A COMPARISON OF M-ESTIMATION AND S-ESTIMATION ON THE FACTORS AFFECTING IR DHF IN EAST JAVA IN 2017

Mardiana [1], Arief Wibowo[2], Mahmudah[2], Pipit Festi W[3]
[1] Bachelor Degree of Public Health Department, Faculty of Health and Pharmacy, Universitas Muhammadiyah of East Borneo, Borneo, Indonesia
[2] Biostatistics and Demography Department, Master's degree of Public Health Department, Faculty of Public Health, Universitas Airlangga, Surabaya, Indonesia
[3] Faculty of Health Sciences, Universitas Muhammadiyah of Surabaya, Surabaya, Indonesia
Correspondence Addres : Mardiana
Email : mar348@umkt.ac.id, Phone : +6281351414752

ABSTRACT

Robust regression on M estimation and S estimation is the Ordinary Least Square (OLS) regression on the data outlier. East Java is one of the provinces in Indonesia with a high case fatalitiy rate (1.34%). The raising of Dengue Haemoragic Fever (DHF) in East Java has been influenced by climate, population density, human behavior, and environmental sanitation. This study aimed to compare robust regression research by using M estimation and S estimation on the factors that affect IR DHF. This was done to get the best model regression on the data outlier based on the biggest R^2 adjusted and the smallest MSE. This study utitlized observational research with a non-reactive research design using secondary data. The independent variable consisted of population density, healthy behavior, healthy living environment house, and precipitation in East Java in 2017. The dependent variable was incident rate of DHF in 2017. The population included 38 regencies in East Java, while the sample was 35 regencies/cities selected using simple random sampling. The analysis used robust regression on M estimation and S estimation weighting by Tukey's Bisquare. Robust regression on S estimation was found to be the best robust regression on data outlier with R^2 adjusted (0.996) and MSE (0.229). Robust regression on S estimation was $\hat{y}$ = 54.826 + 0.011 (population density) – 0.136 (% healthy behavior) - 0,404 (% healthy house) - 0,005 (precipitation). Some factors that affect IR DHF can be the main focus for the prevention and control of DHF for the government and society.

Keywords: robust regression, outlier, estimation, estimation, DHF

INTRODUCTION

The analysis of linier regression is analysis used for the prediction of correlation between one or some independent variables and dependent variables (Pramana *et al.*, 2017). The Ordinary Least Square (OLS) method is the general method for parameter estimation regression by minimizing the number of residual on regression model. One of conditions that need to be filled on OLS method is the data distribution must be normal (Gujarati, 2013). If the data are abnormal, it could have been influenced by the condition of data outlier (Herawati *et al*, 2011). The OLS method has the weakness of being sensitive towards the data outlier (Lainun and Tinungki, 2018).

Outlier is a condition when the data are abnormal with different characters than the general data (Draper and Smith, 1992). The existence of outliers is important because they accommodate the information absent from other data (Candraningtyas *et al.*, 2013). One data outlier identification method is the Difference in Fit Standarized (DFFITS) (Neter, Wasserman and Kutner M, 1997).

Robust regression on M estimation aximum of likelihood (M) and S estimation cale (S) is an alternative regression method to resolve outlier data (Yohai, 1987). The strength of M estimation has an efficiency up to 95%, while S estimation has the

 License doi: 10.20473/ijph.vl16il.2021.349-362 Received 7 May 2020, received in revised form 19 September 2020, Accepted 21 September 2020, Published online: December 2021

highest break down point of 50% (Yohai, 1987). The best model from the comparison of robust regression on M estimation and S estimation was chosen based on the biggest adjusted R^2 and the lowest Mean Square Error (Rahman and Widodo, 2018). Previous research has found the S estimation to be the most optimal over M estimation (Susanti, Pratiwi, and Sulistijowati, 2013). Other research has said that M estimation can produce the biggest adjusted R^2 with the smallest MSE (Cahyandari and Hisani, 2012).

Dengue Haemoragic Fever (DHF) is a global health problem, and it is also prevalent in Indonesia. DHF originates from the dengue virus and is spread by Aedes sp (Irianto, 2014). 2.5 billion (40%) of the population lives at risk of DHF (CDC, 2010). Cases of DHF have risen with wider distribution in Indonesia (CDC, 2010)). In Indonesia, dengue cases are increasing with wider distribution (Masriadi, 2017). In 2017, there were 59,047 cases (IR: 22.55% per 100,000) with a decease of as many as 444 people (CFR : 0.75%). The Case Fatality Rate (CFR) of DHF has become >1%, which is in the high category (Indonesian Ministry of Health, 2017).

East Java had a 1.34% CFR in 2017. This means that East Java was one of the areas that had high DHF case in Indonesia (East Java Provincial Health Office, 2017). There have been 7,866 cases in 2017 (IR: 20% per 100,000) with up to 106 fatalities (CFR: 1.34%) (East Java Provincial Health Office, 2017). The spread of DHF cases in East Java can be attributed to factors such as climate, population density, community behavior, and environmental sanitation (East Java Provincial Health Office, 2017).

The effort of DHF control program needs to get more attention from the government and society. One of things that can be done is research about the factors that influence IR DHF. It is hoped that the result of this research, such as the risk factors model on IR DHF, can be used as consideration for DHF prevention-control in East Java in the future.

METHODS

This research had an observational, non-reactive (Unobstrusif) design using secondary data analysis. This research focused on secondary data without direct individuals.

