Basic and Applied Science Conference (BASC) 2021
Volume 2021
http://dx.doi.org/10.11594/nstp.2021.1110

NST PROCEEDINGS

OPEN ACCESS

Conference Paper

The Factors Affecting Soybean Production in Indonesia Using Robust Regression with Least Median of Squares (LMS) Estimation

Aninda Puspa Ratri, Yuliana Susanti*, Isnandar Slamet

Department of Statistics, Sebelas Maret University, Indonesia

Corresponding author:
E-mail: yuliana-susanti@staff.uns.ac.id

ABSTRACT

Soybean is a product with a source of protein that improves the nutrition of Indonesians. The demand for soybeans is increasing, but the domestic production is not sufficient, so that the soybean production in Indonesia must be increased. This study aims to determine the influential factors on soybean production in Indonesia. The data of soybean production in Indonesia had outliers. Outliers cause the residual is not normally distributed so that the assumption of normality is violated. This problem was solved using robust regression. The estimation used was the Least Median of Squares because this estimation has a quite large breakdown point value. The results of the study show that the soybean production in Indonesia was influenced by the field area, the number of soybean seeds, and rainfall. The most influential factor on soybean production is the number of soybean seeds, field area, and rainfall. The attempts that must be conducted by the government to increase soybean production are by having socialization about soybean cultivation and ensuring the availability of soybean seeds in Indonesia.

Keywords: Soybean, robust regression, Least Median of Squares (LMS)

Introduction

Soybean is a plant used as the source of vegetable protein to improve the nutrition of Indonesians. In 2018 it is known that the total soybeans production in Indonesia is 82.598 thousand tons while the demand for soybeans reaches 2.5 tons (Subaedah et al., 2019). So far, the domestic production of soybean has not been able to meet the demand for soybean in Indonesia. Therefore, Indonesia imports soybeans from other countries to fulfill the demand for soybeans in Indonesia. If the consumption increases, the import, and production will increase as well. If import increases, then the consumption increases but production decreases (Ningrum et al., 2018). The domestic production of soybean is able to meet the demand for soybeans in Indonesia of around 30%, while 70% of soybean demands are met by import from other countries. This indicates the need to increase soybean production in Indonesia. Before increasing the soybean production, the factors influencing the soybean production in Indonesia are must be known first. The factors influencing soybean production in Indonesia are the field area, number of soybean seeds, and rainfall (Herawan, 2015).

Based on research conducted by Nasution (2017) using the regression method, it was concluded that the area of rice, corn, and soybeans had a strong influence on production. Based on (Mustikawati et al., 2018), the productivity of soybeans in dry land is 64.25% lower than the productivity of soybeans in the field. Productivity in dry land will be lower if there is a drought factor when the pod is forming and filling. According to (Suhartini, 2018) the area of paddy fields has decreased by 0.24% each year (Ruminta et al., 2020). Explained that there is a significant relationship between the changes in rainfall and soybean production, productivity, and planting

How to cite:
Ratri, A. P., Susanti, Y., & Slamet, I. (2020). The factors affecting soybean production in indonesia using robust regression with Least Median of Squares (LMS) estimation. *Basic and Applied Science Conference (BASC) 2021*. NST Proceedings. pages 70-78. doi: 10.11594/ nstp.2021.1110

area. According to (Rasyid, 2013) the soybean seeds will affect the plant height and the number of soybeans produced.

The regression analysis is a mathematical model that can be used to find out the relationship pattern between two or more variables (Montgomery, 2009). Ordinary Least Square (OLS) is one method that is often used to obtain parameter estimation values in regression modeling. OLS the classical assumption requirements are required so that the model can be considered as a good model. Some cases of classical assumption requirements were not fulfilled. One of the causes is outliers. Outliers cannot be simply discarded because they might include important information. In order to resolve this problem, a robust method is required. This method is known as robust regression (Febrianto et al., 2015).

