BAREKENG: Jurnal Ilmu Matematika dan Terapan
March 2022 Volume 16 Issue 1 Page 243-252
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss1pp243-252

COMPARISONS BETWEEN ROBUST REGRESSION APPROACHES IN THE PRESENCE OF OUTLIERS AND HIGH LEVERAGE POINTS

Anwar Fitrianto [1*], **Sim Hui Xin**[2]

[1]*Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University*
Dramaga Campus, Bogor 16680, West Java, Indonesia
[2]*Department of Mathematics, Faculty of Sciences, Universiti,Putra Malaysia, Malaysia*
UPM Serdang, 43400 Seri Kembangan, Selangor, Malaysia

Corresponding author e-mail: ¹ anwarstat@gmail.com*

Abstract. *The study aimed to compare a few robust approaches in linear regression in the presence of outlier and high leverage points. Ordinary least square (OLS) estimation of parameters is the most basic approach practiced widely in regression analysis. However, some fundamental assumptions must be fulfilled to provide good parameter estimates for the OLS estimation. The error term in the regression model must be identically and independently comes from a Normal distribution. The failure to fulfill the assumptions will result in a poor estimation of parameters. The violation of assumptions may occur due to the presence of unusual observations (which is known as outliers or high leverage points. Even in the case of only one single extreme value appearing in the set of data, the result of the OLS estimation will be affected. The parameter estimates may become bias and unreliable if the data contains outlier or high leverage point. In order to solve the consequences due to unusual observations, robust regression is suggested to help in reducing the effect of unusual observation to the result of estimation. Four types of robust regression estimations employed in this paper: M estimation, LTS estimation, S estimation, and MM estimation, respectively. Comparisons of the result among different types of robust estimator and the classical least square estimator have been carried out. M estimation works well when the data is only contaminated in response variable. But in the case of presence of high leverage point, the M estimation doesn't have good performance.*

Keywords: *breakdown points, influential, leverage, outlier, robust regression,*

Article info:
Submitted: 11th January 2022 *Accepted: 4th March 2022*

How to cite this article:
A. Fitrianto and S. H. Xin, "COMPARISONS BETWEEN ROBUST REGRESSION APPROACHES IN THE PRESENCE OF OUTLIERS AND HIGH LEVERAGE POINTS", *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, iss. 1, pp. 243-252, Mar. 2022.

1. INTRODUCTION

Regression analysis is a statistical method practiced widely to explain the relevant between variables of interest. The first type and very popular regression analysis is linear regression. In linear regression, a linear model is fitted to the response variable, Y with one or more explanatory or predictor variable, X. This is the simplest model where it assumes a linear equation can explain the relationship among the variables. When there is only one Y and one X involved in the linear model, it is called a simple linear model while multiple linear model include one Y with at least two X. The classical approach in linear regression models is the Ordinary Least Square (OLS) technique where the sum of square of errors is minimized. Minimization of this sum means that the deviation of observation Y from the fitted regression line is minimized.

Three underlying assumptions must be fulfilled by the errors term in linear regression analysis. These assumptions are normality, constant variance, and independence assumptions, [1]. In general, the errors are assumed to be independently and identically distributed random variables from normal distribution with mean zero and constant variances. The violation of these assumptions will cause a misleading analysis or even disturb the validity of the linear regression model fitted to the variables [2]. The estimations computed by linear regression will become unreliable and cannot provide useful information about the data if assumptions violate, ([3]-[4]).

In real-life data, unusual observations are a widespread issue in data analysis, [5]. One of the possible unusual observations is observed in the response variable's direction. In that case, it is called an outlier while the extreme value contained in the predictor variables is called a high leverage point. In regression, an outlier is defined as observation that does not follow the general pattern of the whole data set, [6]. Meanwhile, a high leverage point represents a x-value that lies far away from the rest of the data, [7]. Not all outliers and high leverage points are influential points, [8]. An observation is categorized as influential if removing that particular point singly or in combination will cause changes to the fitted model and hence the parameters of estimation. An influential point will affect the analysis's precision based on the OLS regression method. The situation worsens when such a point's presence causes the violation in linear regression assumptions, [9].