Independent variable included population density, percentage of healthy behavior, and percentage of healthy housing and precipitation. The dependent variable was the IR of DHF in East Java in 2017.

Population density data consisted of a comparison of the total population in regencies/cities in 2017 with the total area of districts/cities in 2017 in units of people/km^2.

Data on the percentage of healthy behavior consisted of a comparison of the number of households with healthy behavior in districts/cities in 2017 with the number of households monitored in the districts/cities in 2017 in percentages.

The percentage of healthy houses was the ratio between the number of healthy houses and the total number of houses in each district/city in East Java in 2017 in percentages. The rainfall was the average amount of rainfall collected in each district/city in East Java in 2017 in millimeters (mm).

The IR of DHF was the number of DHF cases that occurred in districts/cities in 2017 divided by the total population of districts/cities in 2017 per 100,000 inhabitants. All independent and dependent variables are presented in a ratio scale.

The population was all of East Java, meaning 38 regencies. The sample consisted of 35 regencies/cities chosen by simple random sampling. The identification of outliers was done using the DFFITS method. Analysis of the data was done using robust regression on M estimation and S estimation weighting by Tukey's Bisquare. The best regression was chosen

based on the biggest adjusted R^2 and the smallest MSE.

This study has a Certificate of Ethical conduct with Certificate Number 111/EA/KEPK/2019 from the Health Research Ethics Commission of the Faculty of Health, Universitas Airlangga, Indonesia.

RESULTS

Descriptive analysis

East Java is one of the provinces located on Java Island. East Java has two major seasons: rain season and summer. East Java has a width of up to 47,799.75 km^2 that consists of 38 regencies/cities with 20 regencies and 9 cities (Central Bureau of Statistics of East Java province, 2018).

Table 1. Description of Population Density, Percentage of Healthy Behavior, Percentage of Healthy Housing, and Precipitation in 35 Regencies in East Java Province 2017

No	Variable	Mean	Standard Deviation	Minimum	Maximum
1	Incidence Rate of DHF	34.28	45.55	1.62	199.29
2	Population Density (X_1)	172.44	1,978.07	277.55	8,200.77
3	% Healthy Behavior (X_2)	52.98	15.38	24.22	100.00
4	% Healthy Housing (X_3)	70.24	16.90	25.08	96.84
5	Precipitation (X_4)	2,006.60	457.60	920.45	2,884.60

Table 1 shows that of the 35 regencies in East Java, the highest IR DHF in 2017 was found in Blitar (199.29 per 100,000) whereas the Madiun regency has the lowest IR (1.62 per 100,000). Surabaya had the biggest population density (8,200.77 inhabitants/km^2) while Banyuwangi had the smallest population density (277.55 inhabitants/km^2).

The highest percentage of healthy behavior was in Ngawi (100%) while Probolinggo had the lowest healthy behavior percentage (24.22%). The highest percentage of healthy housing was found in Batu (96.84%), whereas the lowest percentage was in Sampang (25.08%). In addition, the highest precipitation was found in Pacitan (2,884.60 mm^3) and the lowest was in Situbondo (920.45 mm^3).

Analysis of OLS Regression and Outlier

The first step was to analyze the regression by using OLS to identify the data outlier. The results obtained an F count of 1.074 with a p-value of 0.387. On the significant level of 0.05, it was found that the p-value was 0.87 > 0.05. Based on the analysis of the regression with the OLS method, population density, percentage of healthy behavior, percentage of healthy housing, and precipitation had no significant effect toward the IR of DHF. In addition, the adjusted R^2 was small (0.009) and the MSE was big (2,056.321).

Table 2. Outlier Identification Results Based on Value of |DFFITS|

No	Regency	Value of |DFFITS|
1.	Ponorogo	1.0193
2.	Trenggalek	2.78512
3.	Tulungagung	2.64808
4.	Blitar	1.30444
5.	Malang	0.96786
6.	Lumajang	4.32659
7.	Banyuwangi	1.74405
8.	Bondowoso	1.63577

No	Regency	Value of |DFFITS|
9.	Situbondo	1.34405
10.	Probolinggo	5.82698
11.	Pasuruan	0.94938
12.	Madiun	1.76515
13.	Magetan	0.82441
14.	Ngawi	22.45565
15.	Bojonegoro	0.87157
16.	Tuban	1.47327
17.	Lamongan	3.30937
18.	Gresik	1.04495
19.	Bangkalan	2.76082
20.	Sampang	19.76643
21.	Pamekasan	1.32436
22.	Sumenep	1.42717
23.	Kota blitar	17.13597
24.	Kota malang	18.08846
25.	Kota pasuruan	10.34168
26.	Kota madiun	20.30211
27.	Kota surabaya	34.29463
28.	Kota batu	28.46445

The existence of outliers was confirmed by DFFITS. If the result of |DFFITS| > $2\sqrt{p/n}$, it meant that the data included the outlier. The value of p = k + 1 meant the independent variable plus 1. There were four independent variables, putting the value of p at 5. N was the number of observation in this research. It was known that $2\sqrt{p/n}$ was 0.75592. Estimation using OLS is the best estimate when the percentage of outliers is 0%. If the data is contaminated with outliers up to 10%, the bias value generated by OLS will be higher. Based on DFFITS analysis in 35 regencies/cities, there were 28 regencies that had |DFFITS| > 0.75592. Based on the oulier identification by using DFFITS, it was known that from 35 observations, more than 50% had |DFFITS| > 0.75592. It can be concluded that the data included the outlier.

