Robust regression with MM-estimator for modelling the number maternal mortality of pregnancy in Central Java, Indonesia

A Prahutama, A Rusgiyono

Statistics Department, Faculty of Sciences and Mathematics, Diponegoro University, Semarang, Indonesia

Corresponding author: alan.prahutama@gmail.com

Abstract. Regression is a method of statistics analysis for modeling between the dependent variable to independent variables. Estimation in the regression model uses the Ordinary Least Square (OLS) method. In the regression model, several assumptions are required. One of the reasons for the violation of the assumption is that there are outliers in the data. There are several methods that can be used to overcome outliers, one of which is using the Robust MM-estimator regression. The MM-estimator has high efficiency when the error is normally distributed and also has a high breakdown value. To implement the SDG’s program, one program is health. It has objectiveness to reduce maternal mortality rate. Maternal mortality consist of pregnancy, give of birth and postpartum mother. In this study the data used was the number of maternal mortality of pregnancy in Central Java in 2019 as the dependent variable, while the independent variables used included the percentage of pregnant women consuming Fe3 tablet (X1), percentage of households that have a clean and healthy lifestyle (X2); percentage of pregnant women who made the first visit (X3); The results obtained in multiple regression modeling produce R square value of 0.222679 and Mean Square Error (MSE) of 6.894871. The assumptions in the regression model unfulfilled, may be caused by outlier in data. In MM robust regression modeling, the R square value is 0.2695, and the MSE value is 4.989. It shows that robust regression modeling with MM estimator is better than multiple regression.

1. Introduction

Sustainable Development Goals (SDGs) are a global action plan agreed upon by world leaders, including Indonesia, to end poverty, reduce inequality and protect the environment. The SDGs contain 17 Goals and 169 Targets that are expected to be achieved by 2030. One of the SDGs goals is to ensure a healthy life and support the welfare of all ages. The goal has a target, namely by 2030, reducing the maternal mortality ratio to 70 per 100,000 births. Maternal mortality is the number of maternal mortality resulting from pregnancy, childbirth, and postpartum per 100,000 live births at a certain time [1].

Figure 1 shows the maternal mortality rate in Central Java from 2013 to 2017 in Central Java. This figure indicates a decrease in maternal mortality from 2013 to 2017, although there was an increase in 2014. The reduction in infant mortality from 2013 to 2017 has not fully reached the SDGs target, which is below 70. Based on data on maternal mortality, the characteristics can be divided into three namely, maternal mortality of pregnant, maternal mortality of give birth and maternal mortality childbed / postpartum.
According to BPS ("Badan Pusat Statistika") or central bureau statistics [2], the highest number of maternal mortality is the number of postpartum or postpartum maternal mortality, while the number of maternal mortality of pregnant is the second highest. In this study, it will focus on the factors that affect the maternal mortality of pregnant. Maternal mortality is maternal death that occurs during pregnancy, starting from conception to birth of the fetus. This death does not depend on the length and place of pregnancy, caused by anything related to pregnancy or aggravated by the pregnancy or its handling, but not death caused by accident or accident. One of research about maternal mortality is modelling maternal mortality and infant mortality in East Kalimantan using Poisson regression [3]. In addition, modelling maternal and infant mortality in North Sumatera using Geographically Weighted Bivariate Gamma Regression. In this research, they use health care behavior dimension such as the percentage of households with a clean and healthy lifestyle and percentage of pregnant women carrying out the First Visit and also they use percentage of pregnant women receiving Fe3 tablets as a predictor variables [4].

Regression analysis is an analysis used to model the independent variable and the dependent variable. Estimation of parameters in the regression model uses Ordinary Least Square (OLS). In the regression model, the significance testing of the parameters simultaneously and partial testing of the parameters is carried out [5]. In addition, the regression model requires several assumptions, including normally distributed residuals, homoscedastic residuals and no autocorrelation of the residuals. There are some model using regression analysis, but some assumptions are not fulfilled. One of the reasons is the outliers in the data. In relation to the regression analysis, outliers can cause the residuals to be not normally distributed, the variance is not homogeneous, and the interval estimates have a wide range. The method that can be used to analyze remote data is robust regression.

