحدار الخطي المتعددة ومعالجة القيم المتطرفة لكل مجموعة بيانات. تم استخدام المحاكاة والبيانات الحقيقية لمقارنة كفاءة النماذج المقدرة بناءً على الطريقة الحصينة الكلاسيكية للدالة التربيعية الثنائية والخوارزمية المقترحة من خلال برنامج MATLAB المخصص لهذا الغرض. أظهرت نتائج البحث أن الخوارزمية المقترحة فعالة في معالجة القيم المتطرفة وتقدير معلمات الضبط المثلى وتقدير معاملات نموذج الانحدار الخطي المتعدد بدقة.

الكلمات المفتاحية: نموذج الانحدار الحصين، القيم الشاذة، الدالة المرجحة الثنائية، ومعلمة الضبط.

1. Introduction:

The purpose of regression analysis is to construct a mathematical model that describes the relationship between two types of variables (independent & dependent), can predict how changes in the independent variable will affect dependent variables and comprehend their relationship with each other thanks to this model. The primary goals of regression analysis involve defining the data, analyzing the relationship between variables and estimating the model and its parameters to better explain the data. Changes in the explanatory variables affect and control the dependent variable. Using the regression model in predictions, decision-making, and future planning.

Regression equations are commonly assessed using methods such as maximum likelihood and least squares due to their computational capacity and ease of usage. However, these methods may lose their effectiveness if the data contains outliers. Robust regression techniques can be used in these situations to produce more accurate estimates. (Cohen et al. 2002: 13) (Ismael & Albairmani, 2021: 514)

Outliers are points in the data that show significant variation from the data set's average values and may result from mistakes or natural variations. Outliers have the potential to seriously impair an analysis, even if they are

small. Robust regression techniques are commonly employed to obtain more accurate estimates in such situations. (Ipek & Keser, 2023: 10).

The use of robust estimators often aims to reduce the impact of outliers and accurately estimate parameters even when the basic assumptions of OLS regression are violated due to the presence of outliers. A common robust estimate method that modifies the loss function to decrease the impact of outlier observations is binary quadrature functions. This method ensures that outliers are down-weighted during the estimation process, resulting in more accurate and trustworthy parameter estimates. (Wilcox, 2011: 20)

Tuning parameters are adjusted to raise statistical models' efficiency when observations or outliers differ from established assumptions. Specifically, tuning parameters are intended to strike a balance between robustness (lessening the influence of outliers) and efficiency (ensuring accurate estimation when data follows known patterns). (Chen & Paschalidis, 2018: 15)

The research's objective is to create a more efficient and robust regression technique by using binary quadrature functions and improving tuning parameter estimates. The proposed method aims to maintain the efficiency and interpretability of the estimates while enhancing the reliability of MLR models in the presence of outliers. By improving the estimation of the tuning parameter and expanding on established robust methods, this paper reviews common robust estimation techniques, the role of tuning parameters is examined, popular robust estimating techniques are reviewed, and approaches for identifying optimal values are assessed. This article aims to present a robust method with an optimal tuning parameter that addresses the outlier problem compared to the traditional robust method and has more accurate regression coefficients.

2. **Multiple Linear Regression**: Multiple linear regression (MLR) is an analysis of statistics used to provide a mathematical explanation for the numerous connections between the response (dependent variable) y and the independent variables (x_1, x_2,..., x_k). (Rencher & Schaalje, 2008:243; Ali & Awaz, 2017: 42).

$$y = \beta 0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \ldots (1)$$

Included in the regression model is an error term (ε) to account for unobserved factors. The regression coefficients (β_0, β_1,..., β_k) represent the parameters of the model, with each β_j indicating the change in the expected

value of y for a unit change in x_j, holding other variables constant. The errors are assumed to be normally distributed with a zero mean and constant variance ($\boldsymbol{\varepsilon \sim N (0, \sigma^2)}$). Under these conditions, OLS estimation provides unbiased, consistent, and efficient estimates of the unknown parameters. (Doane & Seward, 2021; Rawlings, Pantula, & Dickey, 1998:75)

MLR is applied models used for different purposes:

A. Prediction. Estimates of the individual parameters b_0, b_1,..., b_k have little significance for prediction compared to the overall effects of x variables on y. To achieve good model performance, we need good estimates.
B. Describing the data: It is important to explain and summarize the data and the variables used in the estimation.
C. Estimation of parameters. There may be theoretical ramifications for a proposed model based on the predicted parameter values.
D. Selecting or screening variables. Determining and focusing the importance of each variable in modelling the variance in y. Variables that explain a large variance in y are retained; Delete small contributions.
E. Output control: The causal relationship that is assumed to exist between y and x variables. After estimation, the model can be used to control the process's output when the inputs are changed. Systematic experiments can give the best results.

