Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 2 No. 3 2022, 24 - 31
DOI: *https://doi.org/10.22487/27765660.2022.v2.i3.16138*
eISSN: 2776-5660

UNPACKING OUTLIER WITH WEIGHT LEAST SQUARE (IMPLEMENTED ON PEPPER PLANTATIONS DATA)

Rizka Pradita Prasetya[1*]
[1]Magister Program in Applied Statistics, Faculty of Mathematics and Natural Science, Padjadjaran University, Indonesia

***e-mail**: *rizka21010@mail.unpad.ac.id*

ABSTRACT

Outliers in regression analysis can cause large residuals, the diversity of the data becomes greater, causing the data to be heterogenous. If an outlier is caused by an error in recording observations or an error in preparing equipment, the outlier can be ignored or discarded before data analysis is carried out. However, if outliers exist not because of the researcher's error, but are indeed information that cannot be provided by other data, then the outlier data cannot be ignored and must be included in data analysis. There are several methods to deal with outliers. The Weight Least Square method produces good results and is quite resistive to outliers. The WLS method is used to overcome the regression model with non-constant error variance, because WLS has the ability to neutralize the consequences of violating the normality assumption caused by the presence of outliers and can eliminate the nature of unusualness and consistency of the OLS estimate. To compare the level of estimator accuracy between regression models, the mean absolute percentage error (MAPE) is used. Based on the results of this study, it was concluded that the WLS method produced a smaller Mean Absolute Percentage Error value so that the use of this method was more appropriate because it was not susceptible to the effect of outliers.

Keywords: *Outlier, Weight Least Square, Mean Absolute Percentage Error.*

Cite: Prasetya, R. P. (2023). Unpacking Outlier with Weight Least Square (Implemented on Pepper Plantations Data). *Parameter: Journal of Statistics*, *2*(3), 24-31. https://doi.org/10.22487/27765660.2022.v2.i3.16138

[*]Received: Dec 2022; Reviewed: Jan 2023; Published: Jan 2023

INTRODUCTION

Outlier, is an observation in a data set that has a different pattern or value from other observations in the data set. An outlier is a rare or unusual observation that occurs at one of the extremes of most data. The extreme point in the observation is a value that is far or completely different from most of the other values in the group, for example, the value is too small or too large. Outliers are data that do not follow the general pattern or the overall data pattern (Sanford 2005). Outliers can affect the estimation results of regression parameters, and can also cause a violation of the normality assumption of the data. Outliers in regression analysis can cause large residuals from the formed model, the diversity of the data becomes greater, causing the data to be inhomogeneous. If an outlier is caused by an error in recording observations or an error in preparing equipment, the outlier can be ignored or discarded before data analysis is carried out. However, if outliers exist not because of the researcher's error, but are indeed information that cannot be provided by other data, then the outlier data cannot be ignored and must be included in data analysis. However, the analysis of the data obtained can be biased and inefficient so that the regression model obtained does not fit the data modeled with Ordinary Least Square (OLS).

OLS is a method that is often used in estimating the parameters of linear regression models. The estimator generated by OLS will be unbiased and efficient (Best Linear Unbiased Estimator/BLUE) if the residual or error component meets several classical assumptions, namely: normality, homogeneity of variance, and no autocorrelation (Myers 1986). If there is a violation of these assumptions, it is obtained that it is biased and inefficient so that the regression model obtained does not fit the modeled data. There are several methods to overcome outliers, handling outlier can be done by correcting data, containing outliers, eliminating outliers, transforming, modifying, etc (Aguinis, H., Gottfredson, R. K., & Joo 2013). The WLS method is used to overcome the regression model with non-constant error variance, because WLS has the ability to neutralize the consequences of violating the normality assumption caused by the presence of outliers and can eliminate the nature of unusualness and consistency of the OLS estimate. Based on the description above, inspires us to compared which one is better estimation, between the OLS regression estimation method and the Weighted Least Square when an outliers happened.

MATERIALS AND METHODS

Data Source

The data used in this study is data from the Strategic Commodity Survey of Pepper Plants, Bangka Belitung Islands Province, data sourced from the Statistic Indonesia with the data period used is 2021 (BPS 2021).