Robust regression is a method used when the distribution of residual data is not normal, or several outliers affect the model. In the robust regression, there are several estimations, such as M estimation, S estimation, MM estimation, LMS estimation, and LTS estimation. The least Median of Squares (LMS) is a robust regression estimation method by minimizing the median and residual squares (Rousseeuw & Leroy, 1987). LMS has a high breakdown point of 50%. The breakdown point is one method to measure the robustness of an estimator. The greater the breakdown point value, the estimator will be more robust. According to Daniel (2019) the LMS estimator shows better result when compared to the OLS estimator because the resulting regression equation has a smaller error value.

The data of soybean production in Indonesia in 2014 have outliers. Outliers cannot be simply discarded because it is important information. Therefore, this study used robust regression with LMS estimation to determine the factors influencing soybean production in Indonesia.

Material and Methods

Regression analysis

The regression analysis is a mathematical model that can be used to find out the relationship pattern between two or more variables. Simple linear regression is a regression model that has only one indepemdent variable (Neter et al., 1983). The multiple linear regression model is the extension of a simple linear regression model. In general, it can be written as:

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_j \quad (1)$$

where:

Y_i : dependent variable of i-data.

X_j : independent variable of j-data.

B_k : regression coefficient parameter.

F-Test and t-Test

F-Test is used to find out whether the independent variable has a linear relationship with the dependent variable or not. F_{value} is compared to F_{table} at a significant level of 5%. If F_{value} is greater than F_{table}, so there is at least one independent variable has an affect. The formula of the F-test is:

$$F = \frac{\text{Mean square regression}}{\text{Mean square error}}. \quad (2)$$

$F_{table} = F_{(k-1,n-k-1\,;\,\alpha),}$ k: the number of independent varibale, n: the number of data. (Sugiono, 2008)
The t-test is used to find out the significance among dependent variables. The formula of the t-test is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (3)$$

$t_{table} = t_{(n-k-1\,;\,\alpha)}$, k: the number of independent variable, n: the number of data.

where
t : t-test value.
r : correlation coefficient.
n : the number of data.

The conclusion was drawn by comparing t_{value} with t_{table}. If t_{value} is greater or equal to t_{table} with a significance level of 5%, then the variable has a significant effect (Sugiono, 2008).

Classical assumption test

The classical assumption test is a requirement for a model considered as a good model, and it can be used in an analysis. The classical assumption are normality test, non-multicollinearity test, homoscedasticity test, and non-autocorrelation test.

Normality Test

The normality test aims to determine if the residual in a regression model is normally distributed (Ghozali, 2011). The most frequently used statistical test is by Kolmogorov-Smirnov Test. Where statistical test:

$$D = \max_x[F_n(x) - F_0(x)] \quad (4)$$

where:
$F_n(x)$: Normal Cumulative Probability.
$F_0(x)$: Empirical Cumulative Probability.
The critical area is H_0, which is rejected if $D > W_{(1-\alpha)}$ value or sig< α value, where $W_{(1-\alpha)}$ is a quantile 1- α on the two-sided Kolmogorov-Smirnov table.

Non-multicollinearity test

The non-multicollinearity test is used to find out whether there is a correlation among independent variables in the regression model or not. A good regression model is a regression that does not contain multicollinearity (Ghozali, 2011). To measure the collinearity is by using Variance Inflation Factors (VIF) with the formula:

$$VIF_j = c_{jj} = \frac{1}{\left(1-R_j^2\right)} \quad (5)$$

where:
j : 1,2, … , k.
k : the number of the independent variable.
R_j^2 : the coefficient of determination R^2.
If the VIF value > 10, it shows strong multicollinearity.

