Hacettepe Journal of Mathematics and Statistics
Volume 41 (1) (2012), 147 – 155

OUTLIER DETECTION BY REGRESSION DIAGNOSTICS BASED ON ROBUST PARAMETER ESTIMATES

Semra Türkan*†, Meral Candan Çetin* and Öniz Toktamış*

Received 06 : 01 : 2011 : Accepted 23 : 09 : 2011

Abstract

In this article, robust versions of some of the frequently used diagnostics are considered to identify outliers instead of the diagnostics based on the least square method. These diagnostics are Cook's distance, the Welsch-Kuh distance and the Hadi measure. A simulation study is performed to compare the performance of the classical diagnostics with the proposed diagnostics based on robust M estimation to identify outliers.

Keywords: Diagnostics, Linear regression, Robust estimate, Outliers.

2000 AMS Classification: 62 J 20, 62 J 05.

1. Introduction

The usual multiple regression models can be defined as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{X}$ is an $(n \times p)$ full rank matrix of known constants, $\boldsymbol{y}$ an $(n \times 1)$ vector of observable responses, $\boldsymbol{\beta}$ a $(p \times 1)$ vector of unknown parameters and $\boldsymbol{\varepsilon}$ an $(n \times 1)$ vector of random errors with $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ and $\mathrm{V}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$ where σ^2 is an unknown parameter and $\boldsymbol{I}_n$ is the identify matrix of order n. The ordinary least square (OLS) estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2}$$

and the vector of fitted values is

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y} \tag{3}$$

*Department of Statistics, Faculty of Science, Hacettepe University, 06532 Beytepe, Ankara,Turkey.
E-mail: (A. Türkan) sturkan@hacettepe.edu.tr (M. C. Çetin) meral@hacettepe.edu.tr
(Ö. Toktamış) oniz@hacettepe.edu.tr
†Corresponding Author.

where

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T. \tag{4}$$

The residual vector is denoted by

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y} \tag{5}$$

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

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

The regression coefficients, fitted values, goodness of fit statistics, etc. can be considerably affected by a single data point. Namely, not all data points in a data set have the same significance in determining estimates, test and other statistics. It is important that the data analyst should be aware of such kind of points (Ullah and Pasha [16]). Observations which individually or collectively, unduly influence the fitted regression equation as compared to other observations are generally called *outliers*. There are two kinds of outlier in regression analysis: the $\boldsymbol{X}$*-outliers*, classically known as *high leverage points*, and the $\boldsymbol{y}$*-outliers*. In the literature various diagnostics have been developed to detect outliers. Since outliers have become part of any serious statistical analysis, there are numerous articles and books on these diagnostics, for example, Cook [8], Cook and Weisberg [9], Belsley *et al.* [4], Welsch and Kuh [17], Andrews and Pregibon [1], Hoaglin and Welsch [11], Atkinson [2] and Chatterjee and Hadi [6]. Some of the major outlier diagnostics are Cook's Distance, the Welsch-Kuh Distance (DFFITS), Modified Cook's Distance, the Andrew and Pregibon (AP_i) measure and the Hadi Measure (H_i^2). However, Cook's distance is one of the most often used outlier diagnostics in classical linear models.

Robust regression 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 square estimates. There are many robust methods in the literature. The M estimator, introduced by Huber [12], is one of them. Although the M estimator is not robust with respect to leverage points, it is extensively used in analyzing data for which the contamination is mainly in the $\boldsymbol{y}$ direction (Chen [7]).

Huber's M estimator minimizes a sum of less rapidly increasing functions of the residuals instead of minimizing a sum of the squares:

$$Q(\beta) = \sum_{i=1}^{n} \rho\Big(\frac{e_i}{\sigma}\Big), \tag{7}$$

where $e_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$. For the OLS estimate, ρ is the quadratic function. If σ is known, by taking derivatives with respect to $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_r$ is also a solution of the system of p equations:

$$\sum_{i=1}^{n} \psi\Big(\frac{e_i}{\sigma}\Big)x_{ij} = 0,\ j = 1, \ldots, p, \tag{8}$$

where $\psi = \rho'$ is the derivative of ρ with respect to $\boldsymbol{\beta}$. The Newton method could be used to estimate the parameter. When convergence is not achieved due to large residuals, the Levenberg-Marquardt is utilized. When robust regression is viewed as iteratively reweighted least squares, the weight from the final iteration may be used to identify cases for further study; the cases with smaller weights receiving the most attention. Also, the residuals from a robust fit can be inspected for anomalies in much the same way as residuals from a least squares fit (Beckman and Cook [3]; Chen [8]; Riazoshams *et al.* [15]).

