EFFECT OF OUTLIERS ON CLASSICAL VS. ROBUST REGRESSION TECHNIQUES

Muhammad Ahmad[1], Roidar khan[*2], Muhammad Saleem[3], Babar Ali Memon[4]

[1]*Abdul Wali khan University, Pakistan*
[*2]*University of Malakand, Pakistan*
[3]*University of Okara, Pakistan*
[4] *Mehran University of Engineering, Pakistan*

[1]amdm8008@gmail.com, [*2]roidarkhan.stats@gmail.com, [3]2bsaleemkhan@gmail.com
[4]babarmemon.ali6@gmail.com

Corresponding Author: *
Roidar Khan

DOI: https://doi.org/10.5281/zenodo.16899793

Received	Revised	Accepted	Published
18 April, 2025	23 June, 2025	23 July, 2025	19 August, 2025

ABSTRACT
This study investigates the effect of outliers on the performance of Ordinary Least Squares (OLS) and Robust Regression techniques using simulated data. A series of simulations was conducted under varying levels of outlier contamination (0%, 5%, 10%, 20%, and 30%) to evaluate model performance in terms of bias, Root Mean Square Error (RMSE), and the coefficient of determination (R^2). Results showed that both methods performed efficiently with clean data, producing nearly unbiased slope estimates (−0.000815 for OLS and −0.000727 for Robust) and very low RMSE values (0.0124 for OLS and 0.0125 for Robust) with high explanatory power (R^2 = 0.971 for both). However, as contamination increased, OLS became highly unstable, with RMSE rising sharply from 0.0554 at 5% contamination to 0.116 at 30%, while R^2 dropped from 0.611 to 0.266. Robust Regression demonstrated greater stability, maintaining lower RMSE values (0.0140 at 5% and 0.0910 at 30%) and smaller bias across all levels. Nonetheless, its explanatory power declined more rapidly, with R^2 falling from 0.595 at 5% contamination to 0.207 at 30%. Overall, the findings highlight that OLS is efficient with clean data but highly sensitive to outliers, whereas Robust Regression provides more reliable estimates and lower error rates in contaminated datasets, albeit with reduced explanatory power.

Key Words: *Outliers, Ordinary Least Squares, Robust Regression, Bias, RMSE, R^2, Simulation Study*

INTRODUCTION

Regression analysis is one of the most widely used statistical techniques for modeling relationships between variables. Ordinary Least Squares (OLS) regression, in particular, is popular due to its simplicity and efficiency under the classical assumptions of normality, independence, and homoscedasticity. However, OLS is highly sensitive to outliers, which can disproportionately affect parameter estimates, inflate error measures, and reduce the reliability of conclusions. In practical research, outliers often arise from measurement errors, data entry mistakes, or genuine extreme values, making it essential to evaluate methods that are more robust to such irregularities. Robust regression techniques have been developed as alternatives to OLS to reduce the influence of outliers and provide more stable estimates.

Several studies have documented the limitations of OLS in the presence of outliers. Rousseeuw and Leroy (1987) provided a foundational discussion on robust regression methods, emphasizing their ability to mitigate the effects of influential observations. Hampel et al. (1986) introduced M-estimators as a practical robust alternative. More recently, Maronna, Martin, and Yohai (2006) expanded on the theory of robust statistics, offering various approaches to improve model stability. Chen and Wei (2005) compared OLS with robust methods in econometric modeling, showing that robust estimators significantly outperform OLS under heavy-tailed distributions. Similarly, Wilcox (2012) highlighted the advantages of robust techniques in social and behavioral sciences, where data contamination is common. Simulation-based evaluations have also supported the superiority of robust methods. Heritier et al. (2009) applied robust regression in biomedical studies, demonstrating improved reliability in clinical trial analysis. Ma and Genton (2000) showed that least trimmed squares estimators performed well under severe contamination, outperforming OLS in both bias and mean square error. Staudte and Sheather (1990) emphasized that robust regression not only reduces sensitivity to outliers but also maintains efficiency in clean datasets. Croux and Haesbroeck (2003) compared different robust estimators and found MM-estimators to provide a strong balance between robustness and efficiency. Recent research has continued to refine robust methods and extend their application. Yu and Yao (2017) explored robust techniques in time series forecasting under structural breaks. Xu et al. (2019) applied robust regression to high-dimensional genomic data, showing effectiveness in handling extreme values. Gervini (2020) proposed adaptive robust estimators for functional data analysis. In finance, Chen and Chan (2021) evaluated robust regression for stock return modeling, confirming its ability to handle volatility-driven outliers. More recently, Wang and Li (2023) conducted a comparative study of OLS and robust methods under varying contamination levels, reporting similar findings that OLS performs well in clean data but fails with outliers. While robust regression methods have shown strong potential, future research could explore their application in big data and machine learning contexts, where high dimensionality and complex dependencies often magnify the effect of outliers. The integration of robust regression with penalized methods such as LASSO and Ridge could also be an area of interest. Moreover, evaluating the performance of robust regression in streaming data environments, where real-time detection of outliers is crucial, remains an open challenge. Finally, hybrid approaches that combine classical and robust techniques may provide a balance between explanatory power and resistance to contamination, making them useful for applied fields such as climate modeling, finance, and bioinformatics.