The Analysis of M Estimation

The next step was using the M estimation weighting function of Tukey's Bisquare (tunning constant = 4.685). The steps of parameter estimation by using M estimation are as follows: (1) interpret β first, that is $\hat{\beta}_0$ by using OLS to get the $\hat{y}_i$; count the residual $e_i = y_i - \hat{y}_i$; (2) count the $\hat{\sigma}_i = \frac{median|e_i - median(e_i)|}{0{,}6745}$. 0,6745 make $\hat{\sigma}$ an estimator approach not refraction from σ;
(3) count $u_i = \frac{e_i}{\hat{\sigma}_i}$; (4) count the weighting function W_i by using Tukey's bisquare (tunning constant c = 4.685 so that efficiency is 95%); (5) count $\hat{\beta}_m$ by using the OLS method based on W_i; (6) $\hat{\beta}_M = (X'WX)^{-1} X'Wy$; and then (7) repeat the second until sixth step to get the convergen $\hat{\beta}_M$ (the difference of β_{Mj}^{l+1} dan β_{Mj}^{l} approaching 0).

Table 3. Iteration Results Robust Regression Analysis of M Estimation Weighting by Tukey's Bisquare

Iteration	B_0	B_1	B_2	B_3	B_4
1.	27.228	+ 0.005	– 0.080	- 0.351	+0.004
2.	-13.140	+ 0.010	– 0.363	- 0.489	+ 0.019
3.	6.076	+ 0.009	– 0.261	- 0.480	+ 0.011
4.	16.147	+ 0.009	– 0.217	- 0.496	+ 0.008
5.	20.861	+ 0.008	– 0.205	- 0.492	+ 0.007
6.	24.754	+ 0.006	– 0.193	- 0.471	+ 0.005
7.	29.231	+ 0.004	– 0.174	- 0.445	+ 0.004
8.	32.078	+ 0.003	– 0.163	- 0.437	+ 0.003
9.	33.105	+ 0.002	– 0.162	- 0.439	+ 0.003
10.	33.490	+ 0.002	– 0.163	- 0.445	+ 0.003
11.	33.524	+ 0.002	– 0.163	- 0.445	+ 0.003
12.	33.539	+ 0.002	– 0.163	- 0.445	+ 0.003
13.	33.533	+ 0.002	– 0.163	- 0.445	+ 0.003
14.	33.530	+ 0.002	– 0.163	- 0.445	+ 0.003
15.	33.529	+ 0.002	– 0.163	- 0.445	+ 0.003
16.	33.529	+ 0.002	– 0.163	- 0.445	+ 0.003

Table 4. Coefficient Result Robust Regression Analysis of M Estimation Weighting By Tukey's Bisquare

Model	B	T count	Significance	MSE	R^2 *adjusted*
Constant	33.529	2.136	0.041	147.415	0.135
Population Density (X_1)	0.002	1.470	0.152		
% Healthy Behavior (X_2)	0.163	1.082	0.288		
% Healthy Housing (X_3)	-0.445	-2.923	0.007		
Precipitation (X_4)	0.003	0.478	0.636		

Based on the result analysis of robust regression on M estimation, the convergent parameter estimation was obtrained on the 16th iteration. The agreement of convergent robust regression was:

$$\hat{y} = 33{,}529 + 0{,}002\ X_1 - 0{,}163\ X_2 - 0{,}445\ X_3 + 0{,}003\ X_4 \quad (1)$$

The agreement model (1) had an F of 2.328 with a p-value of 0.079. The significance standard was 0.05, and the p-value was 0.079 > 0.05. This means that based on robust regression model on M estimation, there was no effect of population density, healthy behavior, healthy house, and precipitation toward IR of DHF in East Java in 2017.

Analysis of S Estimation

The next regression analysis was done by using S estimation with Tukey's Bisquare (tunning constant = 1.547). The steps on estimating the parameter by using S estimation are as follows: (1) interpret β first, that is $\hat{\beta}_0$ by using OLS to get the $\hat{y}_i$; (2) count the residual $e_i = y_i - \hat{y}_i$; (3) count the robust S estimation cale, k + 0.199 , then find the

$\sigma_s =$ $\frac{median|e_i - median(e_i)|}{0{,}6745}$, for the first iteration

$\sqrt{\frac{1}{nK}\sum_{i=1}^{n} w_i {e_i}^2}$, others

(4) count $\mu_i = \frac{e_i}{\sigma_s}$; (5) count the weighting function W_i by using Tukey's bisquare

(tunning constant c = 1,547 so that the breakdown point is 50%); (6) count $\hat{\beta}_S$ by using the smallest quadrate method based on quality of W_i; (7) $\hat{\beta}_S = (X'WX)^{-1} X'Wy$; and then (8) repeat the second to seventh step to get $\hat{\beta}_S$ that is convergent (the difference of β_{sj}^{l+1} and β_{sj}^{l} approaching 0).