2. Outliers diagnostics

There are several diagnostics used to identify outliers but we have taken three of the frequently used influence diagnostics which are given below.

2.1. Cook's distance ($\mathbf{D}_i$). The Cook distance D_i (Cook [8]), measures the distance between the estimates of the regression coefficients with the i-th observation $\hat{\boldsymbol{\beta}}$ and without the i-th observation $\hat{\boldsymbol{\beta}}_{-i}$ for the metric $\frac{1}{p\hat{\sigma}^2}(\boldsymbol{X}^T\boldsymbol{X})$. So, D_i is defined as below

$$\mathrm{D}_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^T(\boldsymbol{X}^T\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \hat{\beta}_{-i})}{\hat{\sigma}^2 p}. \tag{9}$$

Cook suggests that D_i be compared to a central F distribution with p and $n-p$ degrees of freedom. This gives however exaggeratedly high cutoff values. Practically a cutoff value of $\frac{4}{n-p}$ seems more reasonable. Cook's distance can be expressed as follows in terms of the studentized residual and the classical leverage indicator

$$\mathrm{D}_i = \frac{r_i^2}{k}\frac{h_{ii}}{1-h_{ii}}, \tag{10}$$

where r_i is ith internally studentized residual and h_{ii} the ith diagonal element of the hat matrix given in (4). High D_i requires large values of both (r_i^2) and h_{ii}. Thus, for instance, the Cook distance cannot detect high leverage points standing on the hyperplane. Furthermore, like other classical diagnostic measures, it becomes unreliable in the case of multiple data. It is considered that an observation is an influential observation when D_i exceeds the cut-off point of $\frac{4}{n-p}$ (Cook [8]; Rawlings *et al.* [14]).

2.2. The Welsch-Kuh distance (DFFITS). The impact of the i-th observation on the i-th predicted value can be measured by scaling the change in prediction at x_i when the i-th observation is omitted, that is,

$$\mathrm{DFFITS}_i = \frac{|\hat{y}_i - \hat{y}_{i,\,-i}|}{\hat{\sigma}_{-i}\sqrt{h_{ii}}}$$

$$= \frac{\left|x_i'(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})\right|}{\hat{\sigma}_{-i}\sqrt{h_{ii}}} \tag{11}$$

$$= |r_i^*|\sqrt{\frac{h_{ii}}{1-h_{ii}}}, \tag{12}$$

where r_i^* is the ith externally studentized residual and the h_{ii} are the i-th diagonal elements of the matrix given in (4). Belsley *et al.* [4] recommend using $2\sqrt{\frac{p}{n}}$ as a cut-off point for DFFITS (Chatterjee and Hadi [6]).

2.3. Hadi measure ($\boldsymbol{H_i^2}$). Hadi's measure, which is a measure to detect overall potential influence, can be defined as

$$\mathrm{H}_i^2 = \frac{p}{1-h_{ii}}\frac{d_i^2}{1-d_i^2} + \frac{h_{ii}}{1-h_{ii}},\ i = 1, \ldots, n, \tag{13}$$

where $d_i^2 = \frac{e_i^2}{\boldsymbol{e}^T\boldsymbol{e}}$ is the square of the ith normalized residual and h_{ii} is the i-th diagonal element of $\boldsymbol{H}$ defined in (4). Hadi's measure is based on the simple fact that potentially influential observations are outliers as either $\boldsymbol{X}$-outliers, $\boldsymbol{y}$-outliers, or both. Hadi [10] recommends using "$\mathrm{mean}(\mathrm{H}_i^2) + c\sqrt{\mathrm{Var}(\mathrm{H}_i^2)}$" as a cut-off point for Hadi's measure, where c is an appropriately chosen constant such as 2 or 3. Alternatively, Hadi [10] recommends to use "$\mathrm{median}(\mathrm{H}_i^2) + c\mathrm{MAD}(H_i^2)$" as a cut off point where $\mathrm{MAD}(\mathrm{H}_i^2) = \mathrm{median}\left\{|H_i - \mathrm{median}(\mathrm{H}_i)|\right\}/0.674$ (Nurunnabi and Nasser [13]).