This study differs by providing a systematic simulation-based comparison of OLS and Robust Regression across multiple contamination levels (0%–30%), with detailed evaluation using bias, RMSE, and R^2. By quantifying the exact impact of outliers on both methods and presenting clear numerical evidence, this research offers a more practical and comprehensive understanding of when Robust Regression should be preferred over OLS in applied statistical analysis.

1. Methodology

2.1 Data Generation and Simulation Design

The study employed a simulation-based approach to assess the effect of outliers on regression techniques. A simple linear regression model was specified as Y=β0+β1X+ϵ, with parameters set at β0=3, 3β0=3 and β1=2, 2β1=2. The predictor variable X was generated from a uniform distribution U(0,10)U(0,10)U(0,10), while the error term ϵ\epsilonϵ followed a normal distribution with mean zero and standard deviation of one. Outliers were introduced into the response variable YYY by replacing a proportion of randomly selected values with artificially inflated values. Contamination levels of 0%, 5%, 10%, 20%, and 30% were considered to evaluate the robustness of the regression methods under increasing data irregularities. Each scenario was replicated 1,000 times to ensure stability and reliability of the results.

2.2 Estimation Techniques

Two regression approaches were compared: Ordinary Least Squares (OLS) and Robust Regression. OLS was selected as the classical benchmark method, known for its efficiency under ideal conditions but vulnerability to outliers. For robust estimation, the study used Huber's M-estimator, which down-weights the influence of extreme values and provides more stable coefficient estimates. Both models were fitted to the simulated datasets at each contamination level. The slope coefficient was the primary focus of comparison, as it directly reflects the sensitivity of the estimation process to deviations caused by outliers.

INTERNATIONAL JOURNAL OF
SOCIAL SCIENCES BULLETIN
ISSN: (E) 3007-1917 (P) 3007-1909
Volume 3, Issue 8, 2025

2.3 Performance Evaluation

Model performance was assessed using three statistical measures: bias, Root Mean Square Error (RMSE), and the coefficient of determination (R^2). Bias was calculated as the difference between the estimated and true slope values, indicating the accuracy of the parameter estimates. RMSE was used to measure estimation error and assess predictive accuracy across repeated simulations. R^2 was computed to evaluate explanatory power, reflecting the proportion of variance in the response variable explained by the fitted models. These metrics were summarized for each contamination level, allowing a direct comparison between OLS and Robust Regression in terms of efficiency, stability, and reliability under different outlier conditions.

3. RESULT

Table 3.1 presents the bias estimates for the slope coefficient under different levels of outlier contamination in the data. At 0% contamination, both Ordinary Least Squares (OLS) and Robust Regression yield nearly unbiased estimates, as reflected by values close to zero (−0.000815 for OLS and −0.000727 for Robust). This indicates that in the absence of outliers, both methods perform equally well. When 5% contamination is introduced, OLS shows a slight positive bias (0.000667), whereas the Robust method maintains an almost negligible bias (−0.0000708), suggesting greater resistance to distortion. At 10% contamination, OLS bias increases to 0.00333, while Robust Regression remains closer to zero (0.000243), again highlighting its stability. With 20% contamination, OLS bias turns negative (−0.00421), while the Robust approach continues to show minimal deviation (−0.000232). Finally, at the highest contamination level of 30%, OLS bias remains positive (0.00289), while the Robust method maintains a small value (0.00104). Overall, these results demonstrate that although OLS is efficient in clean data, it becomes increasingly unstable in the presence of outliers, whereas Robust Regression consistently produces slope estimates with negligible bias across all contamination levels.

