M estimation, S estimation, and MM estimation in robust regression

Article · March 2014

DOI: 10.12732/ijpam.v91i3.7

CITATIONS
87

READS
6,821

4 authors, including:

Yuliana Susanti
Universitas Sebelas Maret
18 PUBLICATIONS **121** CITATIONS
SEE PROFILE

Hasih Pratiwi
Universitas Sebelas Maret
38 PUBLICATIONS **145** CITATIONS
SEE PROFILE

Twenty Liana
BPTP Kalimantan Tengah
4 PUBLICATIONS **90** CITATIONS
SEE PROFILE

International Journal of Pure and Applied Mathematics
Volume 91 No. 3 2014, 349-360
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: http://dx.doi.org/10.12732/ijpam.v91i3.7

M ESTIMATION, S ESTIMATION, AND MM ESTIMATION IN ROBUST REGRESSION

Yuliana Susanti[1 §], Hasih Pratiwi[2],
Sri Sulistijowati H.[3], Twenty Liana[4]
[1,2,3]Sebelas Maret University
Jl. Ir. Sutami 36A Surakarta, INDONESIA
[4]Assessment Institute for Agricultural Technology of
Kalimantan Tengah, Jl. G. Obos Km. 5
Palangkaraya, INDONESIA

Abstract: In regression analysis the use of least squares method would not be appropriate in solving problem containing outlier or extreme observations. So we need a parameter estimation method which is robust where the value of the estimation is not much affected by small changes in the data. In this paper we present M estimation, S estimation and MM estimation in robust regression to determine a regression model. M estimation is an extension of the maximum likelihood method and is a robust estimation, while S estimation and MM estimation are the development of M estimation method. The algorithm of these methods is presented and then we apply them on the maize production data.

AMS Subject Classification: 62J05, 62G35
Key Words: robust regression, M estimation, S estimation, MM estimation

1. Introduction

Robust regression analysis provides an alternative to a least squares regres-

Received: November 14, 2013

§Correspondence author

sion model when fundamental assumptions are unfulfilled by the nature of the data. When the analyst estimates his statistical regression models and tests his assumptions, he frequently finds that the assumptions are substantially violated. Sometimes the analyst can transform his variables to conform to those assumptions. Often, however, a transformation will not eliminate or attenuate the leverage of influential outliers that bias the prediction and distort the significance of parameter estimates. Under these circumstances, robust regression that is resistant to the influence of outliers may be the only reasonable recourse.

The well-known methods of robust estimation are M estimation, S estimation and MM estimation. M estimation is an extension of the maximum likelihood method and is a robust estimation [11], while the S estimation and MM estimation is the development of M estimation method. By using these methods it is possible to eliminate some of the data, which in some cases it could not always be done additionally if that data is important, such as those which often be found on agriculture field [10], [9]. In this paper we present a robust regression method to determine the optimum regression model.

2. Linear Regression Model

Linear regression is an approach to model the relationship between a scalar response or dependent variable Y and one or more explanatory or independent variables denoted X. In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. A linear regression model involving p independent variables can be expressed as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, i = 1, 2, \cdots, n.$$

Y_i is the response variable on the i-th observation, $\beta_0, \beta_1, \cdots, \beta p$ are parameters, X_i is the value of the independent variable on the i-th observation, and ε_i is a normally distributed random variable. The error $\varepsilon_i \sim N(0, \sigma^2)$ is not mutually correlated [5].

The most commonly used regression method is the method of ordinary least squares (OLS). The OLS estimate is obtained as the solution of the problem

$$\min J = \min \sum_{i=1}^{n} \varepsilon_i^2$$

Taking the partial derivatives of J with respect to $\beta_j, j = 0, 1, \cdots, p$ and setting them equal to zero yields the normal equations and obtains the estimated

regression model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$$

To judge how well the estimated regression model fits the data, we can look at the size of the residuals

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}).$$

A point which lies far from the line (and thus has a large residual value) is known as an outlier. Such points may represent erroneous data, or may indicate a poorly fitting regression line. The ordinary or simple residuals (observed - predicted values) are the most commonly used measures for detecting outliers. Standardized residuals are the residuals divided by the estimates of their standard errors. They have mean 0 and standard deviation 1.

