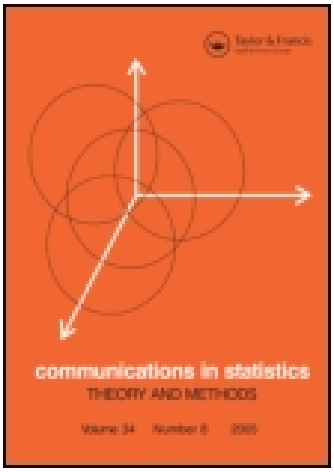

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

A Diagnostic Measure for Influential Observations in Linear Regression

A. A. M. Nurunnabi [a] , A. H. M. Rahmatullah Imon [b] & M. Nasser [c]
[a] Department of Business Administration , Uttara University , Dhaka, Bangladesh
[b] Department of Mathematical Sciences , Ball State University , Muncie, Indiana, USA
[c] Department of Statistics , University of Rajshahi , Rajshahi, Bangladesh
Published online: 08 Feb 2011.

To cite this article: A. A. M. Nurunnabi , A. H. M. Rahmatullah Imon & M. Nasser (2011) A Diagnostic Measure for Influential Observations in Linear Regression, Communications in Statistics - Theory and Methods, 40:7, 1169-1183, DOI: 10.1080/03610920903564727

To link to this article: http://dx.doi.org/10.1080/03610920903564727

Communications in Statistics—Theory and Methods, 40: 1169–1183, 2011

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920903564727

A Diagnostic Measure for Influential Observations in Linear Regression

A. A. M. NURUNNABI[1], A. H. M. RAHMATULLAH IMON[2], AND M. NASSER[3]

[1]Department of Business Administration, Uttara University, Dhaka, Bangladesh
[2]Department of Mathematical Sciences, Ball State University, Muncie, Indiana, USA
[3]Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh

In linear regression it is a common practice of measuring influence of an observation is to delete the case from the analysis and to investigate the change in the parameters or in the vector of forecasts resulting from this deletion. Pena (2005) introduced a new idea to measure the influence of an observation based on how this observation is being influenced by the rest of the data. In this article we propose a new influence measure extending the idea of Pena to group deletion for identifying multiple influential observations in linear regression. We investigate the usefulness of the proposed technique by two well-referred data sets, an artificial large data with high-dimension and heterogeneous sample points and by reporting a Monte Carlo simulation experiment.

Keywords Group deletion; High leverage point; Masking; Outlier; Swamping.

Mathematics Subject Classification 62J05; 62J20.

1. Introduction

The identification of influential observations has drawn a great deal of attention in the last few decades. Since the seminal article of Cook (1977), most of the ideas of finding influential observations in regression are developed on the basis of "deleting the observations one after another and measuring their effects on various aspects of the analyses." Popular diagnostic techniques like Cook's distance, and DFFITS (Belsley et al., 1980) are based on single-case deletion, but it is now evident that the single-case deletion techniques may fail to detect multiple influential observations mainly because of masking and/or swamping problems.

Consider the customary linear regression model

$$Y = X\beta + \varepsilon, \tag{1.1}$$

where Y is an $n \times 1$ vector of response, X is an $n \times k$ ($n > k$; $k = p + 1$) full rank matrix of explanatory variables including one constant predictor, β is a $k \times 1$ vector of

Received August 7, 2008; Accepted December 14, 2009
Address correspondence to A. A. M. Nurunnabi, School of Business, Uttara University, Uttara-6, Dhaka-1230, Bangladesh; E-mail: pyal1471@yahoo.com

unknown finite parameters, and ε is an $n \times 1$ vector of i.i.d. random disturbances each of which follows $N(0, \sigma^2)$. We can reexpress the above model as

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{1.2}$$

When the ordinary least squares (OLS) method is employed to estimate the regression parameters, we obtain

$$\widehat{\beta} = (X^T X)^{-1} X^T Y. \tag{1.3}$$

The ith residual is given by $\widehat{\varepsilon}_i = y_i - x_i^T \widehat{\beta}$ while the general form of residuals is $\widehat{\varepsilon} = (I - H)Y$. The matrix $H = X\,(X^T\,X)^{-1}\,X^T$ is generally known as the leverage or weight matrix. The diagonal elements of H denoted by h_{ii} and defined as

$$h_{ii} = x_i^T (X^T X)^{-1} x_i, \quad i = 1, 2, \ldots, n \tag{1.4}$$

are called leverages. Observations corresponding to excessively large h_{ii} values are termed as high leverage points.