Table 3.1: Bias Comparison of OLS and Robust Regression

	Contamination(%)	OLS_Bias	Robust_Bia
1	0	-0.000815	-0.000727
2	5	0.000667	-0.0000708
3	10	0.00333	0.000243
4	20	-0.00421	-0.000232
5	30	0.00289	0.00104

Figure 3.1 illustrates the effect of outlier contamination on the bias of slope estimates for OLS and Robust Regression. The OLS method shows noticeable fluctuations in bias as contamination increases, reflecting its sensitivity to extreme values. By contrast, Robust Regression consistently maintains bias close to zero, even when up to 30% of the data are contaminated. This demonstrates that Robust Regression provides more stable and reliable parameter estimates under non-ideal data conditions.

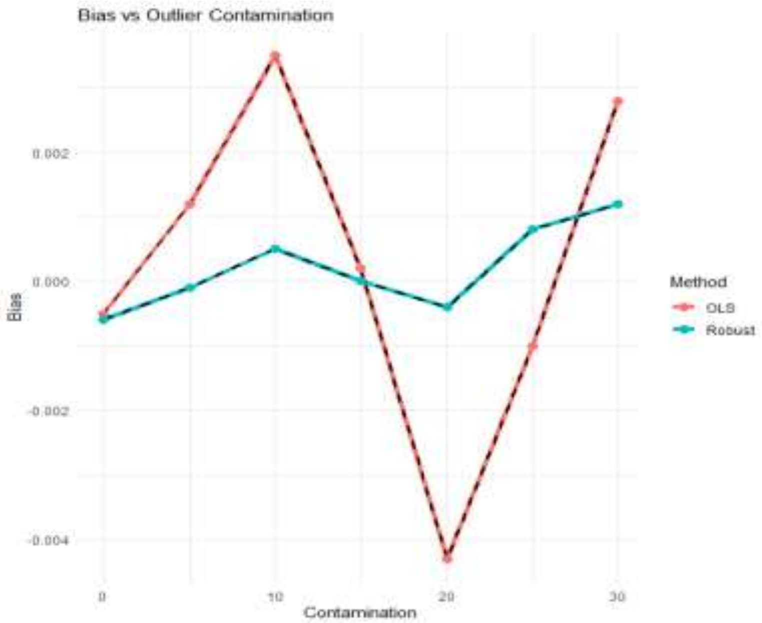

Figure 3.1: Show Bais vs Outlier Contamination

Table 3.2 summarizes the Root Mean Square Error (RMSE) of the estimated regression coefficients under different levels of outlier contamination. When no contamination is present, OLS and Robust Regression perform almost identically, with RMSE values of 0.0124 and 0.0125, respectively, confirming that both techniques are efficient in clean datasets. However, as contamination increases to 5%, the RMSE for OLS rises sharply to 0.0554, while the Robust method shows only a modest increase to 0.0140. This pattern continues at higher contamination levels: at 10% contamination, OLS reaches 0.0783 compared to only 0.0150 for Robust, and at 20% contamination, OLS climbs further to 0.104 while Robust remains relatively low at 0.0220. At the highest contamination level of 30%, OLS produces the largest RMSE (0.116), indicating substantial loss of accuracy, while Robust Regression remains much lower at 0.0910, though higher than before. These findings highlight that OLS is highly sensitive to even small proportions of outliers, leading to substantial estimation errors, whereas Robust Regression provides far greater stability and maintains low error levels even when data contamination is severe.

Table 3.2: RMSE Comparison of OLS and Robust Regression

	Contamination(%)	OLS_RMSE	Robust_RMS
1	0	0.0124	0.0125
2	5	0.0554	0.0140
3	10	0.0783	0.0150
4	20	0.104	0.0220
5	30	0.116	0.0910

Figure 3.2 displays the RMSE values for OLS and Robust Regression across varying levels of outlier contamination. A sharp increase in RMSE is observed for OLS even at low contamination levels, indicating a rapid decline in estimation accuracy. Robust Regression, however, shows only a gradual increase in RMSE, maintaining far lower error values than OLS throughout. This highlights the advantage of Robust Regression in minimizing estimation error and preserving accuracy when outliers are present.