Cut-off points should be used with caution. Diagnostic methods are not designed to be formal tests of a hypothesis. 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 best be done using graphical displays such as a stem-and-leave display, index plot, or P-R plot (Hadi [10]).

3. Suggested method

As is well known, a large number of diagnostics have been proposed to detect outliers. The diagnostics which are based on the ordinary least squares estimates are not efficient and cannot detect correctly swamping and masking effects. In this paper, robust versions of diagnostics have been proposed to identify the outliers. So, the robust version of Cook's Distance, DFFITS and Hadi's measure is used to detect outliers in the data. We propose the use of the Huber-M estimator of $\boldsymbol{\beta}$ instead of $\hat{\boldsymbol{\beta}}$, 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 in (6) to obtain a robust Cook's Distance. The robust version of Cook's Distance can be expressed as

$$\text{RD}_i = \frac{(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_{r,\,-i})^T(\boldsymbol{X}^T\boldsymbol{X})(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_{r,\,-i})}{\hat{\sigma}_r^2\, p}, \tag{14}$$

where $\hat{\boldsymbol{\beta}}_r$ is the robust estimation of $\boldsymbol{\beta}$ and $\hat{\sigma}_r$ the robust scale estimation of σ. In the same way, a robust DFFITS is obtained by

$$\text{RDFFITS}_i = \frac{\left|\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_{r,\,-i})\right|}{\hat{\sigma}_{r,\,-i}\sqrt{h_{ii}}}. \tag{15}$$

In order to obtain robust version of Hadi's measure, robust normalized residuals, which are calculated after a robust fit, are used instead of normalized residuals in (13). Our robust version of Hadi's measure can also be expressed as follows:

$$\text{RH}_i^2 = \frac{p}{1-h_{ii}}\frac{d_{r_i}^2}{1-d_{r_i}^2} + \frac{h_{ii}}{1-h_{ii}},\ i = 1, \ldots, n, \tag{16}$$

where $d_{r_i}^2 = \dfrac{e_{r_i}^2}{\boldsymbol{e}_r^T\boldsymbol{e}_r}$ and e_{r_i} is the i-th robust residual.

For the comparison of the classical 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.