Variable Definitions

Based on previous research the factors of harvested area, spacing and fertilization affect the productivity of pepper plants and the factors that affect pepper production are land area, number of productive pepper plants and height of pepper plants(Yuhan 2017; Zahara, Rangkuti, and Asnawi 2014). The variables used in this study are as follows:

- Papper Production (Y)
  Pepper production is the amount of black pepper/white pepper harvested in kilograms during January-December 2020.
- Productive Pepper Plants (X_1)
  Productive Plants are plants that are producing and/or have has produced, and is currently producing or is not producing because it is not yet in season.
- Harvested Area (X_2)
  Harvested area is the area of plants that are harvested. If the respondent can only answer in local units, the officer must convert it into square meters (m^2) in accordance with the conversion applicable in the local area.
- Pepper Plant Spacing (X_3)
  Pepper plant spacing at the beginning of planting in cm in the form of multiplication of length and width.

Outlier Identification

The presence of outliers can interfere with the data analysis process, so it is necessary to identify the presence of outliers using a diagnostic method so that outliers can be identified.There are several methods in identifying outliers, namely (1) single-construct techniques; such as boxplot, stem and leaf plot, schematic plot, standard deviation analysis, percentage analysis, (2) multiple-construct techniques; such as scatter plot, qq-plot, standardized residual, studentized residual, leverage value and others, (3) influence techniques; such as cook's distance, DFFITS, DFBETAS and others. In statistical analysis, the box plot is an easy graphic method to find out the extreme data (outliers) of a data suggests that the presence of outliers can be detected by the following methods (Soemartini 2007):

- Box Plot

  The box plot is the most common method using quartile and range values. Quartiles 1,2, and 3 will divide a data sequence into four parts. Range (IQR, Interquartile Range) is defined as the difference between quartile 1 and quartile 3, or IQR=Q3-Q1. Outlier data can be determined, namely values that are less than 1.5*IQR to the 1st quartile and values more than 1.5*IQR to the 3rd quartile. Boxplot is a summary of the sample distribution presented graphically that can describe the shape of the data distribution (skewness), size central tendency and the size of the spread (diversity) of observational data (Williamson, Parker, and Kendrick 1989).

  There are 5 statistical measures that we can read from the boxplot:

  - minimum value: smallest observation value
  - Q1: lowest quartile or first quartile
  - Q2: median or middle value
  - Q3: highest quartile or third quartile
  - maximum value: the largest observation value.

  The boxplot can also show the presence or absence of outliers and extreme values from the observation data.

- DFFITSi and Cook's Distance values can be used to identify whether an observation are an influence observation or not.

DFFITSi

DFFITSi is an influential measure caused by the i-th observation on the estimated value of $\hat{y}$. The equation is given as follows:

$$DFFITS_i = \frac{\hat{y_\imath} - \hat{y}_{\imath-1}}{s_{-i}\sqrt{h_{ii}}} \tag{1}$$

And y_i is the estimated test value, y_{i-1} is the estimated test value without the i-th observation, s_{-i} is the estimated standard error without the i-th observation. h_{ii} are i-th element of the diagonal matrix H. An i-th observation will affect the regression equation if the value of $|DFFITS_i| > 1$, for n $\leq$ 30 and $|DFFITS_i| > 2\left(\frac{p}{n}\right)^{1/2}$for n>30.

Where p represents the number of parameters including intercepts and n represents the number of observations.

Cook's distance

Is a measure of the effect of the ith observation on all estimated regression coefficients. In Cooks Distance the effect of the i th observation is measured by the distance D, the distance is obtained from the following equation:

$$D_i = \frac{(b-b_{-i})^T(X^TY)(b-b_{-i})}{ps^2} = \frac{e_i^2}{ps^2}\frac{h_{ii}}{(1-h_{ii}^2)} \tag{2}$$

With b vector estimation of regression coefficients including observation to i, b_{-i} vector of estimated regression coefficients without observation i-th , e_i residual value at observation i-th, h_{ii} element of i-th, from the diagonal matrix H, p number of parameters including intercepts in the model . Cooks suggests calculating the percentile value of this F-distribution and stating the observation that matters if this percentile is 50%. Cook's 50th percentile recommendation is equivalent to DFFITS > $\sqrt{p}$ (Baltagi 2008).