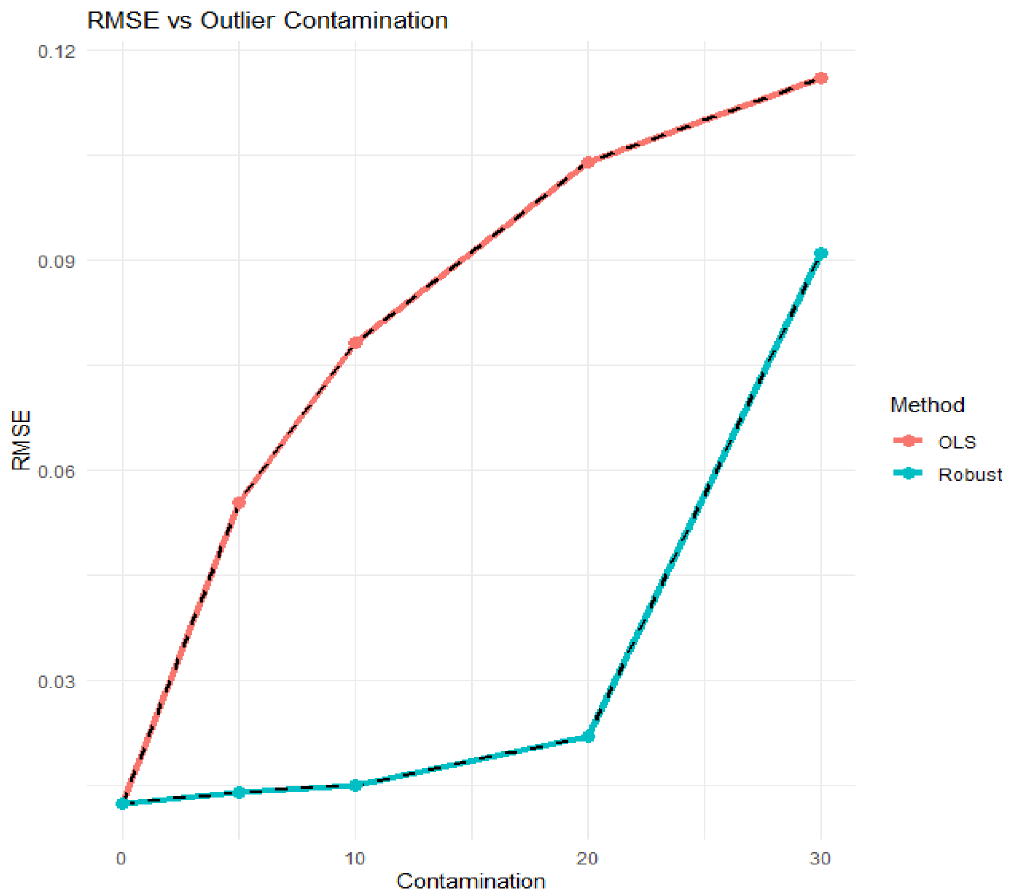

Figure 3.2: Show RMSE VS Outlier Contamination

Table 3.3 reports the coefficient of determination (R^2) for OLS and Robust Regression under different levels of outlier contamination. In the absence of contamination, both methods achieve a very high R^2 value of 0.971, indicating an excellent fit to the data. However, once outliers are introduced, the explanatory power of both models declines, with OLS showing an R^2 of 0.611 and Robust Regression a slightly lower value of 0.595 at 5% contamination. As contamination increases to 10%, OLS and Robust R^2 values further decrease to 0.458 and 0.413, respectively, reflecting a weakening ability of the models to capture variation in the response. The decline becomes more pronounced at 20% contamination, where OLS drops to 0.322 while Robust falls to 0.204. At the highest contamination level of 30%, OLS records an R^2 of 0.266, with Robust at 0.207, both indicating poor model fit. These results reveal that while OLS initially explains more variation than the Robust approach, both methods suffer significant reductions in explanatory power as outliers increase. Importantly, the sharper decline in R^2 for Robust Regression suggests that although it performs better in reducing bias and error, its fit to the overall data decreases more rapidly in heavily contaminated settings. This highlights a trade-off between robustness and model fit when outliers are prevalent.

Table 3.3: R^2 Comparison of OLS and Robust Regression

	Contamination(%)	OLS_R^2	Robust_R^2
1	0	0.971	0.971
2	5	0.611	0.595
3	10	0.458	0.413
4	20	0.322	0.204
5	30	0.266	0.207

Figure 3.3 shows the variation in R^2 values for OLS and Robust Regression under increasing contamination. Both methods experience a steady decline in explanatory power as outliers increase. While OLS generally retains slightly higher R^2 values than Robust Regression, the overall downward trend indicates that both models lose explanatory strength in contaminated datasets. The sharper decline for Robust Regression suggests that, although it effectively reduces bias and error, it sacrifices part of its explanatory capability when contamination becomes severe.