Homoscedasticity test

Homoscedasticity Test is used to examine whether regression occurs an inequality of variance of the residual from one data to another. The requirement that must be met in the regression model is the absence of heteroscedasticity indications (Ghozali, 2011). This test used Breusch-Pagan.

$$\phi = \frac{1}{2}(ESS) \quad (6)$$

where:
ESS : Explained sum of squares.
H_0 rejected if $\phi_{value} > \chi^2_{(\alpha;p-1)}$

Non- Autocorrelation test

The non-autocorrelation test aims to examine whether there is a correlation between the error in periodt and the error in the previous period in the linear regression model. If there is autocorrelation, then it is called autocorrelation problem (Ghozali, 2011). One method that can be used to detect the presence of autocorrelation is Durbin-Watson (DW Test) Test. The decision of whether there is autocorrelation is:

Table 1. Decision of Durbin Watson test

Critical Area	Result
When the DW value is between DU and 4 – DU	no autocorrelation.
When the DW value is smaller than DL	positive autocorrelation.
When the DW value is greater than 4 – DL	negative autocorrelation.

Outlier detection

Outlier data is data that is far (extreme) from other data. The method used to identify can be by DFFITS (Difference fitted value FITS) method. The difference fitted value FITS is a method showing the value of change in the predicted value if a particular case is released and has been standardized (Dewi et al., 2016). The formula of DFFITS is as follows:

$$\mathrm{DFFITS_i} = \mathrm{t_i}\sqrt{\frac{\mathrm{h_{ii}}}{1-\mathrm{h_{ii}}}} \tag{7}$$

where $\mathrm{t_i} = \mathrm{e_i}\sqrt{\frac{\mathrm{n-k-1}}{\mathrm{SSR}(1-\mathrm{h_{ii}})-\mathrm{e_i^2}}}$

$\mathrm{e_i}$ is residual i, SSR is the sum of squared residuals, and $\mathrm{h_{ii}}$ is leverage value. Data is considered an outlier if the $|\mathrm{DFFITS}| > 2\sqrt{\frac{\mathrm{k}}{\mathrm{n}}}$ value and k are the numbers of parameters in the model, and n is the amount of data.

Robust Regression

Robust regression is a method used to overcome the problem of outliers. This method is an importand tool to analyzing data affected by outliers. When the data in linear regression has outliers and result in an abnormally distributed model, the robust regression model can be used (Rousseeuw & Leroy, 1987).

LMS Estimation

The basic principle of the robust regression LMS estimation method is to match most of the data after the outlier has been identified as a point that is not related to data (Rousseeuw & Leroy, 1987). In OLS, the thing that needs to be done is by minimizing the residual squares ($\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{e_i^2}$), then in LMS, the thing that needs to be done is by minimizing the median of residual squares:

$$\mathrm{M_j} = \min\left\{\mathrm{med}\ \mathrm{e_i^2}\right\} = \min\{\mathrm{M_1}, \mathrm{M_2}, \dots, \mathrm{M_s}\} \tag{8}$$

where $\mathrm{e_i^2}$ is residual squares of the estimated results by OLS.
The method to obtain the $\mathrm{M_1}$ value is by finding the subsets of matrix X of $\mathrm{h_I}$ data is:

$$\mathrm{h_i} = \mathrm{h_1} = \left[\frac{\mathrm{n}}{2}\right] + \left[\frac{\mathrm{p+1}}{2}\right] \tag{9}$$

where n is the amount of data, and p is the number of parameters.
According to (Rousseuw, 1984), w_{ii} weight is formulated with the following conditions:

$$w_{ii} = \frac{\psi(\varepsilon_i^*)}{\varepsilon_i^*} \tag{10}$$

W_{ii} weighter is determined based on the function of a weighter:

$$\psi(\varepsilon_i^*) = \begin{cases} \varepsilon_i^* \text{ if } |\varepsilon_i^*| \leq 2.5 \\ 2.5 \text{ if } \varepsilon_i^* > 2.5 \\ -2.5 \text{ if } \varepsilon_i^* < -2.5 \end{cases} \tag{11}$$

where $\varepsilon_i^* = \frac{e_i}{\hat{\sigma}}$ and $\hat{\sigma} = 1.4826 \left[1 + \frac{5}{n-p}\right] \sqrt{M_j}$.
After the w_{ii} weight is calculated, the W matrix can be formed as follows:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} \tag{12}$$

where the w_{ij} matrix entry= 0, where $i \neq j$.

