A SIMULATION STUDY ON EFFECT OF OUTLIERS IN REGRESSION MODEL WITH DUMMY VARIABLE

Maw Maw Khin[1]

Abstract

This paper aimed to study the effect of outliers in the regression model with a dummy variable based on simulated data. The two robust methods namely Robust Distance Least Absolute Value (RDL_1) and Least Trimmed Squares (LTS) and classical method like Ordinary Least Squares (OLS) estimation method were applied to these data. Simulation study showed that the RDL_1 and LTS methods detected several outliers whereas the OLS residuals did not reveal any outliers. Based on the mean squared error (MSE) criterion, the RDL_1 estimator is more resistant, but it suffers from the swamping effect. The LTS estimator has the second smallest MSE and the OLS method has the largest MSE in this case. The OLS regression is the best when data are free from outliers.

Keywords: outliers, OLS regression, RDL_1, LTS, MSE

Introduction

Regression analysis is an important tool for any quantitative research. For the regression analysis, the Ordinary Least Squares (OLS) method can produce bad estimates when the error distribution is not normal, particularly when the errors are heavy-tailed. It explores the relationship between dependent and explanatory variables. Many hypotheses claimed by economic theories can be tested by applying a regression model on real data. The OLS method is mostly applied in the regression technique. The application of this specific method requires several assumptions. A researcher should be aware of the fact that the OLS method performs poorly if these assumptions are not fulfilled.

In the last two centuries, various strategies were introduced to test whether the model assumptions are fulfilled or not. Besides, various more general regression techniques are available based on less stringent conditions. Until the mid-20th century, violations of the model assumptions were treated independently of any common error source. But in particular, outlying observations within the data can cause violations of model assumptions, and thereby it can have a huge impact on regression results.

Even if one outlying observation can destroy OLS estimation, resulting in parameter estimates that do not provide useful information for the majority of the data. Outliers will make the error variance inflate, the confidence interval becomes stretched, and the estimation cannot become asymptotically consistent. When outliers inflate the error variance, they damage the model of power to detect the outliers. Rousseeuw and Van Zomeren (1990) proposed a vertical outlier, good leverage point, and bad leverage point.

Rousseeuw and Van Zomeren (1990) pointed out that high leverages can affect the estimated slope of the regression line in OLS, thus they may cause more serious problems than the vertical outlier. Moreover, their occurrence in regression models may move to some low leverage as well as high leverage and it can turn in vice versa. These two concepts are called masking and swamping in linear regression. Furthermore, the range of explanatory variables increases when they exist in regression analysis. Thus, the multiple coefficient of determination (R^2) which is a well-known and popular measure of goodness-of-fit in the regression models will increase even by any changes of a single *X* variable. Besides, high leverages may be the prime source of collinearity-influential observations whose presence can make collinearity and can destroy the existing

[1] Dr, Professor and Head of Department of Statistics, Yangon University of Economics

collinearity pattern among the *X* variables. In this respect, to eliminate these outliers' effects on linear regression the role of robust method becomes necessary.

Robust regression methods have been developed as an improvement to OLS estimation in the presence of outliers and provide information about what a valid observation is and whether this should be thrown out. The primary purpose of robust regression analysis is to fit a model that represents the information in the majority of the data. In this context, robust regression is to employ a fitting criterion that is not as vulnerable as OLS to unusual data. One remedy is to remove influential observations before using the OLS fit.

The robust methods provide an alternative to an OLS regression model when fundamental assumptions are unfulfilled by the nature of the data. When the estimates of the parameters of statistical regression models and test assumptions, it is frequently found that assumptions are substantially violated. Sometimes, the variables can be transformed to confirm those assumptions. Often, however, a transformation will not eliminate or satisfy 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 remedy. The objectives of the study are (i) robust regression methods are better than the OLS estimation methods when data contain outlying observations and (ii) if data were free from outliers, the OLS estimation method outperforms the robust regression methods.