Pregibon (1981) pointed out residuals, standardized residuals, and leverage values are useful for detecting extreme points, but not for assessing their impact on various aspects of the fit. To assess the impact of extreme points on fit we draw our attention to the influential cases. According to Belsley et al. (1980), "An influential observation is one which either individual or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates ... than is the case for most of the other observations". Draper and John (1981) mentioned that the observation with the largest residual is not the most influential; however, deletion of observation, which has a small residual, may has a marked effect on the parameter estimation. Welsch (1982) pointed out that neither the leverage nor the Studentized residual alone will usually be sufficient to identify influential case. A large body of literature is now available (see Atkinson and Riani, 2000; Montgomery et al., 2001; Chatterjee and Hadi, 2006) for the identification of influential observations. The general idea of influence analysis is to introduce small perturbations in the sample and see how these perturbations affect the model. The most common approach is to delete one data point and see how this deletion affects the vector of parameters or the vector of forecasts. Cook's distance (Cook, 1977) and DFFITS (Belsley et al., 1980) are two most popular single-case deletion techniques to the practitioners. The ith Cook's distance is introduced as

$$CD_i = \frac{\left(\widehat{\beta}^{(-i)} - \widehat{\beta}\right)^T (X^T X) \left(\widehat{\beta}^{(-i)} - \widehat{\beta}\right)}{k\widehat{\sigma}^2}, \tag{1.5}$$

where $\widehat{\beta}^{(-i)}$ is the estimated parameter of β with the ith observation deleted. The ith difference in fits (DFFITS) is defined as

$$\text{DFFITS}_i = \frac{\widehat{y}_i - \widehat{y}_i^{(-i)}}{\widehat{\sigma}_{(i)}\sqrt{h_{ii}}}, \quad i = 1, 2, \ldots, n, \tag{1.6}$$

where $\widehat{y}_i^{(-i)}$ and $\widehat{\sigma}_{(i)}$ are, respectively, the ith fitted response and the estimated standard error with the ith observation deleted. Welsch (1982) considered DFFITS as a better choice because, it is more informative about σ^2 than CD_i and it will calculate simultaneous effect on both the parameter estimates and the estimate of variance.

Atkinson (1986) pointed out that, in presence of masking single deletion diagnostic methods often fail to reveal outliers and influential observations. We anticipate that the single-case deletion measures discussed above ineffective in the identification of multiple influential observations because of masking and/or swamping effects. A lot of literature has been written based on group deletion measures (see Imon, 2005) to remedy the problem of masking and/or swamping. Cook and Weisberg (1982) suggested generalized Cook's distance in this regard. But this type of measure is defined only for the deletion of an influential group and consequently cannot be applied for the identification of the influential cases for the entire data set. Imon (2005) proposed generalized DFFITS (GDFFITS) for the entire data set in a linear model and defined as

$$\text{GDFFITS}_i = \begin{cases} \dfrac{\widehat{y}_i^{(-D)} - \widehat{y}_i^{(-D-i)}}{\widehat{\sigma}^{(-D-i)}\sqrt{h_{ii}^{(-D)}}} & \text{for} \quad i \in R \\ \dfrac{\widehat{y}_i^{(-D+i)} - \widehat{y}_i^{(-D)}}{\widehat{\sigma}^{(-D)}\sqrt{h_{ii}^{(-D+i)}}} & \text{for} \quad i \in D \end{cases} \tag{1.7}$$

where the set of cases "remaining" in the analysis is indexed by R after the omission of suspected cases indexed by D, $h_{ii}^{(-D)} = x_i^T (X_R^T X_R)^{-1} x_i$, and $h_{ii}^{(-D+i)} = x_i^T (X_R^T X_R + x_i x_i^T) x_i = h_{ii}^{(-D)}/1 + h_{ii}^{(-D)}$.