Multiple Linear Regression Analysis

Regression is a tool to measure the presence of a relationship between the independent variable and the dependent variable. In multiple linear regression there are more than one independent variable. In general can be expressed as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{3}$$

where

Y : dependent variable
x_k : independent variable k-th
β_0 : constant parameter
β_k : independent variable parameter variable k-th
ε : error

Multiple linear regression model requires a linear relationship between independent and dependent variables, as well as non-multicollinearity between independent variables.

Linearity

The linearity test is intended to determine whether there is a linear relationship between the dependent variable and the independent variable. To test linearity, you can look at the QQ Plot or the results of the Ramsey Resettest linearity test. If the p-value of the Ramsey Resettes test is more than alpha = 0.05, it can be concluded that there is a linear relationship between the dependent and independent variables.

Non Multicolinearity

Multicollinearity testing is intended to determine whether there is a correlation between independent variables. The independent variables have no multicollinearity problem if the VIF value is less than 10.

Ordinary Least Square (OLS)

Ordinary least squares (OLS) method is the most widely used method for estimating regression parameters. OLS is a parameter estimation method that minimizes sum square error. The estimator generated by OLS will be unbiased and efficient (Best Linear Unbiased Estimator/BLUE) if the residual or error components meet several classical assumptions, namely: normality, homogeneity of variance/homoscedasticity, and no autocorrelation

Weighted Least Square (WLS)

The WLS method is used to overcome the regression model with non-constant error variance, because WLS has the ability to neutralize the consequences of violating the normality assumption caused by the presence of outliers and can eliminate the nature of unusualness and consistency of the OLS estimate. Some of the observations used in the regression analysis have outliers that lead to deviations from the assumption error. The deviation can be observed from $\mathrm{var}(\varepsilon)\neq I\sigma^2$ but a diagonal matrix with elements on the main diagonal are not the same. For example the regression model as follows:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{4}$$

$E(\varepsilon) = 0, var(\varepsilon) = \boldsymbol{W}\sigma^2$ and $\varepsilon \sim N(0, \boldsymbol{W}\sigma^2)$. Based on this equation, it can be seen that the deviation form causes the OLS formula to be invalid, thus changing the procedure for obtaining the estimated value by using the WLS estimator. The principle of WLS is to find the value of the parameter by adding the weight value, so that (Strutz 2011):

$$\boldsymbol{\beta} = \left(\boldsymbol{X^T W^{-1} X}\right)^{-1} \boldsymbol{X^T W Y} \tag{5}$$

If the the Residuals vs Fitted plot has a non-constant variance pattern so that the amount of variation corresponds to the mean. Use the following procedure to determine the appropriate weights:

- Store residuals and fitted values from OLS regression
- Calculate the absolute value of the OLS residual.
- Regress the absolute value of the OLS residual versus the fitted value of the OLS and save the fitted value of the regression. The fitted value is the estimated value of the standard error.
- Calculate weight is equal to 1/fits^2, where fits is the fitted value of the regression in the last step.
- Then regress using weight

Mean Absolute Percentage Error

To compare the level of estimator accuracy between regression models, the Mean Absolute Percentage Error (MAPE) was used. MAPE is defined by(Khair et al. 2017):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y_i}}{y_i}\right| x\ 100\% \qquad (6)$$

Where y_i is the actual value and $\hat{y_i}$ is the estimated value. The MAPE norm range is [0,100]. The smaller the MAPE value, the better the model is rated. This study aims to obtain a better regression equation than the previous regression equation using OLS for data containing outliers. There are MAPE value categories to see performance or accuracy of models. The category are:

Tabel 1. Mean Absolute Percentage Error (MAPE) category

MAPE	Accuracy
< 10%	The accuracy level is very good
10%-20%	The accuracy level is good
20%-50%	The accuracy level is quite good
>50%	The accuracy level is bad

Data processing in this study using the R studio. Scatter plot and running multiple regression.

RESULTS AND DISCUSSION

Outlier Indentification

Outlier data can be determined, namely values less than 1.5*IQR to quartile 1 and values more than 1.5*IQR to quartile 3.