In regression analysis, in order to deal with the effect of an outlier or high leverage point which is influential, robust regression is introduced, ([10]-[11]). Robust regression is one type of regression approach that is robust towards the presence of outlier and influential points. It is an alternative way of analysis where the computation of the parameter of estimates is not much affected by extreme values in the data set. Different types of robust regression have their advantages and disadvantages. The efficiency of the robust estimation depends on the types and percentage of contamination in the data set. Breakdown point can also be used to compare the performance of robust regressions. According to [9], the breakdown point (BP) indicates the proportion of contaminated observation that estimation can resist and still maintain its robustness. The study's objective is to compare a few robust approaches in linear regression in the presence of outlier and high leverage points.

2. RESEARCH METHODS

2.1 Data Description

The dataset used in this research is obtained from [3]. There are 191 observations of milk production are collected by Dairy Herd Improvement Cooperative (DHI) in New York. The response variable, Y is the current month milk production in pounds (CURRENT), while the predictors X_1 and X_2 are the previous month of the milk production in pounds (PREVIOUS) and the number of days since the presence of lactation (DAYS), respectively. In the data set, lactation means the period that a cow will produce milk.

2.2 The Contamination Methods

The heading at the third level follows the style of the second level heading. Avoid using headings more than three levels. After the OLS analysis is conducted to the original data, some modifications are made to the data sets' observations. In this research, we only consider in contaminating one observation. The value of first observation in each data set is changed to very extreme value in two separate cases. The first case is to make the first observation in the response variable very large while the second case is to alter the first observation in the predictor variable to a very large value. These two cases are then followed with OLS regression analysis respectively after each type of contamination. The following equations are applied in making the modification which is called contamination in the later part of this research.

Case 1: First observation of the data is replaced by the following value of response variable:

$$y_1 = \bar{y} + 10\sigma_{\boldsymbol{Y}},$$

where $\bar{y}$ is the mean of variable Y and $\sigma_{\boldsymbol{Y}}$ is the standard deviation of variable Y.

Case 2: Value of x of the first observation is replaced by

$$x_1 = \bar{x} + 10\sigma_{\mathbf{X}},$$

where $\bar{x}$ is the mean of variable X and $\sigma_{\boldsymbol{X}}$ is the standard deviation of variable X.

2.3 Identification of Outlier, High Leverage Point, and Influential Point

After the contaminations, the detection of outlier, high leverage point and the influential point will be carried out to confirm the presence of unusual observation in the data. The value of standardized residuals can be calculated. [3] suggested a cutoff value of 2 or 3 for this outlier detection method. In this research, a point with an absolute standardized residual greater than 3 is considered as an outlier in our regression model. The hat matrix, **H** in multiple regression is as follows:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \quad (1)$$

The leverage value of an observation is a measure for identifying outlier in the predictor variable, also known as high leverage point. An point is declared as high leverage point if its leverage value, h_{ii} is larger than 2 times of the mean of leverage, [12]. The h_{ii} is the obtained diagonal element of the hat matrix. A leverage point could be an influential point or not. In practice, a point with Cook's Distance or CD_i value higher than the value of 1 is defined as an influential point. The CD_i value can be computed based on Equation (2). Another approach as mentioned in [3], which is by observing graph of the CD_i versus observations. Points with the value of CD_i that stand out from the rest in the graph should be tested as a potential influence to the model. [13] mentioned that CD_i is a very useful method to detect both presence of outliers and high leverage point. The CD_i is given by

$$CD_i = \frac{(e_i^*)^2}{p+1} \times \frac{h_{ii}}{1-h_{ii}},\ i = 1,.2,..,n, \quad (2)$$

where e_i^* is the standardized residual, p is the number of parameters in the regression model and h_{ii} is the leverage value of the i^{th} observation.

2.4 M Estimation

M estimation is a maximum likelihood type estimates which is very sensitive to the presence of outlier in response variable with high efficiency. This is one of the disadvantages of M estimation because it cannot differentiate between good leverage point and bad leverage point as mentioned in SAS User Guide of ROBUSTREG procedure. Due to this drawback, many researches have been carried out to improve the existing formulation of M estimation. Huber-type M estimator denoted as $\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}$ can be obtained by minimizing $Q(\boldsymbol{\beta})$, sum of an increasing function of residuals which increases less rapidly than the function used in OLS estimation. The mathematical formula of $Q(\boldsymbol{\beta})$ as stated by [9] is as follow:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho\left(\frac{e_i}{\sigma}\right), \quad (3)$$

where $\boldsymbol{e_i}$ is the i^{th} residual while σ and ρ represent the standard deviation of residuals and the increasing function which is usually convex.