Pena (2005) introduced a new statistic totally in a different way to measure the influence of observations. To quote him, "instead of looking at how the deletion of a point or the introduction of same perturbation affects the parameters, the forecasts, or the likelihood function, we look at how each point is influenced by the others in the sample. That is, for each sample point we measure the forecasted change when each other point in the sample is deleted". He outlined a procedure to measures how each sample point is being influenced by the rest of the data. He considered the vector

$$s_i = \left(\widehat{y}_i - \widehat{y}_i^{(-1)}, \ldots, \widehat{y}_i - \widehat{y}_i^{(-n)} \right)^T,$$

where $\widehat{y}_i - \widehat{y}_i^{(-j)}$ is the difference between the ith estimated value of y in presence of all observations and with jth observation deleted. Pena (2005) defined his statistic for the ith observation as

$$S_i = \frac{s_i^T s_i}{pV(\widehat{y}_i)}, \quad i = 1, 2, \ldots, n, \tag{1.8}$$

which also can be reexpressed as

$$S_i = \frac{1}{p\widehat{\sigma}^2 h_{ii}} \sum_{j=1}^{n} \frac{h_{ji}^2 \widehat{e}_j^2}{(1 - h_{jj})^2}, \tag{1.9}$$

where h_{ji} is the jith element of the leverage matrix H, $\widehat{y}_i - \widehat{y}_i^{(-j)} = h_{ji}\widehat{\varepsilon}_j/1 - h_{jj}$ and $V(\widehat{y}_i) = \widehat{\sigma}^2 h_{ii}$.

Observations with the values of the statistic sufficiently larger than

$$\frac{S_i - E(S_i)}{\text{std}(S_i)} \tag{1.10}$$

may be considered as influential. Since both the mean and standard deviation of S_i are affected in the presence of influential cases, Pena (2005) called an observation influential for which

$$|S_i| \geq \text{med } (\text{S}_\text{i}) + 4.5 \text{ MAD } (\text{S}_\text{i}), \tag{1.11}$$

where med (S_i) and MAD (S_i) are the median, and the median absolute deviation of the values of the statistic, respectively.

2. Proposed Measure

We observe from the Eq. (1.9) that in Pena's S_i statistic the leverage values get more importance than the conventional influence measure and that is why this statistic can be very useful for identifying high leverage outliers, which are usually considered the most difficult type of heterogeneity to detect in regression problems. But there is evidence that both residuals and leverages could break down easily in the presence of multiple influential observations (see Imon, 2005) especially when these points are high leverage outliers and a single case deletion measure like Pena (2005) may not be able to focus on the real influence of these observations. We know that group deletion helps us to reduce the maximum disturbance by deleting the suspect group of influential cases at a time (see Hadi and Simonoff, 1993). It helps to make the data more homogeneous than before. For this reason, we propose a new influence measure extending the idea of Pena (2005) to a group deletion study. Our proposed method consists of two steps. In the first step we try to identify the suspect influential cases. There is enough evidence that it is not easy to suspect/identify all influential cases at the first time because of masking and/or swamping and if any genuine influential cases are left out in the data set, the identification procedure becomes very cumbersome. So we want to make sure that all potential influential cases are flagged as suspects before the application of any diagnostic measure. At the same time we also want to make sure that no innocent observations are wrongly deleted because the deletion of such points especially when they are good leverage ones may adversely affect (see Habshah et al., 2009) the entire inferential procedure. For this reason in the second step we use a group deletion version of the S_i statistic to confirm whether our suspected cases are genuinely influential or not.

Sometimes graphical display like index plot, scatter plot, and character plot of explanatory and response variables could give us some ideas about the influential observations, but these plots are not useful for higher dimension of regressors. There are some suggestions in the literature for using robust regression techniques like the least median of squares (LMS) or least trimmed of squares (LTS) (Rousseeuw, 1984), reweighted least squares (RLS) (see Rousseeuw and Leroy, 1987), block adaptive computationally effective outlier nominator (BACON) (Billor et al.,

2000), or best omitted from the ordinary least squares (BOFOLS) (Davies et al., 2004) for finding the group of suspect influential observations. Pena and Yohai (1995) introduced a method to identify influential subsets in linear regression by analyzing the eigen vectors corresponding to the non-null eigen values of an influence matrix. Clustering based backward-stepping methods (see Simonoff, 1991) are also suggested in this regard. In our proposed method, we try to find all suspect influential cases at the first step. Any suitable graphical display and/or robust regression techniques mentioned above can be used to flag the suspected group of influential cases. In our study we have employed BACON because it gives a perfect focus on residuals and leverage components.

