Indonesian Journal of Applied Statistics
Volume I No.1 May 2018 ISSN. 2621-086X

A Robust Regression by Using Huber Estimator and Tukey Bisquare Estimator for Predicting Availability of Corn in Karanganyar Regency, Indonesia

Hasih Pratiwi[1], Yuliana Susanti[2], and Sri Sulistijowati Handajani[3]

[1,2,3]Statistics Study Program, Universitas Sebelas Maret, Jl. Ir. Sutami 36A, Surakarta 57126, Indonesia
[1]hpratiwi@mipa.uns.ac.id
[2]yulianasusanti@staff.uns.ac.id
[3]rr_ssh@staff.uns.ac.id

Abstract. Linear least-squares estimates can behave badly when the error distribution is not normal, particularly when the errors are heavy-tailed. One remedy is to remove influential observations from the least-squares fit. Another approach, robust regression, is to use a fitting criterion that is not as vulnerable as least squares to unusual data. The most common general method of robust regression is M-estimation. This class of estimators can be regarded as a generalization of maximum-likelihood estimation. In this paper we discuss robust regression model for corn production by using two popular estimators; i.e. Huber estimator and Tukey bisquare estimator.

Keywords : robust regression, M-estimation, Huber estimator, Tukey bisquare estimator

1. Introduction

Corn is one of the important food crops besides rice and wheat. Some people in Indonesia such as in Madura use corn as a staple food which has advantages and benefits as the highest source of carbohydrates [1,4,5,12]. Because of the importance of addressing the needs of the food, we require an effort to predict production in the future. There are several methods that can be used to predict corn production as well as to investigation several factors that influence it, such as regression analysis [6,10].

The problems that arise in the regression analysis is to determine the best estimators for model parameters, which is heavily influenced by the use of the method. For example, using the least squares method would not be appropriate in solving problems contains several outliers or extreme observations, or the assumption of normality can not be met. By using regression analysis, the production of which go far beyond the production can generally be categorized as an outlier, so using the least squares method to estimate the regression parameters is less precise [2,13]. To overcome this problem, we require a parameter estimation method which is robust. Robust interpreted as insensitivity or resilience to small changes of assumptions. Estimation using the maximum likelihood estimate (MLE) will produce an estimator of the same

nature as the least squares method, so MLE is also not robust to the influence of outliers. A robust technique that is often used is the M-estimation which is an extension of the MLE [7,11]. The purpose of this study is to determine the regression model for predicting corn production in Karanganyar regency using M with Huber estimator and Tukey bisquare estimator.

2. Robust Regression

When the observations $\boldsymbol{y}$ in the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ are normally distributed, the method of least squares works well in the sense that it produces an estimate of $\boldsymbol{\beta}$ that has good statistical properties. However, when the observations follow some non normal distribution, particularly one that has no longer or heavier tails than the normal, the method of least squares may not be appropriate [3].

A number of authors have proposed robust regression procedures designed to dampen the effect of observations that would be highly influential if least squares method were used [7,8,13]. A robust procedure tend to leave the residuals associated with large deviation, thereby making the identification of influential points much easier. In addition to sensitivity to outliers, a robust estimation procedure should be $90 - 95$ percent as efficient as least squares when the underlying distribution is normal [9].

We consider the linear model

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$
$$= \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon$$

for the ith of n observations. If an estimator for $\boldsymbol{\beta}$ is $\boldsymbol{b}$ and the residuals are given by

$$e_i = y_i - \hat{y}_i$$

we may define a class of robust estimators that minimize a particular objective function

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(\boldsymbol{y_i} - \boldsymbol{x}_i'\boldsymbol{b})$$

where the function ρ gives the contribution of each residual e to the objective function. A reasonable ρ should have the following properties:

1. always non negative, $\rho(e) \geq 0$,
2. $\rho(0) = 0$,
3. symmetry, $\rho(e) = \rho(-e)$,
4. monotone in $|e_i|, \rho(e_i) \geq \rho(e_i{}')$ for $|e_i| > |ei'|$.

Table 1 compares the objective functions and the weight functions for three estimators, i.e. the least square estimator, the Huber estimator, and the Tukey bisquare (or biweight) estimator.

Table 1: Objective function and weight function for least square, Huber, and Tukey bisquare estimators

Method	Objective function	Weight function
Least square	$\rho_{LS}(e) = e^2$	$w_{LS}(e) = 1$
Huber	$\rho_H(e) = \begin{cases} \frac{1}{2}e^2, \text{for} \mid e \mid \le k \\ k \mid e \mid - \frac{1}{2}, \text{for} \mid e \mid > k \end{cases}$	$w_H(e) = \begin{cases} 1, \text{for} \mid e \mid \le k \\ \frac{k}{\mid e \mid}, \text{for} \mid e \mid > k \end{cases}$
Tukey bisquare	$\rho_T(e) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{e}{k}\right)^2\right]^3\right\}, \text{for} \mid e \mid \le k \\ \frac{k^2}{6}, \text{for} \mid e \mid > k \end{cases}$	$w_T(e) = \begin{cases} \left[1 - \left(\frac{e}{k}\right)^2\right]^2, \text{for} \mid e \mid \le k \\ 0, \text{for} \mid e \mid > k \end{cases}$

e: residual, *k:* constant.