Data source and variables of the study

The data of the study were obtained from the publication of Badan Pusat Statistik: *Statistik Indonesia 2020* and the publication of Setjen Pertanian RI: *Statistik Sarana Pertanian 2019.* The variables used in this study were the number of soybean productions according to the province of 2019 (Y), field area (X_1), the number of soybean seeds (X_2), and rainfall (X_3).

Analysis method

The steps for modeling the soybean production in Indonesia with LMS estimation were by:

1. Modeling with multiple linear regression.
2. Conducting classical assumption test.
3. Detecting outlier.
4. Estimating the regression coefficient $\widehat{\beta_i}$ with the least square method.
5. Calculating the value of the residual square (e^2).
6. Calculating the median of residual squares.
7. Calculating the $h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right]$.
8. Sorting the variables according to the least residual squares, then cut them according to h value.
9. Estimating β_{new} with h data.
10. Doing the first to sixth steps until obtaining the convergent β value.
11. Calculating w_{ii}.
12. Estimating the regression coefficient $\hat{\beta}_{LTS}$ with the w_{ii}.

Results and Discussion

The modeling of soybean production in Indonesia of 2019 with OLS was:

$$\hat{Y} = 29523 + 0.0184\, X_1 + 84.3\, X_2 - 12.5\, X_3.$$

Based on the model above, it can be known that the field area value (X_1) increased by one unit, and other variables were considered constant, then it will increase soybean production by 0.0184. If the value of the number of soybean seeds (X_2) increased by one unit and other variables were considered constant, it would increase the soybean production by 84.3. If the rainfall (X_3) increased by one unit and other variables were considered constant, it would decrease the soybean production by 12.5. This modeling had an r-square of 91.9%. The variables of soybean production

(Y) can be explained by field area variable (X_1), the number of soybean seeds (X_2), and rainfall (X_3) of 91.9%, while the remaining 8.1% was explained by other variables.

The classical assumption test is a requirement for a model considered as a good model, and it can be used in an analysis. In this study, normality test used Kolmogorov-Smirnov with the following results:

Table 2. The results of normality test

Kolmogorof Smirnov Value	**P-value**
0.58824	0.001588

Based on the table 1, it was obtained the p-value = 0.001588 < 0.05. Therefore, it can be concluded that residual data was not normally distributed. Residual data that was not normally distributed must be conducted an outlier detection. The next assumption test was a non-multicollinearity test. The measureent of multicollinearity used VIF.

Table 3. The results of non-multicollinearity test

Variables	**VIF**
Field area (X_1)	3.488
The number of soybean seeds (X_2)	3.476
Rainfall (X_3)	1.016

Based on table 2, it was obtained the VIF value < 10. Therefore, it can be concluded that there was no multicollinearity. The next assumption test was homoscedasticity using Breusch-Pagan.

Table 4. The results of homoscedasticity test

Breusch-Pagan Value	**P-Value**
3.9507	0.2668

On the table 3, it was obtained the p-value = 0.2668 < 0.05. Therefore, it can be concluded th at the data was homogeneous or no heteroscedasticity indications. The last assumption test was non-autocorrelation test using Durbin Watson.

Table 5. The results of non-autocorrelation test

Durbin Watson Value	**dL**	**dU**
2.117002	1.27074	1.65189

DW value = 2.117002 was between dU = 1.65189 and 4 – dU = 2.34811. Therefore, it can be concluded that there was no autocorrelation.

When tested for the classical assumption of normality, it can be concluded that the residual data was not normal. Not normal residual data can be caused by the outlier. The outlier was detected using DFFITS. Based on the data with p = 4 and n = 34. Then, $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{4}{34}} = 0.6859$.

Table 6. Outlier data

No	Data	DFFITS Value
1.	3	0.79066
2.	12	3.30507
3.	15	4.13025
4.	18	1.04707

The results of identification showed that there were four outliers. Data on table 5 was outlier because | DFFITS | > $2\sqrt{\frac{p}{n}}$.