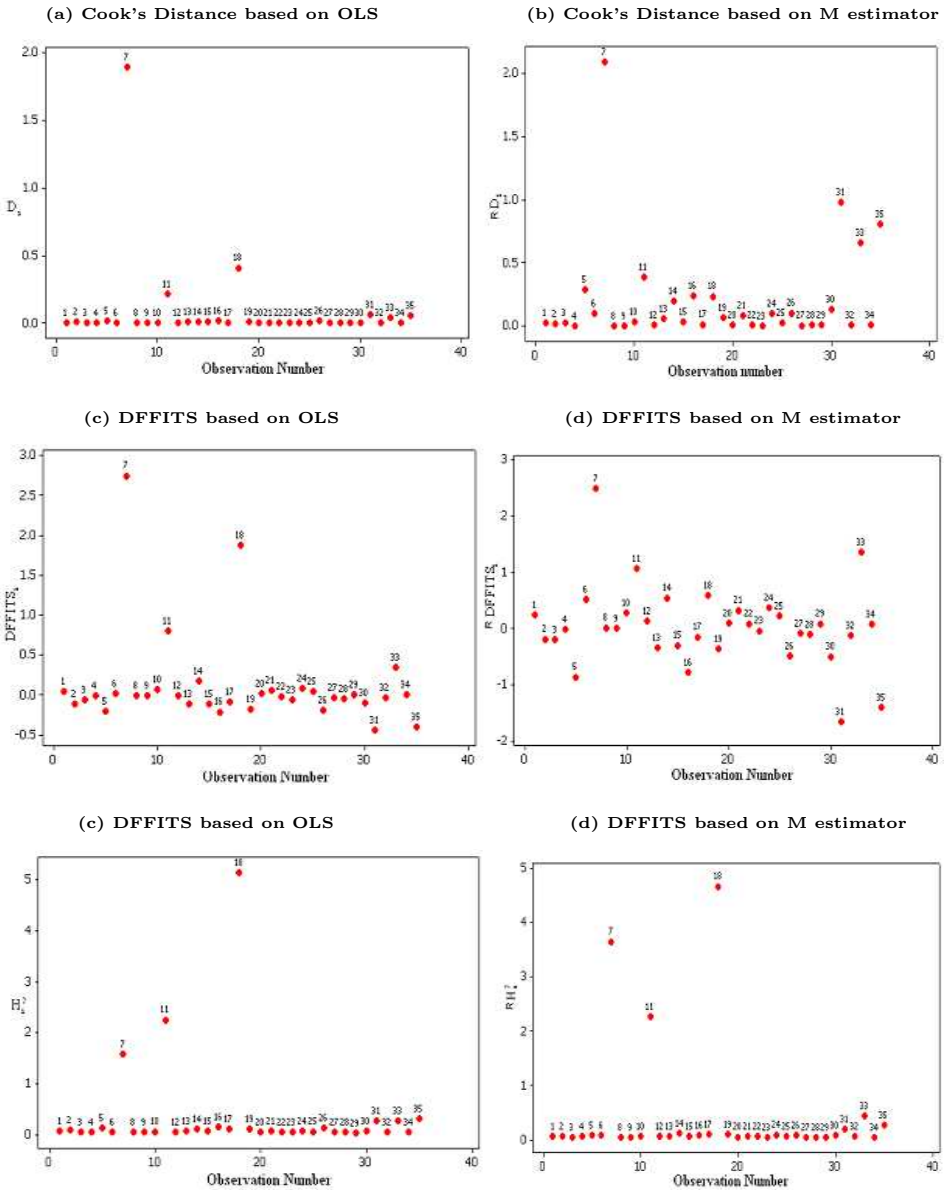
4. Numerical example

In this section, a set of real data which is referred to as the *Scottish Hill Races data* is used to see how well the diagnostic statistics with robust estimator perform for the regression model. This data has been used many times and here it is taken to compare against. The data set was collected from 35 Scottish Hill races in 1984. The data presents the relationship between the record times (in seconds) as the response variable and the distance raced (in miles) and climb (in feet) as predictors. The model associated with the data is:

$$\text{Time} = \beta_0 + \beta_1\text{Distance} + \beta_2\text{Climb} + \varepsilon. \tag{17}$$

The corresponding index plots of the diagnostic measures that are based on the OLS and Huber- M estimator are given in Figure 1.

Figure 1. Index plots of diagnostic measures that are based on the OLS and robust estimates



Examination of the index plots shows that D_i , $DFFITS_i$, H_i^2 and RH_i^2 could correctly identify observations 7, 11 and 18 as outliers, however, RD_i and $RDFFITS_i$ could correctly identify observations 7 and 33 as outliers. In addition, RD_i and $RDFFITS_i$ also identify observations 31 and 35 as outliers. In addition, a pair-wise comparison of diagnostics could be made. Accordingly, the former version of Cook's distance failed to detect the observations 31, 33 and 35 as outliers. However, Cook's distance with plugged-in robust estimates can do so, but the observation 18 is no longer detected as an outlier.

In terms of the two DFFITS statistics, the new statistic detected the observations 7, 31, 33 and 35 as outliers, but the former version of this statistic missed out the last 3 observations. The two versions of Hadi's measure lead to the same conclusion on outlier detection. For this data, Hadi [10] also identified observations 7, 18 and 33 as outliers and observation 11 as of high leverage.

The Procedure ROBUSTREG in SAS Version 9 provides code to fit robust regression and displays outliers and leverage points in robust regression estimates. Hence, the Scottish Hill Races data is analyzed by the procedure ROBUSTREG to detect outliers and leverage points in robust estimates for the data. As seen from Table 1, the results of the procedure substantially improve the results of the proposed robust diagnostics in Figure 1.

