IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 11, Issue 6 Ver. V (Nov. – Dec. 2015), PP 60-67
www.iosrjournals.org

Identification of Outliers in Time Series Data via Simulation Study

[1]ADEWALE Asiata. O, [2]Nurul Sima MOHAMAD SHARIFF
[1,2,]Faculty of Science & Technology University Sains Islam Malaysia (USIM) Bandar Baru Nilai, 71800 Nilai Negeri Sembilan Malaysia

Abstract: *This paper compares the performance of regression diagnostics techniques based on Ordinary Least Squares (OLS) estimators and four types of robust regression based on robust estimators to detect and identify outliers. It is known that OLS is not robust in the presence of multiple high leverage points. Thus several robust regression models are used as alternative and its approach is more reliable and appropriate method for solving this problem. The comparisons are made via simulation studies. Our results have shown that in some cases diagnostics based on the OLS and some robust estimators give similar outcomes, they detect the same percentage of correct outlier detection. And under small sample size OLS and M-estimation perform best for innovative outliers. The results also shows that Least Trimmed Square is the best among all its counterparts under large sample size.*

Keywords: *Outliers, Ordinary Least Squares (OLS), Regression diagnostics, Robust regression, Simulation Studies.*

I. Introduction

Outliers are usually encountered in time series data analysis. The presence of outliers in time series analysis can seriously has negative impact in the analysis because they may stimulate substantial biasness in parameter estimation, model misspecification and incorrect inference, [1]. Outliers has been defined by Abd. Mutalib and Ibrahim [2] as data points or observations that deviate distinctly from other observations or data points which are abnormally large or small from the other observations. The relevancies of outlier detection and identification in time series have been used in fraud detection, financial institute, public health and Telecommunication Company. According to Lopez-de-Lacalle[3], there are five types of outliers that are commonly found, namely, innovation outlier (IO), additive outlier (AO), level shift (LS), temporary change (TC)and seasonal level shift (SLS).

The most popular way to analyse time series regression model is to use Ordinary Least Square (OLS) method. It is the best technique if all the statistical assumptions are valid but when the data or the series are contaminated with outliers, these statistical assumptions are invalid. There arise the needs of regression diagnostics tools or techniques to detect and identify the outliers or influential observations. There are many types of regression diagnostics tools in the literature, among them are: The welsch-kuh distances, Cook's Distance and Hadis influence Measure**.** However, these methods are not robust in presence of multiple high leverage point, which can cause masking and swamping effects [4]. According to Widodo et al [5], robust regression approach is more reliable and appropriate method for solving this problem. The robust estimators are relatively unaltered by large changes in a small series of data and also small changes in a large part of the series. Yafee [6] discusses that there are several kinds of robust estimators in the literature among which are Least Absolute Deviations (LAD or L1), Least Median Squares (LMS), Least Trimmed Squares (LTS) and Huber M-Estimation. These robust estimators will be used in this research work by S plus statistical software through simulation study. Their performance will be compared to one another and the best technique among them will be determined.

II. Materials and Method

OLS Estimation

Consider the time series model of simple autoregressive, AR(p)

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t \quad (1)$$

The time series model of simple autoregressive, AR(p) as the sameform as (1), can be written as

$$y_i = \beta_0 + \beta_1 x_{i2} + ... + \beta_p + \varepsilon, \qquad i = 1,2,...,n \quad (2)$$

Model (2) can be written in matrix form as,

$$Y = X\beta + \varepsilon \quad (3)$$

Where Y = vector of nx1 response, X is the n x p matrix of explanatory variables, β is the vector of parameter (regression coefficient), ε is the random error distributed as normal distribution with mean zero and σ^2 . Here n is number of observations and p is number of regressors.
The final estimates of β is then,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{4}$$

The residual is obtained as follows:

$$e = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X^T X)^{-1} X^T Y = (1 - H)y \tag{5}$$

Where $H = X(X^T X)^{-1} X^T$ is leverage / hat matrix.
The i-th elements of H is,

$$h_{ii} = x_i (X^T X)^{-1} x_i^T \tag{6}$$

Regression Diagnostics Methods

There is numerous regression diagnostics methods used to identify outliers. However, this study will only consider five methods that are listed below.