2.5 Least Trimmed Squares (LTS) Estimation

Another robust is the least trimmed squares (LTS) regression which acts as an extension based on the concept of the trimmed mean, [14]. A trimmed mean is a mean after some percentage of observations in the upper or lower bound of the data is removed. The LTS estimator found in [15] has a breakdown point of 0.5, which is considered very resistant to extreme value as the maximum possible value for the breakdown point to be meaningful is 0.5. LTS estimation has shown a low relative efficiency, although the estimation is highly resistant, [16]. The value of relative efficiency is very low and hence LTS estimator is not desirable as a stand-alone estimator.

However, LTS estimation still has its own advantages and usefulness in helping the extension of other estimators. The LTS estimator, as stated in [15], is a p -vector which is given by the formula:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{LTS}} = arg\min_{\boldsymbol{\beta}} Q_{LTS}(\boldsymbol{\beta}),(4)$$

where $Q_{LTS}(\boldsymbol{\beta}) = \sum_{i=1}^{h} e_{(i)}^2$.The $e_{(1)}^2 \leq e_{(2)}^2 \leq \ldots \leq e_{(n)}^2$ are the ordered squared residuals with $e_{(i)}^2 = (y_i - x_i^T\boldsymbol{\beta})^2$ where $i = 1,2,\ldots,n$and $x_i^T$is the i^{th} row of the $\boldsymbol{X}$ matrix while the trimming constant, h is defined as $\frac{n}{2} + 1 \leq h \leq \frac{3n+p+1}{4}$, and p is number of predictor variable.

2.6 S Estimation

S estimation is another type of robust regression estimation introduced by [17] for investigation with multiple regression cases. The S estimation has the following mathematical expression:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{s}} = arg\min_{\beta} \boldsymbol{S}(\boldsymbol{\beta}),(5)$$

where the dispersion of the estimator, $S(\boldsymbol{\beta})$ is the result of solving the following expression:

$$\frac{\mathbf{1}}{\boldsymbol{n-p}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\chi}\left(\frac{y_i - x_i^T\boldsymbol{\beta}}{\boldsymbol{S}}\right) = \boldsymbol{\theta}.(6)$$

The θ is set as$\int \chi(s)\, d\Phi(s)$ where χ is one type of function such that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{s}}$and $S(\widehat{\boldsymbol{\beta}}_{\boldsymbol{s}})$ are asymptotically consistent estimates of $\boldsymbol{\beta}$ and σ.

The estimate results are obtained by minimizing the residual scale in the M-estimates. The good property of this estimator is that the estimates have high breakdown point and better efficiency if compared to LTS estimation. The drawback of S estimation is that the estimates do not simultaneously have a high breakdown point and efficiency. This estimation method was developed later by [18], who defined S-estimator in a one-way random effects model, improving the upper bound of breakdown point.

2.7 MM Estimation

MM works by utilizing high breakdown estimation such as LTS estimation to obtain the initial estimates before continuing with the procedures of M estimation. [19] recommended S estimation as the initial computation for MM estimation. MM estimation is proven to be an appropriate robust method when dealing with the problem caused by both outlier and high leverage point in regression analysis, as presented in [13].

[20] demonstrated that if the ratio of the number of parameter to the size of the sample is large, the efficiency of MM estimate will be lower than the nominal result. Some solutions to fix the problem are suggested by correcting the scale and making the tuning constant larger. The way to perform MM estimation includes the following procedures:

1. Compute an initial high breakdown value estimate, $\widehat{\boldsymbol{\beta}}'$ by using high breakdown estimation,
2. Find $\hat{\sigma}'$ such that

$$\frac{1}{n-p} \sum_{i=1}^{n} \chi\left(\frac{y_i - x_i^T\widehat{\boldsymbol{\beta}}'}{\hat{\sigma}'}\right) = \theta,(7)$$

where $\theta = \int \chi(s) d\Phi(s)$.

3. Find a local minimum $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$

$$Q_{MM} = \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^T\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}}{\hat{\sigma}'}\right),(8)$$

such that $Q_{MM}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}) \leq Q_{MM}(\widehat{\boldsymbol{\beta}}')$. The algorithm of M estimation is then applied.