3. Robust Regression

Robust regression is a regression method that is used when the distribution of residual is not normal or there are some outliers that affect the model. This method is an important tool for analyzing the data which is affected by outliers so that the resulting models are stout against outliers [4]. When researchers set of regression models and to test the common assumption that the regression assumptions are violated, the transformation seemed unlikely to eliminate or weaken the influence of outliers which eventually became biased predictions. Under these circumstances, robust regression is resistant to the influence of outliers is the best method. Robust regression is used to detect outliers and provide results that are resistant to the outliers [3].

3.1. M Estimation

One of the robust regression estimation methods is the M estimation. The letter M indicates that M estimation is an estimation of the maximum likelihood type. If estimator at M estimation is $\hat{\beta} = \beta_n(x_1, x_2, \cdots, x_n)$ then

$$E[\beta_n(x_1, x_2, \cdots, x_n)] = \beta. \tag{1}$$

Equation (1) shows that the estimator $\hat{\beta} = \beta_n(x_1, x_2, \cdots, x_n)$ is unbiased and has minimum variance, so M-estimator has the smallest variance estimator compared to other estimators of variance:

$$var(\hat{\beta}) \geq \frac{[\tilde{\beta}']^2}{nE\Big(\frac{d}{d\beta} \ln f(x_i; \beta)\Big)^2}$$

where $\hat{\hat{\beta}}$ is other linear and unbiased estimator for β.

M estimation is an extension of the maximum likelihood estimate method and a robust estimation [11]. In this method it is possible to eliminate some of the data , which in some cases is not always appropriate to do especially if it is eliminated is an important data or seed, whose case often encountered in agriculture [10], [9]. M estimation principle is to minimize the residual function ρ:

$$\hat{\beta}_M = \min_{\beta} \rho(y_i - \sum_{j=0}^{k} x_{ij}\beta_j). \tag{2}$$

We have to solve

$$\min_{\beta} \sum_{i=1}^{n} \rho(u_i) = \min_{\beta} \sum_{i=1}^{n} \rho(\frac{e_i}{\sigma}) = \min_{\beta} \sum_{i=1}^{n} \rho(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{\sigma})$$

to obtain (2), and we set estimator for σ:

$$\hat{\sigma} = \frac{MAD}{0.6745} = \frac{median|e_i - median(e_i)|}{0.6745}.$$

For ρ function we use the Tukey's bisquare objective function:

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^4}, & |u_i| \le c \\ \frac{c^2}{6}, & |u_i| > c. \end{cases}$$

Furthermore we look for first partial derivative $\hat{\beta}_M$ to β so that

$$\sum_{i=1}^{n} x_{ij}\psi(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}}) = 0, j = 0, 1, \ldots, k \tag{3}$$

where $\psi = \rho', x_{ij}$ is i-th observation on the j-th independent variable and $x_{i0} = 1$.

Draper and Smith [4] give a solution for equation (3) by defining a weighted function

$$w(e_i) = \frac{\psi(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}})}{(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}})} \tag{4}$$

Because $u_i = \frac{e_i}{\hat{\sigma}}$, we can rewrite equation (4) with

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

We take $c = 4.685$ for Tukey's bisquare weighted function. So equation (3) becomes

$$\sum_{i=1}^{n} x_{ij} w_i (y_i - \sum_{j=0}^{k} x_{ij}\beta) = 0, j = 0, 1, \ldots, k. \tag{5}$$

Equation (5) can be solved by iteratively reweighted least squares (IRLS) method. In this method we assume that there is an initial estimate $\hat{\beta}^0$ and $\hat{\sigma}_i$ is a scale estimate. If j is numbers of parameters then

$$\sum_{i=1}^{n} x_{ij} w_i^0 (y_i - \sum_{j=0}^{k} x_{ij}\beta^0) = 0, j = 0, 1, \ldots, k. \tag{6}$$

In matrix notation, equation (6) can be written as

$$\mathbf{X'W_iX}\beta = \mathbf{X'W_iY} \tag{7}$$

where $\mathbf{W_i}$ is a $n \times n$ matrix with its diagonal elements are the weighted. Equation (7) is known as weighted least squares (WLS) equation. Solution for this equation gives an estimator for β, i.e. $\hat{\beta} = (\mathbf{X'W_iX})^{-1}(\mathbf{X'W_iY})$. A detailed description of M estimation is presented in Algorithm 1.

