Linear regression based on Minimum Covariance Determinant (MCD) and TELBS methods on the productivity of phytoplankton

N Gusriani*, Firdaniza
Department of Mathematics, Faculty of Mathematics and Science, Padjadjaran University, Jalan Raya Bandung-Sumedang Km 21 Jatinangor, Indonesia

*Corresponding author: nurul.gusriani@unpad.ac.id

Abstract. The existence of outliers on multiple linear regression analysis causes the Gaussian assumption to be unfulfilled. If the Least Square method is forcedly used on these data, it will produce a model that cannot represent most data. For that, we need a robust regression method against outliers. This paper will compare the Minimum Covariance Determinant (MCD) method and the TELBS method on secondary data on the productivity of phytoplankton, which contains outliers. Based on the robust determinant coefficient value, MCD method produces a better model compared to TELBS method.

1. Introduction
Outlier is one of the problems in linear regression analysis that gives a very big influence on the model if it is used the Least Squares method (LS) [8]. The existence of the outliers in the data will cause the regression line to shift toward the outliers. This will cause the resulting error to be large so that the Gaussian assumption of normality may not be met. In LS method, outliers are usually set aside from the data set. Setting aside this data is not a wise procedure because sometimes the data provides information that other data cannot provide. To deal with this, the robust method is used as an alternative of LS method. The robust procedure is intended to accommodate the data weirdness, as well as eliminate the identification of data outliers and also automatic in tackling outliers [2].

Some of the robust methods that have been introduced include M estimator proposed by Huber in 1973, Least Median Square (LMS) and Least Trimmed Square (LTS) proposed by Rousseeuw in 1984 [6]. Birch introduced the robust method that was a generalization of the M estimator [1] called the Bounded-Influence estimator (B-I estimator). Furthermore, the regression of Minimum Covariance Determinant (MCD) proposed by Rousseeuw in 2004 and further deepened by Hubert [6]. The latest method is the TELBS method [11]. Similar to the MCD method, the TELBS method can deal with the problem of calling either in space X or space Y.

Based on the similarity in its ability to handle the problem of outliers, this paper examines TELBS and MCD methods applied to secondary data on the productivity of phytoplankton. The model obtained from both of these methods will be compared based on the robust coefficient of determination to determine the best model. The data centralize measurement used in robust determinant coefficients is median in accordance with the nature of data containing the outliers.
2. Linear regression analysis

Regression analysis is a statistical application that can build the relationship model between the independent variable \(X\) and the dependent variable \(Y\). In the linear model, the commonly used method for estimating the parameter is the Least Squares method (LS). The LS method is the easiest method to perform with certain assumptions called the Gaussian assumption [4]. If all Gaussian assumptions are met then \(\hat{\beta}\) is a linear, non-biased and has minimum variance estimator.

Model of linear regression sample in matrix form, expressed by:

\[
y = X\beta + e
\]  

where \(y\) is the dependent variable observation vector (\(n \times 1\)), \(X\) is the independent variable observation matrix (\(n \times (k + 1)\)), \(\beta\) is the parameter to be estimated ((\(k + 1) \times 1\)) and \(e\) is a vector residual (\(n \times 1\)). Using the LS method, the component estimated by the principle is to minimize the sum of the residual squares to produce the following estimates:

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

The use of the LS method in estimating parameter parameters is a very well known method because it is easy to understand and to calculate. However, the use of this method is usually influenced by the presence of outliers. Outlier is the furthest observation of data centre that may have a major effect on regression coefficients [10]. The condition will arise if the sample data presented is varied and extreme. Every time there is an outlier, the model will usually change, so if used as a prediction model will get poor results and give a wrong interpretation. The existence of outliers can cause errors and variance of data to be large. Consequently, the parameter interval becomes large.

Outlier is rejected if after searching it results in errors in measurement or analysis, inaccuracy of data recording, and damage to the measuring tool. If it is not the result of such mistakes, careful investigation must be made [3]. An investigation or diagnosis in regression analysis is one way to monitor problems arising in relation to data or models. One way of diagnosing data related to outlier is to use a hat matrix [7] defined as follows:

\[
H = X(X'X)^{-1}X'
\]

The matrix hat essentially transforms the observed value \(y\) vector to the estimated value \(\hat{y}\) vector. The main diagonal of the hat matrix is that \(h_{ii}\) will show a value greater than \(2p / n\), if the data is the outlier where \(p = k + 1\).