2. Data and Method

To show the fact that the robust procedure outperforms the classical method in the dummy variable regression model, simulation with data set contaminated by different types of outliers was carried out. Observations used in this analysis were to be classified into four categories namely regular data, good leverage points, vertical outliers, and bad leverage points. The model included two types of regressors namely continuous and discrete regressors was expressed in the following form.

$$y_i = 9 + x_{i1} + x_{i2} + I_{i1} + \varepsilon_i , \qquad i = 1,\ldots,30, \tag{1}$$

Where both x_{i1} and x_{i2} follow a standard normal distribution, I_{i1} is a binomial distribution with a success rate of 0.5, and ε_i is a normal distribution with mean zero and standard deviation 0.5. The regressand variable was generated by the model stated in Equation (1). Once these 30 observations have been generated, cases 25 and 26 were then transformed to be vertical outliers by doubling their *y* values and keeping the others. Cases 27 and 28 were bad leverage points by adding 9 to their x_1 values and keeping the others as well. Case 29 and 30 were good leverage points by adding 9 to both x_1 and x_2 values and reproducing the corresponding *y* values as the model (1). The resulting simulated data are presented in Appendix Table (1).

Result and Discussion

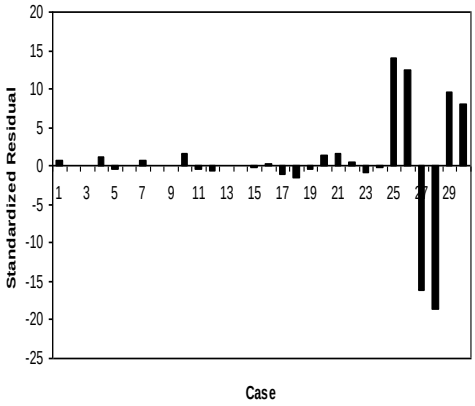
The Robust Distance Least Absolute Value (RDL_1), Least Trimmed Squares (LTS), and OLS estimation methods were applied to these simulated data and the results were summarized in Table (1). First, the RDL_1 estimation method was applied to these simulated data. For checking outliers, the standardized residuals were shown in Figure 1(a). The case 25, 26, 27, 28, 29, and 30 were revealed as outliers. This is because the weight was calculated by the continuous design matrix without considering the model fitting. The resulting weights were shown in Figure 1(b). Therefore, case 27, 28, 29, and 30 were outlying observations occurred from ***X*** space and will be given relatively small weights as shown in part (b) of Figures (1). These make case 29 and 30 become bad leverage points in the diagnostic plot of Figure 1(c). The cutoff values were indicated

± 2.5 and $\sqrt{\chi^2_{2,0.975}}$ by horizontal and vertical lines. These results pointed out that the RDL_1 method results in the swamping effect due to its weights for the L_1 procedure obtained from the application of a continuous design matrix.

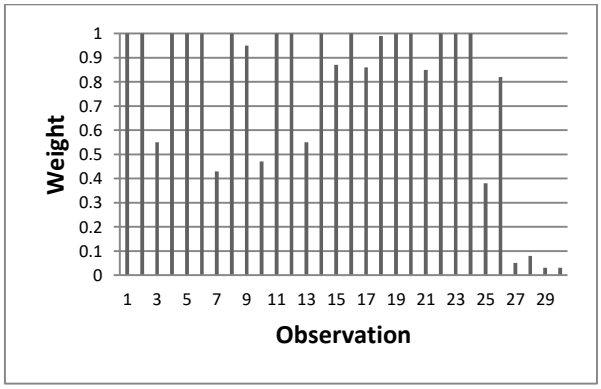
Thus, the same data set was used to apply the LTS estimation method and the results were displayed in Table (1). Parts (a), and (b) of Figure (2) show the results of robust standardized residuals and the diagnostic plot obtained from LTS regression analysis, respectively. From part (a), the LTS procedure detected cases 25, 26, 27, and 28 to be outliers. It gave a weight 1 for both cases 29 and 30. The corresponding diagnostic plot also divided all points into the right categories as the original configuration of these data being generated. The fit from the LTS method ignored outlying observations, which gave the MSE of 0.3042 shown in Table (1).

Then, the OLS estimation method was applied using the simulated data again. The large MSE of OLS for the simulated data set argued that data were highly influenced by outliers and Figure 3(a) shows the vertical outliers or bad leverage points. The OLS regression estimators often break down in the presence of those outliers. It was evident from the graphical sketch of data as the OLS line was pulled towards the middle of the two groups of the data points rendering it was an unrepresentative line. A Gaussian Q-Q plot was shown in Figure 3(b) confirms that the residuals were roughly normally distributed. Only a few outliers can cause the distribution to be heavier-tailed.