After finding a group of suspected cases we would like to employ an influence measure to check whether all of the suspect cases are genuine influential points or not. Here, we develop a new statistic extending the idea of Pena (2005) to a group deletion study. Let us assume that d observations among a set of n observations are suspected as influential observations. Let us denote a set of cases "remaining" in the analysis by R and a set of cases "deleted" by D. Hence, without loss of generality, assume that these observations are the last d rows of X and Y so that

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix} \qquad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix}.$$

After formation of the deletion set indexed by D, we would compute the fitted values $\widehat{Y}^{(-D)}$. Let $\widehat{\beta}^{(-D)}$ be the corresponding vector of estimated coefficients when a group of observations indexed by D is omitted. We define the vector of difference between $\widehat{y}_j^{(-D)}$ and $\widehat{y}_{j(i)}^{(-D)}$ as

$$t_{(i)}^{(-D)} = \left(\widehat{y}_1^{(-D)} - \widehat{y}_{1(i)}^{(-D)}, \ldots, \widehat{y}_n^{(-D)} - \widehat{y}_{n(i)}^{(-D)}\right)^T \tag{2.1}$$

$$= \left(t_{1(i)}^{(-D)}, \ldots, t_{n(i)}^{(-D)}\right)^T \tag{2.2}$$

and

$$t_{j(i)}^{(-D)} = \widehat{y}_j^{(-D)} - \widehat{y}_{j(i)}^{(-D)} = \frac{h_{ji}\widehat{\varepsilon}_i^{(-D)}}{1 - h_{ii}}, \quad j = 1, 2, \ldots, n, \tag{2.3}$$

where

$$h_{ji} = x_j^T (X^T X)^{-1} x_i \quad \text{and} \quad \widehat{\varepsilon}_i^{(-D)} = y_i - \widehat{y}_i^{(-D)}.$$

Finally, we introduce our new measure as squared standardized norm,

$$M_i = \frac{t_{(i)}^{(-D)^T} t_{(i)}^{(-D)}}{kV\left(\widehat{y}_i^{(-D)}\right)}, \tag{2.4}$$

where

$$V\left(\widehat{y}_i^{(-D)}\right) = s^2 h_{ii} \text{ and } s^2 = \frac{\widehat{\varepsilon}^{(-D)^T}\widehat{\varepsilon}^{(-D)}}{n - k}.$$

Using (2.1)–(2.4), we obtain

$$M_i = \frac{1}{ks^2 h_{ii}} \sum_{j=1}^{n} h_{ji}^2 \frac{\hat{\varepsilon}_i^{(-D)2}}{(1-h_{ii})^2}. \tag{2.5}$$

The statistic M_i is a generalization of the statistic S_i defined in (1.9). Pena (2005) established different properties of the statistic S_i based on some properties of the hat matrices which do not usually change (see Imon, 2002) when a group of observations are omitted from the design. Hence following the same argument given by (Pena, 2005), we consider the ith observation to be influential if it satisfies the rule

$$|M_i| \geq \text{med } (M_i) + 4.5 \text{ MAD } (M_i). \tag{2.6}$$

Likewise, the Pena's S_i statistic the leverage values get an extra importance in the M_i statistic and that is why this statistic can be very useful for identifying high leverage outliers.

3. Examples

In this section, we compare the performance of our newly proposed measure with Cook's distance, DFFITS, and Pena's (2005) statistic for the identification of influential observations in linear regression through several well-referred data sets and a high-dimensional large artificial data.

3.1. *Hertzsprung-Rusell Diagram Data*

Our first example is the well-known Hertzsprung-Rusell Diagram (HRD) data taken from Rousseeuw and Leroy (1987). This two-variable data set contains 47 observations with 5 outliers (cases 7, 11, 20, 30, and 34).

Table 1 shows different influence measures for the HRD data. We observe from this table that the Cook's distance fails to identify any of the influential cases. DFFITS can successfully identify 3 cases (20, 30, and 34) as influential but masks one (case 7) and swamps one (case 14). Pena's statistic (S_i) can identify 5 cases correctly but swamps 5 more observations (cases 14, 17, 19, 29, and 35). Now we apply our newly proposed method to this data. When the BACON is employed to this data, 6 observations (cases 7, 11, 14, 20, 30, and 34) are flagged as suspect influential observations. We keep computing the statistic M_i for the entire data set until all suspect influential cases individually satisfy the rule (2.6) and finally identify six observations (cases 7, 9, 11, 20, 30, and 34) as influential. It is interesting to note that our method puts back the suspect case 14 to the estimation set but identifies a new one (case 9) as influential that got masked before.

Figure 1(a) displays a scatter plot of this data with the LS and the robust LMS lines in it. It is clear from this plot that the case 14 is closer to LMS line than LS line, which supports our finding that the case 14 is less influential than case 9.