Table 1. The results of the procedure ROBUSTREG

	The ROBUSTREG Procedure Diagnostics				
Obs	Mahalanobis Distance	Robust MCD Distance	Leverage	Standardized Robust Residual	Outlier
7	3.6501	7.4596	*	12.0848	*
11	4.7416	13.8685	*	0.2658	
17	1.3621	4.8308	*	-0.1455	
18	0.9543	1.3231		13.5073	*
31	1.7783	3.6923	*	-0.2669	
33	2.2051	6.5284	*	5.2469	*
35	2.3499	7.4945	*	0.5633	

Diagnostics Summary		
Observation Type	Proportion	Cutoff
Outlier	0.0857	3.0000
Leverage	0.1714	2.7162

Table 1 shows that observations 7, 18 and 33 are outliers and that observation 11 has high leverage. This result confirms the results of the proposed diagnostics.

5. Simulation study

In this section, a simulation study is conducted to compare regression diagnostics with robust versions of them to reveal outliers. In the simulation study, data having 15 and 45 observations and 3 independent variables such as x_1, x_2, x_3 are generated from a uniform distribution. The residuals are generated from a normal distribution with mean 0 and variance $\sigma^2 = 1$ and $\sigma^2 = 10$, respectively. These variables are added to a regression model and in this model, we take the values of the coefficients as $(5, 3, \sqrt{6})$ (Çetin [5]). In order to see the effects of outliers on the results of the analysis, diagnostics based on OLS and the robust M estimator are examined in the case of one outlier and two outliers for $\boldsymbol{e}$. 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 OLS and M estimator are applied to the sets A and B of data. 500 repetitions have been made and the results are shown in Tables 2 and 3. The values in these tables show the percentage of correctly detected outliers for the diagnostics.

Table 2. Simulation results for Case A, n = 15 and n = 45

	$\sigma^2 = 1$		$\sigma^2 = 10$	
n = 15	One Outlier	Two Outliers	One Outlier	Two Outliers
D_i	100%	99%	100%	100%
$DFFITS_i$	100%	100%	100%	100%
H_i^2	99%	60%	98%	78%
RD_i	92%	87%	42%	100%
$RDFFITS_i$	100%	97%	70%	100%
RH_i^2	100%	100%	100%	100%
n = 45	One Outlier	Two Outliers	One Outlier	Two Outliers
D_i	100%	100%	100%	100%
$DFFITS_i$	100%	100%	100%	100%
H_i^2	100%	100%	100%	100%
RD_i	99%	97%	100%	97%
$RDFFITS_i$	85%	58%	86%	61%
RH_i^2	100%	100%	100%	100%

As seen from Table 2, in the case of $\sigma^2 = 1$, one outlier and $n = 15$ and 45, diagnostics based on the OLS and M estimator give similar results. In addition, in the case of $\sigma^2 = 1$, two outliers, $n = 15$, all diagnostics except H_i^2 can correctly identify outliers, however in the case of $\sigma^2 = 1$ and $\sigma^2 = 10$, two outliers and $n = 45$, all diagnostics except $RDFFITS_i$ can correctly identify outliers. In the case $\sigma^2 = 10$, two outliers, $n = 15$ and one outlier, $n = 45$, diagnostics are successful in correctly detecting outliers. However, in the case of $\sigma^2 = 10$, one outlier and $n = 15$, RD_i could correctly identify 42% of the outliers.

