IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 3, Issue 8, August 2016

A Comparative Study on Robust Regression Methods

Guem Mi Lee[1], Kyupil Yeon[2], Hyeuk Kim[2]

Manager, Lucis Co., Seoul, Korea [1]

Assistant Professor, Department of Applied Statistics, Hoseo University, Asan, Korea [2]

Abstract: The research should analyse data after removing the outliers which have high influences or reducing the effects of influential points. The paper introduces the robust estimation methods to reduce the influences of outliers in regression modelling. We describe LTS estimator, LMS estimator, M-estimator, S-estimator, and MM-estimator among various robust estimation methods. Then, we make an experiment for real data and investigate the performances for several methods. The result shows that the robust estimation methods with reduction of influential points perform better than ordinary least squares method in regression analysis.

Keywords: Robust regression, Influential point, Least trimmed of squares, MM-estimator

I. INTRODUCTION

We want that the regression model is not excessively determined by a few observations when fitting data into a model. In other words, it is difficult to build a robust model if a few points influence a fitted model such as the estimated regression coefficients, the fitted value, and t-value. There are two methods to solve the phenomenon: removal of the points and reduction of the effects of the points [1]. The least trimmed of squares, the least median of squares, M-estimation, S-estimation, and MM-estimation are described in the paper and the performances of the two methods are compared with real data. We explain several methods in the second section and compare several robust regression methods with real data in the third section. In the last section, we draw a conclusion.

II. ROBUST REGRESSION

There is the linear regression model.

$$y_i = x'_i\beta + \varepsilon_i \quad (\mathrm{i} = 1, \dots, \mathrm{n})$$

Where x_i is the p-dimensional explanatory vector, y_i is the response variable and $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$isthe regression coefficient. ε_i's the error terms that are independent and follows a standard normal distribution respectively. The general method in regression analysis is to minimize the sum of the squared error terms, which is the least squares method, namely LS. In the approach, the estimates of the regression coefficients are as follows

$$\hat{\beta}_{LS} = argmin_\beta \sum_{i=1}^{n} r_i(\beta)^2 \quad , \qquad r_i(\beta) = y_i - x'_i\beta$$
$$= argmin_\beta (y_i - x'_i\beta)'(y_i - x'_i\beta)$$
$$= (\mathrm{X}'\mathrm{X})^{-1}X'y$$

The above estimates are known as BLUE(Best Linear Unbiased Estimator).

The squared loss function is not bounded. Therefore, it is possible that the regression line has a severe problem if there are a few outliers which have very large absolute residuals. Figure 1 is an example for the situation. The right picture in Figure 1 has one outlier, but it is a big influential point and biases a regression line. In other words, only one observation can break down a regression line and the regression line which is fitted by such observations has a tendency to make the residuals of normal points large.

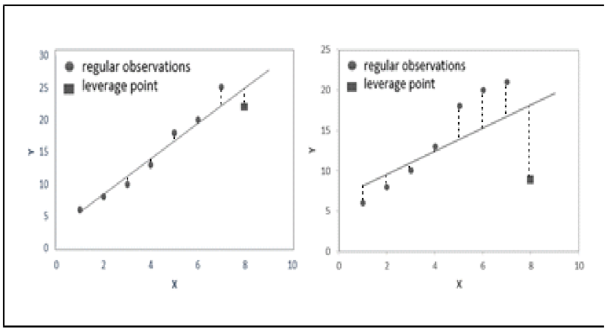

Fig. 1Regression lines for regular situation and the situation with an outlier

Several robust regression methods are developed to improve defect, which is described above. Above.

A. Least Trimmed of Squares(LTS)

The method of the least squares is to minimize the variance of the residuals. The demerit of the approach is that it can be affected by the observation with large residual. A robust estimate is developed to solve the demerit. The least trimmed of squares [2] uses a part of observations instead of all observations to calculate the variance of the residuals. The estimate of the least trimmed of squares minimizes the following objective function.