Hadi's influence measure

$$H_i^2 = \frac{h_{ii}}{1 - h_{ii}} + \frac{p}{1 - h_{ii}} \frac{d_i^2}{1 - d_i^2}, \quad i = 1,2,3,..., n \tag{7}$$

Where $d_i^2 = \frac{e_i}{\sqrt{SSE}}$ the normalized, *SSE* is sum squares residual residuals, It identify influential observations.

The cut-off point for H_i^2 measure is $\left\lfloor mean(H_i^2) + C\sqrt{\text{var}(H_i^2)} \right\rfloor$ C is constant which take the value of 2.

The welsch-kuh distances

$$DFFits = \frac{e_i * \sqrt{h_{ii}}}{\hat{\sigma}_{(-i)}(1 - h_{ii})}, \; i = 1,2,3,..., n \tag{8}$$

has the cuff-off value of $2 * \sqrt{\frac{p}{n}}$

Cook's Distance

$$D_i = \frac{r_i^2 * h_{ii}}{(p+1)(1 - h_{ii})}, \; i = 1,2,3,..., n \tag{9}$$

Where $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$, Any observation that exceeds the cut-off of $\frac{4}{n - p}$ is considered as influential observation.

Robust Regression

Robust regression methods are designed not to be wholly affected by violations of assumptions by the core data generating process. A robust regression is performed on a high breakdown point and high efficiency regression, [9].

Huber M-Estimation

Huber-M estimation uses Huber weight function as its weight function. The Huber M-Estimator scale estimate of $\hat{\sigma}_m$ and the Huber M-Estimator error e_m are used instead of $\hat{\sigma}$, and e_i which are based on OLS method.

Least absolute deviations (LAD or L1)
LAD obtains a higher effectiveness than OLS through minimizing the sum of the absolute errors, [7]. It scale estimate denoted by $\hat{\sigma}_{L1}$ and its error e_{L1} are used instead of $\hat{\sigma}$, and e_i which are based on OLS.

Least median squares (LMS)
LMS is a robust estimator that has been hypothetically has breakdown point of 50%, [8]. This means the LMS provides reliable outcomes even if up to 50% of contaminated data or series exist. It has the characteristics of solving the liner model by minimising the median of the weighted squared. LMS scale estimate denoted by $\hat{\sigma}_{LMS}$ and its error e_{LMS} are used instead of $\hat{\sigma}$, and e_i which are based on OLS.

Least trimmed squares (LTS)
This robust regression techniques minimizes the sum of squared residuals over *n* observations and subset k of those observations, thus the *n-k* observations which are excluded does not has effect on the fit.The LAD scale estimate is denoted by $\hat{\sigma}_{LAD}$ and its error e_{LAD} are used instead of $\hat{\sigma}$, and e_i which are based on OLS.

Note that their cuff-value remains the same as the OLS.

Incorporating outliers into the series.

From the original series, model 1:
$q_t = y_t + \omega I_t(\tau)$, the series is contaminated with outliers.
ω is the magnitude of the outliers, τ is the time that the outliers occur and $I_t\ (\tau)$ is a dummy variable which has zero value at all lags except when time $t = T$

$$I_t = \begin{cases} 0 & , t \neq \tau \\ 1 & , t = \tau \end{cases}$$

when contamination occurs at $t = \tau$, $I_t = 1$ otherwise 0.

III. Result and Discussion

The sample size used is *n=30 & 200,* parameter is set to be $\phi = 0.7$, $\omega = 5$, *standard deviation* $\sigma = 1$, *300 replications* for *n=30* and *500 replications* for *n = 200*. To assess the power of the procedure, the following case will be considered;

i. Single outlier of AO / IO
ii. Multiple outliers AOs / IOs
iii. Both multiple outliers AOs and IOs

Fifthteen measures will be run for each cases,
Under *n=30,* the location for single outlier is set to be $\tau = 12$ and for the multiple outlier, the location is set to be k1 =12, and k= 20. For *n=200,* the location for single outlier is set to be $\tau = 26$, and multiple outliers; k1 =26, k2 =62 and k3 =99.