Table 3. Simulation results for Case B, n = 15 and n = 45

	$\sigma^2 = 1$		$\sigma^2 = 10$	
n=15	One Outlier	Two Outliers	One Outlier	Two Outliers
D_i	100%	100%	100%	100%
$DFFITS_i$	100%	100%	100%	100%
H_i^2	100%	52%	100%	61%
RD_i	100%	98%	100%	99%
$RDFFITS_i$	100%	99%	100%	100%
RH_i^2	100%	100%	100%	100%

Table 3. Continued

	$\sigma^2 = 1$		$\sigma^2 = 10$	
n=45	One Outlier	Two Outliers	One Outlier	Two Outliers
D_i	100%	100%	100%	100%
$DFFITS_i$	100%	100%	100%	100%
H_i^2	100%	50%	100%	50%
RD_i	93%	97%	93%	96%
$RDFFITS_i$	46%	55%	43%	50%
RH_i^2	100%	100%	100%	100%

As seen from Table 3, in the case $\sigma^2 = 1$ and $\sigma^2 = 10$, one outlier and $n = 15$, all diagnostics exactly identify outliers. However, in the case of $\sigma^2 = 1$ and $\sigma^2 = 10$, two outliers and $n = 15$, all diagnostics except H_i^2 can correctly identify outliers. In the case of $n = 45$, $RDFFITS_i$ fails to identify outliers. In addition, in the case $\sigma^2 = 1$ and $\sigma^2 = 10$, two outliers and $n = 45$, H_i^2 fails to identify outliers.

6. Conclusion

In this paper, robust versions of Cook's distance (D_i), Welsch-Kuh distance (DFFITS) and the Hadi measure (H_i^2) are proposed to detect outliers. As seen from results of the Scottish Hill Races data, the diagnostics based on OLS and M-estimators detect the same observations as outliers. In addition to these observations, as stated in the study of Hadi [10], diagnostics based on the M estimator detect other observations as outliers. Robust versions of the diagnostics support the study of Hadi [10]. Hadi [10], detect all outliers in the Scottish Hill Races data in two steps, while robust versions of the diagnostics detect all outliers in the data in one step. Also, the data is analyzed using the ROBUSTREG procedure in SAS Version 9 and the results of this procedure support the results of the proposed diagnostics based on robust estimates. The results of the simulation study agree well with the real data.

References

[1] Andrews, D. F. and Pregibon, D. *Finding the outliers that matter*, J. Roy. Statist. Soc. Ser. B **40**, 85–93, 1978.

[2] Atkinson, A. C. *Two graphical displays for outlying and influential observations in regression*, Biometrika **68**, 13–20, 1981.

[3] Beckman, R. J. and Cook, R. D. *Outlier. . . s*, Technometrics **25** (2), 119–149, 1983.

[4] Belsley, D. A., Kuh, E. and Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (Wiley, New York, 1980).

[5] Cetin, M. *Robust model selection criteria for robust Liu estimator*, European Journal of Operational Research **199**, 21–24, 2009.

[6] Chatterjee, S.and Hadi, A. S. *Sensitivity Analysis in Linear Regression* (Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1988).

[7] Chen, C. *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, 2002).

[8] Cook, R. D. *Detection of influential observations in linear regression*, Technometrics **19**, 15–18, 1977.

[9] Cook, R. D. and Weisberg, S. *Residuals and Influence in Regression* (Chapman & Hall, New York, 1982).

[10] Hadi, A.S. *A new measure of overall potential influence in linear regression*, Computational Statistics and Data Analysis **14**, 1–27, 1992.
[11] Hoaglin, D. C, and Welsch, R. E. *The Hat matrix in regression and ANOVA*, The Amer. Statist. **32**, 17–22, 1978.
[12] Huber, P. J. *Robust Statistics* (John Wiley & Sons, New York, 1981).
[13] Nurunnabi, A. A. M. and Nasser, M. *Outlier detection by regression diagnostics in large data* (International Conference on Future Computer and Communication, 2009), 246-250.
[14] Rawlings, J. O., Pantula, S. G. and Dickey, D. A. *Applied Regression Analysis: A Research Tool* (Springer, New York, 1998).
[15] Riazoshams, A. H., Habshah, B. and Adam, C. M. B. *On the outlier detection in nonlinear regression*, Engineering and Technology **60**, 264–270, 2009.
[16] Ullah, M. A. and Pasha, G. R. *The origin and developments of influence measures in regression*, Pakistan Journal of Statistics **25** (3), 295–307, 2009.
[17] Welsch, R. E. and Kuh, E. *Linear Regression Diagnostics* (Technical report 923-77, Solan School of Management, Massachusetts Institute of Technology, 1977).