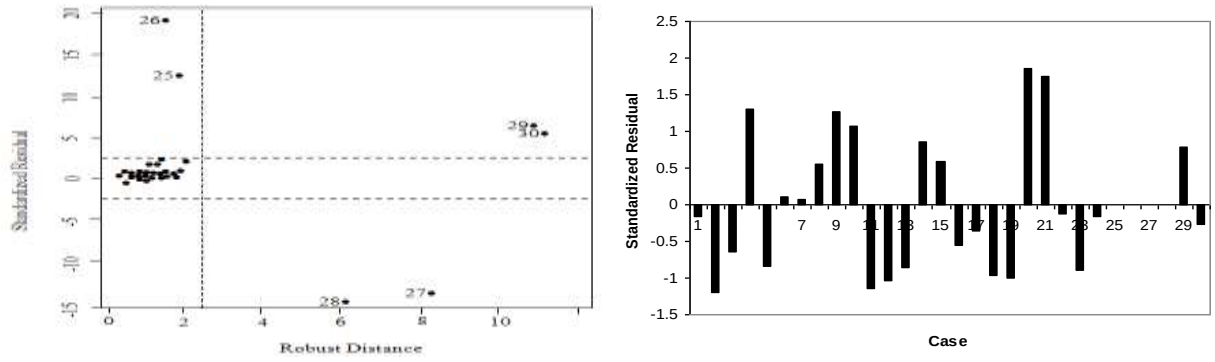
According to the result of LTS analysis, the observations (25, 26, 27, and 28) gained from the simulated data were excluded and the remaining data were rerun using the OLS estimation method. Table (2) presents the regression results for the two data sets (contaminated and non-contaminated). New OLS regression represents the results after eliminating the outlier found through the LTS method. The intercept and slope coefficients changed and all were statistically significant at a 1% level. The fact that the F and R^2 values increased indicates that the new OLS regression is well-matched with those remaining data. The OLS line fits the simulated (non-contaminated) data well with a reasonable MSE of 0.1230. Figure 4(a) shows the OLS residuals without considering the cases 25, 26, 27, and 28. The cases 29 and 30 were located near the regression surface. Figure 4(b) suggests that the residuals were approximately normally distributed. Based on the results from non-contaminated data, it can be concluded that the OLS regression (New) method outperforms than two robust methods.



(a)



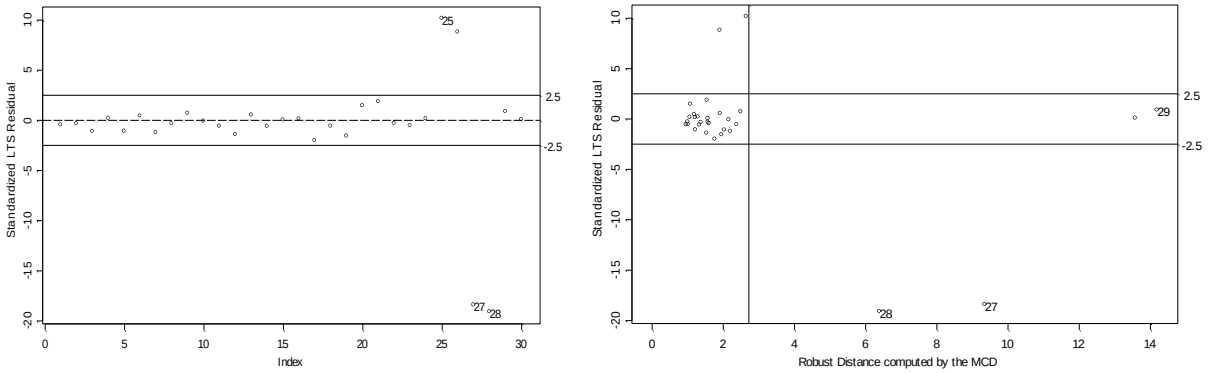
(b)



(c) (d)

Source: Appendix Table (1)

Figure 1 Simulated Data Set Using the RDL_1 Procedure: (a) plot of the standardized residuals; (b) plot of weights; (c) diagnostic plot and (d) least squares residuals without cases 25, 26, 27, and 28



(a) (b)

Source: Appendix Table (1)

Figure 2 Simulated Data Set Using the LTS Robust Procedure: (a) plot of the standardized residuals; and (b) diagnostic plot

Table 1 OLS, LTS and RDL_1 Regression Models Fitted to the Simulated Data

Method	Coefficients				MSE
	Constant	x_1	x_2	I_{i1}	
OLS	8.84^{***}	0.43^{**}	1.34^{***}	2.47^{**}	6.1504
LTS	9.22^{***}	1.16^{**}	0.72^{***}	1.37^{***}	0.3042
RDL_1	8.94^{***}	0.82^{***}	0.77^{***}	1.40^{***}	0.1849