MLR analysis of real-world data commonly encounters outliers or observations that deviate significantly from the overall pattern. The accuracy of the regression results can be significantly harmed by even one outlier. (Cohen et al. 2002:60; Omer et al., 2024:115)

3. Outlier: Least squares estimates are the most accurate choice for estimating regression coefficients when errors are normally distributed; however, they may not be as precise when errors are skewed or contain significant outliers. Outliers, or data points with unusually large errors, can significantly affect the least squares fit. To address this issue, alternative fitting methods have been developed that are less sensitive to outliers. (Seber,2003:265)

In regression models, outliers can occur in the dependent variable (response variable) and are simply called outliers, or they can occur in the explanatory variables, where they are known as leverage points. Leverage points are classified as either good, which does not significantly affect the regression equation, or bad, which affects the model coefficients. The presence of outliers in both the dependent and explanatory variables can

complicate the estimation process. Traditional estimation methods, such as the least squares method, become ineffective if the model errors do not follow a normal distribution. (Rousseeuw & Leroy,1987:216)

Dealing with known data contamination is relatively straightforward: you can correct, delete, or replace contaminated data points. However, dealing with outliers, which are rare and unusual data points, requires a different approach. Here are three common strategies: (Belsley et al. 2004:15)

Delete outliers: Simply remove the outliers from the dataset and move on with the analysis.

Revise the model: Modify the regression model to include terms that could explain the outliers, such as interaction terms or higher-order terms.

Transform the data: Transform the data to normalize the distribution and reduce the influence of outliers.

Robust regression: Use alternative regression methods that are less sensitive to outliers.

There are various methods for detecting outliers in linear regression. These methods can be categorized into graphical (Scatter plots, Box plots, Williams plots, Q-Q plots and Graphs of predicted residuals) and analytical approaches (Predicted residuals, Standardized residuals, Jackknife residuals, DFFITS, Cook's distance and Atkinson's measure). These methods help us identify data points that are unusual or don't fit the expected pattern, suggesting they might be outliers.

4. **Robust Linear Regression Model**: Robust regression is an important alternative to parameter estimation in regression models, especially when outliers affect the accuracy of traditional methods such as ordinary least squares (OLS). These challenges arise when the assumptions on which the model is built do not match the available information, leading to inaccurate results and thus making traditional methods less effective. Unlike OLS, robust regression takes outlier values into account and uses statistical techniques to reduce the influence of these values on parameter estimates, thus obtaining more reliable results. We can accurately detect outliers by analyzing the residuals from a robust regression, allowing us to make appropriate adjustments that traditional methods often fail to achieve. This highlights the significance of robust regression for getting more precise estimation results when dealing with difficult data that contains outliers. The

main robust methods for estimating a robust regression model include M-estimation (Reduces the influence of outliers by weighting residuals differently (e.g., Huber loss)), S-estimation (Focuses on minimizing the spread or dispersion of residuals (e.g., MAD)) and Mm-estimation (Uses a two-step process: an S-estimate for scale, followed by an M-estimate for parameters). (Rousseeuw & Leroy,1987:248)

5. **Binary Weighted Functions:** The binary-weighted function uses different weights to analyze data points if they are intrinsic or outliers. The weight value (w_i) for each observation (i) is given as follows:

$$w_i = \begin{cases} 1 & \text{if } |r_i| \leq T \\ 0 & \text{if } |r_i| > T \end{cases} \quad \dots (2)$$

In equation (2), (r_i) represents the residual for observation i, and (T) is a threshold used to identify outliers. This threshold is frequently established based on the distribution of residuals, such as a multiple of the standard deviation or the median absolute deviation. Observations that fall within the minimum residuals are given a weight of one and are considered identical, while observations that exceed this limit and have a weight of zero are excluded and are considered outliers. (Kotlyarova & Zinde-Walsh, 2009:270)

6. **Role of Tuning Parameters**: Tuning parameters act like thresholds or weights, giving each data point relative importance. These weights determine how much each point, especially outliers, will affect the final value of the estimate. These parameters play a direct role in determining the strength of the model and its ability to work efficiently. Robustness is the ability of a model to achieve its main performance and resist when exposed to extreme values in the data. Larger tuning parameters lead to greater power because they reduce the influence of outliers through weights. (Gábor & Banga, 2015:12; Sedeeq et al. 2023: 715)

Efficiency refers to the ability of a model to extract the true shape of the data when the data follows its expected pattern (expected distribution). The tuning parameters require attention and care to avoid loss of efficiency on normal data (non-outlier data). Models that are robust and achieve consistent performance may lose some of their efficiency, while models that are tuned to high efficiency may be more sensitive to outliers. (Yonekura & Sugasawa, 2023: 5)

7. **Mean Square Error of parameters:** In statistics, use parameter estimation to build models that predict future outcomes. To gauge the accuracy of these estimates, employ error metrics. A common metric is Mean Squared Error (MSE).