Table 1. A simulation study on the power of the outlier detection in regression diagnostics tools based on ols.

	$\tau = 12$	$\tau = 12$	**(n=30,ai=0.7,nsimul=300,p=1,k1=12,k2=20)**					
Case	**Single AO**	**Single IO**	**2 AOs**		**2IOs**		**Both AO and IO**	
			1st outlier	**2nd outlier**	**1st outlier**	**2nd outlier**	**1st outlier (IO)**	**2nd outlier (AO)**
Ordinary Least Square (OLS)								
H_i^2	94.6	99.3	80	80	86.3	87.3	99.3	0.3
Dffits	87.7	99.7	70.7	79	97.7	98	99.6	0.3

Di	82.7	99	54	61	92.7	93	99	0.3

Summary of the outliers detection performance. Note that the numbers are in percentage.

Table 1. Present the result of the performance of 300 replications on outlier-detection method for each single and multiple outlier specifications using the cut-off *C* value of each classical diagnostic techniques as stated earlier respectively. The numbers and percentage of the correct detection and identification are given under each location of the outliers. Under multiple outlier and "Both AO and IO", the label k1 and k2, tells us the location of 1st outlier, outlier and 2nd outlier respectively, and τ for single outlier.

The result shows that OLS based diagnostic technique, H_i^2 detect 94.6% of correct number of outlier, follow by $Dffits$(87.7%) and D_i, (82.7%) under single AO while in single IO, the percentage detection for $Dffits, H_i^2$ and D_i is very powerful i.e (99.7%, 99.3% and 99%). For multiple IO (2IOs), there seems to be a better correct detection percentage ranging from 86.3% to 99.6%, and for AO (2AOs), 54% to 80%. The result outcome of "Both AO and IO" seems to favour additive outliers (IO) and perform woefully under the additive outlier (AO), this may be that OLS based diagnostic technique can only detect correctly the first outlier that comes on it way and swamp the rest of the outlier.

Table 2. A simulation study on the power of the outlier detection based on robust versions of regression diagnostics tools.

	$\tau = 12$	$\tau = 12$	**(n=30,ai=0.7,nsimul=300,p=1,k1=12,k2=20)**					
case	**Single AO**	**Single IO**	**2 AOs**		**2IOs**		**Both AO and IO**	
			1st outlier	**2nd outlier**	**1st outlier**	**2nd outlier**	**1st outlier (IO)**	**2nd outlier (AO)**
M estimation								
H_i^2	94.6	99	80.3	82.3	88.3	91.7	99	0.3
Dffits	86.3	99.7	71.2	78.3	97.7	98.7	99.6	0.3
Di	84	99.3	53.3	61	92.3	92.7	99.3	0.3
L1/Least Absolute Deviation / Least Absolute value								
H_i^2	84.3	98.3	76.7	77.3	88.3	92	98.3	0.3
Dffits	77.7	99.3	58	64.3	95.3	95	96.3	0.3
Di	71	95.7	45	52	90	90.3	95.7	0.3
Least Median Square								
H_i^2	87.3	96	82.7	86	84	85	96	0.3
Dffits	75.3	83	63.3	68.3	83	84.3	82	0.3
Di	69	79.7	56.3	56	79.3	81	79.7	0.3
Least Trimmed Square								
H_i^2	87.3	98	73.7	73	89.7	90.7	98	0.3
Dffits	79.3	91.7	63	65.7	91.3	92.7	91.7	0.3
Di	71	85.7	43	48.3	84.7	85.7	85.7	0.3

Summary of the outliers detection performance. Note that the numbers are in percentage.

Table 2. present the results of the proposed methods based on robust version which contains four different kinds of robust regression namely M-estimation, LAD/L1, LMS and LTS respectively, the table shows the comparison between their power of performance accordingly in the correctly detection and identification in percentage.