Table 5. Iteration Results Robust Regression Analysis of S Estimation Weighting by Tukey's Bisquare

Iteration	B_0	B_1	B_2	B_3	B_4
1.	37.637	+ 0.010	– 0.268	- 0.089	- 0.002
2.	43.075	+ 0.011	– 0.223	- 0.274	- 0.001
3.	46.938	+ 0.012	– 0.211	- 0.361	- 0.001
4.	50.770	+ 0.012	– 0.208	- 0.422	- 0.001
5.	54.476	+ 0.012	– 0.198	- 0.464	- 0.001
6.	56.666	+ 0.012	– 0.119	- 0.505	- 0.004
7.	56.377	+ 0.012	– 0.050	- 0.494	- 0.006
8.	56.732	+ 0.011	– 0.090	- 0.436	- 0.006
9.	55.318	+ 0.011	– 0.122	- 0.398	- 0.006
10.	55.320	+ 0.011	– 0.125	- 0.398	- 0.006
11.	55.322	+ 0.011	– 0.128	- 0.399	- 0.006
12.	55.312	+ 0.011	– 0.130	- 0.400	- 0.005
13.	55.279	+ 0.011	– 0.132	- 0.401	- 0.005
14.	55.219	+ 0.011	– 0.133	- 0.401	- 0.005
15.	55.166	+ 0.011	– 0.134	- 0.402	- 0.005
16.	55.115	+ 0.011	– 0.134	- 0.402	- 0.005
17.	55.064	+ 0.011	– 0.135	- 0.402	- 0.005
18.	55.008	+ 0.011	– 0.135	- 0.403	- 0.005
19.	54.948	+ 0.011	– 0.135	- 0.403	- 0.005
20.	54.887	+ 0.011	– 0.136	- 0.403	- 0.005
21.	54.842	+ 0.011	– 0.136	- 0.404	- 0.005
22.	54.826	+ 0.011	– 0.136	- 0.404	- 0.005
23.	54.826	+ 0.011	– 0.136	- 0.404	- 0.005

Table 6. Coefficient Result Robust Regression Analysis of S Estimation Weighting by Tukey's Bisquare

Model	B	T count	Significance	MSE	R^2 *adjusted*
Constant	54.826	62.407	0.000	0.229	0.996
Population Density (X_1)	0.011	79.607	0.000		
% Healthy Behavior (X_2)	-0.136	-6.836	0.000		
% Healthy Housing (X_3)	-0.404	-20.186	0.000		
Precipitation (X_4)	-0.005	-8.457	0.000		

Based on the analysis result of robust regression on S estimation, the convergent parameters were obtained on the 22nd iteration. The agreements of convergent robust regression were as follows:

$$\hat{y} = 54{,}826 + 0{,}011\ X_1 - 0{,}136\ X_2 - 0{,}404\ X_3 - 0{,}005\ X_4 \quad (2)$$

The agreement model (2) had an F of 1,985.256 and a p-value of 0.000. On the significant standard of 0.05, the p-value was 0.000 < 0.05. This means that based on robust regression on S estimation, there was a significant effect of population density, healthy behavior, healthy housing, and precipitation toward IR of DHF in East Java in 2017.

The Best Robust Regression Model

The best robust regression should fulfill the criteria of the biggest adjusted R^2 and the smallest MSE. Based on the result analysis of robust regression by using M estimation and S estimation, it was found that R^2 adjusted the robust regression model on S estimation (0.996) > R^2 adjusted on M estimation (0.135). While based on MSE value, the robust regression on S estimation (0.229) < MSE M estimation (147.415). The conclusion is the analysis of robust regression by using the S estimation method produced the best agreement model because it fulfilled the criteria of the biggest value from adjusted R^2 on the smallest MSE.

DISCUSSION

The Analysis of M Estimation

Regression methods using OLS for modeling minimizes the quadrate from residual fit regression (Rahmadeni and Anggreni, 2014). That model is not effective because of the refraction caused by data outlier. The data outlier will make the analysis of OLS regression become refracted on the interpretation results and inefficient (Herawati, Nisa, and Setiawan, 2011). This is because the smallest quadrate is sensitive to the outlier (Lainun and Tinungki, 2018).

The DFFITS method identifies the outlier multivariately (Shodiqin, Aini, and Rubowo, 2018). This method is used for identifying the outlier scale on observation data overall showing the changing of the value that predicts if the observation is out of standard (Neter, Wasserman, and Kutner M, 1997). The outlier criterion is if that data observation has an absolute DFFITS > $2\sqrt{\frac{p}{n}}$ (Rahman and Widodo, 2018).

The analysis of robust regression is the regression analysis for resolving the data outlier. The use of M estimation like OLS is the alternative M estimation method, but especially for outlier data (Rousseeuw and Leroy, 1987). The analysis of robust regression on M estimation was introduced by Huber on 1964 with efficiency characteristics, but also by minimizing residual objective functions (Alma, 2011).

The first step of using robust regression analysis on M estimation is analyzing the linear OLS regression method. The residual unstandardized result from the OLS method is then used for weighting on M estimation. The strength of M estimation is the count process is simpler than the other robust estimation and it has the highest efficiency of up to 95%, but it is not effective enough on the proportion outlier 25%-30% (Herawati, Nisa, and Setiawan, 2011).

In this research, weighting Tukey's Bisquare was chosen because weighting Tukey's Bisquare using a tunning constant (c) of up to 4.568 could achieve 95% efficiency. As the explanation of previous research, the use of c = 4.568 will make for 95% efficiency (Lainun and Tinungki, 2018).

Based on the analysis, the agreement of robust regression on M estimation gives no significant effect for IR of DHF. Simultantly of population density, healthy behavior, healthy housing, and precipitation give have effect on IR of DHF. This could be caused by the characteristic of M estimation.