2.1. Minimum Covariance Determinant (MCD) method

The Minimum Covariance Determinant (MCD) estimator is one of the first affine equivariant and highly robust estimators of multivariate location and scatters [6]. MCD method begins by determining the subset of matrix \(X\) as much as \(h\) observations are:

\[
h = \frac{n+k+1}{2}
\]

So there will be a combination of observations as much as \(a = C^n_h\). The subsets of the matrix \(X\) are denoted by:
From each $H_b$, call as $H_{bm}$, the average vector and the covariance matrix are determined by:

$$t_{bm} = \frac{1}{h} (H_{bm})' v$$  \hspace{1cm} (6)$$

$$C_{bm} = \frac{1}{h} [H_{bm} - vt_{bm}'] [H_{bm} - vt_{bm}]$$  \hspace{1cm} (7)$$

where $v$ is the vector column is worth 1 ($h \times 1$). Next for each $b$ is calculated the value of $\text{det}(C_{bm})$. For $m = 1$, if $\text{det}(C_{bm}) \neq 0$, then the distance of the Mahalanobis, is the element in the diagonal of the matrix, is determined by:

$$d_{MD}^2 = \left| X' - v' \bar{X} \right| C^{-1} \left( X' - v' \bar{X} \right)'$$  \hspace{1cm} (8)$$

where $X'$ is the matrix of the observed independent variables ($n \times k$), $v'$ is column vector is worth 1 ($n \times 1$). $C^{-1}$ is the inverse of the covariance matrix ($k \times k$), and $\bar{X}$ is the average matrix of each row ($1 \times k$). The distance of the mahalanobis is sorted from the smallest value and taken as $h$, then the selected number of observations is identified. After that, a new $X'$ matrix is formed as $H_{bm}$ ($m = m + 1$) and the new $t_{bm}$ and $C_{bm}$ values are determined. The next step is to compare the value of $\text{det}(C_{bm})$ and $\text{det}(C_{bm-1})$.

1. If $\text{det}(C_{bm}) \neq \text{det}(C_{bm-1})$, using equation (8) recovered the value of the distance of the new mahalanobis, then new $X'$ is formed to obtain new $H_{bm}$ matrix;
2. If $\text{det}(C_{bm}) = \text{det}(C_{bm-1})$, data processing is done from by taking another subset.

This process is done until all of the $H_b$ subsets are out. Further, the smallest determinant of covariance of all determinants of the subsets of $H_b$ is selected. The subset with the smallest covariance determinant is called $H_{MCD}$, so further can be also determined $t_{MCD}$ and $C_{MCD}$. The $H_{MCD}$ matrix element is the independent variable of the selected sample unit based on the smallest order of the mahalanobis distance so that it can be determined also the pair of the dependent variable.

The variance-covariance matrix of the variables $X$ and $Y$ variables whose chosen probes can be obtained by the following formula:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{XX} & \hat{\Sigma}_{XY} \\ \hat{\Sigma}_{YX} & \hat{\Sigma}_{YY} \end{bmatrix}$$  \hspace{1cm} (9)$$

Then we can determine covariance of error from:

$$\hat{\Sigma}_e = \hat{\Sigma}_{YY} - \hat{\theta}_{MCD}' \hat{\Sigma}_{XY} \hat{\theta}_{MCD}$$  \hspace{1cm} (10)$$

with $\hat{\theta}_{MCD} = \hat{\Sigma}_{YY} \hat{\Sigma}_{XY}$. 
The main principle in the MCD method is to calculate the weights $w_{ii} = 1$ if $d(e_{MCD}) \leq \sqrt{x_{1-\alpha}}$ with $d(e_{MCD}) = \sqrt{e_{MCD}(\sum_{e_{MCD}})^{-1}(e_{MCD})'}$ so $W_{MCD}$ matrix $(n \times n)$ is:

$$W_{MCD} = \begin{bmatrix}
    w_{11} & w_{12} & \cdots & w_{1n} \\
    w_{21} & w_{22} & \cdots & w_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    w_{n1} & w_{n2} & \cdots & w_{nn}
\end{bmatrix}$$ (11)

with the matrix entries $w_{ij} = 0$ for $i \neq j$. This matrix is the robust weight matrix and is used to obtain sensitivity parameter estimates with the following equations:

$$\hat{\beta}_{MCD} = \left(X'W_{MCD}X\right)^{-1}X'W_{MCD}y$$ (12)