Note: (1) Significant at *** 1%, **5%, * 10%

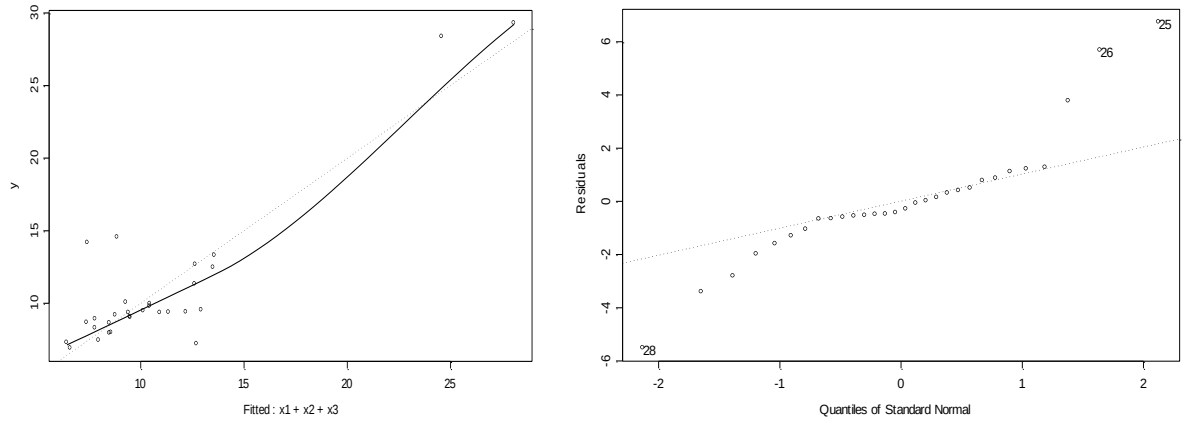
Source: Appendix Table (1)

Table 2 Two Ordinary Least Squares Regression Results

OLS Regression Based on Contaminated Data				New OLS Regression Based on Non-contaminated Data		
Variable	**Coefficient**	**Standard Errors of Coefficients**	***t* Statistics**	**Coefficient**	**Standard Errors of Coefficients**	***t* Statistics**
Constant	8.8410 ***	0.5926	14.9203	9.1344^{***}	0.0872	104.7172
x_1	0.4276 **	0.1737	2.4611	1.0139^{***}	0.0412	24.6214
x_2	1.3386 ***	0.2355	5.6847	0.9480^{***}	0.0504	18.8205
I_{i1}	2.4729 **	0.9456	2.6153	0.9954^{***}	0.1484	6.7095
MSE = 6.1504, R^2 = 0.7995, F = 34.55^{***}, n=30				MSE = 0.1230, R^2 = 0.996, F = 2046.865^{***}, n=26		

Note: (1) Absolute value of *t* statistics in parentheses
(2) Significant at *** 1%, **5%, * 10%

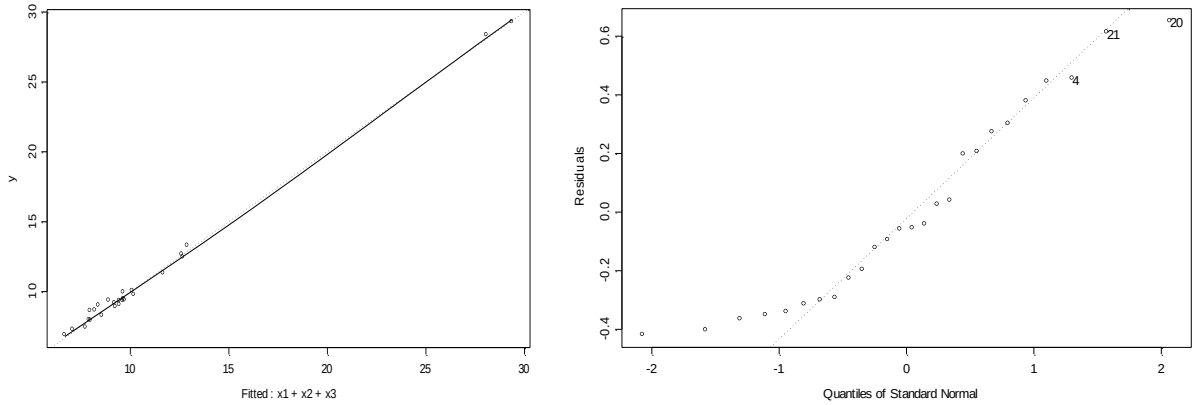
Source: Appendix Table (1)