This research used M estimation using the Tukey Bisquare with a tunning constant of up to 4.685. Tunning constant was symbolized by C as the value that was set to determine the level of robust weighting (Setiarini and Listyani, 2017). The use of c = 4.685 make the efficiency of

M estimation become 95%. The efficiency explains how a robust technique like OLS can produce no outliers (Setiarini and Listyani, 2017). The high efficiency makes a derivation of breakdown point value (Rousseeuw and Leroy, 1987).

The higher efficiency of an estimator, the lower its breakdown point value. The lower breakdown point value of an estimator, the lower its ability in resolving outlier (Lainun and Tinungki, 2018). From the identification of the outlier result, 80% of the research data included outliers. The characteristic of M estimation having high efficiency makes M estimation produce the wrong estimation value because of the lack of a breakdown point on outlier existence.

Previous research explained that when the outlier proportion is 25% and 30%, the M estimation becomes ineffective. Finally, the residual that was produced became bigger. This means that M estimation was very influenced by the outlier (Herawati, Nisa, and Setiawan, 2011).

The result of robust regression research by using M estimation showed that the agreement of convergent robust regression on M estimation produce the MSE value of 147.415. This MSE value is smaller than the MSE from the OLS regression method analysis. Moreover, the adusted R^2 on M estimation (0.135) was bigger than the adjusted R^2 OLS (0.009). This means that the M estimation method is better than the OLS method.

The preliminary research explained that M estimation produced the smallest residual value compared to residual from the OLS method (Susanti, Pratiwi, and Sulistijowati, 2013). The OLS method produced the bigger MSE because it has the refraction to outliers that make the credibility of the interval become bigger, while M estimation is resolves that by giving a small weight on the data outlier (Lainun and Tinungki, 2018).

The Analysis of S Estimation

The other estimation method used in the research was the regression analysis on scale estimation. Scale estimation is a coefficient regression estimation on the outlier data. The strength of S estimation is the highest breakdown (BD) estimator of up to 50% (Alma, 2011). That BD value was gotten by using Tukey's Bisquare with a tunning constant of up to 1.547 (Hidayatulloh, Yuniarti, and Wahyuningsih, 2015). The weakness of S estimation is the efficiency value is low at 28% (Alma, 2011).

Based on the results, the trial on the model of robust regression of S estimation is simultant and partially showed significant effects. S estimation still produced influential independent variables significantly with outlier data condition. This happened because S estimation can resolve up to 50% of data outliers (Pitselis, 2013).

Outlier identification results by using DFFITS explained that 28 of 35 data or up to 80% had outliers. This means that the characteristic of S estimation can resolve up to 50% of outliers, meaning 50% of 80% outliers from research data can be resolved by the S estimation. S estimation can be used for coefficient regression estimation for research data.

The result of robust regression on S estimation research showed that the MSE value (0.229) was smaller than MSE with the OLS method (2,056.321). Moreover, the adjusted R^2 (0.996) was bigger than the adjusted R^2 with the OLS method (0.009). The conclusion is the S estimation method was better than the OLS method.

The preliminary research explained the S estimation as the method with the best regression because it produced a determinant coefficient (R^2 adjusted) that was bigger with MSE that was smaller than the OLS method (Hidayatulloh, Yuniarti, and Wahyuningsih, 2015).

The result of regression parameter with the OLS method become worse because the estimation value was influenced by the outlier that caused deviation of regression to the real parameter value (Herawati, Nisa, and Setiawan,

2011). S estimation had a breakdown point value of 50%, meaning it can handle up to 50% of outlier data and still produce the better coefficient regression (Alma, 2011).

The Best Robust Regression Model

The robust regression analysis was needed as the M estimation method with high credibility (Pitselis, 2013). The choosing of the best robust regression was based on which produced the smallest MSE value and the biggest adjusted R^2 for every independent variable considered in the model (Rahman and Widodo, 2018).

Mean Square Error (MSE) is the indicator of error measurement estimation on the regression model that was produced. The smaller the residual produced, the better that model (Yuniastari and Wirawan, 2014). Adjusted R^2 is the measurement of variation value that can explain the corrected robust regression model. The bigger the regression coefficient produced, the better the regression model (Dahlan, 2012).

Based on the comparison of MSE value and adjusted R^2, the best robust regression was S estimation with Tuqey's bisquare. That happened because S estimation fulfilled the regression criteria of smallest MSE and biggest adjusted R^2 compared with the smallest MSE and adjusted R^2 on the M estimation, while the best robust regression agreement based on the S estimation was as follows:

$\hat{y} = 54{,}826 + 0{,}011\ X_1 - 0{,}136\ X_2 - 0{,}404\ X_3 - 0{,}005\ X_4$

The preliminary research explained that S estimation produced an adjusted R^2 adjusted of up to 99.9%. The coefficient value determination was bigger than the M estimation (99.6%). Meanwhile, the MSE produced by S estimation (287.345) was smaller than the MSE of M estimation (1,016.53) (Susanti, Pratiwi, and Sulistijowati, 2013).

The factor that made the S estimation the best estimation based on the criteria of adjusted R^2 and MSE was the breakdown point. The breakdown point estimator was influenced by the use of the tunning constant (Alma, 2011).

Theoretically, S estimation has the strength as the estimator with the highest breakdown point (50%) by using the tunning constant (1.547), but it has low efficiency (8%) (Shodiqin, Aini, and Rubowo, 2018). This means that S estimation can resolve up to 50% of outliers from the observation data overall. If the data of outliers are up to 50%, the model on the regression S estimation can be used better (Alma, 2011). The tunning constant will be influential on the ability of the estimator to be more specific and minimize the number of residual quadrate (Pradewi and Sudarno, 2012).