Table 1. Data Quartile

Quartile	Y	X_1	X_2	X_3
Q1	60	300	1250	2.25
Q2	146	500	2500	2.56
Q3	250	800	5000	2.89
IQR	190	500	3750	0.64
1,5 IQR	285	750	5625	2.085

Table 1. shows that in production (y), productive plants (x_1), harvested area (x_2), and distance (x_3) there are outliers, with data that is a bit far from the distribution above for the third quartile value. The boxplot describes the shape of the data distribution (skewness), a measure of central tendency and a measure of the spread (diversity) of observational data. Based on the boxplot image above, it shows the existence of skewness. If we have a very large range of data, then the smaller values can be attracted by the larger values. Data transformation is the process of taking mathematical functions and applying them to the data. In this section we discuss a common transformation known as the log transformation. Typical log transformations use base 10, base 2 and natural logs. The idea of performing a log transformation is to restore the symmetry of the data.

The plot below is a boxplot for all variables. Based on the boxplot below, it can be seen that the variables, production (y), productive plants (x_1), harvested area (x_2) and pepper planting space (x_3), in the boxplot showed outliers.

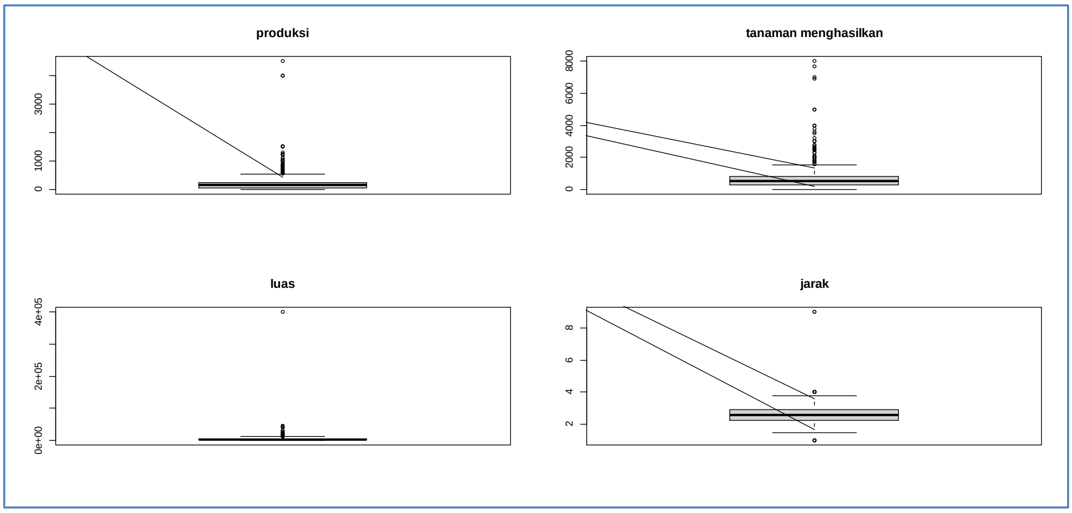

Figure 1. Boxplot of all variables

For this reason, the next analysis uses data variables that have been transformed into natural algorithms.

DFFITS

Next, detect whether the outlier observations have an effect or not. In this study n> 30 then the observation will have an effect if $|DFFITS_i| > 2\left(\frac{p}{n}\right)^{1/2}$. After processing the data with R, the results of the data processing show that the critical value is $2\left(\frac{p}{n}\right)^{1/2} = 0.09652342$. There are observations that have the value $|DFFITS_i|$exceeds the critical value of 93 observations. This means that if the observation is removed from the data set, it will have an effect on the estimated value of $\hat{y}$.

Cook's Distance

Cook's 50th percentile recommendation is equivalent to DFFITS > $\sqrt{p}$ (Velleman and Welsch, 1981). Based on the results of the data processing, the critical cook distance value is $\sqrt{p}$ = 1.732051. There are 6 observations whose DFFITS value is greater than the critical value of cooks distance. This means that if the observation is removed from the data set, it will have an effect on the estimated value of $\hat{y}$.

Multiple Linear Regression Analysis

Ordinary least squares (OLS) method is the most widely used method for estimating regression parameters. The following shows the results of processing with R.

