Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 2, pp. 429～444

Simultaneous Identification of Multiple Outliers and High Leverage Points in Linear Regression

A.H.M. Rahmatullah Imon[1] · M. Masoom Ali[2]

Abstract

The identification of unusual observations such as outliers and high leverage points has drawn a great deal of attention for many years. Most of these identifications techniques are based on case deletion that focuses more on the outliers than the high leverage points. But residuals together with leverage values may cause masking and swamping for which a good number of unusual observations remain undetected in the presence of multiple outliers and multiple high leverage points. In this paper we propose a new procedure to identify outliers and high leverage points simultaneously. We suggest an additive form of the residuals and the leverages that gives almost an equal focus on outliers and leverages. We analyzed several well-referred data set and discover few outliers and high leverage points that were undetected by the existing diagnostic techniques.

Keywords : Added residual and leverage, Generalized potentials, Generalized Studentized residuals, High leverage points, Outliers

1. Introduction

The ordinary least squares (OLS) technique is the most popular and commonly used regression techniques despite all of its shortcomings. Under usual assumptions OLS estimators have some nice and desirable properties, but the violation of these assumptions has drastic consequences on the presence of one or more unusual observations in the data set. In a regression problem

[1] First Author : Department of Statistics, Univ. of Rajshahi, Rajshahi-6205,Bangladesh. Currently, Visiting Professor, Department of Mathematical Sciences, Ball State Univ., Muncie, IN 47306-0490 USA.
E-mail : imon_ru@yahoo.com

[2] Corresponding Author : Department of Mathematical Sciences, Ball State University, Muncie, IN 47306-0490 USA.
E-mail : mali@bsu.edu

observationsthat fail to match with the fitted model are termed as outliers and highly unusual observations among the explanatory variables are known as high leverage points. The identification of outliers is really necessary because the presence of a single outlier may break down the entire OLS analysis. On the other hand the identification of high leverage points is really necessary because the presence of these points makes the identification of outliers very difficult.

A good number of detection methods are now available in the literature for the identification of outliers and high leverage points. In section 2, we briefly discuss different issues and techniques for the detection of outliers and high leverage points. Most of the diagnostic methods deal the issues of outliers and high leverage points separately. We propose a new technique in section 3 based on the added form of residuals and leverages so that observations unusual in either direction would be easily detected. We present few examples in section 4 to show how this newly proposed technique works to identify outliers and high leverage points simultaneously.

2. Outliers and High Leverage Points

We write a standard regression model as

$$Y = X\beta + \in \qquad (2.\ 1)$$

where Y is an $n\times1$ vector of response or dependent variables, X is an $n\times k\ (n>k)$ matrix of explanatory variables including one constant, β is a $k\times1$ vector of unknown finite parameters and $\in$ is an $n\times1$ vector of random disturbances. We can re-express the general linear model (2.1) by

$$y_i = x_i^T\beta + \in_i, \qquad i = 1, 2, ..., n$$

where yi is the i-th observed response and xi is a $k\times1$ vector of explanatory variables. The OLS estimates of the regression parameters are $\hat{\beta} = \left(X^TX\right)^{-1}X^TY$. Thus the i-th residual is given by

$$\hat{\in}_i = y_i - x_i^T\hat{\beta}, \qquad i = 1, 2, ..., n \qquad (2.2)$$

In matrix notation (2.2) becomes

$$\hat{\in} = Y - X\ \hat{\beta}$$

which can also be expressed as

$$\hat{\in} = (I - W)\in$$

where $W = X\left(X^T X\right)^{-1} X^T$ which is generally known as weight matrix or leverage matrix.

In regression analysis it is sometimes very important to know whether any set of X-values are exerting too much influence on the fitting of the model. A set of influential X-values is known as a high leverage point. The diagonal elements of W, denoted as w_{ii} and defined by

$$w_{ii} = x_i^T \left(X^T X\right)^{-1} x_i, \qquad i = 1, 2, \ldots, n \tag{2.3}$$

are called the leverage values. Observations corresponding to excessively large wii values are termed as high leverage points.

Much work has been done on the identification of high leverage points in linear regression. Most of them are based on the examination of wii values as defined in (2.3) [see Imon (2002)]. Well known Mahalanobis distances are also suggested to use as measures of leverages in the literature, but Mahalanobis distance for each of the points has a one-to-one relationship with w_{ii} [see Rousseeuw and Leroy (1987)]. Hadi (1992) introduced a single case deleted measure of leverages known as potentials. The i-th potential is defined as

$$p_{ii} = x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i \tag{2.4}$$

where $X_{(i)}$ is the data matrix X with the i-th row deleted. Observations corresponding to excessively large potential values are considered as high leverage points.