2.8 Data Analysis Procedures

Let consider the model of regression be

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, (9)$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ is a vector consists of n responses, $\boldsymbol{X} = (x_{ij})$ is a matrix with $i = 1, \ldots, n$ and $j = 1, \ldots, p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is an unknown vector which contains p parameters or coefficients to be estimated and $\boldsymbol{\varepsilon} = (e_1, \ldots, e_n)^T$ is an unknown error vector include n error terms. The regression model for the Milk Production data is CURRENT= $\beta_0 + \beta_1$ PREVIOUS+ β_3 DAYS+ ε. The Statistical Analysis Software (SAS) is used to assist in the computation. The discussion of the formula that will be presented in the following subsections is the basic complimentary computational method available in SAS.

Some steps are followed in order to carry out the analysis to the Milk Production data. First, the OLS regression analysis is performed on the original data before two different types of contamination are done. Then diagnostics of error term is investigated to make sure the original data fulfill the regression assumptions before we carry out the contaminations to the data set. The contamination onto the first observation in variable Y and variable X_1 are then being carried out separately in two different cases after the diagnostics. Then, the OLS estimation is performed again to the contaminated data and the detection is conducted to confirm the presence of outlier or HLP is carried out. Afterward, four types of Robust Regression, including M, LTS, S, and MM, are conducted to analyze the contaminated data.

3. RESULTS AND DISCUSSION

3.1 The OLS for the Milk Production Data

In Table 1, p-values of *t*-test for each term are all less than 0.05. Applying OLS regression on the original data, both predictor variables significantly affected the response variable. The linear model can explain the linear relationship among these variables Y, X_1, and X_2. Assuming X_2 is fixed, then a unit rise in variable X_1 will cause Y to raise 0.8534 units. On the other hand, when X_1 is fixed, one unit increase in X_2 will decrease Y by 0.03912 units. The value of R^2 indicates that the following fitted linear model can explain 75.65% of the variation in Y.

$$\widehat{\text{CURRENT}} = 16.6439 + 0.8534\text{PREVIOUS} - 0.03912\text{DAYS}.$$

Table 1. Summary of OLS Regression for Original Data, Data with High Leverage Point and Data with Outlier in Milk Production Data

Source	Parameter Estimates	Value of Results	t-test (*p-value*)
Original Data	b_0	16.6439	<.0001
	b_1	0.8534	<.0001
	b_2	-0.03912	0.0001
	RMSE	8.6610	
	R^2	0.7565	
Data with a High Leverage Point	b_0	43.4437	<.0001
	b_1	0.4767	<.0001
	b_2	-0.0607	<.0001
	RMSE	12.8628	
	R^2	0.4630	
Data with an Outlier	b_0	33.3133	<.0001
	b_1	0.7143	<.0001
	b_2	-0.0825	<.0001
	RMSE	15.6959	
	R^2	0.4725	

Compared to the contaminated data set, the RMSE for original data was the smallest (Table 1). The smaller the value of RMSE indicates smaller estimation error. Hence, when the data was not contaminated with HLP or outlier, the OLS regression provides a better result to describe and predict the Milk Production Data. When there is a high leverage point or outlier in the data, the result of the OLS method may provide misleading information. The model becomes inappropriate to express the relationship between predictors and response variables. Both contaminations in response and predictor variables provided lower R^2 value and higher RMSE than the original data. Although all parameter estimates were significant to the fitted model, it is evident that the value of estimation when contamination occurs deviates from the value of original data. The effect of extreme value to the estimated coefficients was clearly shown.

Effects of HLP or outlier on the normality test were displayed in Table 2. The results were obtained based on residuals for the data contaminated with high leverage points or outliers. The normality assumptions for the data in both cases were violated because of the contamination. This finding has shown the effect of unusual point not only to the unreliable parameter of estimates but also disturb the validation of the regression model. Moreover, some diagnostics plots in the presence of HLP or outliers were provided. In Figure 1, the scatter plot of Rstudent versus Leverage displayed a point with a leverage value greater than 0.3, which was classified as HLP as it exceeds the cutoff value of $2\left(\frac{p+1}{n}\right) = 0.04$.

