International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 10 Issue 7 July - 2026, Pages: 256-259

Handling Assumption Violations In A Simple Regression Model

Hasnita

Faculty of Islamic economics and business, State Islamic institute Palangkaraya, Indonesia
Hasnita@uin-palangkaraya.ac.id

Abstract*: Rice is the staple food in Indonesia, and the majority of the population relies on rice as their primary source of carbohydrates. Consequently, rice production is closely associated with population growth and demand. To examine the relationship between rice production and population, a simple linear regression model was developed. However, when the assumptions of linear regression are violated due to the presence of outliers or influential observations, the resulting model may produce biased and unreliable estimates. This study discusses several approaches to addressing assumption violations in simple linear regression. The methods employed include the Box–Cox transformation, logarithmic transformation, and the Weighted Least Squares (WLS) method to improve model performance and ensure more reliable regression results.*

Keywords ***Simple Linear Regression, Violation Of Assumptions, Box-Cox Transformation, Weighted Least Squares.***

1. Introduction

Rice is an important food commodity in Indonesia. According to Prasetiyani and Widiyanto [4], population projections are calculations of the population in the future. One of the uses of population projections is rice supply planning, so that population projections can influence predictions of the amount of rice production in Indonesia. Rice production can be predicted using regression, which is a method to determine the relationship between dependent variables and independent variables. In regression, there are three assumptions that must be met, namely the assumptions of normality, homoscedasticity, non-autocorrelation, and non-multicollinearity (multiple linear regression). In some cases, assumptions are often found that are not met

Effectively addressing these assumption violations in simple linear regression is crucial for enhancing the quality and accuracy of predictions. Recognizing the need for more effective methods to confront these challenges, this research aims to explore strategies and techniques for handling assumption violations in simple linear regression models.

By delving into a deeper understanding of how to address these assumption violations, we can optimize the performance of simple linear regression models and enhance their accuracy. This research not only seeks to provide a more profound comprehension of assumption violations but also offers practical solutions that researchers and practitioners across various fields can implement. The goal is to have a positive impact on the quality of statistical analyses in the context of using simple linear regression models. Several methods applied to this data are the Box-Cox transformation method, logarithmic transformation, and the Weighted Least

Square method. Of the three methods, we will examine which method can correct violations of assumptions in the classical regression model using the least squares method.

2. LITERATUR REVIEW

2.1 SIMPLE LINEAR REGRESSION MODEL

Extreme According to Montgomery et al. [3], the general form of a simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \ ,$$

Where y_i is the dependent variable at the observation i, β_0 is the intercept between the regression line and the axis y, $\beta_1$is the regression coefficient parameter, x_{i1} is the independent variable, n is the number of observations, and ε_i is the remainder i with $\varepsilon_i \widetilde{bsi}\ NID(0,\sigma^2)$ which means that it is independent, there is no autocorrelation between error components in different observations , or in other words the error in one particular period does not depend on the error in other periods $E(\ \varepsilon_i,\ \varepsilon_j) = 0, i \neq j$. Stochastic means that the error (ε_i) is a random real number. Which ε_i can be a positive, negative or 0 value. Each error value has a certain probability, so the error is a random variable. Identical means that the variance of the error (ε_i) is constant for any value of X. The variance ε_i is the same for all observations.$E(\ \varepsilon_i\ )^2 = \sigma^2$

The simple linear regression model has assumptions for the residuals of the resulting model. These assumptions are as follows according to their degree of importance:

a. There is no autocorrelation or independent residuals. Because if the remainder ε_i is influenced by ε_{i-1} then it illustrates Y_i that it is not only influenced by the variable X_t but also by ε_{i-1} where $\varepsilon_{i-1}$to determine the value. If the violation of this assumption is ignored, then the parameter estimator β value is not a BLUE estimator. In addition, the variance of the residuals will cause the variance of the actual observations (σ^2) to be underestimated. Meanwhile, the value R^2 will be overestimated. The test will t and F be invalid and if applied it will produce misleading/misleading conclusions.

b. Homoscedasticity or residual variance is homogeneous. If there is heteroscedasticity it will affect the test t and F make it inaccurate where the var $(\hat{\beta}_1)$ will be very large. Then this will also cause the confidence interval $\hat{\beta}_1$ to widen. So if residuals with different variations are still used, the inference or conclusion will be misleading.

