© 2018 IJSRSET | Volume 4 | Issue 9 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

M-Estimation Use Bisquare, Hampel, Huber, and Welsch Weight Functions in Robust Regression

Tiyas Yulita[1], Khairil Anwar Notodiputro[2], Kusman Sadik[3]

[1]Department of Statistics, Faculty of Science and Technology, Nahdlatul Ulama Lampung University, Indonesia

[2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Indonesia

ABSTRACT

The estimation by the least squares method (LSM) is often used in simple or multiple regression model. However, it was not uncommon for the response variables in model which contain contamination or outliers. LSM is known will be very sensitive to these problem, so if LSM is still used in regression then parameter estimate can be bias. Robust regression is well known as a method that robust from effect of outliers in order to obtain better result from LSM. The paper will discuss the methods of M-estimation to model the response data which contain the outliers using Bisquare, Hampel, Huber, and Welsch weight function using simulation data and HDI (Human Development Index) data in West Java Province. On the HDI data, the M estimations prediction method with the Welsch weight function yields $R^2 = 89{,}28\%$, the best of other weight functions.

Keywords: M-Estimator, Robust Regression, Weight Function.

I. INTRODUCTION

Linear regression is one of popular method in modeling data. Regression model is used to know the relationship between predictor and response variables. To model data use linear regression, we often use estimator method because it is unbiased and has minimum varians. The prediction of LSM is obtained by minimizing the sum of squares of error between the actual data and its estimate. Modeling with LSM requires several assumptions that must be fulfilled, one of these is the error normality. If this assumption is not fulfilled then the LSM estimator may be biased and no longer efficient (minimum variety [8]. LSE is known to be sensitive to the presence of outliers, one outlier has a considerable effect on the predicted outcome. This paper will discuss about the method of M estimator which is known as capable method to produce a robust regression model on the existence of outliers.

The robustness of M estimator will be evaluated by simulation data which is conditioned in some situations, response variable is consist of 5%, 15%, and 20% outliers in data. We use various weight functions to find out the function which is give best result. In addition to simulation data, robust regression is also applied to model HDI (Human Development Index) in West Java Province (Indonesia).

II. METHODS AND MATERIAL

A. M- Estimator

M-estimator was introduce by Huber (1964)as a method which is used in robust regression model. Robust regression is used when error's distribution is not normal so it will affect the model. This method is used to find out which data is suspected as a outliers and then analyze it to be a robust model. A point that is far from the line of conjecture (with a big error) can be considered as an outlier[6]. The most commonly way to detect an outlier is standardized residual or standardized error obtained from the error divided by the estimates of standard error. Many researchers conduct data transformations to eliminate or reduce the effect of the outliers which obtain bias parameter estimator, but this is considered less practical because we have to determine the appropriate transformation first. So robust regression is present to provide an alternative way of overcoming outliers of data [7] . The M estimator uses a simple approach between computation and theory. The basic principle of estimator M is to minimize the objective function:

$$\sum_{i=1}^{n}\rho(\varepsilon_i^*) = \sum_{i=1}^{n}\rho\left(\frac{\varepsilon_i}{\hat{\sigma}}\right) = \sum_{i=1}^{n}\rho\left(\frac{y_i - x_i\hat{\beta}}{\hat{\sigma}}\right) \qquad (5)$$

where :

ε_i^*: scale of *i*-th residuals ;

$\rho(\varepsilon_i)$: a function which is give contribution in every residuals to objective function;

$\hat{\sigma}$: scale parameter.

Fox (2002) said that a good function ρ is:

1. $\rho(\varepsilon) \geq 0$
2. $\rho(0) = 0$
3. $\rho(\varepsilon) = \rho(-\varepsilon)$
4. $\rho(\varepsilon) \geq \rho(\varepsilon_{i'})$ for $|\varepsilon_i| > |\rho_{i'}|$

As example, to least square estimation $\rho(\varepsilon_i) = \varepsilon_i^2$. If $\psi = \rho'$ is a differential of ρ, the minimization of equation (5) is :

$$\sum_{i=1}^{n}\psi\left(\frac{y_i - x_i\hat{\beta}}{\hat{\sigma}}\right)x_i = 0 \qquad (6)$$