Algorithm 1

1. Estimate regression coefficients on the data using OLS.
2. Test assumptions of the regression model
3. Detect the presence of outliers in the data.
4. Calculate estimated parameter $\hat{\beta}^0$ with OLS.
5. Calculate residual value $e_i = y_i - \hat{y}_i$.
6. Calculate value $\hat{\sigma}_i = 1.4826$ MAD.
7. Calculate value $u_i = e_i/\hat{\sigma}_i$.
8. Calculate the weighted value

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

9. Calculate $\hat{\beta}_M$ using weighted least squares (WLS) method with weighted w_i.
10. Repeat steps 5-8 to obtain a convergent value of $\hat{\beta}_M$.
11. Test to determine whether independent variables have significant effect on the dependent variable.

3.2. S Estimation

The regression estimates associated with M-scales is the S-estimators which proposed by Rousseeuw and Yohai [6]. S estimation is based on residual scale of M estimation. The weakness of M estimation is the lack of consideration on the data distribution and not a function of the overall data because only using the median as the weighted value. This method uses the residual standard deviation to overcome the weaknesses of median. According to Salibian and Yohai [7], the S-estimator is defined by $\hat{\beta}_s = \min_\beta \hat{\sigma}_s(e_1, e_2, \ldots, e_n)$ with determining minimum robust scale estimator $\hat{\sigma}_s$ and satisfying

$$\min \sum_{i=1}^{n} \rho\Big(\frac{y_i - \sum_{i=1}^{n} x_{ij}\beta}{\hat{\sigma}_s}\Big)$$

where

$$\hat{\sigma}_s = \sqrt{\frac{1}{nK}\sum_{i=1}^{n} w_i e_i^2}$$

$K = 0.199, w_i = w_\sigma(u_i) = \frac{\rho(u_i)}{u_i^2}$, and the initial estimate is

$$\hat{\sigma}_s = \frac{median|e_i - median(e_i)|}{0.6745}.$$

The solution is obtained by differentiating to β so that

$$\sum_{i=1}^{n} x_{ij}\psi(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}_s}) = 0, j = 0, 1, \ldots, k \tag{8}$$

ψ is a function as derivative of ρ:

$$\psi(u_i) = \rho'(u_i) = \begin{cases} u_i\Big[1 - \Big(\frac{u_i}{c}\Big)^2\Big]^2, & |u_i| \le c \\ 0, & |u_i| > c. \end{cases}$$

where w_i is an IRLS weighted function:

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

$u_i = \frac{e_i}{\sigma_s}$ and $c = 1.547$. We can solve equation (8) by using IRLS method. Algorithm 2 shows several stages in S estimation.

Algorithm 2

1. Estimate regression coefficients on the data using OLS.
2. Test assumptions of the classical regression model
3. Detect the presence of outliers in the data.
4. Calculate $\hat{\beta}^0$ with OLS.
5. Calculate residual value $e_i = y_i - \hat{y}_i$.
6. Calculate value

$$\hat{\sigma}_i = \begin{cases} \frac{median|e_i - median\ e_i|}{0.6745}, & \text{iteration} = 1; \\ \sqrt{\frac{1}{nK}\sum_{i=1}^{n} w_i e_i^2}, & \text{iteration} > 1. \end{cases}$$

7. Calculating value $u_i = \frac{e_i}{\hat{\sigma}_i}$
8. Calculate weighted value

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

9. Calculate $\hat{\beta}_S$ with WLS method with weighted w_i.
10. Repeat steps 5-8 to obtain a convergent value of $\hat{\beta}_S$.
11. Test to determine whether independent variables have significant effect on the dependent variable.