c. Residual normality. Whether the residuals are normal or not does not directly affect the accuracy of the Least Squares Method because the central limit postulate applies which states that random samples n are taken from a very large population with μ and σ^2 then $\bar{X}$ will be normal distribution. However, this assumption is important when making predictions because if there are outliers, the error is not normally distributed and can affect the confidence interval for $\hat{\beta}_1$. (Gujarati) [2]

Assumptions are also needed to guessβincluding: (a) the expected value of the residual is 0, (b) the variance of the residual is homogeneous for each value X, (c) the errors are independent of each other $E(\varepsilon_i, \varepsilon_j) = 0, i \neq j$. Meanwhile, to test the hypothesis β, the assumption ε_i of normal distribution is required $\varepsilon_i \sim N(0, \sigma^2)$. So there are the same assumptions for guessing and testing, namely .β the residuals are normally distributed

The presence of outliers can cause inhomogeneity in the residual variance. So if there is heteroscedasticity in the residual assumption test then this is a result of the presence of outlier observations. If these outlier observations are ignored, they can change the results of the regression analysis substantially. (Gujarati) [2].

2.2 Least Square Method

The least squares method (MKT) has the principle of minimizing the residual sum of squares (JKS). The MKT parameter estimator in matrix form is

$$\hat{\beta} = (X^t X)^{-1} X^t Y \tag{3.1}$$

with

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

2.3 Box-cox Method

Based on a random sample In simple linear regression analysis it is necessary to pay attention to several assumptions regarding the residuals. If these assumptions are not met, then transformations can be carried out on the dependent variables. One of the transformations that can be carried out is the Box-Cox Transformation. This transformation can be carried out on dependent variables that have positive values. This transformation is a power transformation with a single parameter (λ), to Y become $Y\lambda$. The selection of these parameters is λ the one that produces the smallest residual sum of squares (Yati, et al) [7]. In practice, using the Minitab program can help show the most optimal lambda value.

2.4 Weighted Regression Method

Residuals that are not identical $var\ (\varepsilon_i)$ result in not being the same for each i, denoted $var(\varepsilon_i) = \sigma_i^2$ (Winahju, WS) [5], to overcome this, each observation is given a weight. Observations that are far from the average are given little weight. Determining this weighting is not simple, it must be adjusted case by case, a formula cannot be determined for all cases. In general, weighted least squares regression analysis will produce smaller variance for each regression coefficient (Draper and Smith) [1].

3. RESEARCH METHODES

3.1 Research Data

The data used in this research is rice production in 33 provinces (n=33) in Indonesia 2015, and population projections as a factor that can influence it. Data includes rice production in Indonesia (Y) in units of tonnes and projected population in each province (X_1) in units of thousands of people.

3.2 Research Step

The first step, estimate the regression parameters with MKT. The second step, testing the classic assumptions of the regression model. The third step, detect outliers with the studentized detected residual and leverage point values, as well as detecting observations that influence the DFITS value. The fourth step, constructing the model using the methods: (i) Box-Cox transformation, (ii) logarithmic transformation, and (iii) Weighted Least Square. The fifth step, evaluate the model by creating a model comparison table based on the resulting statistics to obtain the best model

4. RESULT AND DISCUSSION

4.1 Regression Model with MKT

The simple linear regression equation obtained using MKT is

$$\hat{y}_i = 5409 + 294x_{i1}$$

4.2 Classic Assumption Test

The classic assumptions that must be met are the normality assumption, homoscedasticity assumption, and non-autocorrelation assumption. The results of testing classical assumptions on the regression model using MKT are shown in the following table.

Figure Plot of residuals checking classical assumptions

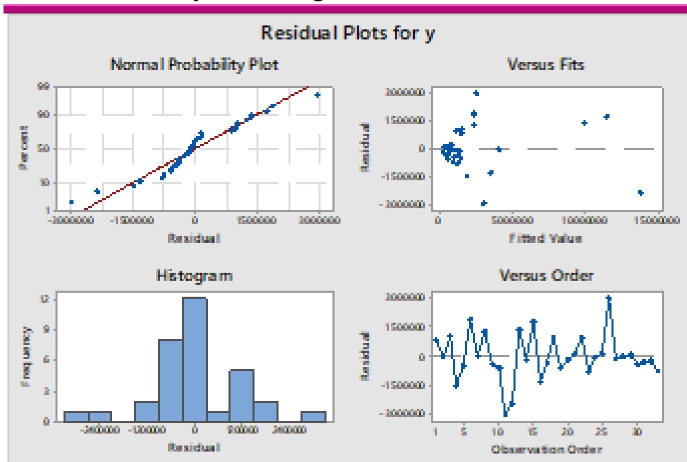

Table 4.1. Classic assumption testing results and conclusions

Classical Assumptions	Conclusion
Normality	Not fulfilled
Homoscedasticity	Not fulfilled
Non Autocorrelation	Fulfilled

Based on Table 4.1, it is found that the assumptions of normality and homoscedasticity are not met. Next, detection is carried out on data that has the potential to become outliers and influential observations.