$\psi(\cdot)$ is a function to obtain weight function $w_i = \psi(\varepsilon_i^*)/(\varepsilon_i^*)$, equation (6) will be:

$$\sum_{i=1}^{n}w_i\left(\frac{y_i - x_i\hat{\beta}}{\hat{\sigma}}\right)x_i = 0 \qquad (7)$$

or

$$X^TWX\hat{\beta} = X^TWy \qquad (8)$$

Equation (8) minimize $\sum_{i=1}^{n}w_i(yi - \hat{y}_i)^2$, This equation is said as weighted least square. This method will be used to obtain M estimator. So the estimation of parameters are:

$$\hat{\beta} = (X^TWX)^{-1}X^TWy \qquad (9)$$

The solution of equation (8) is obtained by IRLS (Iteratively reweighteed least square) method. [4]

B. Methods

We use $\psi(\cdot)$ function to obtain Bisquare, Hampel, Huber dan Welsch weight functions. First, we use West Java's HDM (Human Development Index) data in 2016 with contamination effects to evaluate the robust regression. Second, we generate simulation data which contain 5%, 15%, and 20% outliers to evaluate $\psi(\cdot)$ functions. The simulation's steps are:

1. Generate residuals vector $(\varepsilon^{(0)})$ ~ $(N(0,1))$, the size is 100×1.
2. Generate the elements of vector $\boldsymbol{x}$ ~ $(U(0,5))$, the size is 100×1.
3. Generate outliers ventor ($\boldsymbol{p}^{(\boldsymbol{n})}$) with size 100×1 , and element value are 0 except element as outliers. We give 5%, 15%, dan 20% outliers on data.
4. Sum of $\boldsymbol{p}^{(n)}$ and $\boldsymbol{\varepsilon}^{(\mathbf{0})}$ as contamination residuals (vektor $\boldsymbol{\varepsilon}$**),** where $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^{(0)} + \boldsymbol{p}^{(n)}$
5. Set $\beta_0 = 1$ and $\beta_1 = 1$, generate vector $\boldsymbol{y}$, where $\boldsymbol{y} = 1 + \boldsymbol{x} + \boldsymbol{\varepsilon}$. Then $\boldsymbol{x}, \boldsymbol{y}$ as sample data.
6. Estimate regression coefficients base on model $\boldsymbol{y} = \hat{\boldsymbol{\beta}}\boldsymbol{x} + \boldsymbol{\varepsilon}$ with use $\boldsymbol{x}, \boldsymbol{y}$ from (5) then use M method with functions $\psi(\cdot)$, these are Bisquare, Hampel, Huber, and Welsch.
7. Step (1) to (6) is repeated 10 times then save the result of $\hat{\beta}_0$, $\hat{\beta}_1$.
8. Obtain value of $KTG(\hat{\beta}_0) = \frac{1}{m}\sum_{i=1}^{m}(1 - \hat{\beta}_0)^2$, and $KTG(\hat{\beta}_0) = \frac{1}{m}\sum_{i=1}^{m}(1 - \hat{\beta}_1)^2$; where m = 10.

The scale estimate $\hat{\sigma}$ in this paper is obtained by Huber function from iteration on follow equation:

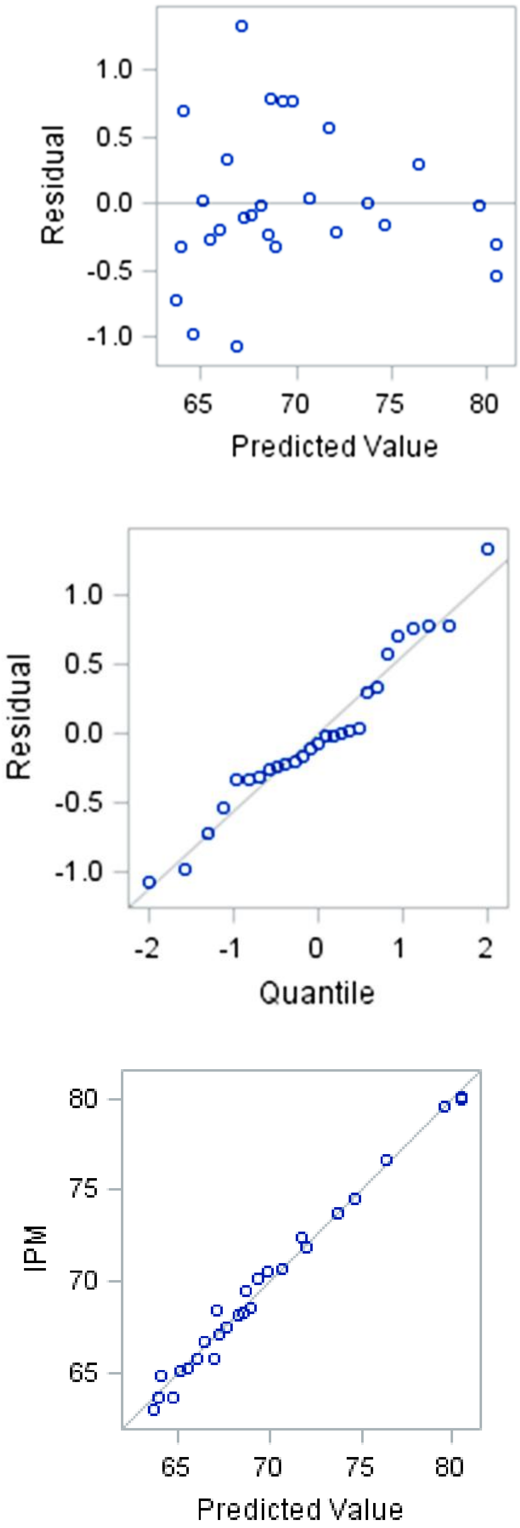
$$(\hat{\sigma}^{(m+1)})^2 = \frac{1}{nh}\sum_{i=1}^{n}\chi_d\left(\frac{y_i - x_i\hat{\beta}}{\hat{\sigma}^{(m)}}\right)\left(\hat{\sigma}^{(m)}\right)^2 \qquad (10)$$

where $\chi_d(x) = \begin{cases} \frac{x^2}{2} & if\ |x| < d \\ \frac{d^2}{2} & others \end{cases}$

is Huber function and

$$h = \frac{n-p}{n}(d^2 + (1 - d^2)\Phi(d) - 0{,}5 - d\sqrt{2\pi}e^{-\frac{1}{2}d^2} \qquad (11)$$

is Huber constant, where $d = 2{,}5$. [6], and the weight function $W(\cdot)$ are:



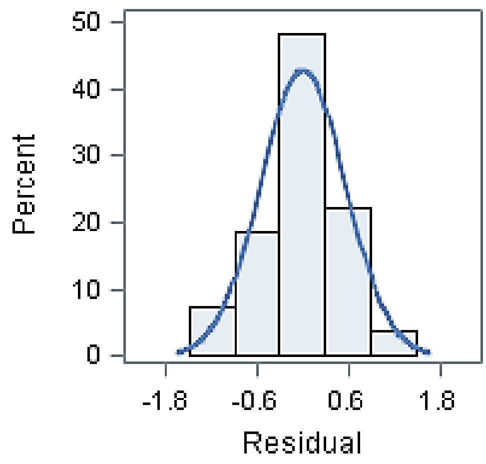


Figure 1: Plot of residuals by Least Square Method from HDI data

1. Bisquare

$$W(x,c) = \begin{cases} (1-(x/c)^2)^2, & if|x| < c \\ 0, & others \end{cases}$$

2. Hampel

$$W(x,a,b,c) = \begin{cases} 1, & |x| < a \\ \frac{a}{|x|}, & a < |x| \leq b \\ \frac{a}{|x|}\frac{c-|x|}{c-b}, & b < |x| < c \\ 0, & others \end{cases}$$

3. Huber

$$W(x,c) = \begin{cases} 1, & if|x| < c \\ \frac{c}{|x|}, & others \end{cases}$$

4. Welsch

$$W(x,c) = exp\left(-\frac{(x/c)^2}{2}\right) \quad [5]$$

R^2 value will be obtained by formula:

$$R^2 = \frac{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right) - \sum \rho\left(\frac{y_i - x_i\hat{\beta}}{\hat{s}}\right)}{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right)}$$

where $\rho' = \psi$, $\hat{\mu}$ is location estimate on M estimator, and $\hat{s}$ is parameter estimate of scale on full model [1].