3.2. *Hawkins-Bradu-Kass (1984) Data*

Hawkins et al. (1984) presented an artificial data set with three regressors containing 75 observations with 14 unusual observations among them first ten cases

Table 1
Measures of influences for Hertzsprung-Rusell Diagram data

Index	CD (1.00)	DFFITS |0.412|	S_i (0.68)	M_i (0.23)
1	0.002	0.065	0.463	0.0569
2	0.044	0.300	0.486	0.0431
3	0.000	−0.027	0.627	0.0653
4	0.044	0.300	0.486	0.0431
5	0.001	0.045	0.554	0.1154
6	0.012	0.152	0.437	0.0469
7	0.045	−0.299	**1.039**	**0.7934**
8	0.009	0.131	0.493	0.0095
9	0.010	0.144	0.627	**0.4091**
10	0.001	0.035	0.463	0.0298
11	0.067	0.365	**1.044**	**6.5866**
12	0.010	0.140	0.435	0.0697
13	0.011	0.147	0.442	0.0245
14	0.090	**−0.439**	**0.979**	0.0321
15	0.020	−0.203	0.572	0.0401
16	0.006	−0.109	0.437	0.0584
17	0.046	−0.314	**0.686**	0.0814
18	0.025	−0.226	0.437	0.2327
19	0.028	−0.241	**0.686**	0.0199
20	0.136	**0.523**	**1.044**	**7.1984**
21	0.014	−0.170	0.572	0.0165
22	0.022	−0.214	0.572	0.0503
23	0.012	−0.155	0.437	0.1143
24	0.000	−0.027	0.446	0.0425
25	0.000	0.010	0.455	0.0090
26	0.004	−0.086	0.437	0.0375
27	0.005	−0.095	0.572	0.0016
28	0.000	−0.022	0.455	0.0005
29	0.017	−0.184	**0.705**	0.0000
30	0.234	**0.691**	**1.044**	**8.0440**
31	0.012	−0.153	0.455	0.0706
32	0.002	0.067	0.486	0.0334
33	0.003	0.078	0.435	0.0080
34	0.413	**0.935**	**1.044**	**8.8467**
35	0.019	−0.195	**0.686**	0.0020
36	0.043	0.296	0.528	0.0006
37	0.002	0.060	0.467	0.0163
38	0.003	0.078	0.435	0.0080
39	0.004	0.086	0.467	0.0062
40	0.015	0.175	0.435	0.1130
41	0.005	−0.098	0.455	0.0212
42	0.000	0.031	0.435	0.0000
43	0.008	0.129	0.451	0.0052

(*Continued*)

Table 1
Continued

Index	CD (1.00)	DFFITS |0.412|	S_i (0.68)	M_i (0.23)
44	0.006	0.113	0.435	0.0262
45	0.024	0.220	0.479	0.0108
46	0.000	0.008	0.435	0.0030
47	0.009	−0.132	0.437	0.0840

(1–10) are high leverage outliers and next four cases (11–14) are high leverage points. Most of the single case deletion techniques fail to detect these influential observations. Some of them identify four high leverage points wrongly as outliers. On the other hand, robust regression techniques like LMS and RLS identify outliers correctly, but they do not focus on the high leverage points (see Rousseeuw and Leroy, 1987).

We compute different influence measures for this data and results are presented in Table 2. We observe from this table that Cook's distance identifies only one