DOI 10.17148/IARJSET.2016.3830

$$\hat{\beta}_{LTS} = argmin_{\beta} \sum_{i \in H} r_i(\beta)^2$$

where $\mathrm{H} \in \{1, \dots, \mathrm{n}\}$ and $|\mathrm{H}| = \mathrm{h} < n$. h is the number of the observations which are considered in the least trimmed of squares. It is described in other form as follows.

$$\hat{\beta}_{LTS} = argmin_{\beta} \sum_{i=1}^{h} r_{(i)}^2$$

where $r_{(1)}^2 \le r_{(2)}^2 \le \cdots \le r_{(n)}^2$ are the order statistics of the squares of residuals.

The breakdown point of $\hat{\beta}_{LTS}$ is $\frac{[n/2]-p+1}{n}$ when $\mathrm{h} = [n/2] + 1$.The estimate of least trimmed of squares is difficult to be used as the independent estimate since it has a high breakdown point. That is, it is very robust for an outlier but a low relative efficiency. However, it is useful for an initial value for GM-estimate [3] or diagnostic plot for an outlier.

B. Least Median of Squares(LMS)

The least median of squares is to minimize the median of observations instead of the mean. The median is the more robust measure than the mean in statistics which describe the central tendency of observations. The estimate of the least median of squares minimizes the following objective function.

$$\hat{\beta}_{LTS} = argmin_{\beta} med\{r_i(\beta)^2 : i = 1, \dots, n\}$$

The breakdown point of the estimate of the least median squares is 50 percent like the estimate's of the least trimmed of squares. However, a relative gradual efficiency is at most 37 percent and the speed of the convergence is $n^{-1/3}$ in the estimate of the least trimmed of squares. It plays an important role in the calculation of an MM-estimator since it provides the initial estimate of an residual.

C. M-estimator

Huber(1973) [4] proposed M-estimator which minimizes the objective function.

$$\hat{\beta}_M = argmin_{\beta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\beta)}{\hat{\sigma}}\right)$$

We derive the p equations if we differentiate the above objective function about β_j and set the corresponding result to 0.

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\beta)}{\hat{\sigma}}\right) x_{ij} = 0 \;\; (j = 1, \dots, p)$$

Moreover, we get the following equation if the weight w_i is defined as $w_i = \psi(r_i(\beta)/\hat{\sigma})/(r_i(\beta)/\hat{\sigma})$.

$$\sum_{i=1}^{n} w_i(y_i - x_i^{'}\beta)x_i = 0$$

The iteratively reweighted least squares (IRWLS) is applied to find the solution of the above equation and it's algorithm is described below.

Iteratively Reweighted Least Squares (IRWLS):

Step 1: Calculate the initial estimate $\hat{\beta}^{(0)}$through the least squares.
Step 2: Calculate the initial residual $r_i^{(0)}$ by using $\hat{\beta}^{(0)}$
Step 3: Calculate the weight $w_i^{(0)}$ by using $r_i^{(0)}$.
Step 4: Calculate the estimate of the regression coefficient, $\hat{\beta}^{(1)}$, by weight $w_i^{(0)}$through the weighted least squares. In matrix notation, $\hat{\beta}^{(1)}$ is $\hat{\beta}^{(1)} = (X^T W X)^{-1} X^T W$ when $W = \mathrm{diag}(w_i)$the diagonal matrix whose elements are the corresponding weights is.
Step 5: Calculate $\hat{\beta}^{(2)}$ by using $\hat{\beta}^{(1)}$ from Step 1 to Step 4.
Step 6: Repeat the above procedures until $\hat{\beta}$ is stabilized.
Gschwandtner and Filzmoser [5] describes ρ function, ψ, and the weight w which are used generally.The breakdown point of M-estimator is known as $1/\rho$ and it has a very small breakdown point when there are many explanatory variables.

D. S-estimator

S-estimator has been developed by Rousseeuw and Yohai in 1984 [6]. It is the value which minimizes the robust scale parameters about the residuals.

$$\hat{\beta}_S = argmin_{\beta} \hat{\sigma}(r_1(\beta), \dots, r_n(\beta))$$

$\hat{\sigma}$ is the solution of the following equation.