Table 7. The results of F-Test

F Value	P-value
113.14	0.000

Based on table 6, it was known that F = 113.14 > $F_{(2;30;0.05)} = 0.302$ or p-value < 0.05, so there is at least one independent variable has an affect rainfall affects the soybean production in Indonesia.

Table 8. The results of t-Test

Variables	t Value	P-Value
Field area (X_1)	0.89	0.381
The number of soybean seeds (X_2)	9.18	0.000
Rainfall (X_3)	-2.49	0.018

Based on table 7 it was known that $|t_{test}| > t_{(30;0.05)} = 2.042$ or p-value < 0.05 and then the number of soybean seeds and rainfall significantly affect the soybean production in Indonesia. Meanwhile, for $|t_{test}| < t_{(30;0.05)} = 2.042$ or p-value > 0.05, the field area did not significantly affect the soybean production in Indonesia.

After having F-test and t-test, then it was continued by analyzing the modeling with LMS estimation. The modeling of soybean production in Indonesia of 2019 was:

$$\hat{Y} = 16440 + 0.0254X_1 + 85.7X_2 - 6.84X_3.$$

Based on the model above, it can be known that if field area value (X_1), the number of soybean seeds (X_2) and rainfall (X_3) are constant, the soybean production in Indonesia is 16440 tons. If the field area value (X_1) increased by one unit, and other variables were considered constant, then it will increase soybean production by 0.0254. If the value of the number of soybean seeds (X_2) increased by one unit and other variables were considered constant, it would increase the soybean production by 85.7. If the rainfall (X_3) increased by one unit and other variables were considered constant, it would decrease the soybean production by 6.84. This modeling had an R^2adjusted = 96.8% and $R^2 = 97.1\%$ which means that the variable of soybean production (X_2) can be explained by field area variable (Y), the number of soybean seeds (X_1), and rainfall (X_3) of 97.1%, while the remaining 2.9% was explained by other variables.

Table 9. The results of F-Test

F Value	P-value
336.32	0.000

Based on table 8, it was known that F = 336.32 > $F_{(2;30;0.05)} = 0.302$ or p-value < 0.05, then so there is at least one independent variable has an affect the soybean production in Indonesia.

Table 10. The results of t-Test

Variables	t Value	P-Value
Field area (X_1)	3.86	0.001
The number of soybean seeds (X_2)	20.01	0.000
Rainfall (X_3)	-2.65	0.013

Based on table 9, it was known that $|t_{test}| > t_{(30:0.025)} = 2.042$ or p-value < 0.05, then the variables of the field area, the number of soybean seeds, and rainfall significantly affect the soybean production in Indonesia.

The soybean production model in Indonesia is estimated using OLS. The result of classical assumption test are that the normality test is not fulfilled, the homoscedasticity test is fulfilled, the non autocorrelation test is fulfilled and the non multicollinearity test is fulfilled. If the normality test is not fulfilled, outliers will be detected. There are 4 outliers in data. Then the LMS estimation will be used for soybean production data in Indonesia. In table 8, the F-test value of the LMS estimation model is more than F-table, so there is at least one independent variable has an affect. T-test was conducted to find out what variables were influental. In table 9 it can be concluded that the significant independent variables are the field area, the number of soybean seeds and rainfall.

Conclusion

The modeling of soybean production in Indonesia in 2019 with LMS estimation was $\hat{Y} = 16440 + 0.0254X_1 + 85.7X_2 - 6.84X_3$ with R^2 of 97.1%. The most influential factors on the soybean production in Indonesia were the number of soybean seeds, field area, and rainfall.

Acknowledgment

I am especially gratefull for Dra. Yuliana Susanti, M.Si and Drs. Isnandar Slamet, M.Sc. Ph.D. for their contiuous encouregement and give advice for my research. I also thank all to the lecturers of Departement of Statistics, Sebelas Maret University who have help me complete my studies.

References

Badan Pusat Statistik. (2020). *Statistik Indonesia 2020.* Jakarta: Badan Pusat Statistik.

