ALTERNATIVE OUTLIERS DETECTION PROCEDURES IN LINEAR REGRESSION ANALYSIS: A COMPARATIVE STUDY

Moawad El-Fallah Abd El-Salam
Department of statistics &Mathematics and Insurance,
Faculty of commerce, Zagazig University ,
Zagazig City, Egypt ,
Telephone: 00201093093959

ABSTRACT: *A Common problem in linear regression analysis is outliers, which produces undesirable effects on the least squares estimates. Many widely used regression diagnostics procedures have been introduced to detect these outliers. However, such diagnostics, which are based on the least squares estimates, are not efficient and cannot detect correctly swamping and masking effects. In this paper, we attempt to investigate the robustness of some well known diagnostics tools, namely, Cook's distance, the Welsch-Kuh distance and the Hadi measure. The robust version of these diagnostics based on the Huber-M estimation has been proposed to identify the outliers. A simulation study is performed to compare the performance of the classical diagnostics with the proposed versions. The findings of this study indicate that, the proposed alternative versions seem to be reasonable well and should be considered as worthy robust alternative to the least squares method.*

KEYWORDS: Outliers; Diagnostics; Linear regression; least squares method; Huber– M estimation ; Simulation study .

INTRODUCTION

Consider the linear regression model as:

$$Y = X\beta + \in \quad , \qquad (1)$$

where, Y is an $(n \times 1)$ vector of response, X is an $(n \times p)$ design matrix of rank p, β is a $(p \times 1)$ vector of unknown parameters and $\in$ is an $(n \times 1)$ vector of random errors with $E(\in) = 0$ and $V(\in) = \sigma^2 I_n$, where σ^2 is an unknown parameter and I_n is the identity matrix of order n. The Ordinary Least Square (OLS) estimator of β is :

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y \quad , \qquad (2)$$

and the vector of fitted values is :

$$\hat{Y} = X\hat{\beta} = HY , \qquad (3)$$

where,

$$H = X(X'X)^{-1} X' \qquad (4)$$

The residual vector is defined as :

$$e = Y - \hat{Y} = (I_n - H)Y , \qquad (5)$$

and the least square estimate of σ^2 is the residual mean square ,

$$\hat{\sigma}^2 = \frac{e'e}{n-p} \tag{6}$$

It is important to note that , the least square estimates are very sensitive to the outlying data points. Many diagnostics tools have been developed to detect these outlying observations. However, the majority of such diagnostics are developed from the classical least squares, and as a result, these diagnostics are not efficient to detect the correctly swamping and masking effects. In this respect, robust regression approach is an important tool for analyzing data contaminated with outliers. It can be used to detect outliers and provides resistant results in the presence of outliers. In this way , robust methods, which are not easily affected by outliers, are put forward to remedy the effects of outliers on least squares estimates . There are many robust methods in the literature . The M-estimation, introduced by Huber (1981), is the most used one.

The Huber-M estimator can be defined as the solution of the following minimization problem :

$$\min_{\beta} \sum_{i}^{n} \rho\left(\frac{e_i}{S}\right), \tag{7}$$

where , S is a robust estimate of scale, and ρ is related to the likelihood function for an appropriate choice of the error distribution (Abd El-sallam, 2003) .
The minimization problem (7) gives the system of p equations as :

$$\sum_{i}^{n} x_{ij}\, \psi\left(\frac{e_i}{S}\right) = 0, \qquad j = 1,2,\ldots,P \tag{8}$$

where $\psi = \rho'$ is the derivative of ρ with respect to β, and x_{ij} is the i^{th} observation on the j^{th} regressor. The ψ function is nonlinear and equation (8) must be solved by iterative methods.The iteratively reweighted least squares is most widely used to obtain the robust M-estimator ($\hat{\beta}_M$) as : [For more details, see, Beckman and Cook, 1983, and Riatoshams, et al., 2009]

$$\hat{\beta}_M = (X'WX)^{-1} X'WY, \tag{9}$$

where, W is an $(n \times n)$ diagonal matrix of weights, and the robustness of an estimator depends mainly on the equivalent weights derived from their corresponding ψ - functions (Huber, 1981) in order to resist the influence of outliers.