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\beta)}{\hat{\sigma}}\right) = \delta \;,\delta = E_{\Phi}[\rho(r)]$$

where Φ is a standard normal distribution.
S-estimator $\hat{\beta}_S$also satisfies the following equation.

$$\hat{\beta}_S = argmin_{\beta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\beta)}{\hat{\sigma}}\right) , \hat{\sigma} = \hat{\sigma}(r_1(\beta_S), \dots, r_n(\beta_S))$$

We differentiate the equation when ψ = ρ′ and set a differential equation to 0. The following equation for estimation is constructed and is solved by the iteratively reweighted least squares.

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\beta}_S)}{\hat{\sigma}}\right) x_i = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\hat{\beta}_S)}{\hat{\sigma}}\right) = \delta$$

DOI 10.17148/IARJSET.2016.3830

S-estimator is built to have a high breakdown point. It's breakdown point ϵ_n^* is given below when a function ρ satisfies from (R1) to (R3) [6].

(R1) ρ is a symmetric function and differential continuously. Also, $\rho(0) = 0$.

(R2) ρ is an increasing function with the interval $[0, c]$. There exists c such that ρ becomes a constant function in the interval $[c, \infty]$.
(R3) $\frac{E_{\Phi}(\rho)}{\rho(c)} = \frac{1}{2}$

The breakdown point ϵ_n^* of S-estimator is $\epsilon_n^* = \left(\left[\frac{n}{2}\right] - p + 2\right)/2$.

We figure out that the gradual breakdown point is 50 percent when n goes to infinity. S-estimator has a low gradual relative efficiency if the constant c is determined for a high breakdown point, and a high gradual relative efficiency if the constant c is determined for a low breakdown point. Therefore, S-estimator is usually used for the initial estimate of MM-estimator.

E. MM-estimator

Yohai(1987) [7] proposed MM-estimator which was robust and efficient. It is made of combining M-estimator and S-estimator. MM means the procedure for finding M-estimator is used at least two times while searching for the final estimate. Suppose that $\hat{\beta}^{(0)}$isS-estimator and $\hat{\sigma}$is M-estimator about the corresponding scale parameter. MM-estimator is defined as below.

$$\hat{\beta}_{MM} = argmin_{\beta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\beta)}{\hat{\sigma}}\right)$$

The solution of theabove equation is calculated with the initial value $\hat{\beta}^{(0)}$ through the iteratively reweighted least squares. We describe the specific procedure for the solution as follows.

Step 1: Compute an initial regression estimate $\hat{\beta}^{(0)}$ through S-estimator using Huber or Tukey's biweight function.
Step 2: Compute M-estimator σ about scale parameter by $\hat{\beta}^{(0)}$.
Step 3: Compute M-estimator using $\hat{\beta}^{(0)}$ and $\hat{\sigma}$ through method of weighted least squares.

$$\sum_{i=1}^{n} \omega_i \left(\frac{r_i^{(0)}}{\hat{\sigma}}\right) x_i = 0$$

where ω_i is Huber or Tukey's biweight.
Step 4: the weight is recalculated by the residual from Step 3.
Step 5: We fix the scale parameter from Step 2 and repeat Step 3 and Step 4 until the value converges.

III.EXPERIMENTS

The real data is used for comparison of several regression methods. De Long and Summers [8] have studied the growth of 61 countries in 1991. The dataset consists of one response variable which is GDP(growth domestic product) and the four explanatory variables. They are LFG(labor force growth), GAP(GDP per worker gap), EQP(equipment share), and NEQ(non-equipment share).

A. Ordinary Regression by BLUE

Table 1 describes the result of the multiple regression.

Var.	U.R.C.		S.R.C.	t	P	VIF
	B	S.E.				
C.	-0.01	0.01	-	-1.39	0.17	-
LFG	-0.03	0.20	-0.02	-0.15	0.88	1.33
GAP	0.02	0.01	0.29	2.21	0.03	1.41
EQP	0.27	0.07	0.51	4.06	<0.01	1.32
NEQ	0.06	0.04	0.23	1.79	0.08	1.39

Table 1 Result of the multiple regression analysis through the least squares

We explain several abbreviations in the first column in Table 1. Var. is the explanatory variable in regression model. U.R.C. is an Unstandardized regression coefficient and S.R.C. is a standardized regression coefficient. t and P means t statistic and p-value respectively. VIF describes a variance inflation factor, which is available to check the degree of the multicollinearity. C. in the second column is an abbreviation of the constant in the regression model.
The model is valid since F statistic for model is 7.17 and the corresponding p-value is less than 0.001. The explanatory variables describe the model about 33.9% since R^2 is 0.339. The effective variables are GAP and EQP based on 5 percent of the significance level. Growth data also has no collinearity since all VIF's for explanatory variables are less than 10.