Based on the research result that has been conducted, 28 of 35 data were outliers, meaning that >50% of outliers in the research data. Outlier proportions of up to 50% can be resolved better by S estimation, which will produce the regression agreement with the biggest adjusted R^2 and the smallest MSE.

The M estimation produced the wrong value of estimation because the highest efficiency was up to 95% from the use of a tunning constant of 4.685. The breakdown point of M estimation was less than 50% (Rousseeuw and Leroy, 1987). The higher the estimator's efficiency, the lower its breakdown point (Lainun and Tinungki, 2018).

Previous research also explained that when the outlier proportion is up to 25% and 30%, the M estimation becomes ineffective. Finally, MSE that was produced was big. This means that M estimation was very influenced by outliers (Herawati, Nisa, and Setiawan, 2011).

The Factors of IR of DHF in East Java

Based on regression model result on S estimation by using Tukey's Bisquare (tunning constant = 1.547), there were some factors that significantly influenced IR of DHF that include population density, healthy housing, healthy behavior, and precipitation.

The percentage of healthy housing is the main focus that influenced IR of DHF. Healthy housing refers to homes that fulfill the requirement of health, including the component of construction and structuring the house based on physical and biological requirements. The percentage of healthy housing is the comparison percentage of healthy housing with the numbers of all houses in one area (East Java Provincial Health Office, 2017).

The results showed when there was derivation in the percentage of healthy housing, there was an increase of IR of DHF. Every derivation of 1% healthy housing had an IR of up to 0.404 per 100,000. If there was derivation of healthy housing of up to 50%, there would be an increase of IR of DHF of up to 20.2 per 100,000.

This result is in line with with previous research that explained that healthy housing is correlated with the better physical environments. That research concluded that physical environment has significant correlation with DHF (Umaya, Faisya, and Sunarsih, 2013).

Homes with better physical environments have smaller risk of DHF cases than homes that have bad physical environments (Umaya, Faisya, and Sunarsih, 2013). The condition of environment sanitation correlates with the rise of DHF cases. Homes that are a place of propagation for Aedes sp are 3.8 times more likely for DHF infections than homes that are not (Sofia, Suhartono, and Wahyuningsih, 2014).

The percentage of healthy behavior is the second factor that influences IR of DHF. Clean and healthy behavior is a set of behavior that is practiced from the awareness of healthy population programs. The percentage of healthy behavior is the comparison between the households with healthy behavior against all monitored household in one area (East Java Provincial Health Office, 2017).

The result of the research showed that if there is the derivation of healthy behavior percentage, there is a raise of IR of DHF. Every 1% derivation of healthy behavior percentage will increase IR of DHF by up to 0.136 per 100,000 population. If there is derivation of up to 50%, there will be IR of DHF up to 6.8 per 100,000.

The research result is in line with the preliminary research that concluded that healthy behavior has correlation with DHF (Raksanagara and Raksanagara, 2018). Healthy behavior has correlation with the frequency of fever in family members. The higher the healthy behavior level, the lower the frequency of fever in family members (Prabowo, 2016). This is because in healthy behavior there is behavior to clean mosquito larva every week. The cleaning of mosquito larva every week is an effort to eliminate the breeding place of mosquitoes to control the vector of DHF (Hastuti, Dharmawan, and Indarto, 2017). The behavior of cleaning the mosquito larva every week will substract the risk of DHF spreading (Raksanagara and Raksanagara, 2018). The less healthy behavior, the bigger the risk of DHF (Monica, Devianto, and Yanuar, 2012).

The population density is the third factor that influenced IR of DHF. The population density is thee population every km^2. The population density is the comparison between the number of population and the area's capacity in km^2 (Central Bureau of Statistics of East Java Province, 2018).

The research result showed that if there was an increase of population density, there was an increase of IR of DHF. Every increase of 1 inhabitant per km^2 increased IR of DHF by up to 50 inhabitants per km^2, indicating an increase in IR of DHF of up to 0.55 per 100,000.

The research result is in line with the preliminary research that showed population density will influence IR of DHF (Prasetyowati, 2015). Research with the same result was also conducted in Pasuruan, East Java that explained that high population density will increase IR of DHF (Ali and Ma'rufi, 2016).

The other factors that influence IR of DHF is precipitation. Precipitation is the volume of rain that falls in millimeters/mm (Sihombing, Nugraheni, and Sudarsono, 2018). One of factor that cause high IR of DHF is the change of climate. Some borne disease is caused by climate such as precipitation. The increase of vector distribution is influenced by the changing of climate. Diseases spread by vectors (vector borne disease) such as DHF need to be monitored when there is a change of climate (Yushananta and Ahyanti, 2014).

The research result showed that precipitation has a significant effect on the cases of DHF. The derivation of precipitation will increase DHF. Every derivation of 1 mm will increase the number of DHF by up to 0.005 per 100,000 populations. If there is a derivation of precipitation of up to 500 mm, it will increase the number of DHF by up to 0.25 per 100,000.

The research result is in line with preliminary research that explained the precipitation has a negative correlation with DHF. This means that when there are increases of precipitation, it will derivate cases of DHF. The derivation of precipitation will increase DHF cases. The frequency of heavy rain will decrease DHF. That can be caused by the rain that dissolves mosquito larva of Aedes sp in collecting and saving water until the population of Aedes sp decreases (Suhermanto, Tunggul, and Widartono, 2012).