$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$ … (3)

This formula calculates the average difference (error) squares between the estimate (θ) and the true parameter (θ). (Aduol, 2003:271)

MSE is crucial for evaluating the performance of various statistical models, including linear, nonlinear, and machine learning models. Its importance lies in its ability to assess both the systematic error (bias) and the variability (variance) of the estimates. The primary objective of many statistical modelling tasks is to minimize MSE, often requiring a balance between reducing bias and variance

8. **Proposed Method**: The proposed method includes choosing the optimal tune parameter value that produces the minimum mean square error of the multiple linear regression parameters and treating outliers through the following steps:

Step 1: Iterate k from 0.1 to 10 in increments of 0.01. For each iteration, employ a binary square weight function to estimate the parameters of a robust MLR.

Step 2: Determines the optimal tune parameter value that gives the minimum mean square error of the parameters estimated.

Step 3: Use a binary square weight function depending on the value of the optimal tuning parameter in estimating the parameters of the robust regression model.

Note that the proposed algorithm gives the largest value for the tuning parameter in the absence of outliers in the data, which provides estimates of the parameters of the linear regression model completely similar to the ordinary least square method. The program handles heavily contaminated data and multiple outliers by using the following strategies: The binary square weight function reduces the impact of outliers by assigning them lower weights, ensuring that they do not disproportionately influence the regression parameters. The tuning parameter k is optimized to minimize the MSE, ensuring the model adapts to the data and appropriately down-weights outliers. Iterative Proposed: By iterating through different values of k, the method ensures that the optimal level of robustness is chosen for the given

Doi: www.doi.org/10.25130/tjaes.21.71.2.28

dataset. This robust regression approach allows the method to effectively handle heavily contaminated data or multiple outliers, ensuring that the model remains reliable and accurate even in challenging scenarios.

9. **Simulation Study**: Experiment simulation is used to consider the combination and repeat it (500) times for different instances, as seen below:

A. There are three sample sizes used: 30, 50, and 100.

B. The explanatory variables are independent from a normal distribution, where p is the number of the explanatory variables, which can be 1, 2, and 3.

C. Add outliers to generate data using the randomizing function in MATLAB (2 outliers).

Figure 1: Residual case order plot for the first trial in simulation (n=30)

By observing the figure above, we find that the variable Y contains the outlier values.

Table 1 presents the mean square error (MSE) following parameter estimation for 1000 iterations of both the standard and new methods. Because the MSE of the suggested method is lower than that of the classical approach, it outperformed the latter in every simulation situation.

Doi: www.doi.org/10.25130/tjaes.21.71.2.28

Table (1): Results of Simulation experiments for new robust with classical methods

p	Sample size(n)	OLS	Robust	New Robust	Optimal Tune
		MSE	MSE	MSE	
1	30	2.2359	0.1809	0.1794	5.5849
	50	0.7215	0.0988	0.0946	5.6284
	100	0.2326	0.0488	0.0477	5.5724
2	30	1.4676	0.1350	0.1305	5.5647
	50	0.4204	0.0666	0.0647	5.6061
	100	0.1362	0.0337	0.0327	5.5844
3	30	0.6581	0.0840	0.0824	5.6138
	50	0.2505	0.0422	0.0408	5.5728
	100	0.0725	0.0209	0.0202	5.5749

10. **Application Aspect**: The effect of different concentrations of two drugs (two independent variables) at different levels (xi1, xi2) was studied on patients with nephritis (inflammation of the kidneys), the response variable (Yi) represents the number of cases of recovery because of taking the two drugs through (15) observations representing (n = 15) samples (Appendix).

Figure (2): Residual case order plot for real data

Doi: www.doi.org/10.25130/tjaes.21.71.2.28

The outlier values are found in the variable Y, as can be seen in the above figure.

Table 2: Results of real data for new robust with classical methods

OLS	Robust	New Robust	Optimal Tune
MSE	MSE	MSE	
564.6658	10.1862	**8.7345**	4.9650

The calculated regression model is preferred, as evidenced by the criteria comparing Table 2 of the new robust regression model, where the MSE criteria value was lower.

11. Conclusions: The objective is to find a compromise between accuracy and robustness such that the model continues to function properly even in the presence of aberrant data points. To determine the best settings for a given dataset, this approach frequently uses iterative testing. suggested that Results from a tune robust estimation are more dependable and comprehensible when there are anomalies or irregular data distributions. Less biased coefficients, a better fit for the central data, and more precise standard errors are the main distinctions from OLS. Because perfect data conditions are uncommon in real-world data analysis, these results make robust estimation very valuable.