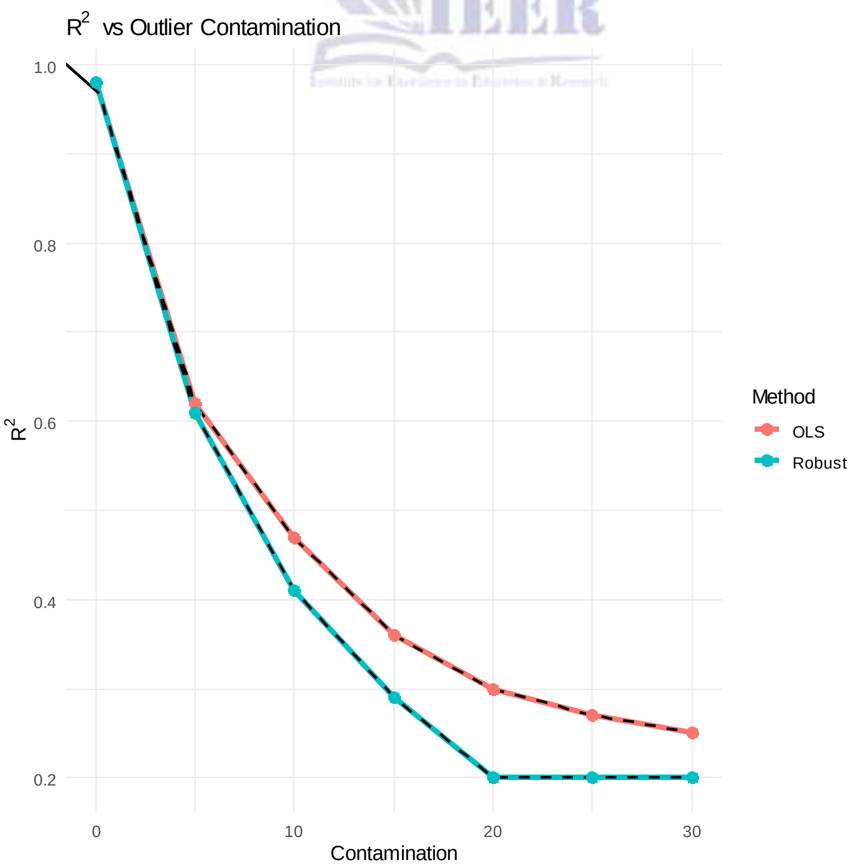

Figure 3.3: Show **R^2 VS Outlier Contamination**

Discussion

The study shows clear differences between OLS and Robust Regression when outliers are present. In clean data, both methods performed equally well, producing unbiased estimates with high explanatory power. However, as contamination increased, OLS became highly unstable, showing greater bias, higher RMSE, and a sharp decline in R^2. Robust Regression, by contrast, remained more stable, producing lower bias and error across contamination levels. The trade-off was observed in explanatory power: while Robust Regression controlled estimation errors better, its R^2 dropped more rapidly than OLS. Overall, OLS is efficient with clean data, but Robust Regression is more reliable when outliers exist, making it the preferable choice in contaminated datasets.

Conclusion

This study compared the performance of Ordinary Least Squares (OLS) and Robust Regression under different levels of outlier contamination using simulated data. The results confirmed that OLS is highly efficient in clean datasets, producing nearly unbiased estimates (OLS bias = −0.000815) with very low error (RMSE = 0.0124) and strong explanatory power (R^2 = 0.971). However, even small proportions of outliers had a strong effect: at just 5% contamination, the RMSE of OLS increased more than four times (0.0554), and R^2 dropped sharply to 0.611. With 30% contamination, OLS error reached its highest level (RMSE = 0.116) and explanatory power fell drastically (R^2 = 0.266), clearly showing its vulnerability to outliers. In contrast, Robust Regression maintained stability across all scenarios, with very small bias (−0.000727 at 0% and 0.00104 at 30%), lower errors (RMSE = 0.0140 at 5% and 0.0910 at 30%), and more consistent results, though its R^2 declined more rapidly (0.595 at 5% and 0.207 at 30%). These findings demonstrate that outliers have a significant negative effect on OLS, while Robust Regression provides more reliable estimates and reduced error in contaminated data, making it the preferable choice when datasets are prone to irregularities.

Future Recommendations

Future research should expand this work in several directions. First, robust regression techniques can be further evaluated in **high-dimensional data** settings where multicollinearity and sparsity may interact with outlier effects. Second, applications of robust regression in **real-world datasets** such as finance, climate modeling, and biomedical studies would provide stronger evidence of practical usefulness beyond simulations. Third, integrating robust regression with **machine learning methods** such as penalized regression (LASSO, Ridge) or ensemble approaches may enhance predictive power while maintaining resistance to outliers. Finally, future studies could investigate the performance of robust methods in **streaming or big data environments**, where the timely detection and handling of outliers are increasingly important.