Some factors that influence IR of DHF are the percentage of healthy housing, percentage of healthy behavior, population density, and precipitation. They should be the main focus for the prevention effort of DHF by society. So far, there has been no vaccine found yet to resolve the dengue virus, meaning the effort to eliminate vectors is the most effective effort to prevent and control DHF in society.

Some efforts to eliminate that can be used are self-protection from the Aedes sp, biological control, chemical control, epidemiology survey, the handling of the environment, and the integrated controlling of the environment to substract the potential of Aedes sp propagation (Soedarto, 2012). This effort needs the active participation of society and cooperation across sectors for optimal results.

The strength of this research is that it is able to prove that the robust regression method using S estimation is better because the ability to handle outlier data is higher than that of M estimation. Meanwhile, the lack of this research is the use of small samples and limited variables.

Further research can modify the tunning constant value and the efficiency value, and use other types of weighting such as the Huber weighting function.

As a form of follow-up to research results, efforts must be made to increase the community's active participation in increasing PHBS and creating healthy homes with environmental conditions to reduce the breeding of the *Aedes sp*. mosquito.

CONCLUSIONS

The robust regression analysis is an alternative regression method for data outliers. The use of Scale (S) estimation is better than Maximum of Likelihood (M) estimation. The M estimation was ineffective when the outlier proportion was up to 25% and 30% and the MSE produced was big. This means that M estimation was greatly influenced by outliers**.**

S estimation produced the biggest adjusted R^2 and the smallest MSE. That was caused by the use of the tunning constant. The smaller the tunning constant of an estimator, the lower its efficiency, meaning a bigger breakdown point and better outlier data resolution. Some factors that influenced IR of DHF in East Java in 2017 include population density, percentage of healthy housing, percentage of healthy behavior, and precipitation. Cooperation between society and the government is important for preventing and tackling DHF in society.

REFERENCES

Ali, K. and Ma'rufi, I. (2016) 'Study of Factors Caused Dengue Haemorrhagic Fever Case Study : Pasuruan , Jawa Timur- Indonesia', *Journal of Medical and Bioengineering*, 5(2), pp. 108–112. doi: 10.18178/jomb.5.2.108-112.

Alma, O. G. (2011) 'Comparison of Robust Regression Methods in Linear Regression', *International Journal Contemporary Mathematical Science*, 6(9), pp. 409–421.

Central Bureau of Statistics of East Java Province (2018) *Jawa Timur Province in Figures 2018*. Surabaya.

Cahyandari, R. and Hisani, N. (2012) 'Model Regresi Berganda Menggunakan Penaksir Parameter Regresi Robust M-Estimator (Studi Kasus : Produksi Padi di Provinsi Jawa Barat Tahun 2009)', VI(1), pp. 85–92.

Candraningtyas, S. *et al.* (2013) 'Regresi Robust Mm- Estimator Untuk Penanganan Pencilan Regresi Robust Mm- Estimator', *Jurnal Gaussian*, 2(4), pp. 395–404.

CDC (2010) *Epidemiology Dengue*.

Dahlan, M. S. (2012) *Regresi Linear Disertai Praktik dengan SPSS*. Jakarta: Epidemiologi Indonesia.

Draper, N. R. and Smith, H. (1992) *Analisis Regresi Terapan (Terjemahan dari Applied Regression Analysis)*. 2nd edn. Jakarta: Gramedia Pustaka Utama.

Gujarati, D. N. (2013) *Basic Econometrics*. 4th edn. Ne: McGraw-Hill/Irwin.

Hastuti, N. M., Dharmawan, R. and Indarto, D. (2017) 'Sanitation-Related Behavior , Container Index , and Their Associations with Dengue Hemorrhagic Fever Incidence in Karanganyar , Central Java', 2, pp. 174–185.

Herawati, N., Nisa, K. and Setiawan, E. (2011) 'Analisis Ketegaran Regresi Robust Terhadap Letak Pencilan: Studi Perbandingan', *Bulletin of Mathematics*, 3(01), pp. 49–60.

Hidayatulloh, F. P., Yuniarti, D. and Wahyuningsih, S. (2015) 'Regresi Robust Dengan Metode Estimasi-S Robust Regression Method To Estimate - S', *Eksponensial*, 6(2), pp. 163–170.

Irianto, K. (2014) *Epidemiologi Penyakit Menular dan Tidak Menular : Panduan Klinis*. Bandung: Alfabeta.

Lainun, H. and Tinungki, G. M. (2018) 'Perbandingan Penduga M , S , dan MM pada Regresi Linier dalam Menangani Keberadaan Outlier', *Jurnal Matematika, Statistika dan Komputasi*, 15(1), pp. 88–96.

Masriadi (2017) *Epidemiologi Penyakit Menular*. Jakarta: Rajawali Pers.

Monica, D., Devianto, D. and Yanuar, F. (2012) 'Pemodelan Faktor-Faktor Yang Mempengaruhi Kejadian Dbd (Demam Berdarah Dengue) Menggunakan Regresi Logistik Biner Untuk Wilayah Regional 2 Indonesia (Sumatera)', *Jurnal Matematika UNAND*, VI(1), pp. 9–16.

Neter, J., Wasserman, W. and Kutner M, H. (1997) *Model Linear Terapan (Terjemahan dari Applied Linear Model) oleh Bambang Sumatri*. Edited by B. Sumatri. Bogor: FMIPA-IPB.