Table 2. Regression Analysis with OLS

Variable	Coefficient	Std. Error	p-value
Intercept	-1.07973	0.17175	4.43e-10
Productive plants (x_1)	0.22171	0.03631	1.35e-09
Harvested area (x_2)	0.69580	0.03355	< 2e-16
Pepper plant spacing (x_3)	0.02155	0.09527	0.821

Linearity

Residuals vs Fitted

This plot is useful for determining whether the residuals show a non-linear pattern. If the red line in the center of the plot is roughly horizontal then we can assume that the residuals follow a linear pattern.

Ramsey Resettest

Based on the Ramsey Resettest test with R, the p-value is 0.9091, greater than alpha = 0.05, so it can be concluded that the linearity assumption is met.

Table 3. Ramsey Resettest

Statistic Test	p-value
Ramsey Resettest	0.9091

Normality

Normal Q-Q

This plot is useful for determining whether the residuals from the regression model are normally distributed. If the points in this plot fall approximately along a straight diagonal line, then we can assume the residuals are normally distributed.

Scale-Location

Plots are useful for checking the assumption of the same variance called homoscedasticity among the residuals in a regression model. If the red line is roughly horizontal across the plot, then the assumption of equal variance is likely met. In the Scale Location above, it can be seen that the assumption of the homoscedastic variance is fulfilled.

Residual vs Leverage

These plots are useful for identifying influential observations. If any point in this plot falls outside of Cook's distance (dashed line) then that is an influential observation. In the Diagnostic Plot above, it can be seen that there is an influential observation.

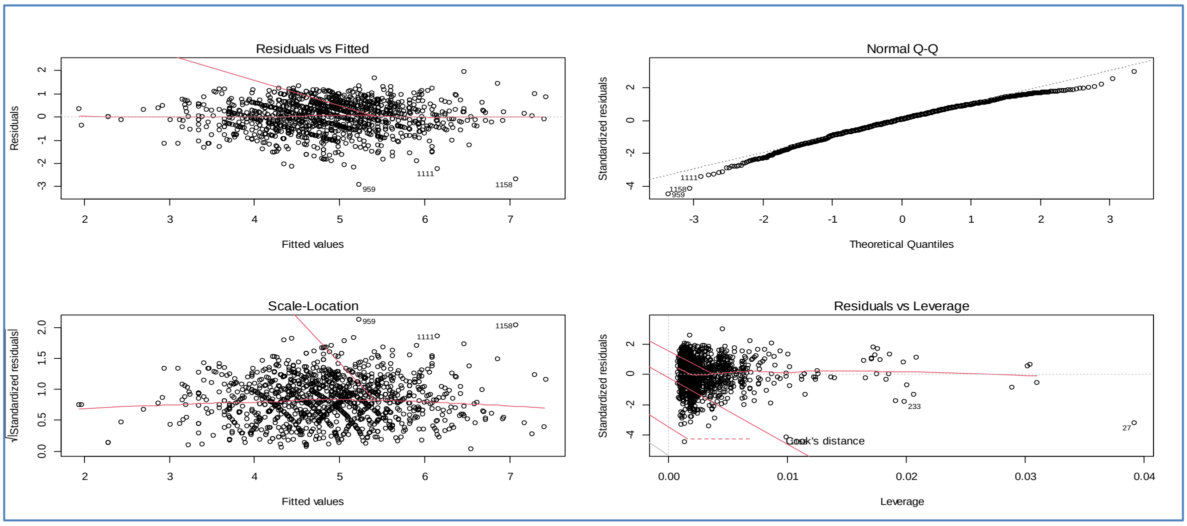

Figure 2. Diagnostic Plot

Multicolinearity

The independent variables do not have multicollinearity problems if the VIF value is less than 10. Based on the VIF value using, there is no variable with a VIF value > 10, so it is concluded that there is no multicollinearity.

Table 5. VIF value

Variable	Coefficient
Productive plants (x_1)	2.468252
Harvested area (x_2)	2.725466
Pepper plant spacing (x_3)	1.199689

Heterogenity dan Autocorrelation

The results of the R output using the Breusch Pagan test showed a p-value less than 0.05, it can be concluded that there are symptoms of heteroscedasticity in the OLS model.