In this paper, we attempt to study the robustness of some well known diagnostics tools to detect outliers, which namely, Cook's distance (D), the Welsch-Kuh distance (FFITS) and the Hadi measure (H_i^2). Based on the Huber-M estimation, the robust version of these diagnostics measures have been introduced. In section (2), a brief review of the classical forms of these diagnostics is presented. In section (3), we define the proposed alternative versions of the diagnostics under concern. Section (4) presents a numerical example to illustrate how the alternative robust outliers diagnostics are better than the classical versions.

Section (5) presents the results of a Monte Carlo simulation study to investigate such outliers diagnostics perform well, and some concluding remarks are presented in section (6).

The Classical Outliers Diagnostics :

There are many diagnostics have frequently used to identify outliers. Three of the most commonly considered diagnostics are : Cook's distance, the Welsch-Kuh distance, and the Hadi measure.

Cook's Distance (D_i) :

The Cook distance, D_i ,(Cook, 1977), measures the distance between the estimates of the regression coefficients with the i^{th} observation $\hat{\beta}$ and without the i^{th} observation $\hat{\beta}_{-i}$ for the metric $\frac{1}{P\hat{\sigma}^2}$ $((X'X)$. So , D_i is defined as :

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{\text{-i}})'(\text{X'X})(\hat{\beta} - \hat{\beta}_{\text{-i}})}{\hat{\sigma}^2 P} \qquad (10)$$

In this case, it is considered that an observation is an influential observation when D_i exceeds the cut-off point of $\frac{4}{n-P}$ (Cook, 1977).

The Welsch-Kuh Distance (DFFITS) :

The Welsch-Kuh distance measures the distance between the estimates of the predicted value with and without the i^{th} observation , ($\hat{y}_i$ and $\hat{y}_{-i}$ respectively), which is defined as :

$$DFFITS_i = \frac{\left| \hat{y}_i - \hat{y}_{-i} \right|}{\hat{\sigma}_{-i}\sqrt{h_{ii}}} , \qquad (11)$$

where, h_{ii} are the i^{th} diagonal elements of the hat matrix given in (4) . Belsley et al. (1980) recommend using $2\sqrt{\frac{P}{n}}$ as a cut-off points for DFFITS.

Hadi Measure (H_i^2) :

Hadi (1992) recommend a measure to detect overall potential influence, which is defined as :

$$H_i^2 = \left(\frac{P}{1\text{-}h_{ii}}\right)\left(\frac{d_i^2}{1-d_i^2}\right) + \left(\frac{h_{ii}}{1-h_{ii}}\right) , \qquad (12)$$

where , $d_i^2 = \frac{e^2}{e'e}$, is the square of the ith normalized residual. Hadi's measure is based on the simple fact that potentially influential observations are outliers as either X-outliers, Y – outliers, or both. Hadi (1992) recommends using $\left[\text{mean}(H_i^2) + C\sqrt{var(H_i^2)} \right]$ as a cut-off point for H_i^2 measure, where C is an appropriately chosen constant such as 2 or 3 .

Chatterjee and Hadi (1988) recommended that, the cut-off points should be used with caution. Because, diagnostics measures are not designed to be formal tests of a hypothesis, but they are designed to detect observations which affects regression results more than other observations in a data set. Thus, the values of a given diagnostics should be compared to each other. This can be done using graphical displays such as a stem and leave display, index plot, or P-R plot.