Pitselis, G. (2013) 'Journal of Computational and Applied', *Journal of Computational and Applied Mathematics*. Elsevier B.V., 239, pp. 231–249. doi: 10.1016/j.cam.2012.09.009.

Prabowo, A. (2016) 'Hubungan Perilaku Hidup Bersih dan Sehat (PHBS) dengan Frekuensi Sakit Anggota Keluarga', *PROFESI*, 13(2), pp. 60–65.

Pradewi, E. D. and Sudarno (2012) 'Kajian Estimasi-M IRLS Menggunakan Fungsi Pembobot Huber dan

Bisquare Tukey pada Data Ketahanan Pangan di Jawa Tengah', *Media Statistika*, 5(1), pp. 1–10.

Pramana, S. *et al.* (2017) *Dasar-Dasar Statistika dengan Software R (Konsep dan Aplikasi)*. Kedua. Bogor: iN Media.

Prasetyowati, I. (2015) 'Kepadatan Penduduk dan Insidens Rate Demam Berdarah Dengue (DBD) Kabupaten Bondowoso, Jawa Timur', *The Indonesia Journal of Health Science*, 5(2), pp. 1–12.

Rahmadeni and Anggreni, D. (2014) 'Analisis Jumlah Tenaga Kerja Terhadap Jumlah Pasien RSUD Arifin Achmad Pekanbaru Menggunakan Metode Regresi Gulud', *Jurnal Sains, Teknologi dan Industri*, 12(1), pp. 48–57.

Rahman, M. B. and Widodo, E. (2018) 'Perbandingan Metode Regresi Robust Estimasi Least Trimmed Square , Estimasi Scale , dan Estimasi Method Of Moment', in *Prosiding Seminar Nasional Matematika (PRISMA)*, pp. 426–433.

Raksanagara, A. S. and Raksanagara, A. (2018) 'Perilaku Hidup Bersih dan Sehat Sebagai Determinan Kesehatan yang Penting pada Tatanan Rumah Tangga di Kota Bandung', *Jurnal Sistem Kesehatan*, 1(1), pp. 1188–1197. doi: 10.1002/j.1875-9114.2012.01178.x.

Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. New York: John Wiley & Sons Inc.

Setiarini, Z. and Listyani, E. (2017) 'Analisis Regresi Robust Estimasi-S Menggunakan Pembobot Welsch dan Tukey Bisquare', *Jurnal Matematika*, 6(1), pp. 48–55.

Shodiqin, A., Aini, A. N. and Rubowo, M. R. (2018) 'Perbanding Dua Metode Regresi Robust Yakni Metode Least Trimmed Squares (LTS) dengan Metode Estimator-MM (Estimasi-MM) (Studi Kasus Data Ujian Tulis Masuk Terhadap Hasil IPK Mahasiswa UPGRIS)', *Jurnal Ilmiah Teknosains*, 4(1), pp. 35–42.

Sihombing, C. G., Nugraheni, E. and Sudarsono, W. (2018) 'the Relationship Between Rainfall, Air Temperature and Wind Speed Effects Dengue Hemorrhagic Fever Case in Bengkulu City At 2009-2014', *Jurnal Kedokteran Diponegoro*, 7(1), pp. 366–380.

Soedarto (2012) *Demam Derdarah Dengue*. Jakarta: Sagung Seto.

Sofia, Suhartono and Wahyuningsih, N. E. (2014) 'Hubungan Kondisi Lingkungan Rumah dan Perilaku Keluarga dengan Kejadian Demam Berdarah Dengue Di Kabupaten Aceh Besar The Relationship of Home Environmental Conditions and Family Behavior with Genesis Dengue In Aceh Besar', *Kesehatan Lingkungan Indonesia*, 13(1), pp. 30–37. doi: 10.14710/JKLI.13.1.30 - 38.

Suhermanto, Tunggul, S. T. B. and Widartono, B. S. (2012) 'Spatial Analysis on Vulnerability to Dengue Hemorrhagic Fever in Kotabaru Subdistrict , Jambi Municipality , Jambi Province', *Tropical Medicine Journal*, 02(1), pp. 45–56.

Susanti, Y., Pratiwi, H. and Sulistijowati, S. (2013) 'Optimasi Model Regresi Robust Untuk Memprediksi Produksi Kedelai Di Indonesia', in *Seminar Nasional Matematika dan Pendidikan Matematika*, pp. 978–979.

Umaya, R., Faisya, A. F. and Sunarsih, E. (2013) 'Hubungan Karakteristik Pejamu, Lingkungan Fisik Dan Pelayanan Kesehatan Dengan Kejadian Demam Berdarah Dengue (Dbd) Di Wilayah Kerja Puskesmas Talang Ubi Pendopo Tahun 2012', *Jurnal Ilmu Kesehatan Masyarakat*,

4(November), pp. 262–269.

Yohai, V. J. (1987) 'High Breakdown-Point and High Efficiency Robust Estimates For Regression', *The Annals of Statistics*, 15(4), pp. 1580–1592.

Yuniastari, N. L. A. K. and Wirawan, I. W. W. (2014) 'Peramalan Permintaan Produk Perak Menggunakan Metode Simple Moving Average Dan Exponential Smoothing', *Sistem dan Informatika*, 9(1), pp. 97–106.

Yushananta, P. and Ahyanti, M. (2014) 'Pengaruh Faktor Iklim dan Kepadatan Jentik Ae.Aegypti terhadap Kejadian DBD', *Jurnal Kesehatan*, 5(1), pp. 1–10.