3.3. MM Estimation

MM estimation procedure is to estimate the regression parameter using S estimation which minimize the scale of the residual from M estimation and then proceed with M estimation. MM estimation aims to obtain estimates that have a high breakdown value and more efficient. Breakdown value is a common measure of the proportion of outliers that can be addressed before these observations affect the model [3]. MM-estimator is the solution of

$$\sum_{i=1}^{n} \rho_1'(ui)X_{ij} = 0 \text{ or } \sum_{i=1}^{n} \rho_1'\Big(\frac{Y_i - \sum_{j=0}^{k} X_{ij}\hat{\beta}_j}{s_{MM}}\Big)X_{ij} = 0$$

where s_{MM} is the standard deviation obtained from the residual of S estimation and ρ is a Tukey's biweight function:

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^2}, & -c \le u_i \le c; \\ \frac{c^2}{6}, & u_i < -c \text{ or } u_i > c. \end{cases}$$

Algorithm 3

1. Estimate regression coefficients on the data using the OLS.
2. Test assumptions of the classical regression model.
3. Detect the presence of outliers in the data.
4. Calculate residual value $e_i = y_i - \hat{y}_i$ of S estimate.
5. Calculate value of $\hat{\sigma}_i = \hat{\sigma}_{sn}$.
6. Calculate value $u_i = \frac{e_i}{\hat{\sigma}_i}$.
7. Calculate weighted value

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

8. Calculate $\hat{\beta}_M M$ using WLS method with weighted w_i .
9. Repeate steps 5-8 to obtain a convergent value of $\hat{\beta}_M M$.
10. Test to determine whether independent variables have significant effect on the dependent variable.

4. Best Regression Model

In application we use a secondary data obtained from the Indonesian Ministry of Agriculture and BPS-Statistics Indonesia in 2011 [1]. To achieve the research objectives, we do the following steps:

1. Identify factors that affect maize production
2. Investigate the correlation among variables.
3. Estimate regression model with all factors using OLS method of least squares and test all assumptions.
4. Determine outliers
5. Estimate regression model using M estimation, S estimation and MM estimation
6. Test whether independent variables have significant effect on the dependent variable.

Not all factors are suspected to affect maize production availability for each province. The complete data associated with maize production (in tons) was harvested area (in hectares), monthly average rainfall (in millimeters), monthly average humidity (in percent), monthly average temperature (in degree Celcius), monthly average long the sun shines (in percent), and numbers of agricultural man power in food crops subsector (in person). We will expressed maize production (Y) as a function of harvested area (X_1), monthly average rainfall (X_2), monthly average humidity (X_3), monthly average temperature (X_4), monthly average long the sun shines (X_5), and numbers of agricultural man power in food crops subsector (X_6). The estimated regression model using OLS is

$$\hat{y} = 623773 + 4.17x_1 + 25x_2 - 8521x_3 + 14097x_4 - 5668x_5 + 0.155x_6 \quad (9)$$

with $R^2 = 98.8\%$, $R^2_{adjusted} = 98.5\%$ and $s = 131501$. The value of $F = 354.81$ with $p-$value $= 0 < 5\%$ indicates that the linear regression model (9) fits with the data. Test of assumptions shows that normality assumption is unfulfilled and there is one outlier, i.e. observation number 19. So we will estimate robust linear regression models using M estimation, S estimation, and MM estimation.

We apply Algorithm 1, Algorithm 2, and Algorithm 3 respectively to obtain robust regression models (10) - (12):

$$\hat{y} = 1483237 + 4.43x_1 + 37x_2 - 5645x_3 - 37101x_4 - 732x_5 + 0.05447x_6 \quad (10)$$

$$\hat{y} = 1885905 + 4.39x_1 + 8.2x_2 - 9306x_3 - 36997x_4 - 2364x_5 + 0.0549x_6 \quad (11)$$

$$\hat{y} = 1201266 + 4.36x_1 + 41x_2 - 4610x_3 - 30916x_4 - 342x_5 + 0.0728x_6 \quad (12)$$

There are some criteria that can be used to determine the best regression model, i.e. the coefficient of determination R^2 or $R^2_{adjusted}$ and standard deviation s. The best model will have largest R^2 or $R^2_{adjusted}$ and smallest s. We can see in Table 1 that the model (11) gives bigger $R^2_{adjusted}$ and smaller s than the model (10) or (12), so the best regression model is the model (11). S estimation can reduce outlier data, even can omit it.