The Alternative Outliers Diagnostics :

This section is devoted to discuss the robust versions of the above outliers diagnostics. In this respect, the Huber-M estimator of β instead of $\hat{\beta}$, which is the least square estimator , and the robust scale estimate of σ^2 instead of $\hat{\sigma}^2$, which is the least square estimator, is used. The robust version of D_i can be obtained by :

$$R\,D_i = \frac{(\tilde{\beta} - \tilde{\beta}_{\text{-i}})'(\text{X}'\text{X})(\tilde{\beta} - \tilde{\beta}_{\text{-i}})}{\tilde{\sigma}_r^2\, P} \qquad (13)$$

where $\tilde{\beta}$ is the robust estimation of β and $\tilde{\sigma}_r$ the robust scale estimation of σ . By this way, the robust version of DFFITS is obtained by :

$$RDFFITS_i = \frac{|\tilde{y}_i - \tilde{y}_{-i}|}{\tilde{\sigma}_{-i}\sqrt{h_{ii}}}\ , \qquad (14)$$

where, $\tilde{y}_i$ and $\tilde{y}_{-i}$ are the robust predicted values of Y with and without the i^{th} observation .

Finally, the robust version of H_i^2 can also be obtained by :

$$RH_i^2 = \left(\frac{P}{1\text{-}h_{ii}}\right)\left(\frac{\tilde{d}_i^2}{1-\tilde{d}_i^2}\right) + \left(\frac{h_{ii}}{1-h_{ii}}\right), \qquad (15)$$

where , $\tilde{d}_i$ is the robust normalized residual, which is calculated after a robust fit, and is used instead of normalized residual in equation (12).

Finally, in order to compare the classical outliers diagnostics with the robust diagnostics, the cut-off points for the robust diagnostics are taken as the cut-off points proposed for Hadi's measure as mentioned in section(2).

Illustrative Example

In this section, a specific set of data taken from Hill (1977) is used to see how well the classical diagnostics with robust version diagnostics perform for the regression model. The data set, which is represented in Appendix consists of 15 observations on six regressors and a response. This data have been used extensively to identify outliers with its effects on the least squares estimators .

For investigating the existence of outlying data points, the classical diagnostics measures with robust version diagnostics are pointed out (Belsley et al. 1980). Table (1) lists the most outlying data points using the least squares and Huber M-estimation fits, while, for the purpose of comparisons, table (2) presents the results of the parameter estimates and the mean squared error (MSE) for the least squares and robust M-estimations.

Table (1): Outlying Data points using LS and M-estimation Diagnostics**.**

Obs.	**Classical Diagnostics**			**Robust Version Diagnostics**		
	D_i	$DFFITS_i$	H_i^2	RD_i	$RDFFITS_i$	RH_i^2
1	2.01*	4.09*	0.72*	2.31*	4.12*	0.81*
2	4.11*	7.98*	1.02*	3.49*	6.11*	0.94*
8	1.72*	1.85*	0.92*	1.82*	2.90*	0.96*
12	0.42	1.11	0.28	1.76*	1.34*	0.36*
15	0.37	1.04	0.30	2.14*	1.99*	0.42*

*denotes the outlying data Points

Table (2) : The Estimates, Outliers and MSE using LS and M-robust estimation.

Estimator	$\hat{\beta}_o$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	MSE	Outlying points
LS	93.08	-0.51	-1.01	-2.32	0.09	-0.36	0.49	10.26	1,2,8
M-Robust	95.1	-0.46	-0.99	-2.14	0.15	-0.35	0.61	4.72	1,2,8,12,15

The results of table (1) show that, D_i , $DFFITS_i$ and H_i^2 could correctly identify observations 1, 2 and 8 are outliers, according to D_i , $DFFITS_i$ and H_i^2 the cut-off point are : 0.44 , 1.18 and 0.36 respectively. However, these classical diagnostics measures failed to detect the observations 12 and 15 as outliers. In addition , a pair-wise comparison of the two formes of diagnostics could be made. Accordingly, the robust version of outliers diagnostics correctly identified the observations : 1, 2, 8, 12 and 15 as outliers. While, the results of table (2) show that, when the robust M-estimation is used, the influence of the outlying data points decreased strongly as be shown by the value of MSE.

Simulation Study

In this section, a Simulation study is conducted to compare the classical regression diagnostics with its robust versions to detect outliers. In the simulation study, data having 20 and 40 observations with 3 independent variables are generated from a uniform distribution. The residual are generated from a normal distribution with mean 0 and variance $\sigma^2 = 1$ and $\sigma^2 = 10$, respectively. These variables, which are added to a regression model, taken the values of the coefficients as $(5,3,\sqrt{6})$. In order to see the effects of outliers on the results of the analysis, diagnostics based on least squares and the robust M-estimation are examined in the case of one outlier and two outliers . In this respect, outliers are generated in two different ways.