It is reported by many authors [see Rousseeuw and Leroy (1987), Imon (2002)] that the presence of multiple high leverage points may cause masking and swamping because of which outliers and/or high leverage points become undetected and some innocent observations may be detected as outliers and/or high leverage points. That is why group deleted leverages are suggested in the literature [see Imon (2002)] when the presence of multiple high leverage points makes the detection of genuine outliers and high leverage points cumbersome.

Multivariate outlier detection methods [see Pea and Prieto (2001), Marona and Zamar (2002)] can be used to identify multiple high leverage points in linear regression.

Excellent reviews of different aspects of outliers in linear regression are available in Rousseeuw and Leroy (1987), Barnett and Lewis (1994), Ryan (1997) and Sengupta and Jammalamadaka (2003). In a regression problem, observations possessing excessively large residuals are simply known as outliers. A variety of identification procedures for single outliers have been suggested in the statistical literature to detect outliers in a regression problem. Most of them are based on the modification of the OLS residuals and they seem to be successful for the identification of a single outlier [see Hawkins, Bradu, and Kass (1984)]. But because of masking and swamping effects, the detection of outliers has become extremely difficult when a group of outliers are present in the data. Most multiple outlier identification methods attempt to separate the data into a 'clean'subset without outliers and a complementary subset that contains all the potential outliers [see Barnett and Lewis (1994)]. Some indirect approaches are available in the statistical literature to identify outliers by robust techniques through a robust regression estimate. Among them least median of squares proposed by Rousseeuw (1984), reweighted least squares and least trimmed squares (LTS) proposed by Rousseeuw and Leroy (1987) have become popular with the statisticians. Some approaches combining diagnostic and robust approaches together [see Hadi and Simonoff (1993), Atkinson (1994), Davies et al. (2004)] are also available in the literature for the identification of multiple outliers in linear regression.

3. Measures as an Additive Form of Residuals and Leverages

Most of the existing diagnostic statistics focus on the issue of the identification of outliers and high leverage points separately. But it is now evident [see Pea and Yohai (1995)] that the presence of one type of observation may cause problem to detect the other type of observations. Some diagnostic measures like Cook's distance or DFFITS focus on both of these two types of cases. But the main problem with this kind of statistics as indicated by Hadi (1992) is that they are expressed as a multiplicative form of residuals and leverages. The values of these statistics could be misleadingly small if either residuals or leverages used in these statistics are small and consequently, they could fail to identify potential outliers or leverage points. Hadi (1992) suggested using a new diagnostic measure, which is an additive form of the residuals and of the leverages and hence assumes large values for observations that have either large residuals or large leverage values, or both. He proposed to use

$$H_i^2 = \frac{k}{(1-w_{ii})} \frac{\hat{\in}_i^{\,2}}{\sum_{i=1}^{n} \hat{\in}_i^{\,2} - \hat{\in}_i^{\,2}} + \frac{w_{ii}}{1-w_{ii}} \tag{3.1}$$

that is likely to focus equally tothe residuals and the leverage values in measuring influence of any observation. As a cut-off point for H_i^2, he suggested using

$$\text{Median } (H_i^2) + \text{c MAD } (H_i^2)$$

where MAD (H_i^2) = Median $\{|H_i^2 -$ Median $(H_i^2)|\}/$ 0.6745 and cis an appropriately chosen constant such as 2 or 3.

Hadi's idea of looking for an additive function of the residuals and of the leverages are intuitively appealing, but since H_i^2 is basically a single case diagnostic measure, because of masking and/or swamping, it may fail to identify multiple outliers and multiple high leverage points. However, it is interesting to note from (3.1) that H_i^2 can be expressed as

$$H_i^2 = \frac{k}{(n-k-1)} \frac{\hat{\in}_i^{\,2}}{\hat{\sigma}_{(i)}^{\,2}\left(1-w_{ii}\right)} + \frac{w_{ii}}{1-w_{ii}}, \qquad i = 1, 2, ..., n \tag{3.2}$$

Here the residual part is multiplied by k / (n k 1) to give an equal weight to residuals and leverages, as the value of the second part of H_i^2 is k / (n k 1) when wii is replaced by its average value k / n.