Table 2. Summary of Normality Test of Residuals for Data with High Leverage Point and Outlier in Milk Production Data

Normality Test	p-value		
	Original Data	Data with HLP	Data with Outlier
Shapiro-Wilk	0.0639	<0.0001	<0.0001
Kolmogorov-Smirnov	0.0870	<0.0100	<0.0100
Cramer-von Mises	0.1417	<0.0050	<0.0050
Anderson-Darling	0.1052	<0.0050	<0.0050

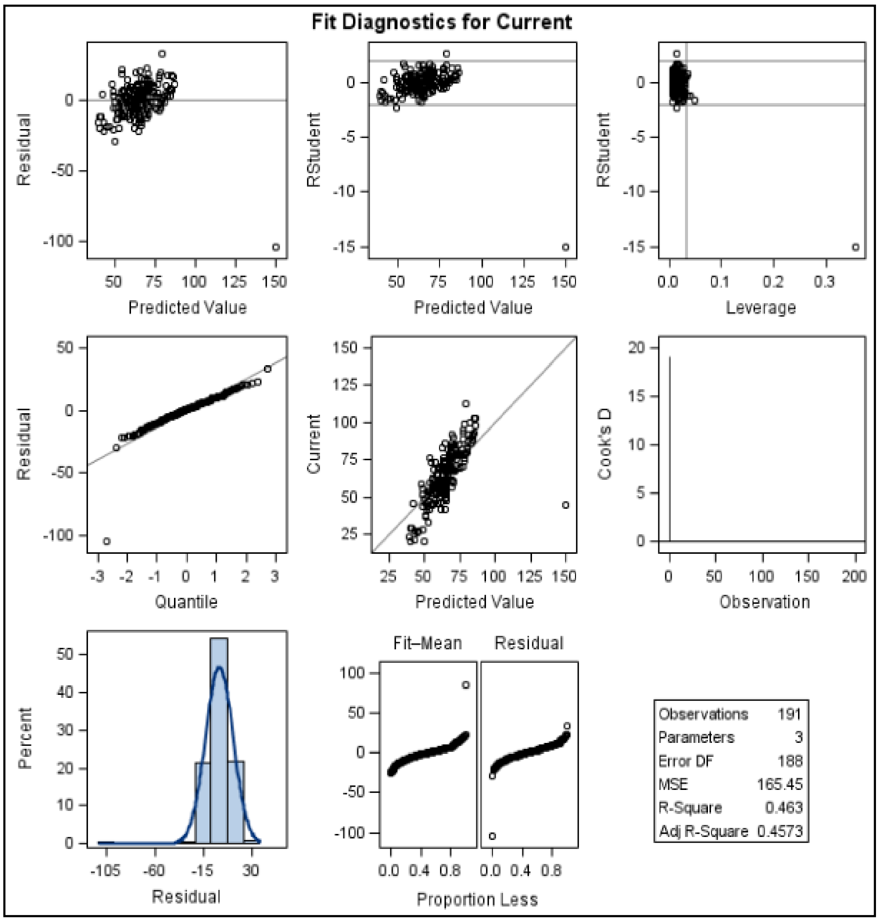

Figure 1. Fit Diagnostics for Milk Production Data when there is a High Leverage Point in the Data

Based on the diagnostic plot in Figure 2, the presence of outlier in the data sets was represented by the point that exceeds the cutoff value of 3 in absolute value for studentized residual versus predicted value plot. The point with a studentized residual value of approximately 20 was identified as an outlier in the response variable. These results have shown that the diagnostics plots efficiently identified the presence of outlier and HLP in the Milk Production Data. However, it could be observed that when contamination occurred only in the response variable, the data also showed up a point that was detected as high leverage point. Therefore, contamination with very extreme value may cause that particular contaminated point to be classified as outlier and high leverage point at the same time. Both Figure 1 and Figure 2 also provided a plot of Cook's Distance values to examine influential points. It was apparent that a point stick out like a sore thumb from the rest of the observations. This indicated an influential point contained inside the data for both the data contaminated with outlier and data contaminated with high leverage point. The presence of influential point may affect the result of regression estimation which is not robust towards extreme values. The estimation may become bias. Therefore, robust methods are used as an alternative way to analyze the data when extreme value is observed from the data set.