Case A : a value in proportion to the variance is added to the largest value of the dependent variable .

Case B : the last observation of the dependent variable has been turned into an outlier by taking too large a value.

The diagnostics based on the LS and M-estimation are applied to the cases A and B of data. 1000 replications have been made and the results are shown in Tables (3) and (4). The values presented in these tables show the percentage of correctly detected outliers for the diagnostics.

Table (3) : Simulation Results for Case A with $n = 20$ and $n = 40$

$n = 20$	$\sigma^2 = 1$		$\sigma^2 = 10$	
	One outlier	Two outliers	One outlier	Two outliers
D_i	93 %	86 %	41 %	100 %
$DFFITS_i$	70 %	97 %	85 %	100 %
H_i^2	98 %	61 %	97 %	77 %
RD_i	100 %	99 %	100 %	100 %
$RDFF_i$	100 %	100 %	100 %	100 %
RH_i^2	100 %	100 %	100 %	100 %
$n = 40$	One outlier	Two outliers	One outlier	Two outliers
D_i	98 %	96 %	100 %	96 %
$DFFITS_i$	85 %	56 %	85 %	60 %
H_i^2	97 %	76 %	100 %	100 %
RD_i	100 %	100 %	100 %	100 %
$RDFF_i$	100 %	100 %	100 %	100 %
RH_i^2	100 %	100 %	100 %	100 %

Table (4) : Simulation Results for Case B with $n = 20$ and $n = 40$

$n=20$	$\sigma^2=1$		$\sigma^2=10$	
	One outlier	**Two outliers**	**One outlier**	**Two outliers**
D_i	100 %	98 %	100 %	100 %
$DFFITS_i$	100 %	100 %	100 %	100 %
H_i^2	100 %	52 %	100 %	60 %
RD_i	100 %	100 %	100 %	99 %
$RDFF_i$	100 %	100 %	100 %	100 %
RH_i^2	100 %	100 %	100 %	100 %
$n=40$	One outlier	Two outliers	One outlier	Two outliers
D_i	92 %	97 %	93 %	95 %
$DFFITS_i$	41 %	54 %	42 %	51 %
H_i^2	100 %	50 %	100 %	50 %
RD_i	100 %	100 %	100 %	100 %
$RDFF_i$	100 %	100 %	100 %	100 %
RH_i^2	100 %	100 %	100 %	100 %

The results of table (3) show the following :

- In the case of $\sigma^2=1$ with one outlier when $n=20$, D_i , $DFFITS_i$ and H_i^2 correctly identify : 93 % , 70 % and 98 % of outliers respectively . In addition, in case of two outliers, D_i correctly identify 86 % , $DFFITS_i$ identify 97 % and H_i^2 identify 61 % of the outliers .
- In the case of $\sigma^2=1$ with one outlier when $n=40$, D_i , $DFFITS_i$ and H_i^2 correctly identify : 98 % , 85 % and 97 % of the outliers . While in case of two outliers, D_i , $DFFITS_i$ and H_i^2 identify : 96 % , 56 % and 76 % of the outliers respectively.
- In Case of $\sigma^2=10$ with one outlier when $n=20$, D_i , $DFFITS_i$ and H_i^2 correctly identify : 41 % , 85 % and 97 % of the outliers respectively , In addition H_i^2 identify 77 % of outliers, in case of two outliers when $\sigma^2=10$ and $n=20$.
- In the case of $\sigma^2=10$ with one outlier when $n=40$, $DFFITS_i$ correctly identify 85 % of outliers, while, for case of two outliers, D_i and $DFFITS_i$ correctly identify 96% and 60 % respectively.