Let us denote a set of cases 'remaining' in the analysis by R and a set of cases 'deleted' by D. Hence R contains (n-d) cases after d < (n-k) cases in Dare deleted. Without loss of generality, assume that these observations are the last of d rows of X and Y. When a group of observations D is omitted, we define

$$w_{ii}^{(-D)} = x_i^T (X_R^T X_R)^{-1} x_i, \qquad i = 1,2,...,n \tag{3.3}$$

where X_R is the data matrix X after the deletion of a group of observation indexed by D. It should be noted that $w_{ii}^{(-D)}$ is the i-th diagonal element of the matrix $X(X_R^T X_R)^{-1} X^T$. We observe that (3.3) is an extension of the potentials defined in (2.4). For the identification of multiple high leverage points, Imon (2002) suggested using the values

$$p_{ii}^{*} = \frac{w_{ii}^{(-D)}}{1-w_{ii}^{(-D)}} \qquad \text{for} \quad \text{iR}$$
$$= w_{ii}^{(-D)} \qquad \text{for} \quad \text{iD} \qquad (3.4)$$

where D is a set points containing suspect high leverage points.

We also define a group-deleted version of residuals

$$t_i^{*} = \frac{y_i - x_i^T \hat{\beta}_{(R)}}{\hat{\sigma}_R \sqrt{1-w_{ii}^{(-D)}}} \qquad \text{for} \quad i \in R$$
$$= \frac{y_i - x_i^T \hat{\beta}_{(R)}}{\hat{\sigma}_R \sqrt{1+w_{ii}^{(-D)}}} \qquad \text{for} \quad i \notin R \qquad (3.5)$$

This type of residuals has been in use [see Hadi and Simonoff (1993), Atkinson (1994), Imon (2005a, 2005b)] for the identification of multiple outliers. Clearly, when $i \in R$, t_i^{*} is the i-th internally Studentized residual and when $i \notin R$, t_i^{*} is the i-th externally Studentized residual based on the subset R. In this paper we shall call t_i^{*} as they are defined in (3.5) generalized Studentized (GS) residuals.

One possible group-deleted version of H_i^2 could be

$$H_i^{(-D)2} = \frac{k}{(n-k-d)} t_i^{*2} + p_{ii}^{*}, \qquad i = 1, 2, ..., n \qquad (3.6)$$

as suggested by Imon (2005a). But the problem with this statistic (3.6) is that both of its residual and leverage parts are unbounded and that is why multiplying the square of t_i^{*} by k/(n-k-d) does not guarantee that the residuals and leverages will get an equal importance. For this reason we suggest using

$$\text{ARLi} = k_1 | t_i^{*} | + k_2 \, p_{ii}^{*}, \qquad i = 1, 2, ..., n \qquad (3.7)$$

where $k_1 = \sum_{i=1}^{n} | t_i^{*} |$ and $k_2 = \sum_{i=1}^{n} p_{ii}^{*}$. Like Hadi's H_i^2 statistic, it may be very complicated to obtain any theoretical distribution for ARLi, but that may not make any problem to obtain suitable cut-off points for them. We propose to use a confidence bound type cut-off value for ARLi given by

$$\text{ARLi} > \text{Median (ARLi)+ c MAD (ARLi)} \qquad (3.8)$$

which we believe should be fairly robust even in many complicated situations. As usual c is any arbitrary chosen value between 2 and 3. The choice of c is analogous to the idea of considering 2σ or 3σ distance of a statistic, which is used quite often [Rousseeuw and Leroy (1987)] in the study of outlier detection.

Here the choice of the deletion set D is really important. For Hadi's H_i^2 we do not have similar choice, each and every observation in turn is deleted to compute H_i^2. But the entire set of ARLi values depend on the selection of D. At first we would like to obtain a data set containing all suspect outliers and high leverage points. We would like to give an equal emphasis on both of these unusual cases. At the initial stage we would alsolike to mark outliers and high leverage points separately. For the identification of suspect outliers we would use the robust reweighted least squares (RLS) residuals proposed by Rousseeuw and Leroy (1987).To compute the RLS residuals, a regression line is fitted without the observations identified as outliers by the LMS technique. The residuals corresponding to outliers are computed externally here and that is why they show their real pictures. The entire set of RLS residuals is computed using the program PROGRESS developed by Rousseeuw and Leroy (1987). For the identification of suspect high leverage points we would consider a method suggested by Imon (2002). For a k variable regression, the j-th point of any regressor X_i can be treated as suspect high leverage points when it falls outside the interval

$$\text{Median } (X_i) \pm \text{ c MAD } (X_i), \qquad i = 1,2, ...,k \qquad j = 1,2, ..., n. \qquad (3.9)$$

Not necessarily, the same data points (if any) of each regressor will satisfy the rule (3.9). We would like to include all data points as suspect high leverage points if they satisfy rule (3.9) for any X_i.