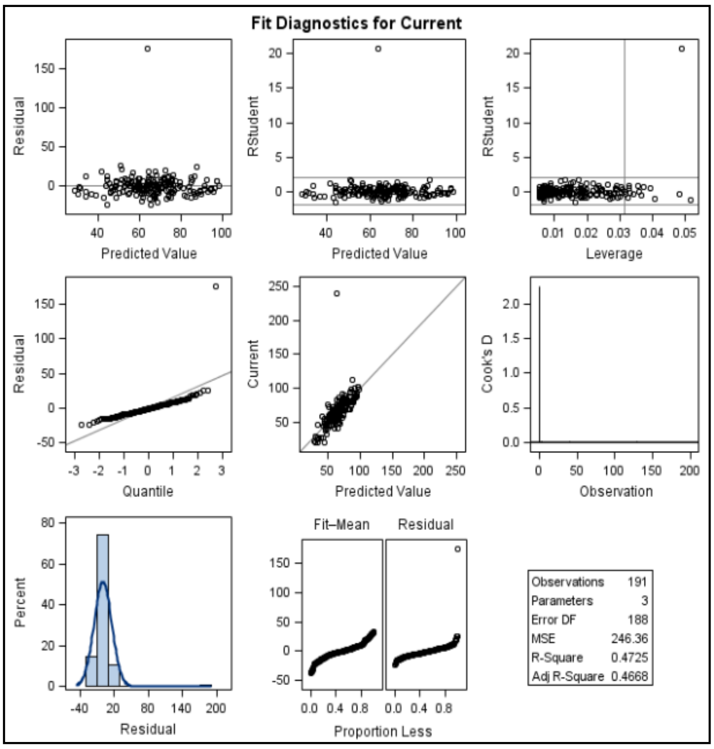

Figure 2. Diagnostics Plot for Milk Production Data when there is an Outlier in the Data

3.2 Robust Regression for the Milk Production Data

In this section, four robust regression approaches were applied to deal with the presence of HLP or outlier in the Milk Production data. Table 3 displays the summary of the results. When there was no outlier or HLP, the value for parameter estimates, b_1 and b_2 of the original data were 0.8534 and -0.03912, respectively. For the parameter estimates in original data, the S estimation results were better than other robust regression approaches. The estimates of the coefficients based on this robust estimation were robust to the presence of HLP as the result of estimation was close to the value obtained in original data. The R^2 was considered favorable as 76.72% of the variation in the response variable, CURRENT can be explained by PREVIOUS and DAYS, which the highest among all regression approaches in the study.

Table 3. Robust Regression in Milk Production Data with Contamination of High Leverage Point (in predictor Variable, X_1)

Parameters Estimate	OLS estimation	M estimation	LTS estimation	S estimation	MM estimation
b_0	16.6439	15.4875	5.4638	12.2736	13.6933
b_1	0.8534	0.8618	0.9669	0.8907	0.8770
b_2	-0.03912	-0.0365	-0.0175	-0.0295	-0.0325
R^2	0.7565	0.6286	0.7878	0.7672	0.5698
Breakdown point			0.2513	0.25	0.2513
Efficiency				0.7589	0.85

Table 4. Robust Regression in Milk Production Data with Contamination of Outlier (in response variable, Y)

Parameters Estimate	OLS estimation	M estimation	LTS estimation	S estimation	MM estimation
b_0	16.6439	15.4875	5.4638	12.2736	13.6933
b_1	0.8534	0.8618	0.9669	0.8907	0.8770
b_2	-0.03912	-0.0365	-0.0175	-0.0295	-0.0325
R^2	0.7565	0.6309	0.7878	0.7715	0.5706
Breakdown point			0.2513	0.25	0.2513
Efficiency				0.7589	0.85

Table 4 displays statistics based on several robust regression approaches for the data set contaminated with outliers. When the contamination was in the response variable, M, S, and MM estimation showed similar results in estimating the regression coefficients. But, S estimation provided the highest R^2.

4. CONCLUSIONS

The contaminations with outlier, high leverage point and influential point are detected well by using studentized residual, leverage and Cook's Distance approach, respectively. The result has shown that the contaminated outlier and high leverage point were influential points. In the presence of the contaminated observations, the OLS method was not robust towards the effect of extreme values. The presence of extreme values will pull the regression line to them and hence the estimated regression line does not represent the whole data well. Therefore, any inference drawn under this situation is not appropriate.