According to Chen [7] robust regression has several estimation methods, including the Maximum likelihood estimator (M-estimator), Scale estimator (S-estimator), and Method of Moments estimator (MM-estimator). Each estimation method has its own strengths and weaknesses. The MM-estimator has high efficiency when the error is normally distributed and has a high breakdown value too. The breakdown value is the minimal proportion of the number of outliers compared to all observed data. The MM-estimator method was first introduced by combining the high breakdown estimation (S-estimator) with high efficiency (M-estimator). The purpose of this research is modelling the number maternal mortality of pregnancy uses robust regression MM-estimator.

2. Literature Review

2.1. Regression Analysis

According to [8], regression analysis is a statistical technique that is often used to determine the relationship between independent and dependent variables. In general, a regression model with k independent variables and n observations can be written as follows

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i \]

for \( i = 1, 2, \ldots, n \)
According to Montgomery and Peck (1992), parameter estimation can be obtained using the least squares method (Ordinary Least Square) by minimizing the number of squares of error.

\[ S(\beta) = (Y - XB)'(Y - XB) \]

so the estimation of parameter \( \hat{\beta} \) can be written as follows:

\[ \hat{\beta} = (X'X)^{-1} X'Y \]

2.2. The testing of significant parameters model

According to [8] in multiple linear regression there are two testing of parameter significant model such as the testing of significant model (F testing) and the testing of the significant parameter (t-testing). The F-testing as follows:

Hypothesis

\[ H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \]

\[ H_1: \text{at least there is } \beta_j \neq 0 , \text{ with } j = 1,2,\ldots,k \]

The testing of statistics is \( F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE} \). Reject \( H_0 \) for \( F > F(a;k;n-k-1) \)

The significant testing of parameters as follows:

Hypothesis:

\[ H_0: \beta_j = 0 \]

\[ H_1: \beta_j \neq 0 \]

The testing of statistics is \( t = \frac{\hat{\beta}_j}{Se(\hat{\beta}_j)} \); with \( Se(\hat{\beta}_j) = \sqrt{var(\hat{\beta}_j)} \). The reject \( H_0 \) for \( |t| > t(a/2; n-k-1) \)

2.3. Robust Regression

According to [8], robust regression is a regression method used when the distribution of the residuals is not normal or there are several outliers that affect the model. This method is an important tool for analyzing data that is infected by outliers, so that the model results become resistant to outliers. Robust regression procedures are intended to accommodate data oddities, while at the same time eliminating the identification of outlier data. Chen [7] mentions several estimation methods in robust regression, three of which are M-estimator (Maximum likelihood type estimator), S-estimator (Scale estimator), and MM-estimator (Method of Moment estimator).

2.4. Robust MM-Estimator

According to Chen [7], the MM-estimator was introduced by Yohai in 1987, where the first step in this estimation is to find the estimated value with the S-estimator, then estimate the final regression parameter estimator with the first step residual using the M-estimator method. The S-estimator guarantees a high breakdown point value, and the M-estimator makes the estimator have high efficiency [9].

According to Susanti (2014) [10] MM-estimator is a solution from:

\[ n \rho_1(u_i)x_{ij} = \sum_{j=1}^{n} \rho_1 \left( \frac{y_i - \sum_{j=0}^{k} x_{ij} \hat{\beta}_j}{s_{mm}} \right) x_{ij} = 0 \]

with \( \frac{y_i - \sum_{j=0}^{k} x_{ij} \hat{\beta}_j}{s_{mm}} \) is residual from parameter estimation of the regression model, \( s_{mm} \) is the standard deviation from S-estimator’s residual, \( \rho_1 \) is objective function of Tukey’s Bisquare.

\[ \rho_1(u_i) = \begin{cases} 
\frac{c^2}{6} \left( \left( \frac{u_i}{c} \right)^2 - 1 \right)^3 , & \text{for } |u_i| \leq c \\
\frac{c^2}{6} , & \text{for } |u_i| > c 
\end{cases} \]

There are three steps in MM-estimator as follows [11]:

\[ \rho_1(u_i) = \begin{cases} 
\frac{c^2}{6} \left( \left( \frac{u_i}{c} \right)^2 - 1 \right)^3 , & \text{for } |u_i| \leq c \\
\frac{c^2}{6} , & \text{for } |u_i| > c 
\end{cases} \]
1. Determine the prior estimation that was showed $\tilde{\beta}$ by estimator whose high breakdown point, the probability is 50%, usually it uses S-estimator.