All observations marked as suspect outliers and high leverage points will now consist the initial deletion set say, D_0. If there is no such observation, we can say at the very beginning that there exist no outliers and/or high leverage points in the data. However we would not allow D_0 to take more than 50% observations, otherwise it will be very difficult to distinguish usual observations from unusual. After the selection of the initial deletion set D_0 we would proceed with testing for multiple outliers and high leverage points using the ARL statistic defined in (3.7). We would fit the regression model by the least squares techniques after deleting the cases belonging to D_0 set. Then the ARL values are computed for the entire data set. If all of the members of the deletion set D_0 individually satisfy the rule (3.8) this set is considered as our final deletion set and all members of this set are declared as unusual.

We anticipate that sometimes the rules for the selection of initial deletion set may be very sensitive and that is why, it is not unlikely that some of the innocent observations are swamped in as outliers in either of the spaces or both. So it may be necessary for checking in swamping before the declaration of any of the observations as outliers or high leverage points. Sometimes we may observe that one or more of the members of D_0 do not satisfy the rule (3.8). So these members are not potential outliers or high leverage points. At this stage, we can put back all the observations that fail to satisfy rule (3.8) together into the estimation subset. But we prefer to put them back sequentially; observations possessing the lowest ARLi values will be the first member coming back to the estimation subset. We will continue this process till all of the members of the revised deletion set individually satisfy the rule (3.8). This set is considered as our final deletion set and all members of this set are declared as outliers or high leverage points or both.

4. Examples

In this section we consider a few well-known data sets, which are frequently referred to in the study of the identification of outliers. We would like to compare our newly proposed diagnostic methods with the other existing ones to identify outliers and high leverage points (if any) using these data sets. For these examples we have considered three different values of c as 2.0, 2.5 and 3.0. We observe that results obtained from different choices of c do not differ significantly. For the purpose of illustration, we consider c = 2.0 in the numerical examples.

4.1 Hawkins-Bradu-Kass (1984) data

Hawkins, Bradu and Kass (1984) constructed an artificial three-predictor data set containing 75 observations with 10 high leverage outliers (cases 1-10), 4 high leverage points (cases 11-14) and 61 low leverage inliers (cases 15-75). Most of the single case deletion identification methods fail to identify the outliers though some of them point out high leverage points as outliers [see Rousseeuw and Leroy (1987)]. On the other hand robust detection techniques like LMS and RLS and the method proposed by Hadi and Simonoff (1993) identify outliers correctly, but do not focus on the high leverage points.

Table 1(a) presents few single case diagnostics including Hadi's H_i^2 statistic. The cut-off values for each of the statistics are presented inside the parantheses. Sometimes Studentized residuals are used to detect outliers and cases having values greater than 2.5 are suspects. For this data only observations 11-14 have significant t values. Cook's CDiidentifies observation 14 as outlier, whereas the DFFITS mark cases 11-14 as outlier. The leverage values w_{ii}'s are not big and if any one considered Velleman and Welsch (1981)'s 'thrice-the-mean' rule only observation 14 appears to be unusual. Hadi's potential values identify cases 11-14 as high leverage points. The added residual leverage measure like H_i^2 identifies case 11, 12 and 14 as unusual.

Table 1(a). Single case diagnostics for the first 14 observations of Hawkins et al. data

Index	t_i (2.50)	w_{ii} (0.16)	p_{ii} (0.11)	CDi (1.0)	DFFITSi (1.0)	H_i^2 (0.72)
1	1.55	0.063	0.067	0.04	0.41	0.21
2	1.83	0.060	0.064	0.05	0.47	0.26
3	1.40	0.086	0.094	0.05	0.43	0.21
4	1.19	0.086	0.088	0.03	0.35	0.17
5	1.41	0.081	0.079	0.04	0.40	0.19
6	1.59	0.073	0.082	0.05	0.46	0.23
7	2.08	0.068	0.073	0.08	5.57	0.33
8	1.76	0.063	0.067	0.05	0.46	0.25
9	1.26	0.080	0.087	0.03	0.37	0.18
10	1.41	0.087	0.095	0.05	0.44	0.21
11	-3.66	0.094	0.104	0.35	-1.30	1.01
12	-4.50	0.144	0.169	0.85	-2.17	1.68
13	-2.88	0.109	0.122	0.25	-1.07	0.64
14	-2.56	0.564	1.292	2.11	-3.03	1.68

