International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp.13429-13434

Outlier Detection based on Robust Parameter Estimates

Nor Azlida Aleng[1], Nyi Nyi Naing[2], Norizan Mohamed[3] and Kasypi Mokhtar[4]

[1,3] School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia.
[2] Institute for Community (Health) Development (i-CODE), Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Terengganu, Malaysia.
[4] School of Maritime Business and Management, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia.
Orcid; 0000-0003-1111-3388

Abstract

Outliers can influence the analysis of data in various different ways. The outliers can lead to model misspecification, incorrect analysis results and can make all estimation procedures meaningless. In regression analysis, ordinary least square estimation is most frequently used for estimation of the parameters in the model. Unfortunately, this estimator is sensitive to outliers. Thus, in this paper we proposed some statistics for detection of outliers based on robust estimation, namely least trimmed squares (LTS). A simulation study was performed to prove that the alternative approach gives a better results than OLS estimation to identify outliers.

Keywords: Outliers, least trimmed squares (LTS) and robust regression.

INTRODUCTION

The linear regression can be expressed in terms of matrices as $y = X\beta + \varepsilon$; where y is an *N*-dimensional column vector, X is an $N \times (K + 1)$ matrix and ε is an *N*-dimensional column vector of error terms, i.e.

$$\underset{N\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix}} = \underset{N\times(K+1)}{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1K} \\ 1 & x_{21} & \cdots & x_{2K} \\ 1 & x_{31} & \cdots & x_{3K} \\ 1 & \vdots & \cdots & \vdots \\ 1 & x_{N1} & \cdots & x_{NK} \end{bmatrix}} \underset{(K+1)\times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}} + \underset{N\times 1}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}} \quad (1)$$

In fitting multiple linear regession model (1.1), the most widely used ordinary least squares (OLS) to find the best estimates of β. Unfortunately, in the presence of outliers, the OLS estimators are still biased. Outliers play an important role in regression. An outliers (observations) that is quite different from most the other values or observations in a data set. Observations in a data set can be outliers in several different ways. According to Barnett and Lewis (1994), an outlier is an observation that is inconsistent with the rest of the data. Even one outlier can effect the regression model. There is an evidence that the outliers can lead to model misspecification, incorrect analysis result and can make all estimation procedures meaningless, (Rousseeuw and Leroy, 1987; Alma, 2011; Zimmerman, 1994, 1995, 1998). Outliers in the response variable represent model failure. Outliers in the regressor variable values are extreme in X-space are called leverage points.

Rousseeuw and Lorey (1987) defined that outliers in three types; 1) vertical outliers, 2) bad leverage point, 3) good leverage point. Vertical outlier is an observation that has influence on the error term of y-dimension (response) but is not influential in the space of x-dimension (regressor). Good leverage point is an observation that are outlying in the independent variables but is not a regression outlier. This point does not affect the least square estimation but it affects statistical inference since this point cut down the estimated standard errors. Good leverage points improve the precision of the regression coefficients. Bad laverage point is an observation that is outlying in independent variables and located far from the true regression line and reduce the precision of regression coefficients.

Generally speaking, there are two techniques for handling the outliers. The first is to use the some robust procedure which resist their influence in the statistical analysis. The second is to remove outliers from the data set. Therefore, in this study it is proposed some kind of robust procedure namely the least trimmed squares (LTS). This method is reliable for identifying the regression outliers both in simple and multivariates situations. The objective of this study was to detect the outliers and leverage points in the data set. The performance of the proposed method was discussed extensively by using medical data.

MATERIALS AND METHODS

This study focused on the blood pressure data. Systolic blood pressure (sbp) is a dependent variable and independent variables namely in Table 1. A total of 100 repondents were selected and diagnosed to have blood pressure problem based on WHO criteria. The explanation

of the variables is shown in Table 1 and the data were collected from Health Centre HUSM in Malaysia.

Table 1: Explanation of the Variables

Code	Variables	Explanation of the variables
y	SBP	Systolic blood pressure
x_1	AGE	Age (year)
x_2	BMI	Body mass index
x_3	TOTCHOLES	Total cholesterol (Mmol/L)
x_4	DIABETES	Diabetes mellitus; 0 = No, 1 = Yes
x_5	DBP	Diastolic blood pressure
x_6	HDL	HDL cholesterol
x_7	HEIGHT	Height (m)
x_8	TRIG	Triglycerides
x_9	WEIGHT	Weight (kg)

As is well known, a large number of diagnostics have been proposed to detect outliers. Practically, the diagnostics which was based on the ordinary least squares estimates were not efficient and biased when outliers existed in the data. Thus, to remedy this problem, least trimmed squres estimamtors (LTS) was proposed. This was an alternative approach in dealing with outliers in regression analysis.