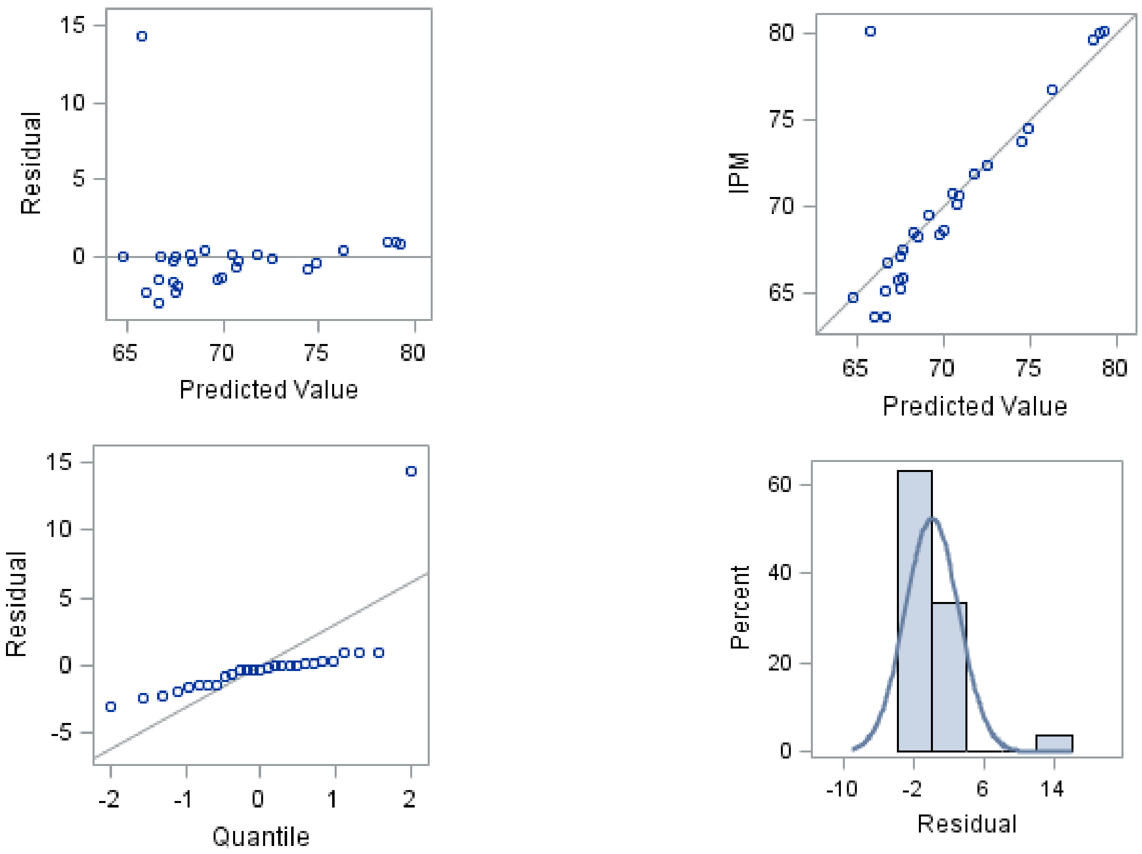


Figure 2 : Plot of residuals by Least Squate Method from HDI data with contaminations

TABLE 1. MEAN SQUARE ERROR (MSE) OF PARAMETER ESTIMATES FROM SIMULATION DATA

Percentage of outliers	Parameter estimate	LSM	Robust Regression use Weight Function			
			Bisquare	Hampel	Huber	Welsch
5%	$\hat{\beta}_0$	10,7553	0,0677	0,0392	0,0662	0,0417
	$\hat{\beta}_1$	25,2679	0,0059	0,0057	0,0100	0,0070
15%	$\hat{\beta}_0$	41,8418	0,1736	0,0935	0,6568	0,0901
	$\hat{\beta}_1$	4,3129	0,0086	0,0127	0,1164	0,0096
20%	$\hat{\beta}_0$	19,3043	0,8659	1,7714	0,3001	2,5519
	$\hat{\beta}_1$	21,1952	0,9771	2,3541	2,9865	2,0190

TABLE 2. PARAMETER ESTIMATES FROM HDI DATA

Parameter estimate	Without contamination (LSM)	With contamination (LSM)	Robust Regression use Weight Function			
			Bisquare	Hampel	Huber	Welsch
$\hat{\beta}_0$	19,7382	63,8005	20,6030	21,8629	23,2906	21,1355
$\hat{\beta}_1$	0,3729	-0,2376	0,3629	0,3435	0,3265	0,3547
$\hat{\beta}_2$	1,7017	1,9586	1,7123	1,7142	1,7289	1,7131
$\hat{\beta}_3$	0,0009	0,0007	0,0009	0,0010	0,0009	0,0009
R^2	98,76%	65,71%	85,68%	70,19%	80,08%	89,28%

In the data analysis of HDI (Human Development Index), the response variable representing HDI from 26 regencies / cities in West Java with three predictor variables, namely LEI (Life Expectancy Index), ALI (Adult Literacy Index), and RpC (Revenue per Capita). The original data is not contain outlier, so the data is given contamination by changing the value of HDI in Cianjur regency from 62,92 to 80,13.