Daniel, F. (2019). Mengatasi Pencilan pada Pemodelan Regresi Linear Berganda dengan Metode Regresi Robust Penaksir LMS. *Jurnal Ilmu Matematika dan Terapan*, 13(3), 145-156. https://doi.org/10.30598/barekengvol13iss3pp145-156ar884

Dewi, E. T., Agoestanto, A., & Sunarmi. (2016). Metode Least Trimmed Square (LTS) dan mm-estimation. *Journal of Mathematics, 5*(1), 1-8. Doi: https://doi.org/10.15294/ujm.v5i1.13104

Febrianto, L. S. (2015). Perbandingan Metode Robust LMS dan Penduga S untuk Menangani Outlier pada Regresi Linear Berganda. *Unnes Journal of Mathematics, 7*(1), 1-7. https://doi.org/10.15294/ujm.v7i1.27381

Ghozali, I. (2011). *Aplikasi analisis dengan program SPSS.* Semarang: Universitas Diponegoro.

Herawan, W. (2015). Model Strategi Percepatan Ketersediaan Kacang Kedelai Melalui Sistem Manajemen Lapangan Terpadu Dalam Mendukung Ketahanan Pangan di Kalbar. *Jurnal Agrosains*, 12(2), 1-8.

Montgomery, D. C. (2009). *Introduction to statistical quality control.* United States of America: John Wiley & Sons.

Mustikawati, D. R., Mulyanti, N., & Arief, R. W. (2018). Productivity of Soybean on Different Agroecosystems. *International Journal of Environtment, Agriculture and Biotechnology (UEAB)*, 3(4), 1154-1159. http://dx.doi.org/10.22161/ijeab/3.4.1

Nasution, A. (2017). Luas Panen dan Produksi Padi, Jagung, Kedelai terhadap ProdukDomestik Regional Bruto Propinsi di Indonesia Pada Program Peningkatan Pangan Pajale. *Junrla Bisnis Tani,* 1(1), 178-192. Doi:10.35308/jbt

Neter, J., Wasserman, W., & Kutner, H. M. (1983). *Applied linear regresiion models.* United States of America: Richard D. Irwin, INC.

Ningrum, I. H., Irianto, H., & Riptanti, E. (2018). Analysis of Soybean Production and Import Trends and Import Factors in Indonesia. *Journal Earth and Environmental Science*, 142.

Pertanian, K. (2019). *Statistik sarana pertanian Tahun 2019.* Jakarta: Pusat Data dan Sistem Informasi Pertanian.

Rasyid, H. (2013). Peningkatan produksi dan mutu benih kedelai varietas hitam unggul nasional sebagai fungsi jarak tanam dan pemberian dosis pupuk P. *Jurnal Gamma*, *49*(1), 1-5.

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

Rousseuw, P. J. (1984). Least Median of Square. *Journal of the American Statistical Association*, *79*(388), 871-880.

Ruminta, Irwan, A. W., Nurmala, T., & Ramadayanty, G. (2020). Analisis Dampak Perubahan Iklim Terhadap Produksi Kedelai dan Pilihan Adaptasi Strategisnya pada Lahan Hujan di Kabupaten Garut. *Jurnal Kultivasi*, 19(2), 115-123. Doi : https://doi.org/10.24198/kultivasi.v19i2.27998

Subaedah, Said, N. S., & Ralle, A. (2019). *Petunjuk teknis budidaya kedelai di lahan Sub Optimal.* Makassar: Fakultas Pertanian, Universitas Muslim Indonesia.

Sugiono. (2008). *Metode penelitian kuantitatif, kualitatif, dan R&D.* Bandung: Alfabeta

Suhartini, S. H. (2018). Analisis sumber sumber pertumbuhan produksi kedelai. *Jurnal Analisis Kebijakan Pertanian*, 16(2), 89-108. Doi: http://dx.doi.org/10.21082/akp.v16n2.2018.89-109

Wulandari, S., Sutarman, & Darnius, O. (2013). Perbandingan Metode Least Trimmed Squares dan Penaksir M dalam Mengatasi Permasalahan Data Pencilan. *Saintia Matematika*, *1*(1), 73-85.