Which one has the most powerful performance among the robust regressions, however their outliers' locations and cut-off values *C* are set as the same in Table 1. The interpretations of the table are as follows:

a. M-estimation

The power of correct detection and identification percentage for single AO under H_i^2 is 94.6% and, $Dffits$ (86.3%) and D_i (84%). There is a very powerful correct percentage detection for single IO i.e. 99% to 99.7%. Under multiple IO, the power of correct percentage detection and identification is between 88.3% to 97.7% for H_i^2 , $Dffits$ and D_i. And for multiple AO, H_i^2, $Dffits$ and D_i has 99% to 99.6% power of correctly detection and identification. The 1st outlier in "both AO and IO", which is IO has 99% to 99.6% and nothing was detected correctly in 2nd outlier which is AO.

b. Least Absolute Deviation (L1/LAD)

All the method has a low percentage detection in multiple AO and little powerful percentage detection on multiple IO of 88.3% to 95.3%. In single IO, correct outlier detection percentage for H_i^2, $Dffits$ and D_i is 95.7% to 99.3%. And single AO is 71% to 84.3%. The 1st outlier which is IO in "both AO and IO" has 95.7% to 98.3% of correct detection in H_i^2, $Dffits$ and D_i. And the 2nd outlier has approximately 0% all through the methods.

c. Least Median Square (LMS).

H_i^2 has 96% power of correct outlier detection in single IO and the rest method has power of 79.7% to 83% while in single AO, all method has power of 69% to 87.3%. For multiple AO, H_i^2 has 82.7% to 86% of correct outlier detection and the rest method has a relatively small percentage of correct detection. H_i^2, $Dffits$ and D_i has a percentage of correct detection of 79.3% to 85% in multiple IO. In "both AO and IO", H_i^2, $Dffits$ and D_i has a percentage of 96%, 82% and 79.7% on the 1st outlier which is IO and 0.3% on 2nd outlier which are IO.

d. Least Trimmed Square (LTS)

H_i^2, $Dffits$ and D_i has 85.7%, 91.7% and 98% power of correct outlier detection in single IO while in single AO, all method has less percentage power of 69% to 87.3%. For multiple AO, all the methods has percentage detection of 43% to 73.7% and for multiple IO 84.7% to 91.3% of correct percentage detection. In "both AO and IO", H_i^2, $Dffits$ and D_i has a powerful percentage of correct outlier detection of 85.7% to 98% under the 1st outlier which is IO, and 0.3% on 2nd outlier which is AO.

Table 3. A simulation study on the power of the outlier detection in regression diagnostics tools based on ordinary least square (ols)

	$\tau = 26$	$\tau = 26$	**(n=100,ai=0.7,nsimul=500,p=1,k1=26,k2=62,k3=99)**								
case	**Single AO**	**Single IO**	**3 AOs**			**3 IOs**			**Both AO and IO**		
			1st outlier	**2nd outlier**	**3rd outlier**	**1st outlier**	**2nd outlier**	**3rd outlier**	**1st outlier (AO)**	**2nd outlier (IO)**	**3rd outlier (IO)**
Ordinary Least Square (OLS)											
H_i^2	97.6	99.4	88.4	88.8	87.2	95	94.4	93.4	99.4	0.2	0.2
$Dffits$	67.8	99.4	56	53.8	53.4	98.6	99.2	98.2	99.4	0	0.2
D_i	55	99.2	38.6	36.6	35.6	96	96.6	95.6	99.2	0	0.2

Summary of the outliers detection performance. Note that the numbers are in percentage.

Table 3. Present the result of the performance of 500 replications on outlier-detection method for each single and multiple outlier specifications using the cut-off *C* value of each classical diagnostic techniques as stated earlier respectively. The numbers and percentage of the correct detection and identification are given under each location of the outliers. Under multiple outlier and "Both AO and IO", the label k1, k2, k3 tells us the location of 1st outlier, 2nd outlier and 3rd outlier respectively.

The result shows that OLS based diagnostic technique, H_i^2 detect 97.6% of correct number of outlier under Single AO. The percentage detection for single IO under H_i^2, *Dffits* and D_i is very powerful i.e (99.4%, 99.4% and 99.2%). For multiple IO (3IOs), there seems to be a better correct detection percentage ranging from 93.4% to 99.2% under the method H_i^2 , *Dffits* and D_i . The result outcome of "both AO and IO" seems to favour additive outliers (AO) and perform woefully under the innovative outlier (IO), this may be that OLS based diagnostic technique can only detect correctly the first outlier that comes on it way and swamp the rest of the outlier.