4.3 Detection of Outlier and Influential Observations

Outliers in the independent variable are detected with the leverage point value (hi) and the critical area is $\left\{hi_i \middle| hi_i > \frac{2(1+1)}{n} = \frac{2(1+1)}{33} = 0.12\right\}$. The results obtained are data 12, 13, and 15 which are outliers in the independent variable. Meanwhile, outliers in dependent variables are detected using the studentized detected residual (TRES) value. The critical area for detecting outliers in the dependent variable is $\left\{TRES_i \middle| |TRES_i| > t_{\frac{\alpha}{2},n-(k+1)-1} = t_{0.025,30} = 2{,}042\right\}$ the results obtained are data 11, 12, and 26 are outliers in the dependent variable. Detection of influential observations uses the DFIT value. The critical area for detecting influential observations is $\left\{DFIT|DFIT| > 2\sqrt{\frac{p+1}{n}} = 2\sqrt{\frac{2}{33}} = 0.492366\right\}$, the results obtained are data 11, 12, 13, 15, and 26.

4.4 Comparison of Regression Equations

Comparison of the resulting Regression Equations is presented in table 4.2.

Table 7.4 Comparison of Regression Equations

	MKT	Box Cox Transformation	Transformation Y'=log Y	WLS
	Y~X	Lambda= 0.5		
Residual assumptions				
Normal	No	Fulfilled	No	No
Homogeneous	No	No	No	No
Non Autocorrelation	Fulfilled	Fulfilled	Fulfilled	Fulfilled
Model				
β_0	Not significant	Significant	Significant	Not significant
β_1	Significant	Significant	Significant	Significant
R	88.26%	71.44%	27.04%	99.28%
PRESS	61468200000000	9589823	21.5856	43895329594037.5
S	1185947	510.127	0.795876	1.00251
Outliers in y	[11],[12],[26]	[11],[26]	[10],[11]	There isn't any
Outliers in x	[12],[13],[15]	[12],[13],[15]	[12],[13],[15]	[2],[28]
Observations matter	[11],[12],[13],[15],[26]	[11],[12]	[10],[11],[12]	[7],[28]

Based on Table 4.2, the three methods that have been tried to overcome assumption violations have not been able to overcome all assumption violations. In Table 4.2 it can be seen that the Box-Cox transformation method can only overcome the assumption of non-normality. Meanwhile, the Y logarithm transformation method and the Weighted Least Square method cannot overcome the two residual assumptions that are violated. According to Williams [4], heteroscedasticity can be caused by inaccurate model specifications or other important variables that have not been included in the model.

5. Conclussion

vthe Box-Cox transform method can only overcome the assumption of abnormalities. Meanwhile, the logarithmic transformation method Y and the Weighted Least Square method have not been able to overcome both assumptions of the violated residuals. According to Williams [4], heteroscedasticity can be caused by improper model specifications or other important variables that have not been incorporated into the model

BIBLIOGRAPHY

[1] Draper, NR and Smith, H. 1992. Applied Regression Analysis. Jakarta.

[2] Gujarati, DN, 2004. Basic Econometrics, 4th edition, The McGraw-Hill Companies, United States.
[3] Montgomery, DC, EA Peck, and GG Vining,. 2007. Introduction to Linear Regression Analysis, 4th edition, Wiley Interscience Publication, New York.
[4] Prasetiyani, I. and D. Widiyanto, Strategy for Facing Food Security (Viewed from Food Needs and Availability) of the Indonesian Population in the Future (2015-2040), 2013. UGM.

[5] Williams, R. Heteroskedasticity. 30 January 2015.https://www3.nd.edu/~rwilliam/.
[6] Winahju, WS, 2013. Weighted Regression (Weighted Least Square). ITS. www.statistics.its.ac.id/wp-content/uploads/2013/09/Regresi-Terbobotiteoririnci.doc
[7] Yati, E. et al. 2013. Box-Cox Transformation in Regression Analysis
Simple Linear. UNAND Mathematics Journal. Vol. 2 No. 2 Pg. 115 -122. Unand.