Now we apply the proposed algorithm to this data. At the initial stage rule (3.9) defined in the previous section identifies 14 observations (cases 1-14) as high leverage points. The robust RLS marks observations 1-10 as outliers. Thus our initial deletion set D_0 contains 14 observations (cases 1-14) as prime suspect. These 14 observations are now omitted to compute the ARLi values for the entire data set. When these observations are omitted from the OLS fit, we observe from Table 1(b) that all of their corresponding ARLi values are significantly higher than the cut-off value and no other observations possesses high ARLi value. Thus we finally declare observations 1-14 as jointly unusual.

It is also interesting to observe from the generalized weights and generalized Studentized residuals presented in Table 1(b) that observations 1-10 have significantly high residuals and observations 1-14 have significantly high leverages. As the ARLi values give an equal focus on leverages and residuals, we observe that cases 1-14 have significantly high ARLi values and they are considered as either outliers or high leverage points.

Table 1(b). Added residual and leverage diagnostics for Hawkins et al. (1984) data

Sl no.	p_{ii}^* (0.152)	t_i^* (2.50)	ARLi (0.024)	Sl no.	p_{ii}^* (0.152)	t_i^* (2.50)	ARLi (0.024)	Sl no.	p_{ii}^* (0.152)	t_i^* (2.50)	ARLi (0.024)
1	14.46	5.35	0.104	26	0.069	-1.13	0.011	51	0.058	1.08	0.010
2	15.22	5.44	0.107	27	0.090	-1.14	0.011	52	0.098	-1.07	0.010
3	16.97	5.32	0.113	28	0.036	0.72	0.007	53	0.139	2.02	0.019
4	18.02	4.89	0.113	29	0.039	0.56	0.005	54	0.083	1.18	0.011
5	17.38	5.15	0.113	30	0.100	-0.02	0.001	55	0.050	0.08	0.001
6	15.61	5.31	0.108	31	0.070	-0.17	0.002	56	0.066	0.10	0.001
7	15.70	5.65	0.111	32	0.073	-0.76	0.007	57	0.050	1.21	0.011
8	14.82	5.59	0.107	33	0.046	-1.01	0.009	58	0.072	-0.22	0.002
9	17.03	5.04	0.111	34	0.094	-1.19	0.011	59	0.046	-0.17	0.002
10	15.97	5.31	0.109	35	0.082	0.75	0.007	60	0.099	-1.07	0.010
11	22.39	0.95	0.094	36	0.040	-1.45	0.013	61	0.111	-0.06	0.001
12	24.03	0.90	0.100	37	0.092	-0.65	0.006	62	0.090	1.08	0.010
13	22.73	1.20	0.097	38	0.056	1.52	0.014	63	0.076	-0.68	0.006
14	28.16	0.87	0.115	39	0.075	-1.28	0.012	64	0.080	-0.71	0.007
15	0.091	-0.76	0.007	40	0.037	-0.75	0.007	65	0.060	1.41	0.013
16	0.104	0.79	0.008	41	0.094	-0.02	0.001	66	0.055	-1.54	0.014
17	0.086	-0.34	0.003	42	0.076	-0.79	0.007	67	0.022	-1.23	0.011
18	0.027	0.11	0.001	43	0.104	1.24	0.012	68	0.099	1.35	0.013
19	0.046	0.36	0.003	44	0.092	-0.86	0.008	69	0.072	0.20	0.002
20	0.096	0.76	0.007	45	0.080	-1.00	0.009	70	0.050	1.72	0.016
21	0.036	1.70	0.016	46	0.081	-0.34	0.003	71	0.034	0.52	0.005
22	0.072	0.83	0.008	47	0.115	-1.90	0.018	72	0.032	-0.16	0.002
23	0.041	-1.57	0.014	48	0.082	0.40	0.004	73	0.048	1.10	0.010
24	0.047	1.19	0.011	49	0.062	1.65	0.015	74	0.058	-1.37	0.013
25	0.090	-0.57	0.006	50	0.056	-0.44	0.004	75	0.096	0.86	0.009