### 2.2. TELBS method

One of the regression methods to estimate regression parameters when there are outliers is the TELBS method. TELBS estimator is better than the least squares method, M and MM estimator and its performed by minimizing objective function [11]:

$$\min_\beta \sum_{i=1}^n \frac{\rho(t_i)}{L_i} = \min_\beta \sum_{i=1}^n \frac{1 - \text{sech}(\omega t_i)}{L_i}$$ (13)

with

$$t_i = \frac{(y_i - x_i'\hat{\beta})(1 - h_{ii})}{\sigma}$$ (14)

$$L_i = \sum_{j=1}^i \max \left\{ M_j, |x_{ij}| \right\}$$ (15)

$$M_j = \text{median} \left\{ |x_{i1}|, |x_{i2}|, \ldots, |x_{in}| \right\}$$ (16)

$\omega$ are positive real numbers called constants of constancy (value 0.628). Estimator value $\sigma$ can be obtained from the equation as follows:

$$\hat{\sigma} = 1.1926 \text{median}_{i,j \in \text{out}} \left( \text{median}_{j \in \text{out}} (r_i - r_j) \right)$$ (17)

The selection of the constant 1.1926 makes $\sigma$ a near-unbiased approximation of the limited sample [9]. To minimize equation (13) it must be derived to the parameter $\hat{\beta}$ and equal to zero, so that equation (13) can be expressed as:
\[
\sum_{i=1}^{n} \frac{\psi(t_i)(1-h_{ii})}{L_i} \frac{\partial(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - ... - \beta_k x_{ik})}{\partial \beta_j} = 0
\]  

(18)

with \(\psi(x) = \frac{\partial \rho(x)}{\partial x} = \omega \text{Sech}(\omega x) \text{Tanh}(\omega x)\).

The weighted function is defined (11):

\[
w_{ii}^* = \frac{\psi(t_i)(1-h_{ii})}{\sigma e_i L_i}
\]

(19)

then equation (18) can be written to be:

\[
\sum_{i=1}^{n} w_{ii}^* e_j \frac{\partial(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - ... - \beta_k x_{ik})}{\partial \beta_j} = 0
\]

(20)

Equation (20) if it is derived to \(\beta_0\) and \(\beta_j\) will produce the matrix form \((X'W^*X)\hat{\beta} = X'W^*y\) so we obtain an estimate of TELBS methods are:

\[
\hat{\beta}_{TELBS} = \left(X'W^*X\right)^{-1} X'W^*y
\]

(21)

where \(W^*\) is the square matrix with the diagonal element is \(w_{ii}^*\) in equation (19) and the matrix entry \(w_{ij}^* = 0\), for \(i \neq j\). In the TELBS method the estimated value obtained by the iteration process until the convergence is achieved.

The robust determination coefficient is then determined to see the suitability of the model as follows [11]:

\[
R^2 = 1 - \left(\frac{\text{Median}_{i \in \text{Sign}} |e_i|}{\text{Median}_{i \in \text{Sign}} \ |y_i - \text{Median}_{i \in \text{Sign}} y_i|} \right)^2
\]

(22)

3. Data

The secondary data used is the result of research on the primary productivity of phytoplankton with the physicochemical factors of aquatic in aquaculture with floating net [5]. From the result of the research, there are twenty data pair, with primary productivity (gr.L/m^2) as an independent variable (\(Y\)) and three independent variables namely light intensity (kcal/m^2) as \(X_1\), PH (\(X_2\)), and phytoplankton density (ind/mL) as \(X_3\).