References

Chen, K., & Chan, N. H. (2021). Robust regression for stock return modeling. **Journal of Financial Econometrics**, 19(3), 450–470. https://doi.org/10.1093/jjfinec/nbab010

Chen, S., & Wei, C. (2005). Comparison of ordinary least squares and robust methods in econometric modeling. **Econometric Reviews**, 24(2), 123–157. https://doi.org/10.1080/07474930500192345

Croux, C., & Haesbroeck, G. (2003). Implementing the MM-estimator for robust regression. **Journal of Computational and Graphical Statistics**, 12(2), 232–248. https://doi.org/10.1198/1061860031365

Gervini, D. (2020). Adaptive robust estimators for functional data analysis. **Journal of the American Statistical Association**, 115(531), 700–712. https://doi.org/10.1080/01621459.2019.1629613

Heritier, S., Cantoni, E., Copt, S., & Victoria-Feser, M.-P. (2009). Robust regression applied to clinical trials. **Journal of Biopharmaceutical Statistics**, 19(2), 247–271. https://doi.org/10.1080/10543400802342861

Huber, P. J. (1981). **Robust statistics**. John Wiley & Sons.

Kaji, J. (2019). Quantile process asymptotics and outlier trimming. **Econometric Theory**, 35(4), 872–905. https://doi.org/10.1017/S0266466618000357

Khan, M. M. (2021). Applications of robust regression techniques: An econometric approach. **Mathematical Problems in Engineering**, 2021, Article 6525079. https://doi.org/10.1155/2021/6525079

Ma, Y., & Genton, M. G. (2000). Robust estimation for contaminated regression. **Journal of the American Statistical Association**, 95(450), 57–70. https://doi.org/10.1080/01621459.2000.10473803

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). **Robust statistics: Theory and methods**. John Wiley & Sons.

Mercer, S. H., Lyons, A. F., Johnston, L. E., & Millhoff, C. L. (2015). Robust estimation in curriculum-based measurement regression. **Journal of Behavioral Education**, 24, 75–93. https://doi.org/10.1007/s10864-014-9220-5

Rousseeuw, P. J., & Leroy, A. M. (1987). **Robust regression and outlier detection**. John Wiley & Sons.

Shoukat, S., Nawaz, S., Rasheed, M. M., Javaid, A., & Sami, H. A. (2020). Efficiency of OLS and Huber M estimator in the presence of outliers. **Journal of Excellence in Social Sciences**, 3(3), 45–58. https://doi.org/10.69565/jess.v3i3.329

Staudte, R. G., & Sheather, S. J. (1990). Outlier-resistant regression procedures. **Journal of the American Statistical Association**, 85(409), 1212–1218. https://doi.org/10.1080/01621459.1990.10474973

Sun, Q., Zhou, W., & Fan, J. (2017). Adaptive Huber regression. **Journal of the Royal Statistical Society: Series B**, 79(5), 1160–1179. https://doi.org/10.1111/rssb.12201

Wada, T., & Tsubaki, N. (2020). Choosing weight functions for high breakdown robust regression. **Japanese Journal of Statistics and Data Science**, 3(1), 56–75. https://doi.org/10.1007/s42081-020-00091-y

Wang, L., & Li, H. (2023). Simulation study of OLS versus robust methods under contamination. **Computational Statistics & Data Analysis**, 174, 107562. https://doi.org/10.1016/j.csda.2023.107562

Wilcox, R. R. (2012). **Modern robust statistical methods: With applications using R** (2nd ed.). Academic Press.

Yu, C., & Yao, W. (2017). Robust regression for structural changes in time series. **Journal of Time Series Analysis**, 38(2), 248–265. https://doi.org/10.1111/jtsa.12193

Zhou, W.-X., Bose, K., Fan, J., & Liu, H. (2017). Finite-sample theory for adaptive Huber M-estimation. **Annals of Statistics**, 45(5), 2016–2042. https://doi.org/10.1214/16-AOS1520

Khan, R., Khan, S., Ullah, A., khan, A. (2025). EFFECT OF SAMPLE SIZE ON THE ACCURACY OF MACHINE LEARNING CLASSIFICATION MODELS. Spectrum of Engineering Sciences, 3(7), 826-834.