Table 4. A simulation study on the power of the outlier detection based on robust versions of regression diagnostics tools.

	$\tau = 26$	$\tau = 26$	**(n=100,ai=0.7,nsimul=500,p=1,k1=26,k2=62,k3=99)**								
Case	**Single AO**	**Single IO**	**3 AOs**			**3 IOs**			**Both AO and IO**		
			1st outlier	**2nd outlier**	**3rd Outlier**	**1st outlier**	**2nd outlier**	**3rd outlier**	**1st outlier (AO)**	**2nd outlier (IO)**	**3rd outlier (IO)**
M estimation											
H_i^2	97.6	99.6	88.6	88.8	87	94.6	94.6	94	99.6	0.2	0.2
Dffits	64.8	99.4	52.4	49.8	48.2	99.2	99	98.4	99.4	0.2	0.2
Di	53	99.2	37.6	34.4	34	96.6	95.6	95.6	99.2	0.2%	0.2
L1/Least Absolute Deviation / Least Absolute value											
H_i^2	98	99.6	90	89.4	87.2	93.8	94	92.8	99.4	0.2	98.4
Dffits	56.4	98.4	49.8	46.2	47	98.8	98.6	98	98.4	0	0.2
D_i	47	98.4	36.2	32.2	0.8	96	96	45.8	98.4	0	0.2
Least Median Square											
H_i^2	97.8	99.4	90.4	90.6	87.8	94.4	94.2	92.2	99.4	0.2	0.2
Dffits	70.4	96.2	55	53	50.6	93.8	92.4	94.2	96.2	0	0.2
D_i	56.2	94.6	42.6	38.8	40.8	91	88	88.8	94.6	0	0.2
Least Trimmed Square											
H_i^2	97.6	99.8	89.2	89.2	87.8	94.6	95	93.8	99.8	0.2	0.2
Dffits	64.6	98.8	53.6	47.2	47.8	98.8	98.4	97.2	98.8	0.2	0.2
D_i	49.6	98.4	36.2	33.6	32.8	97.4	94.8	93.8	98.4	0.2	0.2

Summary of the outliers detection performance. Note that the numbers are in percentage.

Table 4. present the results of the proposed methods based on robust version which contains four different kinds of robust regression namely M-estimation, LAD/L1, LMS and LTS respectively, the table shows the comparison between their power of performance accordingly in the correctly detection and identification in percentage. Which one has the most powerful performance among the robust regressions, however their outliers' locations and cut-off values *C* are set as the same in Table 1. The interpretations of the table are as follows:

a. M-estimation

The power of correct detection and identification percentage for multiple and single AO is very poor under the M-estimation except for H_i^2 single AO that has 99.6% correct detection. Under multiple IO, the power of correct percentage detection and identification is between 94% to 99.6% for H_i^2 , $Dffits$ and D_i which is quiet powerful. H_i^2, $Dffits$ and D_i has 99.2% to 99.6% power of correctly detection and identification under the 1st outlier which is AO in "Both AO and IO" and nothing was detected correctly in 2nd outlier and 3rd outlier which were IOs.

b. Least Absolute Deviation (L1/LAD)

$Dffits$ and D_i has 0.8% to 56.4% power of correct outlier detection which is relatively low, meanwhile H_i^2 has correct power detection of 87.2% to 90% in multiple AO and 98% in single AO. In single and multiple IO, correct outlier detection percentage for H_i^2, $Dffits$ and D_i which has the percentage of 92.8% to 99.6%. The 1st outlier which is AO in "both AO and IO" has 98.4% to 99.4% of correct detection in H_i^2, $Dffits$ and D_i . For 2nd and 3rd outlier, they have approximately 0% correct detection expect the H_i^2 3rd outlier that has 98.4% correct detection.

c. Least Median Square (LMS).