3. Research Method

The data used in this research is secondary data obtained from the Indonesian Ministry of Agriculture and the Central Agency of Statistics (BPS) Karanganyar regency in 2015. The data include the production of corn (*Y*, in ton), harvested area (X_1, in hectare), and rainfall (X_2, in mm) in 15 districts in Karanganyar. The following present an algorithm to estimate parameters of robust regression model:

1. Estimate regression coefficients on the data using the least square method.
2. Test assumptions of the classical regression model.
3. Detect outliers in the data.
4. Calculate initial parameters using least square method.
5. Calculate residuals $e_i = y_i - \hat{y}_i$.
6. Calculate $\hat{\sigma}_i = 1.4826$ MAD, where MAD is median of absolute deviation.
7. Calculate $u_i = e_i/\hat{\sigma}_i$.

8. Calculate Huber weight function:

$$w_i = \begin{cases} \left[1-\left(\frac{u_i}{4.685}\right)^2\right]^2, |u_i| \le 4.685 \\ 0, |u_i| > 4.685 \end{cases}$$

9. Calculate Tukey bisquare weight function:

$$w_i = \begin{cases} \begin{cases} \left[1-\left(\frac{u_i}{1.547}\right)^2\right]^2, |u_i| \le 1.547, \text{iteration} = 1 \\ 0, |u_i| > 1.547 \end{cases} \\ \frac{\rho(e)}{u_i}, \text{iteration} > 1 \end{cases}$$

10. Estimate parameter of the weighted least square method, b_M, with the weight w_i.
11. Repeat steps 5-10 until b_M convergent.
12. Test to determine whether independent variables have significant effect on the dependent variable.

4. Results and Discussion

From corn production data in Karanganyar regency in 2015, the estimated corn production linear regression model (Y) in Karanganyar area based on least square method is

$$\hat{Y} = 2.80 + 6.95X_1 + 0.59X_2 \qquad (1)$$

where adjusted determination coefficient R-square (adj) $= 100\%$ and standard deviation $s = 28.7081$. From the value of $F = 64147.96$ and p-value $= 0 < 5\%$, it indicates that the linear regression model Y with X_1 and X_2 are good. Then we test to see whether the assumptions of linear regression model (1) are met or not. From the test results of assumption (normality, homoscedastic, non-autocorrelation and non-multicolinearity) it turned out tthat only the normality assumption is not fulfilled (Figure1), and there are three outliers of data that is data-2, 11 and 13. Furthermore, since the assumption of normality is violated and there are three outliers, then we estimate a robust regression model by using M-estimation

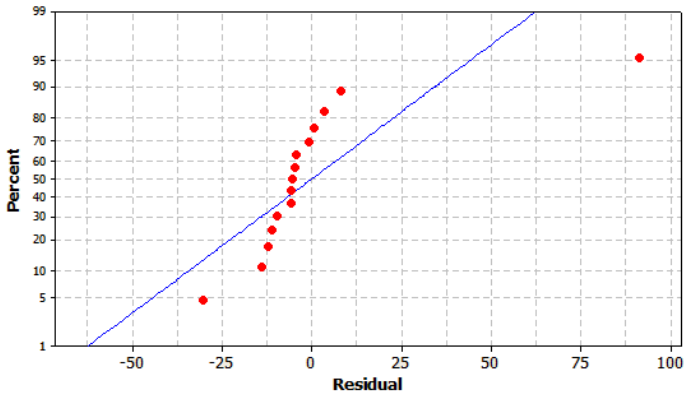

Figure 1. Normality plot of residual for Y

Based on M estimation with Huber function, we obtain the linear regression

$$\hat{Y} = -0.447 + 6.953X_1 + 0.543X_2 \qquad (2)$$

where adjusted R-square $= 100\%$ and standard deviation $s = 9.76$. Furthermore, we test to determine whether the harvest area and rainfall have influence on corn production (Table 2). From Table 2, it can be concluded that the harvested area has significant influence on corn production, while the rainfall does not have significant influence on the production of corn in Karanganyar.

Table 2: Significance Test of Prediction Model with Huber Estimator

Variable	Coefficient	t	p	Conclusion
Constant	$-0{,}447$	-0.060	0.953	Not significant
X_1	6.953	919.580	0.000	Significant
X_2	0.543	1.200	0.254	Not significant

Furthermore, by the Tukey bisquare weighted linear regression model is obtained as follows

$$\hat{Y} = -1.910 + 6.955X_1 + 0.592X_2 \qquad (3)$$

where adjusted R-square $= 100\%$ and $s = 4.69064$. We test to determine whether the harvested area and rainfall have an influence on corn production in Karanganyar or not.