References

1. Aduol, F.W. (2003). Robust geodetic parameter estimation under least squares through weighting based on the mean square error. In Geodesy-the Challenge of the 3rd Millennium (pp. 269-276). Berlin, Heidelberg: Springer Berlin Heidelberg.
2. Ali, Taha Hussein & Awaz Shahab M. (2017) "Uses of Waveshrink in Detection and Treatment of Outlier Values in Linear Regression Analysis and Comparison with Some Robust Methods", Journal of Humanity Sciences 21.5, pp 38-61.
3. Ali, T. H., and Saleh, D.M. (2021). Comparison Between Wavelet Bayesian and Bayesian Estimators to Remedy Contamination in Linear Regression Model, PalArch's Journal of Archaeology of Egypt/Egyptology 18.10, pp 3388-3409
4. Belsley, D.A., Kuh, E. and Welsch, R.E. (2004). Regression Diagnostics; Identifying Influence Data and Source of Collinearity. Wiley, New York.
5. Chen, R. and Paschalidis, I.C., (2018). A robust learning approach for regression models based on distributionally robust optimization. Journal of Machine Learning Research, 19(13), pp.1-48.
6. Cohen,J., Cohen,P., West, S.G. & Aiken, L.S., (2002). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Third edition. Routledge.
7. Doane, D.P. & Seward, L., 2021. Applied Statistics in Business and Economics. McGraw-Hill Education.

8. Gábor, A. and Banga, J.R., 2015. Robust and efficient parameter estimation in dynamic models of biological systems. BMC Systems Biology, 9, pp.1-25.
9. Ismael, A. A. & Albairmani, Z. A. (2021). The effect of some independent variables on granting the basic application certificate using multivariate statistical analysis, Tikrit Journal of Administrative and Economic Sciences, Vol. 17, No. 54,part 3, 509-522
10. Ipek, D. K., Keser I. K. (2023) Outlier detection based on the functional coefficient of variation STATISTICS IN TRANSITION Vol. 24, No. 2, pp. 1–16,
11. Kotlyarova, Y. and Zinde-Walsh, V., 2009. Robust estimation in binary choice models. Communications in Statistics-Theory and Methods, 39(2), pp.266-279.
12. Mahmood. S. H. (2024). Estimating Models and Evaluating their Efficiency under Multicollinearity in Multiple Linear Regression: A Comparative Study, ZANCO Journal of Humanity Sciences, 28(5), 264 – 277
13. Omer, A. W., Sedeeq, B. S., & Ali, T. H. (2024). A proposed hybrid method for Multivariate Linear Regression Model and Multivariate Wavelets (Simulation study). Polytechnic Journal of Humanities and Social Sciences, 5(1), 112-124.
14. Rawlings, O., Pantula, G. & Dickey, D.A., 1998. Applied Regression Analysis: A Research Tool, 2nd Edition. Springer.
15. Rencher, A.C. & Schaalje, G., (2008). Linear models in statistics. 2nd ed. John Wiley & Sons.
16. Rousseeuw, P.J. and Leroy, A.M. (1987). Robust Regression and Outlier Detection. Wiley Series in Probability and Mathematical Statistics, John Wiley, New York.
17. Sedeeq, B.S., Qader, H.M., Aziz, A.A., Saleh, D.M. (2023), Implementing a New Scale Technique in the M-Estimation Method to Estimate Parameters of Multiple Linear Regression: Simulation Study,Tikrit Journal of Administrative and Economic Sciences, Vol. 19, No. 64, Part (1): 712-725, Doi:10.25130/tjaes.19.64.1.38
18. Seber, G.A.F. and Lee, A.J. (2003). Linear Regression Analysis, Second Edition. John Wiley & Sons, Inc., Hoboken, New York
19. Wilcox, R.R., 2011. Introduction to robust estimation and hypothesis testing. Academic press.
20. Yonekura, S. and Sugasawa, S., 2023. Adaptation of the tuning parameter in general Bayesian inference with robust divergence. Statistics and Computing, 33(2), 1-16.

AppendixTable 2. Data of Nephritis patients

Samples	Y_i	x_{i1}	x_{i2}
1	23	.51	.39
2	21	.55	.28
3	80	.67	.18
4	60	.63	.15
5	20	.23	.54
6	25	.33	.44
7	30	.44	.35

Doi: www.doi.org/10.25130/tjaes.21.71.2.28

Samples	Y_i	x_{i1}	x_{i2}
8	35	.60	.33
9	42	.83	.22
10	100	.62	.42
11	29	.70	.69
12	14	.75	.55
13	18	.82	.67
14	34	.77	.72
15	111	.78	.66