4.2 Stack loss data

Here we consider the stack loss data presented by Brownlee (1965) that have been extensively analyzed in the statistical literature. This three-predictor data set (Air flow, Cooling water inlet temperature and Acid concentration) contains 21 observations with 4 outliers (observations 1, 3, 4, and 21). This data set has possibly 4 high leverage points (cases 1, 2, 3 and 21) [see Atkinson (1985)] but those are undetected by very recent diagnostic techniques though some times observation no. 17 is mistakenly declared [see Imon (2003)] as a high leverage point.

When the OLS technique is employed to the data we observe from Table 2(a) that most of the traditional diagnostic methods fail to focus on the unusual cases. Cook's distance does not identify any of the observations as outlier. Studentized residuals, DFFITS and Hadi's H_i^2 can identify only one (case 21) of the four outliers. Leverage values indicate that there is no high leverage point in the data set though potential values identifies case 17 as a high leverage point.

Table 2(a). Single case diagnostics for stack loss data

Index	t_i (2.50)	w_{ii} (0.571)	p_{ii} (0.497)	CDi (1.0)	DFFITSi (1.0)	H_i^2 (1.22)
1	1.19	0.302	0.432	0.154	0.795	0.79
2	-0.72	0.318	0.466	0.060	-0.481	0.59
3	1.55	0.175	0.212	0.126	0.744	0.87
4	1.89	0.129	0.147	0.131	0.788	1.20
5	-0.54	0.052	0.055	0.004	-0.125	0.13
6	-0.97	0.077	0.084	0.020	-0.279	0.32
7	-0.83	0.219	0.281	0.049	-0.438	0.45
8	-0.48	0.219	0.281	0.017	-0.251	0.34
9	-1.05	0.140	0.163	0.045	-0.423	0.44
10	0.44	0.200	0.250	0.012	0.213	0.30
11	0.88	0.155	0.183	0.036	0.376	0.38
12	0.97	0.217	0.277	0.065	0.509	0.51
13	-0.48	0.158	0.187	0.011	-0.203	0.24
14	-0.02	0.206	0.259	0.000	-0.009	0.26
15	0.81	0.190	0.235	0.039	0.388	0.40
16	0.30	0.131	0.151	0.003	0.113	0.17
17	-0.61	0.412	0.701	0.065	-0.502	0.79
18	-0.15	0.161	0.191	0.001	-0.065	0.20
19	-0.20	0.175	0.211	0.002	-0.091	0.22
20	0.45	0.080	0.087	0.004	0.131	0.14
21	-2.64	0.285	0.398	0.692	-2.100	3.17

Table 2(b). Added residual and leverage diagnostics for stack loss data

Index	p_{ii}^* (0.78)	t_i^* (2.5)	ARLi (0.108)	Index	p_{ii}^* (0.78)	t_i^* (2.5)	ARLi (0.108)
1	2.22	4.25	0.309	12	0.41	0.43	0.046
2	2.30	1.81	0.239	13	0.30	-1.64	0.075
3	1.30	4.79	0.253	14	0.31	-1.15	0.061
4	0.28	6.77	0.234	15	0.61	0.37	0.060
5	0.21	-0.02	0.018	16	0.35	-0.42	0.041
6	0.26	-0.47	0.036	17	0.76	-0.26	0.068
7	0.39	-0.05	0.033	18	0.26	-0.18	0.026
8	0.39	0.72	0.054	19	0.32	0.17	0.031
9	0.22	-0.68	0.039	20	0.10	1.77	0.063
10	0.42	0.66	0.054	21	0.91	-4.58	0.216
11	0.25	0.80	0.045				

When we apply newly proposed diagnostic methods to stack loss data we initiallyfind 5 suspect cases. The robust RLS technique identifies cases 1, 3, 4 and 21 as outliers. The rule based on generalized potentials mark observations 1, 2, 3 and 21 as high leverage points. Thus our initial deletion set contains 5 observations, cases 1, 2, 3, 4 and 21. When the model is fitted without these five points we observe from Table 2(b) that all of these observations have significant ARL values that confirms our suspicion that these observations are unusual. Observations 1, 2, 3 and 21 are being detected for possessing high leverages and observations 1, 3, 4 and 21 are being detected for possessing large residuals as indicated by the p_{ii}^* and t_i^* values.