The Least Trimmed Squares Estimators (LTS)

A statistical procedure is regarded as robust if it performs reasonably well even when the assumptions of the statistical model are not true. If we assume our data follows standard linear regression, least squares estimates and test perform quite well, but they are not robust with the presence of outlier observation(s) in the data set (Rousseeuw and Leroy, 1987). In this case we proposed the popular robust technique is the called LTS estimator. LTS estimation is a high breakdown point. The breakdown point is a measure for stability of the estimator when the sample contains a large fraction of outliers (Hampel, 1975). LTS defined as:

$$\hat{\beta}^{(LTS)} = \arg\min = \sum_{i=1}^{h} r^2(\beta) \quad (2)$$

where $r^2_{(1)} \le ... \le r^2_{(n)}$ are ordered squared residuals. Robust regression is extremely useful in identifying outliers. LTS regression is a reliable data analytical tool that may be used to discover regression outliers both in simple and multivariable conditions.

In this paper, three diagnostis are used to identify outliers which are given below.

The robust distance is defined as:

$$RD(x_i) = \sqrt{[x_i - T(X)]^T C(X)^{-1}[X_i - T(X)]} \quad (3)$$

where $T(X)$ and $C(X)$ are the robust locationand scatter matrix for thr multivariable.

The Mahalanobis distances is useful technique for detecting outliers is defined as

$$MD = \sqrt{(x_i - \mu).C^{-1}(x_i - \mu)^T} \quad (4)$$

Where *C* is the classical sample covariance matrix. In classical linear regression,the diagonal element h_{ii}

of the *hat* matrix, $H = X(X^T X)^{-1} X^T$ are used to identify leverage points. Rousseeuw and Van Zomeren (1990) defined the relationship between the h_{ii} and MD_i

$h_{ii} = [((MD_i)^2 / (n-1)] + [1/n]$. Rousseeuw and Lorey (1987) suggest using $h_i > 2p/n$ and $MD_i^2 > \chi^2_{p-1;0.95}$ as benchmarks for leverage and Mahalanobis distances.

The Cook's distance is defined as,

$$CD = (p\sigma^2)^{-1}(\hat{Y}_i - \hat{Y})^T(\hat{Y}_i - \hat{Y}) \quad (5)$$

where σ^2 is estimator of the error variance, $\sigma^2 = \sum_{i=1}^{n} r_i^2 / n - p.$ Cook suggests that *CD* be compared to a central *F* distribution with *p* and *n* – *p* degrees of freedom. This gives the cutoff values is very high. The conventional cutoff value is $4/n - p.$ Generally, when the statistics *CD*, h_i and MD_i are large, case *i* may be an outlier or influential case. Therefore the diagnostic is very important to identify the outliers and provides resistant results in the presence of outliers.

Below is the algorithm in SAS language for the multiple linear regression and robust regression. This algorithm using SAS software which is given as follows:

data BP;

input sbp age bmi totcholes diabetes dbp hdl height trig weight; datalines;

198	81	29.20	224	1	73	47	164.40	106	49.00
166	81	34.44	232	3	80	34	176.90	173	77.87
131	81	22.66	225	1	67	60	168.00	85	34.01
104	81	21.88	200	1	52	55	167.00	83	31.06
123	81	18.48	93	3	66	38	168.0	56	22.21
146	81	31.69	202	1	52	33	162.50	144	53.76
161	81	21.70	211	1	65	57	175.00	107	36.51
133	81	23.71	193	1	64	60	179.00	114	46.04
136	81	26.99	150	3	71	44	170.50	138	48.54
177	81	27.96	206	3	73	42	163.00	209	44.37
166	82	25.02	163	2	76	66	170.30	67	42.64
134	82	23.23	173	1	64	43	172.80	76	39.42
95	82	21.62	211	1	61	62	172.00	98	34.01
122	82	20.98	187	1	69	61	181.00	72	38.78
137	82	27.22	175	1	73	56	177.00	84	55.35
128	83	24.15	213	1	72	52	161.00	85	32.65
167	83	22.52	211	1	74	58	165.50	91	31.74
157	84	28.02	172	1	76	43	169.50	86	50.58
100	84	23.50	156	3	56	35	179.00	192	45.36
198	84	22.46	197	1	87	65	164.50	96	30.84
128	84	25.77	154	2	65	30	174.00	190	48.09
144	84	24.61	230	1	82	47	173.60	148	44.23
121	84	25.68	257	3	71	42	187.00	282	59.89
128	84	22.56	151	2	70	51	172.50	84	37.19
127	84	27.00	143	1	74	46	168.50	59	46.73
194	84	19.48	142	1	76	73	164.00	79	22.44
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
143	85	25.80	160	1	60	65	177.30	111	51.18
128	85	34.15	195	1	68	35	169.00	182	67.61
105	85	33.64	229	1	63	48	176.00	100	74.28
183	85	24.96	177	1	72	44	170.00	263	42.19
176	85	22.54	155	1	64	52	160.50	64	28.11
128	85	23.80	186	1	71	51	161.00	106	31.74
153	85	24.20	234	1	68	40	162.00	279	33.56
142	86	21.19	185	1	68	57	170.00	82	31.29