(a) (b)

Source: Appendix Table (1)

Figure 3 Simulated Contaminated Data Set Using the OLS: (a) scatter plot with OLS line; and (b) quantiles standard normal plot



(a) (b)

Source: Appendix Table(1)

Figure 4 Simulated Non-contaminated Data Set Using the OLS: (a) scatter plot with OLS line; and (b) quantiles standard normal plot

Conclusion

This study illustrated that the RDL_1 and LTS methods outperform the OLS method in the regression model involved with a dummy variable. It was found that based on the contaminated simulation data, the RDL_1 and LTS methods detect several outliers whereas the OLS residuals do not reveal any outliers. Based on the mean squared error (MSE) criterion, the RDL_1 estimator is more resistant but it cannot detect correctly for the cases 29 and 30. It suffers from a swamping effect. The OLS method is much worse in this case.

According to the results of LTS analysis, it is correctly detected the observations 25 and 26 are vertical outliers, 27 and 28 are bad leverage points. It has the second smallest MSE of 0.3042. The cases 25, 26, 27, and 28 are omitted and the remaining data are rerun using the OLS method. In this case, the two robust namely RDL_1 and LTS and OLS methods worked well, indicating that the values of MSE are quite close to each other. Based on this study, it can be concluded that when there are outliers in the data, the robust methods perform better than the OLS method. It is found there is no outlier in the data OLS estimation method is more robust than RDL_1 and LTS methods.

Acknowledgement

I would like to convey my deepest thanks to Rector and Pro-rector of Yangon University of Economics who encouraged me to submit this research paper for the conference organized by MAAS.

References

Hadi, A. S. (1992), Identifying Multiple Outliers in Multivariate Data, *Journal of the Royal Statistical society, Series B*, vol. 54, 761-771.

Huber, P. J. (1973), Robust Regression: Asymptotics, Conjectures and Monte Carlo, *The Annals of Statistics*, vol. 1, 799-821.

Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley and Sons.

Hubert, M., and P. J. Rousseeuw (1996), Robust Regression with a Categorical Covariable, *Robust statistics, Data analysis, and Computer Intensive Methods*, New York: Springer.

Hubert, M., and P.J. Rousseeuw (1997), Robust Regression with Both Continuous and Binary Regressors, *Journal of Statistical Planning and Inference*, vol. 57, 153-163.

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

Rousseeuw, P. J., and B. C. Van Zomeren (1990), Unmasking Multivariate Outliers and Leverage Points, *Journal of American Statistical Association*, vol. 85, 633-639.

Ryan, T. P. (1997), *Modern Regression Methods*, New York: John Wiley and Sons.

Table 1 Simulated Data Set Appendix

Case	x_1	x_2	I_1	y	Case	x_1	x_2	I_1	y
1	0.92	0.04	0	10.05	16	-0.56	1.15	0	9.46
2	-0.11	1.23	0	9.77	17	0.94	-1.72	1	9.33
3	1.39	1.2	1	12.45	18	-1.46	1.13	1	9.38
4	0.13	-1.12	0	8.66	19	1.2	-1.16	0	8.9
5	0.25	-0.86	0	8.28	20	-1.29	0.15	0	8.62
6	-1.44	0.26	0	7.96	21	-1.7	1.04	0	9.01
7	2.18	0.3	1	12.65	22	0.17	-0.1	0	9.17
8	-0.48	-1.65	0	7.28	23	-1.09	-0.3	0	7.43
9	-1.75	0.59	1	9.36	24	-1.27	0.15	0	7.93
10	1.64	1.17	1	13.28	25	0.06	-2.92	1	14.16
11	-0.2	0.55	0	9.05	26	0.45	-1.97	1	14.54
12	0.87	0.7	1	11.31	27	8.89	0.22	0	9.52
13	-1.39	2.01	0	9.33	28	6.87	-1.15	1	7.19
14	0.18	-0.7	1	9.95	29	10.96	8.26	0	28.35
15	-1.21	-1.29	0	6.89	30	10.37	9.21	1	29.28

Source: Simulated Data obtained from Model (1)