4.3 Delivery time data

Now we consider deliverytime data taken from Montgomery and Peck (1992). In this two predictor data we want to explain the time required to service a vending machine (Y) by means of the number of products stocked (X_1) and the distance walked by the route driver (X_2). This data set contains 25 observations.

Most of the detection techniques identify case 9 as an outlier and high leverage point. We observe from Table 3(a) that Studentized residuals, Cook's distance and DFFITS corresponding to this observation are very high which indicates that this is an outlier. This observation also possesses large leverage and potential values. Consequently its H_i^2 value is significantly higher than the cut-off value and thus observation number 9 in all sense is declared as high leverage outlier. However, we find another observation (case 22), as a high leverage point. Although its DIFFITS value is significant, its corresponding residual and Cook's distance are not significantly high and hence this observation is not identified as unusual by Hadi's H_i^2. It is also interesting to note that the robust RLS identifies only observation 9 as an outlier.

Table 3(a). Single case diagnostics for delivery time data

Index	t_i (2.50)	w_{ii} (0.36)	p_{ii} (0.581)	CDi (1.0)	DFFITSi (1.0)	H_i^2 (1.29)
1	-1.63	0.102	0.113	0.100	-0.571	0.41
2	0.36	0.071	0.076	0.003	0.099	0.02
3	-0.02	0.099	0.110	0.000	-0.005	0.00
4	1.58	0.085	0.093	0.078	0.501	0.38
5	-0.14	0.075	0.081	0.001	-0.039	0.00
6	-0.09	0.043	0.045	0.000	-0.019	0.00
7	0.27	0.082	0.089	0.002	0.079	0.01
8	0.37	0.064	0.068	0.003	0.094	0.02
9	3.21	0.498	0.993	3.493	4.330	2.65
10	0.81	0.196	0.244	0.054	0.399	0.09
11	0.72	0.086	0.094	0.016	0.218	0.07
12	-0.19	0.114	0.128	0.002	-0.068	0.01
13	0.33	0.061	0.065	0.002	0.081	0.01
14	0.34	0.078	0.085	0.003	0.097	0.02
15	0.21	0.041	0.043	0.001	0.043	0.01
16	-0.22	0.166	0.199	0.003	-0.097	0.01
17	0.14	0.059	0.063	0.000	0.034	0.00
18	1.11	0.096	0.107	0.044	0.365	0.18
19	0.58	0.096	0.107	0.012	0.186	0.05
20	-1.87	0.102	0.113	0.132	-0.672	0.57
21	-0.88	0.165	0.198	0.051	-0.389	0.11
22	-1.45	0.392	0.643	0.451	-1.195	0.32
23	-1.44	0.041	0.043	0.030	-0.308	0.31
24	-1.50	0.121	0.137	0.102	-0.571	0.34
25	-0.07	0.067	0.071	0.000	-0.018	0.00

Table 3(b). Added residual and leverage diagnostics for delivery time data

Index	p_{ii}^{*} (0.36)	t_i^{*} (2.5)	ARLi (0.124)	Index	p_{ii}^{*} (0.36)	t_i^{*} (2.5)	ARLi (0.124)
1	0.129	-1.88	0.085	14	0.09	0.78	0.040
2	0.115	-0.16	0.020	15	0.07	1.08	0.049
3	0.157	-0.54	0.039	16	0.31	1.34	0.086
4	0.119	1.63	0.074	17	0.07	-0.02	0.009
5	0.089	-0.54	0.030	18	0.12	1.54	0.071
6	0.048	0.05	0.007	19	0.16	-0.19	0.026
7	0.161	-0.71	0.045	20	0.96	-0.08	0.127
8	0.077	0.53	0.029	21	0.32	-1.00	0.075
9	2.372	5.22	0.474	22	1.71	1.54	0.260
10	0.290	1.64	0.095	23	0.07	-1.48	0.063
11	0.406	2.86	0.154	24	0.17	-1.36	0.070
12	0.217	0.23	0.034	25	0.09	-0.78	0.040
13	0.083	0.02	0.011				

We now apply our newly proposed technique to identify unusual observation. As we have already mentioned that the RLS identifies case 9 as outlier, it is a suspect case in our study. But the generalized potential rule identifies cases 9, 11, 20 and 22 as high leverage point. Thus our initial deletion set contains these 4 observations. When these observations are omitted we observe from Table 3(b) that observations 9, 11, 20 and 22 have significant (based on c = 2.0 and c= 2.5) ARL values. Observation 9 possesses large residual and high leverage and observations 20 and 22 are detected for possessing high leverages. However, the observation 20, that possesses very low residual value, may be undetected for the choice of c= 3.0. But we identify observation 11 as an outlier and high leverage point, which, so far as we know, became undetected by the statisticians. This is an interesting example to show that in the presence of multiple outliers and high leverage points unusual cases may get masked in such a way that even robust detection techniques may fail to identify them.