Only H_i^2 attain 97.8% correct percentage detection in single AO and 87.8% to 90.6% correct detection from the 1st outlier to 3rd outlier in multiple AO. H_i^2 also have 87.8% to 90.6% in multiple IO and 99.4% in single IO, however $Dffits$ and D_i has correct percentage detection of 96.2% and 94.6% in single IO and 88% to 94.4% in multiple IO. In "both AO and IO", H_i^2, $Dffits$ and D_i has a powerful percentage of 99.4%,96.2% and 94.6% on the 1st outlier which is AO and 0% to 0.2% on 2nd and 3rd outlier which are IO.

d. Least Trimmed Square (LTS)

H_i^2 has 97.6% power of correct outlier detection in single AO and the rest method has power of 1% to 64.6% while in single IO, all method has power of 98.4% to 99.8%. For multiple AO, H_i^2 has 87.8% to 89.2% of correct outlier detection and the rest method has a relatively small percentage of correct detection. H_i^2, $Dffits$ and D_i has a percentage of correct detection of 93.8% to 98.8% in multiple IO. In "both AO and IO", H_i^2, $Dffits$ and D_i has a powerful percentage of correct outlier detection of 98.4% to 99.8% under the 1st outlier which is AO, and 0.2% on 2nd and 3rd outlier which are IO.

IV. SUMMARY

It is seen from the result that under small sample size, OLS performance is good under innovative outlier for single, and multiple outliers, and also in "both AO and IO", the percentage correction outlier detection is only innovative outlier. However, generally, the power of percentage correct outlier detection only give best result for both OLS based and robust version based under innovative outliers. Robust version base on M-estimation and OLS based perform similar way and the rest robust version method perform less.

Under large sample size, the result also indicate that regression diagnostics tools based on OLS perform similar way to other various kind of robust versions that are based on robust regression. However, in this part of the power of correct outlier detection, LTS performance is the best with the probability percentage of 87.8% to 99.8% followed by L1/LAD (87.2% to 99.6%), LMS (87.8% to 99.4%) and M-estimation (87% to 99.6%). In spite of this Hadi's influence measure H_i^2 perform the best in both OLS based and robust based version. Meanwhile, all diagnostics measure didn't detect any outlier in "both AO and IO" except for the first outlier which is AO, no outlier is detected in second and third outlier under IO.

V. Conclusion

In a small sample size, OLS and M-estimation is suggest to be use for the detection and identification of outlier under innovative outliers (IO). However both method fail to detect any number of correct outliers detection when the sample were mixed by both innovative outliers and additive outliers. Other robust estimation methods performed less. On the other side, large sample size, LTS perform best in the simulation study compared to other measures, however it is not robust to a series that is contaminated with both AO and IO. It can only detect the AO in the series. Also Cook's Distance and The welsch-kuh distance are not robust to multiple AO.

Reference

[1] R. S. Tsay, *Analysis Of Financial Time Series*, pp. 1-448, 2002.
[2] S. S. S. Abd Mutalib, and K. Ibrahim, *Identification Of Outliers: A Simulation Study,* ARPN Journal of Engineering and Applied Sciences, Vol. 10, pp. 326-330, 2015.
[3] J. Lopez-de-Lacalle, *tsoutliers R Package for Detection of Outliers in Time Series,* pp. 1- 28, 2014.
[4] J. Xu,. B. Abraham and S. H. Steiner, *Outlier Detection Methods in Multivariate Regression Models*, pp. 1-27.
[5] E. Widodo, S. Guritno ands.Haryatmi, *Response Surface Models with Data Outliersthrough a Case Study,* Applied Mathematical Sciences, Vol. 9, pp. 1803 – 1812, 2015.
[6] R. A. Yaffee, *Regression Analysis: Some Popular Statistical Package Options,* Social Science, and Mapping Group Academic Computing Services Information Technology Services, pp. 1-12. 2002.
[7] F. H. Thanoon, *Robust Regression by Least Absolute Deviations Method*, International Journal of Statistics and Applications, Vol. 5(3), pp. 109-112, 2015.
[8] S. Bachmann and M. Losler, *IVS Combination Center at BKG - Robust Outlier Detection Weighting Strategies*, IVS General Meeting Proceedings, pp. 266 – 270, 2012.
[9] S-PLUS Language Reference.