2. Counting the residual and M-estimator scale with 50% is breakdown point. $s(e_1(\tilde{\beta}), e_2(\tilde{\beta}), \ldots, e_n(\tilde{\beta}))$ were notated $s_{mm}$. The objective function can be used in this step with $\rho_1$ labelled;

3. Counting the estimation of posterior with M-estimator uses the derivative function effect

$$
\psi(u_i) = \frac{\partial(\rho(u_i))}{\partial(u_i)}
$$

and scale estimation of $s_{mm}$ was obtained from second steps. So the value of last parameter estimation of MM-estimator which was defined as

$$
\hat{\beta}^{(mm)} = (X'WX)^{-1}X'WY
$$

3. Methodology of Research

The data used in this study is secondary data sourced from the 2019 Central Java Provincial Health Profile Report published by the Central Java Provincial Health Office. The research unit used is the district/city in Central Java, consisting of 29 districts and 6 cities, so that the research units are 35 districts/cities. The software that used in this analysis is R.

In this study, the variables to be analyzed are divided into two, namely the dependent variable (Y) and the independent variable (X). The dependent variable used is the number of maternal mortality by pregnant. The independent variable consisted of three variables, namely the percentage of pregnant women receiving Fe3 tablets ($X_1$); the percentage of households with a clean and healthy lifestyle ($X_2$); percentage of pregnant women carrying out the First Visit ($X_3$).

4. Results and Discussion

In this study, there are two steps to model the number of maternal mortality by pregnant, which are modelling by multiple regression and modelling by Robust regression with MM-estimator.

4.1. Modelling the number maternal mortality by pregnant uses multiple regression

The multiple regression model with ordinary least square estimation which for modelling the number maternal mortality by pregnant as follows:

$$
Y = 102.6889 + 0.0285X_1 - 0.066X_2 - 0.9737X_3
$$

Based on those model, we can conclude that the variable of percentage of pregnant that receiving Fe3 tablet has positive impact. It is means that more increase of the pregnant receiving Fe3 tablet, more increase the number of maternal mortality in pregnant. For variable $X_2$ and $X_3$ have negative impact for maternal mortality in pregnant. Its means that more increase the percentage household with clean and healthy lifestyle will more reduce the maternal mortality in pregnant. As well as $X_3$ variable, more increase percentage pregnant women carrying out the first visit, will more reduce the number maternal mortality in pregnant. Even though the $X_1$ variable has miss interpretation with theory, the impact of variable $X_1$ has fewer than others. The following analysis is testing of the significant model using F test as follows:

Hypothesis

- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (the model is not appropriate)
- $H_1 : \text{at least there is } \beta_j \neq 0, j = 1,2,3$ (the model is appropriate)

Earned the value of $F = \frac{SS_R/k}{SS_R/(n-k-1)} = 2.960192$, the criteria of testing is reject $H_0$ for $F > F_{(0.05,231)}$, which the table of $F_{(0.05,231)} = 2.26955$ then $H_0$ rejected. So it can be inferred that $Y = 102.6889 + 0.0285X_1 - 0.066X_2 - 0.9737X_3$ is an appropriate model. The next step is the testing of the parameter. The testing of parameter uses t-test as follows:

Hypothesis

- $H_0 : \beta_j = 0$ (The variable of $X_j$ is not significant)
- $H_1 : \beta_j \neq 0$ (The variable of $X_j$ is significant); for $j=1,2,3$

Criteria of testing is $H_0$ rejected for $|t| > t_{(0.05,31)} = 1.69552$
Table 1. The significant estimation parameter of multiple regression model

| Variable                                      | $|t|$ value | Decision  |
|-----------------------------------------------|-----------|-----------|
| the percentage of pregnant women receiving Fe3 tablets ($X_1$) | 0.330957  | $H_0$ Accepted |
| the percentage of households with a clean and healthy lifestyle ($X_2$) | 1.354963  | $H_0$ Accepted |
| percentage of pregnant women carrying out the First Visit ($X_3$) | 2.700359  | $H_0$ Rejected |

From the significant parameter test, it was found that the percentage of pregnant women who consumed Fe3 tablets and the percentage of households with a clean and healthy lifestyle did not have a significant effect on the number of maternal mortality. Meanwhile, the variable percentage of pregnant women who made the first visit to a midwife/doctor (“Kunjungan Pertama” or K1) had a significant effect on the number of maternal mortality. While the measure of the goodness of the model can be seen with the R-square value is 0.222679, and the Mean Square Error (MSE) is 6.894871.