4. Results and discussion

The hat matrix is used to detect the presence of the outlier. The main diagonal of the hat matrix is shown in Figure 1. It appears that the \(17^\text{th}\), \(18^\text{th}\), and \(19^\text{th}\) data are the outliers represented by a value greater than \(2p/n = 0.3\). If forced by using LS, the value of determination coefficient is 0.57762, meaning that only 57.762% variability of primary production of phytoplankton can be explained by the factors that influence it.
Figure 1. \( h_{ij} \) value of phytoplankton productivity

The small coefficient of determination in LS method becomes the reason for modeling using MCD method and TELBS method. Using Maple 15, the results are presented in the following table:

| Component | MCD method | TELBS method |
|-----------|------------|--------------|
| \( \beta_0 \) | 0.5855     | -0.4259      |
| \( \beta_1 \) | \( 10^{-4} \) | \( 1.75 \times 10^{-4} \) |
| \( \beta_2 \) | 0.0826     | 0.0584       |
| \( \beta_3 \) | \( 3.10^{-4} \) | \( 2.232 \times 10^{-4} \) |
| \( R^2 \) | 0.8994     | 0.8781       |

The results of Table 1 above show that based on the coefficient of determination, MCD produces a better model than TELBS method. This is reinforced by the residues values illustrated in figure below which shows that the residues generated by MCD tend to be less than the TELBS model.

Figure 2. Residues based on MCD and TLBS methods

Thus the best model for the phytoplankton productivity is as follows:

\[
\hat{y}_i = 0.5855 + 10^{-4} x_{i1} + 0.0826 x_{i2} + 3.10^{-4} x_{i3}
\]  

(23)

The coefficient of determination with MCD method gives a value of 0.8994. This means that 89.94% of the primary production variability of phytoplankton can be explained by the factors of light intensity, pH, and phytoplankton density. The remaining 10.06% is determined by other factors, which are not included in the study.

5. Conclusion

In this paper, we develop a linear model of the productivity of phytoplankton. It has been shown, that based on robust coefficient determinant, the MCD model is better than TELBS when data is contaminated by outlier.
Acknowledgments
The authors thank to the Padjadjaran University, due to the funding of this work through the scheme of Riset Fundamental Unpad (RFU) with contract number 872/UN6.3.1/LT/2017.

References
[1] Birch J B 1992 Estimation and Inference in Multiple Regression Using Robust Weight: A Unified Approach Technical Report 92-2, Departemen of Statistics Virginia Polytechnic Institute and State University, Blackburg Virginia.
[2] Cahyawati D 2009 Efektifitas Metode Regresi Robust Penduga Welsch dalam Mengatasi Pencilan pada Pemodelan Regresi Linear Berganda, Jurnal Penelitian Sains, Vol.12, no.1(A), Unsri, Sumatera Selatan.
[3] Chandraningtyas S, dkk. 2013 Regresi Robust MM-Estimators untuk Penanganan Pencilan pada Regresi Linier Berganda. Jurnal Gaussian, Vol. 2, no.4, 395-404, Undip, Semarang.
[4] Gujarati D N 1998 Basic Econometrica, 2nd Edition, New York, McGraw-Hill Inc.
[5] Gusriani, Firdaniza, Ardelina D 2011 Kajian Penaksir Bounded-Influence dan Metode Minimum Covariance Determinant pada Analisis Regresi Linier Berganda untuk Kasus Pencilan, Laporan Penelitian, Jurusan Matematika FMIPA Unpad, Bandung.
[6] Hubert Mia et al. 2008 High-Breakdown Robust Multivariate Methods Statistical Science Vol. 23 No.1 (online: http://arxiv.org/pdf/0808.0657).
[7] Myers R H 1990 Classical and Modern Regression With Applications 2nd edition Boston PSW-KENT Publishing Company.
[8] Rousseeuw P J and Leroy A M 1986 Robust Regression and Outlier Detection England: John Wiley & Sons Ltd.
[9] Rousseeuw P J and Croux 1993 Alternative to the Median Absolute Deviation American Statistical Association. 1993, Vol. 88, No. 424.
[10] Sembiring R K 2003 Analisis Regresi, Edisi 2. Bandung: ITB.
[11] Tabatabai M A et al. 2012 TELBS Robust Linear Regression Method Open Access Medical Statistics, USA, Dove Medical Press.