B. Robust Regression by removing a few influential points

We use Cook's D plots to figure out the influential point.

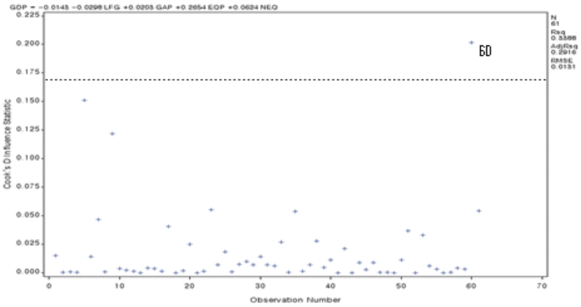

Table 2 The index plots for Cook's D

Cook's D statistic is the measure which standardizes the change of the vector of the least squared estimation when one observation is removed. We consider that the observation is influential if it's Cook's D statistic is high [1]. The standard value for deciding an influential point is

DOI 10.17148/IARJSET.2016.3830

usually $\frac{2(p+1)}{n}$ where p is the number of explanatory variables and n is the number of observations. Therefore, the standard value in this experiments is $\frac{2(4+1)}{61} = 0.16$ and the observation whose index is 60 is an influential point from Table 2.

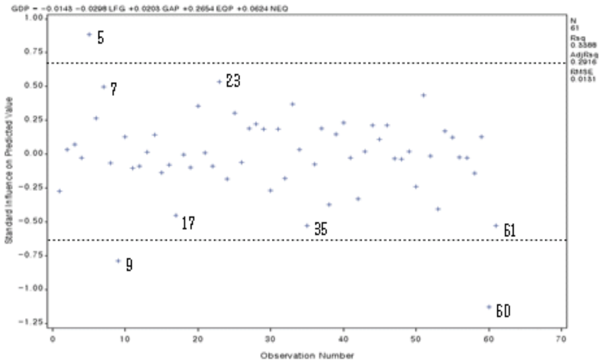

Table 3 The index plots for DFFITS

Table 3 describes one of the statistical methods which diagnoses the influences of observations. DFFITS was developed by Belsley, Kuh and Welsh [9]. DFFITS statistic is the measure which standardizes the change of the fitted value $\hat{y}_i$. An observation is considered that it influences the expected estimate of y_i if it's absolute value of DFFITS is very large. The general standard value is $2\sqrt{\frac{p+1}{n-p-1}}$ where p is the number of explanatory variables and n is the numberof observations. Therefore, the standard value for DFFITS is $2\sqrt{\frac{4+1}{61-4-1}} = 0.60$ and theinfluential points are the observations whose index are 5, 7, 9, 17, 23, 35, 60, and 61.

We determine that the observations whose index are 1, 5, 7, 8, 9, 17, 23, 27, 35, 35, 57, 59, 60, and 61 are influential from two plots. We construct a multiple regression model after removing the influential points. The model is also valid since F statistic is 6.98 and the corresponding significance probability is less than 0.001.

Var.	U.R.C.		S.R.C.	t	P
	B	S.E.			
C.	-0.01	0.01	-	-1.08	0.29
LFG	-0.20	0.21	-0.17	-0.96	0.34
GAP	0.02	0.01	0.35	1.98	0.05
EQP	0.08	0.09	0.16	0.95	0.35
NEQ	0.15	0.03	0.63	4.29	<0.01

Table 4 Result of the multiple regression analysis by removing influential points

$R^2$for the regression model is 0.405. NEQ are statistically significant based on 5 percent of the significance level.

In comparing two multiple regression models, R^2 is increased about 0.066 and the explanatory variables which are statistically significant are changed. GAP and EQP are statistically significant in an ordinary regression model, but are not any more after removing influential points. NEQ is not statistically significant in an ordinary regression model, but is a significant variable after removing influential points.