4.2. **Checking Assumption of multiple regression model for the number maternal mortality in pregnant**

In this study, four classical assumptions are used to see whether the estimator obtained using multiple linear regression with the estimation of the least squares parameter is unbiased with a minimum (efficient) variant. The tests conducted include residual testing for normality, homocedasticity, non-autocorrelation and non-multicollinearity between independent variables. The following are the results of the classic assumption test.

Table 2. The Checking assumption of the number maternal mortality in pregnant model

| Testing            | Method                        | Results                                      | Decision                      |
|--------------------|-------------------------------|----------------------------------------------|-------------------------------|
| Normality          | Kolmogorov Smirnov testing   | $D_{\text{value}} = 0.136$ ; $d_{\text{table}} = 0.202$ | $H_0$ Accepted. It shows that the residual has normal distribution |
| Multicollinearity  | Variance Inflation Factors   | The VIF’s value of $X_1$ is 1.188017; $X_2$ is 1.119417; for $X_3$ is 1.082812 | The VIF’s values are less than 5, so these indicate that no-multicollinearity |
| Autocorrelation    | Durbin Watson test            | $d = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}$ for $k = 3$ and $n = 35$ $\alpha = 0.05$ get $d_U = 1.652B$ and $d_L = 1.2833$. It shows that there is no autocorrelation for residual |
| Homoscedasticity   | Glejser test                  | The $X_1$’s t-value is 0.9488411; The $X_2$’s t-value is 1.7964508; The $X_3$’s t-value is 1.4669356 | The t values are less than $t_{0.5\%,31} = 1.69552$ so it is indicate that the variance of residual is heteroscedasticity. |

In the multiple regression model assumption test, the unfulfilled assumption is the homoscedasticity of the residuals. Therefore, the next step is to check for outliers or not. According to Cohen (2003), a data is called an outlier if the value $|\text{DFFITS}| > 1$ is for small to medium sized data. Due to the number of data ($n$) = 35 including small or medium data sets, the limit used is $|\text{DFFITS}| > 1$. Based on the output results, it is found that the 15th data, namely Grobogan City with a value of $|\text{DFFITS}|$ amounting to 2.776831 which is greater than 1. This indicates that the data is an outlier. The existence of 1 outlier in this case study, resulting in regression with parameter estimation using Ordinary Least Square (OLS) cannot be used to see the effect of the predictor variables. Therefore, robust regression with M-estimate using Andrew's weight and MM-estimate will be used to solve this case.
4.3. Modelling Robust with MM-Estimator for Modelling The Number Maternal Mortality of Pregnant in Central Java

The MM-estimator is a combination of the S-estimator and the M-estimator. For this reason, iteration is carried out using the S-estimator first in setting the scale up to $\hat{\beta}$ convergent. Then proceed with the M-Estimator iteration using the scale obtained in the S-Estimator iteration. Based on the robust MM-estimator output results, it is known that 19 iterations of the S-estimator and 5 iterations of the M-estimator have been made to obtain a convergent estimator. The robust MM-estimator regression model obtained is as follows:

$$Y = 108.0497 + 0.03124X_1 - 0.047X_2 - 1.047X_3$$

The estimator regression model has the same effect as the multiple regression model. Variable $X_1$ has a positive effect, meaning that the higher the percentage value of pregnant women who consume Fe tablets will increase the number of maternal mortality. Meanwhile, the variables $X_2$ and $X_3$ have a negative effect. This means that the higher the percentage of households that have a clean and healthy lifestyle and the higher the percentage of pregnant women who make the first visit, the lower the number of maternal mortality. When viewed from the magnitude of the coefficient, the parameter coefficient on the $X_1$ and $X_3$ variables in the MM regression model is greater than the multiple regression model. This shows that these variables have a greater influence on the maternal mortality model using robust MM estimator regression. The next step is testing the significance of the model using the F test that using Andrew weighted as follows:

Hypothesis

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (the model is not appropriate)

$H_1 : \text{at least there is } \beta_j \neq 0, j = 1,2,3 \text{ (the model is appropriate) }$

The criteria of testing uses F testing with the value of F is 3.8116. Reject $H_0$ for $F > F(\alpha, k, n-p)$ while $F_{(5\%,3,31)} = 2.91$. So $H_0$ rejected, it is indicated that the model of regression robust with MM-estimator is appropriate. The following steps is partial testing using t test.