From Table 2 it can be concluded that the harvested area and rainfall have significant influence on corn production in Karanganyar.

Table 3: Significance Test of Prediction Model with Tukey Bisquare Estimator

Variable	Coefficient	t	p	Conclusion
Constant	-1.910	-0.520	0.615	Not significant
X_1	6.955	1801.850	0.000	Significant
X_2	0.592	2.640	0.025	Significant

Table 4: M-Estimation Result using Huber Estimator and Tukey Bisquare Estimator

	Huber	Tukey bisquare
Adjusted R^2	100%	100%
s	9.760	4.691
Significant variable	X_1	X_1, X_2
Model	(2)	(3)

We determine the best regression model by using two criteria, namely the adjusted determination coefficient R^2 (R^2adjusted) and standard deviation s. The best model would have larger R^2 adjusted and smaller s. Table 4 shows that the model (2) and (3) provide the same adjusted R^2, but model (3) has smaller s than model (2), so that these results support previous residual analysis that the best regression model is the model (3). For model (3), R^2adjusted $= 100\%$ indicates that the total variation Y of 100% is explained by X_1 and X_2 (Figure 2).

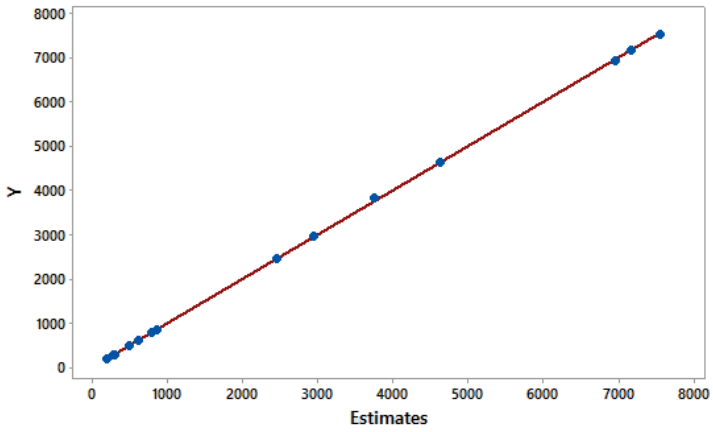

Figure 2. Plot data and estimate

5. Conclusion

Based on the results and discussion, the optimal model for corn production in Karanganyar was obtained by Tukey bisquare estimator. The robust regression model showed that for every increment one hectare of harvested area and one mm of rainfall will increase the corn production as 6.95 tons and 0.592 tons respectively.

Acknowledgements

The author would like to thank the Institution of Research and Community Service (LPPM) Universitas Sebelas Maret (UNS) which provides financial support through Grant of Maintenance Research Group (MRG) 2017.

References

[1] Abdulrachman. Statistik Pertanian (Agricultural Statistics). Departemen Pertanian Indonesia. Jakarta. ISBN: 979-8958-65-9. 2008.
[2] Barnett, J. and Lewis. Outlier in Statistical Data. New York: John Wiley & Sons Inc. 1978.
[3] Birkes, D. and Dodge, Y. Alternative Methods of Regression. New York: John Wiley & Sons Inc. New York. 1993.
[4] Heriawan, R. Statistik Indonesia. Badan Pusat Statistik Indonesia, Katalog BPS: 1101001. Jakarta. 2008.
[5] Heriawan, R. 2008. Produksi Tanaman Pangan. Badan Pusat Statistik Indonesia, Katalog BPS: 5203014. Jakarta. 2008.
[6] Husni, M., Sudi, M., and Mewa, A. Faktor-faktor yang Mempengaruhi Produksi, Konsumsi dan Harga Beras serta Inflasi Bahan Makanan. Jurnal Argo Ekonomi, 2004. 22(2): 119-146.
[7] Koenker, R. and Portnoy, S. M Estimation of Multivariate Regressions. Journal of The American Statistical Association. 1990. 85: 1060-1068.

[8] Li, S.Z., Wang, H., and Soh, W.Y.C. Robust Estimation of Rotation Angle from Image Sequences Using the Annealling M Estimator . Journal of Mathematical Imaging and Vision. 1998. 8: 181-192.
[9] Miguel, A.A. Convergence of the Optimal M-Estimator over a Parametric Family of M-Estimators. 2003.
[10] Montgomery, D.C. and Peck, E.A. Introduction to Linear Regression Analysis. New York: John Wiley & Sons Inc. 2006.
[11] Rousseeuw, P.J., Aelst, S.V., Driessen, K.V., and Gulló, J.A. Robust Multivariate Regression. Journal Technometrics. 2004. 46: 293-305.
[12] Suarni and Yasin, M.. Jagung sebagai Sumber Pangan Fungsional. Iptek Tanaman Pangan. 2011. 6(1): 41-46.
[13] Susanti, Y., Pratiwi, H., and Handajani, S.S. Paddy Availability Modelling in Indonesia Using Spatial Regression. IAENG International Journal on Applied Mathematics (IJAM). 2015. 45(4): 398-403.