149	86	25.15	184	1	52	63	175.00	76	47.09
130	89	26.84	224	1	70	51	171.00	91	48.54
168	90	23.11	216	1	69	44	171.00	131	37.65
161	90	26.67	160	3	82	81	166.50	52	44.00
146	90	24.01	145	1	66	34	165.50	101	35.83
156	90	28.82	176	2	91	51	165.00	88	48.54
95	90	26.21	215	2	60	39	177.00	204	52.17
144	91	24.32	200	1	55	57	163.10	74	34.74
104	91	27.02	185	1	51	61	176.00	78	53.76
125	91	21.92	179	1	50	43	164.00	131	29.02
141	92	28.95	210	1	73	56	173.00	147	56.71
80	92	19.31	98	2	58	52	170.00	70	25.84
163	94	20.55	194	1	77	59	172.00	73	30.84

```
;
run;
ods rtf file='robdunc0.rtf'
style=journal;
ods graphics on;

/* first we do multiple linear regression
*/
proc reg data=BP;
 model sbp = age bmi totcholes diabetes
dbp hdl height trig weight;
run;

/* then we do robust regression, in this
case, LTS-estimation */
proc robustreg data=BP method=LTS;
 model sbp = age bmi totcholes diabetes
dbp hdl height trig weight ;
run;

/* then we do robust weighted regression
*/
ods graphics on;
proc robustreg method=lts(h=33)fwls
data=BP plots=all;
model sbp = age bmi totcholes diabetes
dbp hdl height trig weight /diagnostics
leverage;
output out=robout r=resid sr=stdres;
run;
ods graphics off;
/* QQ plot and histogram */
ods graphics on;
proc robustreg data=BP
plots=(rdplot ddplot reshistogram
resqqplot);
model sbp = age bmi totcholes diabetes
dbp hdl height trig weight;
run;
ods graphics off;

/* Cook's distance plot*/
proc reg data=BP;
plot (only label)=(RStudentByLeverage
CooksD);
run;

ods graphics off;

ods rtf close;
```

RESULTS AND DISCUSSIONS

In this study, a set of real data which is referred to blood pressure data is used to see how well the diagnostic statistics with robust estimator perform for the regression model. The step is the blood pressure data was analyzed using robust regression method LTS with diagnostics tool such as Mahalanobis distance, robust MCD distance and standardized robust residuals. Results of diagnostics of outliers and leverage points are presented in Table 2.

Table 2: Robust diagnostics based on least trimmed squares (LTS)

Obs	*Mahalanobis Distance*	*Robust MCD Distance*	*Leverage*	*Standardized Robust Residual*	*Outlier*
1	4.2932	8.3005	*	-3.9084	*
11	3.5879	4.6341	*	-0.1497	
15	3.4103	7.4316	*	-0.0483	
17	3.6934	7.5759	*	-0.1718	
20	3.0788	3.3079		3.9991	*
21	3.0436	3.0932		-3.7315	*
23	3.3320	3.4903		3.4291	*
26	3.7465	7.9784	*	-4.0062	*
27	3.4066	5.6654	*	-1.8312	
28	4.1170	7.5335	*	-1.4298	
31	3.9957	5.9740	*	1.5210	
32	3.6191	7.3727	*	0.0335	
43	3.9779	4.2679		-4.7390	*
44	2.7220	2.9955		-3.4499	*
49	3.4822	3.6642		-3.7414	*
55	2.8862	3.9204		-3.8469	*
57	2.3094	2.4075		-3.2261	*
62	4.1183	5.8403	*	-0.0572	
65	3.5233	7.3285	*	-1.4330	
68	3.4316	6.4791	*	0.0397	
72	4.3654	5.4444	*	-0.1697	
75	4.3625	8.3595	*	0.1910	
80	3.2016	6.5804	*	0.5169	
81	3.4362	4.6487	*	-0.0485	
93	5.7096	10.4714	*	2.0859	
94	3.7888	7.3084	*	-0.2644	
96	5.1163	7.1482	*	-9.5821	*
97	2.5207	2.6372		-3.6185	*
98	5.5412	8.5083	*	-2.8619	
100	2.9342	3.2646		-3.4534	*

The model of blood pressure is:

$$\hat{y} = 888.873 - 15.423_{age} + 0.427_{dbp} - 4.577_{height} + 5.028_{weight} + e \quad (6)$$

and R squared valus is 0.8523 (85.23%), this indicated the greater the ability of that model to predict a trend. The results of Table 2 show that the existing of outliers in the blood pressure data. 13 observations are considered as outliers and observation 93 has high leverage. Although, the set data presence of outliers, results remain robust.