The results of data processing show that the p-value for the Durbin Watson Test is 2.2e-16 less than 0.05, so it can be concluded that there is an autocorrelation.

Weight Least square

Based on the Residuals vs Fitted plot, the plot has a non-constant variance pattern. The following table shows the results with R linear regression analysis with the Weight Least Square method, the estimates made produce different regression equations.

Table 5. Regression Analysis with WLS

Variable	Coefficient	Std. Error	p-value
Intercept	-1.058012	0.169419	5.75e-10
Productive plants (x_1)	0.214914	0.035493	1.84e-09
Harvested area (x_2)	0.703044	0.032661	< 2e-16
Pepper plant spacing (x_3)	0.006565	0.093762	0.944

The regression coefficient values using the OLS and WLS methods are not much different, however, the standard error with WLS is smaller when compared to OLS.

Mean Absolute Percentage Error

In this case, the researcher is looking for a MAPE value which is smaller than the OLS method. If the MAPE value is smaller than OLS then the regression equation value is better. Based on the results with R, the MAPE value of the regression model with OLS is 11,6427 percent and the MAPE value of the regression model with WLS is 11.64113.

Based on the calculation of the MAPE value in the regression model generated by the two methods on the research data, it can be seen that the WLS method produces a smaller value so that the use of this method is more appropriate because it is not susceptible to the influence of outliers. In the OLS method the estimation is very easy to do, but the estimation of the regression model is affected by outlier data so that the regression equation produces a larger MAPE value.

CONCLUSION

Based on the calculation of the MAPE value in the regression model generated by the OLS and WLS methods on the data, it can be concluded that the WLS method produces a smaller value so that the use of this method is more appropriate because it is not susceptible to the influence of outliers. In the OLS method the estimation is very easy to do, but the estimation is affected by outlier data so that the regression equation produces a larger MAPE value. In the WLS method the estimation will be better than OLS. MAPE value of the regression model with WLS is at good level accuracy. There are several methods to overcome outliers, by correcting data, containing outliers, eliminating outliers, transforming, modifying, least trimmed square, etc. The next researcher can use other methods that may be better according to the state of the data.

REFERENCES

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). "Best-Practice Recommendations for Defining, Identifying, and Handling Outliers." *Organizational Research Methods* 16(2): 270–301.

Baltagi, Badi H. (2008). *Econometrics*. Edited by 4th. Berlin: Springer.

BPS. (2021). *Buku Pedoman Pencacahan Survei Komoditas Strategis Tanaman Perkebunan*. Jakarta: Badan Pusat Statistik.

Khair, Ummul, Hasanul Fahmi, Sarudin Al Hakim, and Robbi Rahim. (2017). "Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error." *Journal of Physics: Conference Series* 930 (1). https://doi.org/10.1088/1742-6596/930/1/012002.

Myers, Raymond H. (1986). *Classical and Modern Regression with Applications*. *Classical and Modern Regression with Applications*. Boston, Mass: Duxbury Press.

Sanford, Weisberg. (2005). *Applied Linear Regression*. 3rd ed. John Wiley and Sons Inc.

Soemartini. (2007). *Pencilan (Outlier)*. Bandung: Universitas Padjadjaran.

Strutz, Tilo. 2011. *Data Fitting and Uncertainty A Practical Introduction to Weighted Least Squares and Beyond*. 2nd ed. Springer.

Williamson, D F, R A Parker, and J S Kendrick. (1989). "The Box Plot: A Simple Visual Method to Interpret Data." *Annals of Internal Medicine* 110 (11): 916–21. https://doi.org/10.7326/0003-4819-110-11-916.

Yuhan, Dely. (2017). "Analisis Faktor-Faktor Produktivitas Tanaman Dan Kelayakan Ekonomi Lada (Piper Nigrum L) Di Kabupaten Belitung Timur." Universitas Muhamadiyah Yogyakarta.

Zahara, Marliana S Rangkuti, and Robert Asnawi. (2014). "Analisis Komparasi Usahatani Lada Dan Faktorfaktor Yang Mempengaruhi Produksi Lada Hitam Di Lampung," 765–72.