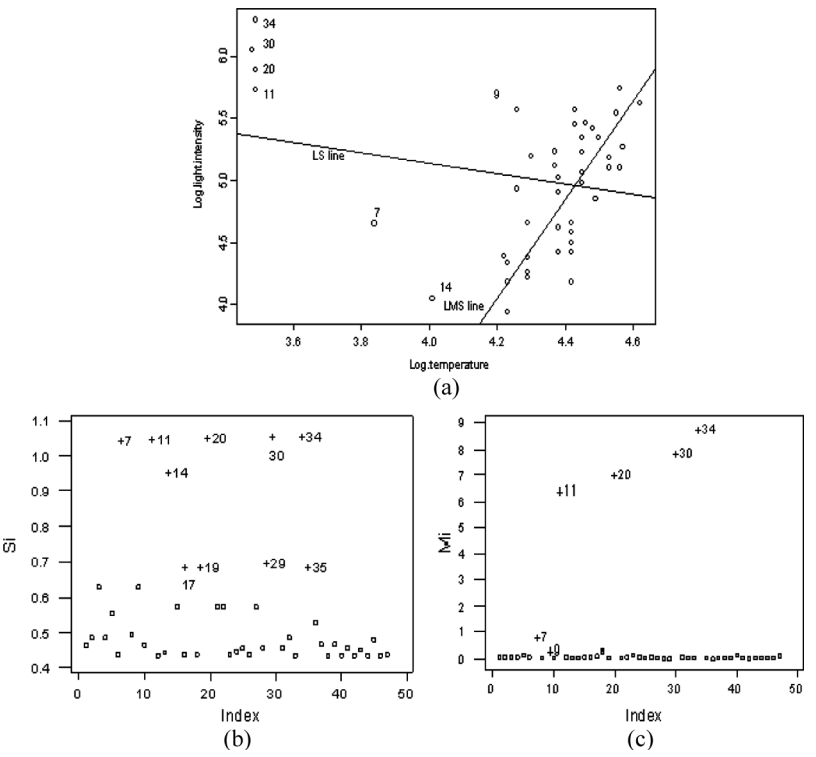

Figure 1. Plots for Hertzsprung-Rusell Diagram data: (a) scatter plot of Log light intensity versus log temperature with LS and LMS lines; (b) index plot of Pena's S_i; and (c) index plot of proposed measure, M_i.

Table 2
Measures of influences for Hawkins-Bradu-Kass data

Index	CD (1.00)	DFFITS |0.462|	S_i (2.286)	M_i (0.027)
1	0.040	0.406	1.6004	**1.8060**
2	0.053	**0.470**	1.8294	**1.9454**
3	0.046	0.430	1.6034	**2.1760**
4	0.031	0.352	1.6260	**1.9243**
5	0.039	0.399	1.7113	**2.0261**
6	0.052	0.459	1.4338	**1.9643**
7	0.079	**0.575**	1.6522	**2.1943**
8	0.052	**0.464**	1.6748	**2.0148**
9	0.034	0.372	1.5926	**1.9373**
10	0.048	0.439	1.5000	**2.0530**
11	0.348	**−1.300**	1.7053	**0.0914**
12	0.851	**−2.168**	1.6328	**0.0994**
13	0.254	**−1.065**	1.9725	**0.1533**
14	**2.114**	**−3.030**	**2.3380**	**0.4167**
15	0.001	−0.074	0.1486	0.0021
16	0.003	0.114	0.2407	0.0024
17	0.001	0.059	0.3105	0.0004
18	0.000	−0.027	0.1740	0.0000
19	0.001	0.053	0.5737	0.0005
20	0.000	0.034	0.3163	0.0021
21	0.001	0.053	0.0830	0.0102
22	0.002	0.093	0.5231	0.0025
23	0.000	−0.033	0.4077	0.0087
24	0.002	0.099	0.6894	0.0051
25	0.000	−0.020	0.4368	0.0011
26	0.000	−0.039	0.4395	0.0046
27	0.004	−0.130	0.1132	0.0046
28	0.000	−0.017	0.6837	0.0019
29	0.000	0.024	0.1010	0.0011
30	0.004	−0.127	1.0067	0.0000
31	0.000	−0.031	0.1057	0.0001
32	0.001	0.049	0.7879	0.0021
33	0.000	−0.008	0.8214	0.0037
34	0.001	−0.055	0.3766	0.0049
35	0.000	−0.034	0.1425	0.0020
36	0.002	−0.081	0.0664	0.0075
37	0.000	−0.026	0.2053	0.0016
38	0.002	0.082	0.1346	0.0080
39	0.003	−0.109	0.7738	0.0058
40	0.000	−0.002	0.2092	0.0021
41	0.003	−0.118	0.3741	0.0000
42	0.004	−0.120	0.3463	0.0023
43	0.010	0.200	0.8728	0.0056

(*Continued*)