	M estimation	S estimation	MM estimation
Model	(10)	(11)	(12)
$R^2_{adjusted}$	99.9%	100%	99.9%
s	36646.5	9824.5	28929.5
Significant variables	X_1, X_6	X_1, X_3, X_4, X_5, X_6	X_1, X_4, X_6
Outlier	no. 16	–	no. 12

Table 1: $R^2_{adjusted}$, s, significant variables, and outlier for models (10)-(12)

Because X_2 is not significant, we estimate regression model without X_2 using S estimation, and we obtain

$$\hat{y} = -1876912 + 4.39x_1 - 9205x_3 - 36855x_4 - 2388x_5 + 0.0553x_6 \quad (13)$$

$R^2_{adjusted} = 100\%$ shows that total variation of Y can be explained by X_1, X_3, X_4, X_5, X_6. The value of $F = 60499.55$ with $p-$value $= 0 < 5\%$ indicates that the linear regression model (13) fits with the data. Test of assumptions shows that all assumptions are fulfilled and there is no outlier so we can use equation (13) for modeling maize production in Indonesia.

5. Conclusion

We have discussed procedures to estimate robust regression model using M estimation, S estimation, and MM estimation. For maize production data in Indonesia we find that the best model for maize production in Indonesia is obtained by S estimation. There is no outlier in this model and all assumption are satisfied. The increment of one hectare of harvested area and one person of

agricultural man power in food crops subsector respectively will increase 4.39 and 0.0553 tons of the maize production. The increment of one percent monthly average humidity, one degree Celsius of monthly average temperature and one percent of monthly average long the sun shines will reduce $9,205; 36,855$ and $2,388$ tons of the maize production respectively.

Acknowledgements

The authors would like to thank the Directorate of Research and Community Service of Higher Education, Indonesian Ministry of Education and Culture and Sebelas Maret University which provide financial support through Research Grant No. 165/UN 27.11/ PN/2013.

References

[1] Badan Pusat Statistik, *Production of Paddy Maize and Soybeans*, www.bps.go.id/release/Production of Paddy Maize and Soybeans, 2012.

[2] D. Birkes and Y. Dodge, *Alternative Methods of Regression*, John Wiley Sons Inc., New York, 1993.

[3] C. Chen, Robust Regression and Outlier Detection with the ROBUSTREG Procedure, *Statistics and Data Analysis*, paper 265-27, SAS Institute Inc., Cary, NC.

[4] N. R. Draper and H. Smith, *Applied Regression Analysis*, Third Edition, Wiley Interscience Publication, United States, 1998.

[5] D. C. Montgomery and E. A. Peck, *An Introduction to Linear Regression Analysis*, John Wiley Sons Inc., New York, 2006.

[6] P. J. Rousseeuw and V. J. Yohai, Robust Regression by Mean of S-Estimators, *Robust and Nonlinear Time Series Analysis*, New York, 1984, 256-274, **doi:** 10.1007/978-1-4615-7821-5-15.

[7] M. Salibian, V.J. Yohai, A Fast Algoritm for S-Regression Estimates, *Journal of Computational and Graphical Statistics*, **15**, No. 2 (2006), 414-427, **doi:** 10.1198/106186006X113629.

[8] V. J. Yohai, High Breakdown Point and High Efficiency Robust Estimates for Regression, *The Annals of Statistics*, **15**, No. 20 (1987), 642-656, **doi:** 10.1214/aos/1176350366.

[9] Y. Susanti and H. Pratiwi, Robust Regression Model for Predicting the Soybean Production in Indonesia, *Canadian Journal on Scientific and Industrial Research*, **2**, No. 9 (2011), 318-328.

[10] Y. Susanti, H. Pratiwi, and T. Liana, *Application of M-estimation to Predict Paddy Production in Indonesia*, presented at IndoMS International Conference on Mathematics and Its Applications (IICMA), Yogyakarta, 2009.

[11] Yuliana and Y. Susanti, Estimasi M dan sifat-sifatnya pada Regresi Linear Robust, *Jurnal Math-Info*, **1**, No. 11 (2008), 8-16.