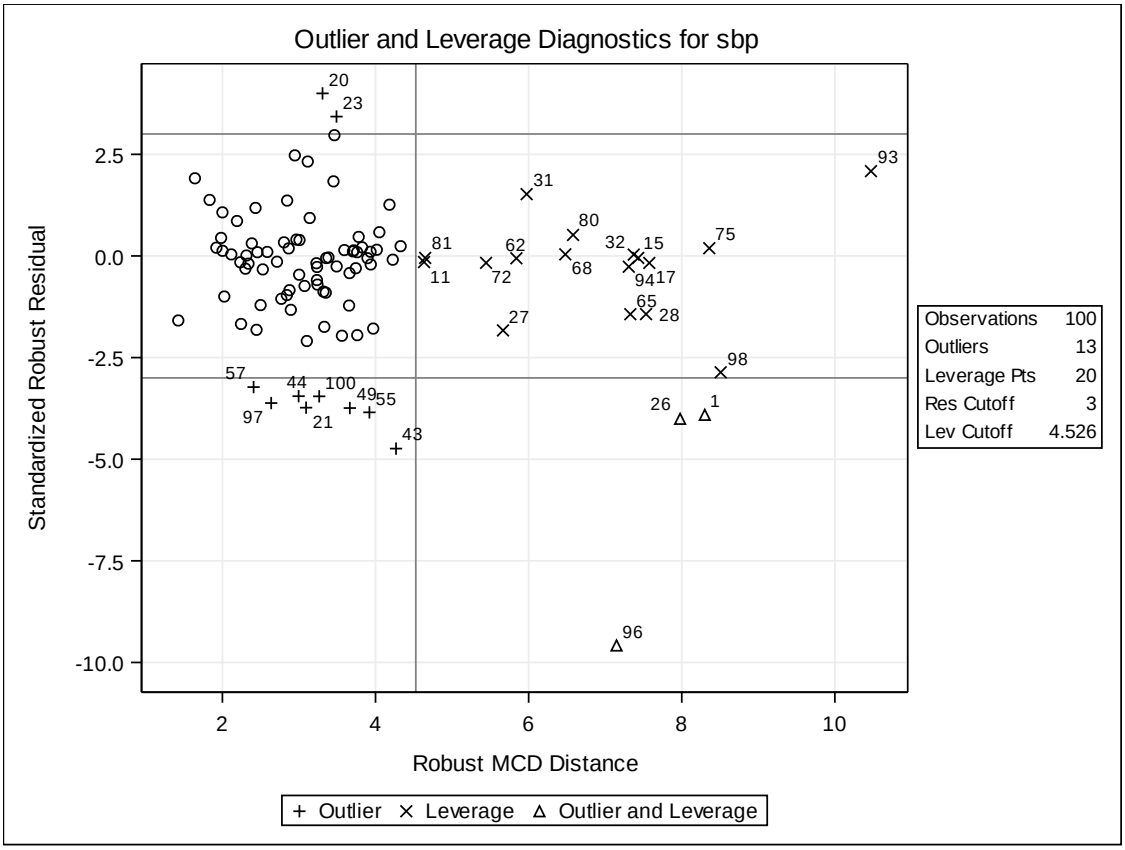

Figure 1: Regression diagnostic plot for systolic blood pressure.

Based on Figure 1, gives evidence of the presence of outlying observations because the poits fall behind the band. Observations (1, 20, 21, 23, 26, 43, 44, 49, 55, 57, 96, 97 and 100) are identified as outliers. We can see that, 20 observations are identified as laverage points and observations 1, 26 and 96 identified as outliers and leverage points at the same time.

CONCLUSION

Least trimmed squres estimamtors (LTS) is an alternative approach in dealing with outliers in regression analysis. The value R^2 gave strong correlation and relation between variables, so it shown that strong good fit model. Robust version of the diagnostics detect all outliers in the data in one step. The results of the simulation study agree well with the real data.

REFERENCES

[1] Alma, O. G. (2011). Comparison of robust regression methods in linear regression. Int.

[2] Journal Contemp. Math. Sciences, 6(9), 409-421.

[3] Barnett, V., & Lewis, T. (1994). Outliers in statistical data. New York: John Wiley and Sons.

[4] Rousseeuw, P. J. and Leory, A. (1987). Robust Regression and Outlier Detection. Wiley Series in

[5] Probability and Statistics.

[6] Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric

[7] tests. Journal of General Psychology. 121(4): 391-401.

[8] Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and

[9] downweighting outliers. Journal of Experimental Education, 64(1), 71-78.

[10] Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by

[11] concurrent violation of two assumptions. Journal of Experimental Education. 67(1): 55-68.