Hypothesis

$H_0 : \beta_j = 0$ (The variable of $X_j$ is not significant)

$H_1 : \beta_j \neq 0$ (The variable of $X_j$ is significant); for $j=1,2,3$

Criteria of testing is $H_0$ rejected for $|t| > t(\alpha/2, n-p) ; t(0.5\%,31) = 1.69552$

Similar to the multiple regression model, in the Robust regression model only the $X_3$ variable is significant. The resulting R square value is 0.2695 and the MSE value is 4.989. It shows that the robust MM regression model is better for modeling the number of maternal mortality than the multiple regression model.

\begin{table}[h]
\centering
\begin{tabular}{c|c|c|c|c|c}
\hline
Variable & Standard Error & t-value & p-value & Decision \\
\hline
$X_1$ & 0.0748 & 0.417 & 0.6792 & $H_0$ Accepted \\
\hline
$X_2$ & 0.0435 & -1.0831 & 0.287 & $H_0$ Accepted \\
\hline
$X_3$ & -0.3244 & -3.2475 & 0.0028 & $H_0$ Rejected \\
\hline
\end{tabular}
\caption{Significant testing of parameter which Regression Robust with MM-Estimator}
\end{table}

Table 3 shows the partial parameter testing. Based on table 3 it can be concluded that at the 5% significance level the variables $X_2$ and $X_3$ do not have a significant effect on maternal mortality modeling. Meanwhile, the significant variable is $X_3$, which is the percentage of pregnant women carrying out the First Visit. The coefficient value on the $X_3$ variable is negative. This shows that the greater the value of the $X_3$ variable, the lower the number of maternal deaths.

5. Conclusions

Multiple regression modeling is quite easy to do, but many assumptions must be fulfilled. In modeling the number of maternal mortality in pregnant, the Robust MM estimator regression model produces a
good measure of the model. However, the variable percentage of pregnant women who consume Fe tablets and the percentage of households that have a proper and clean lifestyle is not significant to the model. So that for further analysis, it is necessary to examine why these two variables are not significant. In addition, the variable percentage of pregnant women who consume Fe tablets has a positive impact on the number of mortality of pregnant women, this is certainly contrary to theory. However, these variables have very little impact if modeled together with variables $X_2$ and $X_3$.

Acknowledgment
This paper is outcome of the research founded by Diponegoro University through Statistics Department Faculty of Sciences and Mathematics with No. of the contract is 2015/UN7.5.8/PP/2020.

References
[1] [Dinkes Jateng] Dinas Kesehatan Provinsi Jawa Tengah 2019 Profil Kesehatan Provinsi Jawa Tengah Tahun 2017 (Semarang: Dinas Kesehatan Provinsi Jawa Tengah)
[2] Badan Pusat Statistika (BPS) Jawa Tengah 2018 Jawa Tengah dalam Angka 2018 (Semarang: Badan Pusat Statistika Provinsi Jawa Tengah)
[3] Darnah, Utoyol M I, Chamidah N 2019 IOP Conf. Series: Earth and Environmental Science 243
[4] Dewi D K, Purhadi and Sutikno 2019 IOP Conf. Series: Material Sciences and engineering 546
[5] Montgomery D C and Peck E A 1992 Introduction To Linier Regression Analysis (New York: John Wiley and Sons, Inc.)
[6] Chen C 2002 Robust Regression and Outlier Detection with The ROBUSTREG Procedure, paper 265-27 (USA: SAS Institute Inc.)
[7] Cohen J, Cohen P, West S G, Aiken L S 2003. Applied Multiple Regression Correlation Analysis for The Behavioral Sciences (New Jersey: Lawrence Erlbaum Associates)
[8] Draper N R and Smith H 1998 Applied Regression Analysis (New York: John Wiley and Sons)
[9] Stuart C 2011 Robust Regression (England: Durham University)
[10] Susanti Y 2014 Int. J. Pure Appl. Math. 91 3 349-360
[11] Samkar H and Alpu O 2010 J. Mod. Appl. Stat. Methods 9 495-501