C. Robust Regression by Least Trimmed of Squares (LTS)

We introduce five robust regression methods which reduce the effects of the influential points. We conduct experiments for LTS and MM-estimator among five methods.

Var.	U.R.C.		χ^2	P
	B	S.E.		
C.	-0.02	0.01	5.65	0.02
LFG	0.05	0.18	0.06	0.80
GAP	0.03	0.01	8.89	0.00
EQP	0.28	0.06	23.60	<0.01
NEQ	0.09	0.03	7.30	0.01

Table 5Result of the multiple regression analysis by Least Trimmed of Squares

Table 5 describes the multiple regression model by Least Trimmed of Squares.$R^2$for the model is 0.741 and the value is much higher than 0.339 and 0.405 for the above models respectively. It is not because the model fits very well but because the formula for r-squared is different. GAP, EQP, and NEQ are statistically significant.

D. Robust Regression by MM-estimator

Table 6 describes the multiple regression model by MM-estimator.The explanatory variables describe the model about 31.1% since R^2 is 0.311. Three variables such as GAP, EQP, NEQ are statistically significant like we apply the robust regression model by Least Trimmed of Squares. There are two differences when we remove influential points and reduce the effects of influential points.

Var.	U.R.C.		χ^2	P
	B	S.E.		
C.	-0.03	0.01	6.70	0.01
LFG	0.12	0.19	0.39	0.53
GAP	0.03	0.01	8.80	0.00
EQP	0.30	0.06	22.37	<0.01
NEQ	0.09	0.03	6.97	0.01

Table 6 Result of the multiple regression analysis by MM-estimator

First, the explanatory variables which are statistically significant are different as we already mention in front. NEQ is the unique explanatory variable which is statistically significant when we remove influential points. On the other hand, there are three statistically significant variables for reducing the effects of influential points. Secondly, the values of the regression coefficients are changed. For example, the regression coefficient for LFG is negative for removing influential points and it becomes positive for reducing the effects of influential points.

IV. CONCLUSION

We introduce the ordinary regression and the robust regression which is useful when there is an outlier in the paper. There are two approaches to handle the influential

DOI 10.17148/IARJSET.2016.3830

points. The first method is to remove influential points and the second method is to reduce the effects of influential points in constructing the regression model. We compare the results of four approaches for the real data. We figure out that the different approaches yield the differences in valid variables and the coefficients of regression through the experiment. From the experiment, we can make a conclusion that the method which reduces the effects of influential points is more proper since it finds the more valid explanatory variables. It is very hard to find influential points or outliers through the least squared method in regression analysis. Therefore, we recommend the robust regression model with reduction of influential points when we handle data without cleansing and suspect that there are a few outliers.

ACKNOWLEDGMENT

The paper summarizes **Guem Mi Lee's** dissertation for master's degree. She graduated Hoseo University in February 2016.

REFERENCES

[1] S. Chatterjee, Regression Analysis by Example, 5th ed., Wiley, 2012.

[2] P. J. Rousseeuw, “Least median of squares regression,” Journal of the American Statistical Association, vol. 79, pp. 871–880, 1984.

[3] Coakley and Hettmansperger, “A bounded influence, high breakdown, efficient regression estimator,” Journal of the American Statistical Association, vol. 88, pp. 872–880, 1993.

[4] Huber, Robust Statistics, 1st ed., New York: John Wiley & Sons, 1973.

[5] Gschwandtner and Filzmoser. (2012) “mvoutlier” on CRAN. [Online]. Available: https://cran.r-project.org/web/packages/mvoutlier/mvoutlier.pdf

[6] P. J. Rousseeuw and V. Yohai, Robust Regression by Means of S estimators, in Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics 26. New York: Springer Verlag, pp. 256–274, 1984.

[7] V. J. Yohai, “High breakdown point and high efficiency robust estimates for regression,” Annals of Statistics, vol. 15, pp. 642–656, 1987.

[8] J. B. De Long and L. H. Summers, “Equipment investment and economic growth,” The Quarterly Journal of Economics, vol. 106(2), pp. 445–502, 1991.

[9] D. A. Belsley, E. Kuh, and R. E. Welsh, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, 1st ed., New York: John Wiley & Sons, 1980.

DOI 10.17148/IARJSET.2016.3830