The results of table (4) show the following :

- In the case of $\sigma^2 = 1$ with one outlier when $n = 40$, D_i and $DFFITS_i$ identify 92 % and 41 % of the outliers. While, in case of two outliers, D_i, $DFFITS_i$ and H_i^2 correctly identify : 97 %, 54 % and 50% respectively. In addition, H_i^2 identify 52 % of the outliers, for the case of two outliers and $n = 20$
- In the case of $\sigma^2 = 10$ with one outlier when $n = 40$, D_i and $DFFITS_i$ correctly identify 93 % and 42 % of the outliers. Also, these measures correctly identify : 95% and 51% of the outliers respectively. In addition, H_i^2 identify 50 % and 60 % only of the outliers, for the case of two outliers when $n = 40$ and $n = 20$ respectively.

CONCLUSION

The potential effects of outliers on the various aspects of least squares analysis are well known. So, a serious problem that often occurs in linear regression analysis is the presence outliers. Many widely used regression diagnostics measures have been presented to detect these outliers. However, such diagnostics, which are based on the least squares estimates, are not efficient to detect correctly swamping and masking effects. In this paper, alternative robust versions of Cook's distance (D_i), Welsch-Kuh distance ($DFFITS_i$) and the Hadi measure (H_i^2) are proposed to detect outliers. As seen from the results of Hill data, the classical diagnostics based on least squares and M-estimation detect the same observations as outliers, In addition to these observations, as stated in Hill (1977) and Belsley et al. (1980), diagnostics based on the Huber M-estimation detect other observations as outliers. A Simulation study is performed, using the ROBUSTREG procedure in SAS version 9, to compare the performance of the classical diagnostics with the proposed versions. The results of this study support the results of the proposed diagnostics based on alternative robust versions. Therefore, the results indicate that , the proposed alternative versions of detection diagnostics see to be reasonable well and should be considered as worthy robust alternatives to the least squares estimation.

REFERENCES

Abd El-Salam M.EL., 2003, Generalized Shrunken-Type M-Estimation, The Egyptian Statistical Journal, 47,(1), 36-49.

Beckman R.J. and Cook R.D. ,1983, Outlier…s, Technometrics, 25 (2), 119-149.

Belsey D.A., Kuh E. and Welsch R.E. ,1980, Regression Diagnostics, New York : John Wiley.

Cetin M. ,2009, Robust model selection criteria for robust Liu estimator, European Journal of operational Research, 199, 21-24.

Chatterjee S. and Hadi A.S. ,1988, Sensitivity Analysis in linear Regression , New York : John Wiley.

Chen C. ,2002, Robust Regression and outlier detection with the ROBUSTREG procedure (proceedings of the Twenty-Seventh Annual SAS Users Group International Conference, Cary, NC. SAS Institute Inc.

Cook R.D. ,1977 , Detection of Influential observations in linear Regression ,Technometrics, 19, 15-18.

Hadi A.S. ,1992,A new measure of overall potential influence in linear Regression, Computantial Statistics and Data Analysis, 14, 1-27.

Hill R.W. ,1977, Robust regression when there are outliers in the Carriers Ph.D. thesis. Harvard university, Cambridge, Masss.

Huber P.J. ,1981, Robust Statistics , New York : John Wiley .

Riazoshams A.H., Habshah B. and Adam C.M.B. ,2009, On the outlier detection in nonlinear Regression , Engineering, and Technology, 60, 264-470.

Ullah M.A. and Pasha G.R ,2009, The origin and developments of influence measures in

Regression, Pakistan Journal of Statistics, 25, (3), 295-307.

ACKNOWLEDGMENTS

I would like to thank the editor and reviewers for helpful comments and discussion to improve the article.

Appendix: Hill Data

.							
	0	0		0	20	0	0
	0	0		0	30	0	0
	0	0		0	50	0	0
	0	0		0	00	0	0
	0	0		0	00	0	0
	0	0		0	80	0	0
	0	0		0	40	0	0
	0	0			20	0	0
	0	0		0	50	0	0
	0	0		0	00	0	0
	0	0		0	00	0	0
	0	0		0	10	0	0
	0	0		0	80	0	0
	0	0		0	40		0
	0	0		0	80		0

Source: Hill (1977).