Table 2
Continued

Index	CD (1.00)	DFFITS |0.462|	S_i (2.286)	M_i (0.027)
44	0.007	−0.168	1.3170	0.0026
45	0.001	−0.059	0.4717	0.0035
46	0.004	−0.127	1.2661	0.0004
47	0.008	−0.182	0.3769	0.0127
48	0.002	−0.081	0.4592	0.0006
49	0.001	0.065	0.1148	0.0098
50	0.000	−0.039	0.0683	0.0007
51	0.002	0.077	0.1369	0.0042
52	0.006	−0.148	0.2823	0.0043
53	0.000	−0.016	0.6842	0.0144
54	0.006	0.159	1.1868	0.0050
55	0.001	0.061	0.5249	0.0000
56	0.001	0.069	0.9839	0.0000
57	0.000	0.042	0.1922	0.0052
58	0.000	0.025	0.7025	0.0002
59	0.001	−0.045	0.1954	0.0001
60	0.006	−0.158	0.2783	0.0042
61	0.000	−0.002	0.2602	0.0000
62	0.001	0.045	0.3167	0.0040
63	0.001	0.056	1.0629	0.0016
64	0.001	−0.065	0.1032	0.0018
65	0.000	−0.023	1.0755	0.0070
66	0.000	−0.023	0.2288	0.0085
67	0.000	−0.030	0.1875	0.0055
68	0.001	0.054	0.3846	0.0064
69	0.000	0.015	0.1470	0.0001
70	0.000	0.035	0.0888	0.0103
71	0.000	0.001	0.0839	0.0010
72	0.000	0.011	0.1358	0.0001
73	0.000	0.042	0.0836	0.0045
74	0.000	−0.040	0.1710	0.0069
75	0.000	−0.041	0.5724	0.0028

(case 14) as influential observation. DFFITS identifies seven observations (cases 2, 7, 8, 11, 12, 13, and 14) correctly but fails to detect the rest of the seven. Figure 2(a,b) shows that both the single case deletion methods are failed to proper identification of influential observations. It is interesting to note that Pena's measure identifies only case 14 as influential; see Table 2 and Fig. 2(c). When we employ BACON to this data set, it flags the first 14 cases as suspects. We compute M_i for the entire data set after the omission of the suspected cases and observe that M_i values corresponding to the first 14 observations individually satisfy the rule (2.6) and hence can be declared as influential observations. Figure 2(d) clearly supports in favor of *Mi*.

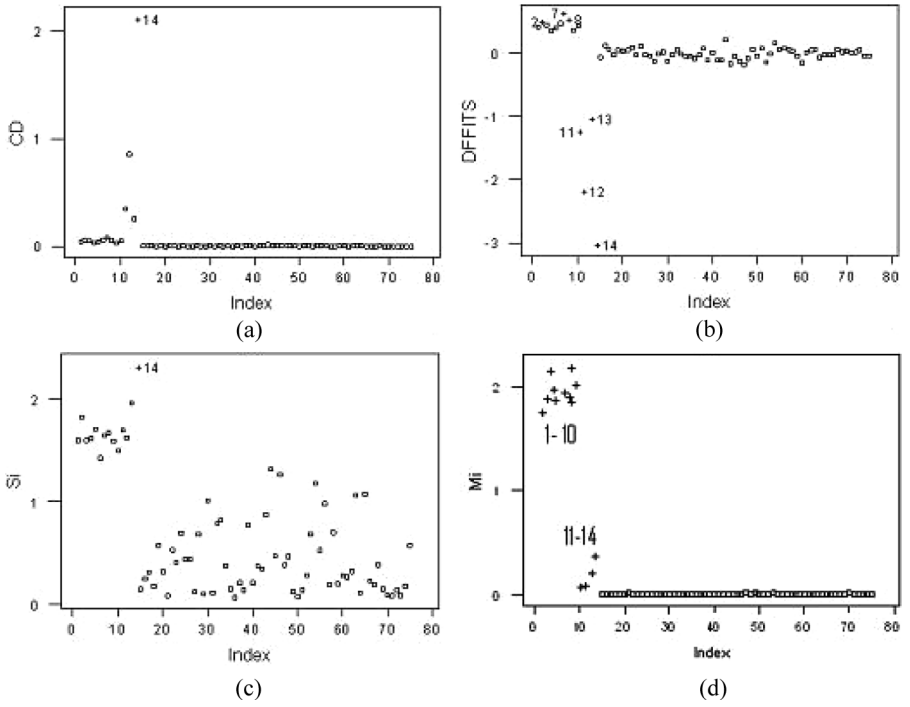

Figure 2. Index plots for Hawkins-Bradu-Kass data: (a) Cooks distance; (b) DFFITS; (c) Pena's Si; and (d) proposed measure, Mi.