It is worth mentioning that for the examples we consider in our study, we have the same initial and final deletion set, but these two deletion sets need not be equal. In many practical situations these two sets could differ because of the inherent feature of the data or the choice of the constant term cin the ARL statistic.

5. Conclusions

In this paper we propose a criterion for the simultaneous identification of outliers and high leverage points in linear regression. We develop a diagnostic procedure, ARL, based on the added form of residuals and leverages giving an equal focus to both of them. We present a few examples that clearly indicate how this method can be effective to identify outliers and high leverage points when all existing diagnostic methods fail to do so.

Acknowledgements

Authors would like to thank the referees for carefully reading the manuscript and gratefully acknowledge their valuable comments and helpful suggestions.

References

1. Atkinson, A.C. (1994). Fast Very Robust Methods for the Detection of Multiple Outliers. *J. Amer. Statist. Assoc.*, 89, 1329-1339.
2. Atkinson, A.C. (1985). *Plots, Transformation, and Regression*, Clarendon Press, Oxford.
3. Barnett, V. and Lewis, T. (1992). *Outliers in Statistical Data*, 3rd ed., Wiley, New York.
4. Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, London.
5. Davies, P., Imon, A.H.M.R. and Ali, M. Masoom (2004). A Conditional Expectation Method for Improved Residual Estimation and Outlier Identification in Linear Regression. *Int. Jour. Statist. Sci.*, 3 (Special volume in honor of Professor M.S. Haq), 191-208.
6. Hadi, A.S. (1992). A New Measure of Overall Potential Influence in Linear Regression. *Comput. Statist. Data Anal.*, 14, 1-27.
7. Hadi, A.S. and Simonoff, J.S. (1993). Procedures for the Identification of Outliers in Linear Models. *J. Amer. Statist. Assoc.*, 88, 1264-1272.
8. Hawkins, D.M., Bradu, D. and Kass, G.V. (1984). Location of Several Outliers in Multiple Regression Data Using Elemental Sets. *Technometrics*, 26, 197-208.
9. Imon, A.H.M.R. (2002). Identifying Multiple High Leverage Points in Linear Regression. *J. Statist. Stud.*, (Special volume in honor of Professor Mir Masoom Ali), 207-218.
10. Imon, A.H.M.R. (2003). Residuals from Deletion in Added Variable Plots. *J. App. Stat.*, 30, 841-855.

11. Imon, A. H. M. R. (2005a). A Stepwise Procedure for the identification of Multiple Outliers and High Leverage Points in Linear Regression, (To appear) *Pak. J. Statist.*, 21, 71-86.
12. Imon, A. H. M. R. (2005b). Identifying Multiple Influential Observations in Linear Regression, (To appear) *J. App. Stat.*
13. Maronna, R.A. and Zamar, R.H. (2002). Robust Estimates of Location and Dispersion for High-Dimensional Datasets. *Technometrics*, 44, 307-317.
14. Montgomery, D.C. and Peck, E. (1992). *An Introduction to Linear Regression Analysis*, 2nd ed., Wiley, New York.
15. Pea, D. and Prieto, F.J. (2001). Multivariate Outlier Detection and Robust Covariance Estimation. *Technometrics*, 43, 286-310.
16. Pea, D. and Yohai, V.J. (1995). The Detection of Influential Subsets in Linear Regression by Using an Influence Matrix. *J. Roy. Stat. Soc. Ser-B*, 57, 18-44.
17. Rousseeuw, P.J. (1984). Least Median of Squares Regression. *J. Amer. Statist. Assoc.*, 79, 871-80.
18. Rousseeuw, P.J. and Leroy, A. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.
19. Sengupta, D. and Jammalamadaka, S. (2003). *Linear Models: An Integrated Approach*, World Scientific, New Jersey.
20. Ryan, T.P. (1997). *Modern Regression Methods*, Wiley, New York.

[received date : Feb. 2005, accepted date Apr. 2005]