Extreme values in the data set may violate assumptions under ordinary least square regression. When fundamental regression analysis assumptions are violated, there should be an alternative way to solve the problem caused by violation of assumptions. If the data contain influential outliers, it is necessary to employ some robust regressions that may help down-weight the influence of those troublesome outliers. The robust methods considered in this paper are LTS estimation, M estimation, S estimation and MM estimation.

Comparisons of the result among different types of robust estimator and the classical least square estimator have been carried out. M estimation works well when the data is only contaminated in response variable. But in the case of presence of high leverage point, M estimation cannot perform well. This is a drawback of M estimation observed in the result of analysis. Both LTS and S estimation have good performance when high leverage point is present. However, S estimation has better efficiency and breakdown points than LTS estimation.

REFERENCES

[1] D. Alita, A. D. Putra, and D. Darwis, "Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 3, 2021.

[2] A. F. Schmidt and C. Finan, "Linear regression and the normality assumption," *Journal of clinical epidemiology*, vol. 98, pp. 146–151, 2018.

[3] S. Chatterjee and A. S. Hadi, *Regression analysis by example*. John Wiley & Sons, 2015.

[4] M. N. Williams, C. A. G. Grajales, and D. Kurkiewicz, "Assumptions of multiple regression: Correcting two misconceptions," *Practical Assessment, Research, and Evaluation*, vol. 18, no. 1, p. 11, 2013.

[5] K. Ayinde, A. F. Lukman, and O. Arowolo, "Robust regression diagnostics of influential observations in linear regression model," *Open Journal of Statistics*, vol. 5, no. 04, p. 273, 2015.

[6] X. Su and C. Tsai, "Outlier detection," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 261–268, 2011.

[7] C. Yu and W. Yao, "Robust linear regression: A review and comparison," *Communications in Statistics-Simulation and Computation*, vol. 46, no. 8, pp. 6261–6282, 2017.

[8] D. B. Flora, C. LaBrish, and R. P. Chalmers, "Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis," *Frontiers in Psychology*, vol. 3, p. 55, 2012.

[9] Ö. G. Alma, "Comparison of robust regression methods in linear regression," *Int. J. Contemp. Math. Sciences*, vol. 6, no. 9, pp. 409–421, 2011.

[10] D. Huang, R. Cabral, and F. de la Torre, "Robust regression," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 363–375, 2015.

[11] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.
[12] J. R. da Silva Cabral de Moraes, G. de Souza Rolim, and L. E. de Oliveira Aparecido, “Software for the detection of outliers and influential points based on the HAT method,” *Australian Journal of Crop Science*, vol. 11, no. 4, pp. 459–463, 2017.
[13] N. Sorokina, D. E. Booth, and J. H. Thornton Jr, “Robust Methods in Event Studies: Empirical Evidence and Theoretical Implications.,” *Journal of Data Science*, vol. 11, 2013.
[14] J. Kalina, “Three contributions to robust regression diagnostics,” *Journal of Applied Mathematics, Statistics and Informatics*, vol. 11, no. 2, pp. 69–78, 2015.
[15] P. J. Rousseeuw, “Least Median of Squares Regression,” *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 871–880, Dec. 1984, doi: 10.1080/01621459.1984.10477105.
[16] D. M. Khan, M. Ali, Z. Ahmad, S. Manzoor, and S. Hussain, “A New Efficient Redescending M-Estimator for Robust Fitting of Linear Regression Models in the Presence of Outliers,” *Mathematical Problems in Engineering*, vol. 2021, 2021.
[17] P. Rousseeuw and V. Yohai, “Robust regression by means of S-estimators,” in *Robust and nonlinear time series analysis*, Springer, 1984, pp. 256–272.
[18] J. Wellmann, “Robustness of an S-Estimator in the One-Way Random Effects Model,” *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 42, no. 2, pp. 215–221, 2000.
[19] R. A. Maronna, “martin, RD, Yohai, VJ (2006). Robust statistics: Theory and Methods.” Wiley, Chichester, West Sussex, UK.
[20] R. A. Maronna and V. J. Yohai, “Correcting MM estimates for ‘fat’ data sets,” *Computational Statistics & Data Analysis*, vol. 54, no. 12, pp. 3168–3173, 2010.