III. RESULTS AND DISCUSSION

A. Simulation Data

Table 1 provides the MSE values for the parameters estimate. TheMSE value for estimate parameter on the generated data with 5% outliers is MSE ($\hat{\beta}_0$) =10,7553 dan MSE $(\hat{\beta}_1) = 25{,}2679$. that parameter estimation with LSE gives result which very different from its parameter. All of $\psi(\cdot)$ functions obtain good enough result. In the 15% outliers, provide results on predictions with LSE that remain ineffective for use with data containing the outliers. Likewise on the data with 20% contamination of outliers. If we look at the four functions of w (·) in the M estimateor, then all four functions are considered good enough to model the data with outliers 5% and 15%, but when 20% , the use of the Hampel, Huber, and Welsch functions no longer effective, so the Bisquare function is recommended to use

B. HDI Data

The use of M estimation method on HDI data by using four functions w (·) obtained the results in table 2. It can be seen that data which has been given contamination that is HDI of Cianjur Regency gives $R^2 = 65{,}71\%$, it's very different from result in uncontaminated data that is $R^2 = 98{,}76\%$,,it's means that in the presence of one contamination can give big influence to result of analysis. For more details we can see in Fig. 1, where there is no outlier when we look at some of plots that are formed. In contrast to Fig. 2 which is a plot with data that has been contaminated. The results of the analysis on the data with contamination are ALI variable obtain significant coefficient at α = 0.05, whereas the coefficient for LEI and RpC variable were not significant. Having used the method of M-estimates on the data with contamination it is known that the use of the Welsch function to form the weights on the M estimator gives the best result with the model :

$$\hat{y}_i = 21{,}1355 + 0{,}3547LEI + 1{,}7131ALI + 0{,}0009RpC$$

with $R^2 = 89{,}28\%$,, followed by bisquare (85.68%), Huber (80.08%), and Hampel (70.19%). On the use of these four functions on the M estimator produces a coefficient value that are all significant.

IV. CONCLUSION

The ability of M predictor to overcome the outliers of the evaluated data using the four weight functions w (·) yields a fairly good estimate of the HDI data, and Welsch function is best to use. While the simulation data, the four functions are considered good enough to model the data with 5% and 15%, but need to be aware when contaminations percentage reach 20% of the data, if we use Hampel, Huber, and Welsch functions considered no longer effective, so the recommended function to use is bisquare.

V. REFERENCES

[1]. Chen, C. 2002. "Robust Regression and Outlier Detection with the ROBUSTREG Procedure." In Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc. http://www2.sas.com/proceedings/sugi27/p265-27.pdf.

[2]. Chun, Y., Weixin, Y., Xue,B. 2014. Robust Linear Regression: A Review and Comparison. arXiv: 1404.6294v`stat.ME24 April 2014.
[3]. Fox,J.2002. Robust Regression. Appendix To an R and S-Plus Companion to Applied Regression.
[4]. Herawati, N., Nisa, K., Setiawan, E. 2011. Analisis Ketegaran Regresi Robust Terhadap Letak Pencilan: Studi Perbandingan. Bulletin of Mathematics. Vol. 03, No. 01 (2011), pp. 49–60.
[5]. Holland, P. W., dan Welsch, R. E. 1977. "Robust Regression Using Iteratively Reweighted Least-Squares." Communications in Statistics—Theory and Methods 6:813–827.
[6]. Huber, P.J.1981. Robust Statistics. New York: John Wiley & Sons
[7]. Ryan, T.P. 1997. Modern Regression Methods. A Wiley-Interscience Publication, New York.
[8]. Yaffee, R.A. 2002. Robust Regression Analysis: Some Popular Statistical Package Options. Statistics, Social Science, and Mapping Group.