3.3. *Artificial High Dimensional Large Data Containing Heterogeneous Points*

Here, we present an artificial data set that is generated in a similar fashion described by Pena (2005). The data set shows the presence of heterogeneity in the sample points and is a mixture of continuous and categorical variables. We generate 500 observations from the model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{20} X_{20} + \beta_{21} Z + \varepsilon, \tag{3.1}$$

where X's have 20 dimensions and they are independent random drawings from uniform distributions. For the categorical explanatory variable Z, the first 400 observations are set at $z = 0$ and the last 100 observations are set at $z = 1$. To get heterogeneous sample points we generate observations for each of the X variables corresponding to $z = 0$ from Uniform (0, 10) distribution while the other X variables are generated independently from Uniform (9, 10) distributions. For the null model we generate errors from Normal (0, 1). The parameter values have been chosen as $\beta_0 = \beta_1 = \cdots = \beta_{20} = 1$ and $\beta_{21} = -100$. We suspect the last 100 cases for influential and construct the deletion set D based on them. When we apply our proposed algorithm it perfectly identifies 100 observations (cases 401–500) as influential.

Figure 3 gives a variety of graphical displays of Pena's S_i and our proposed M_i for this artificial data. The residual versus fitted value scatter plot as given in Fig. 3(a)

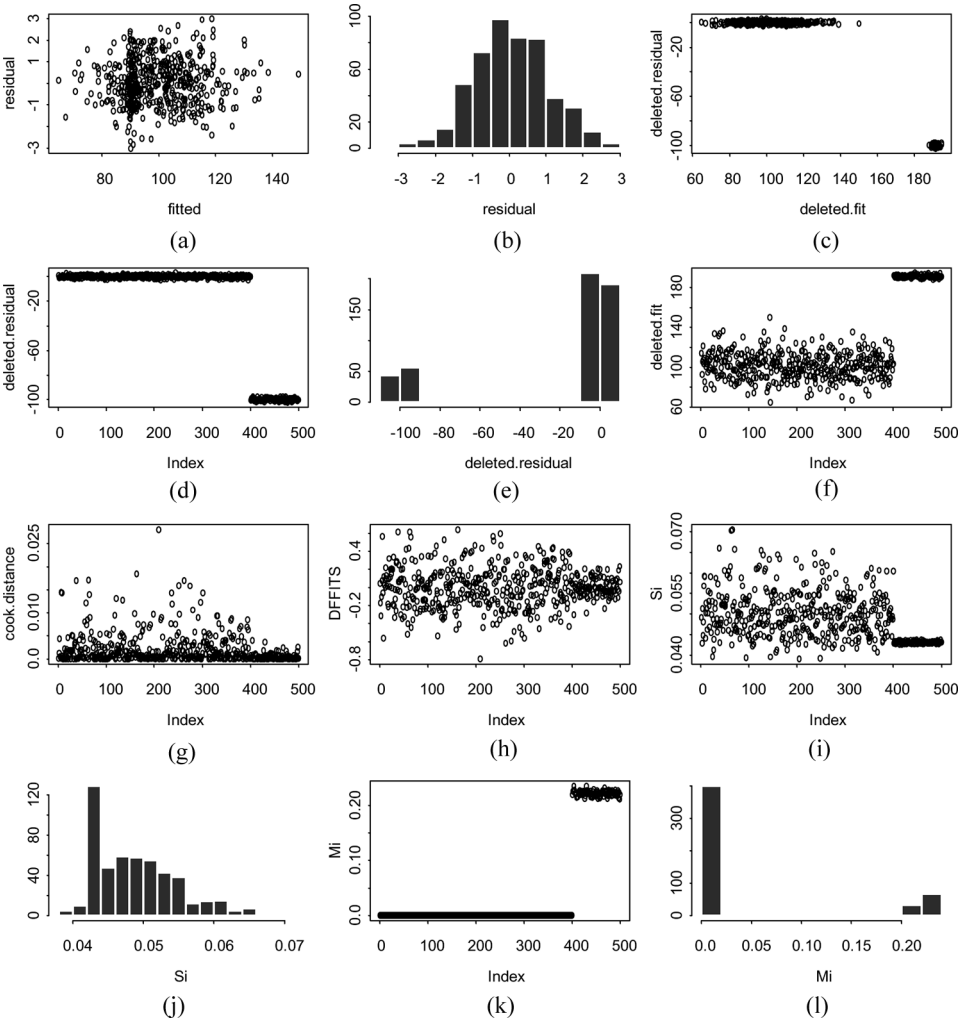

Figure 3. Influence analysis of the artificial large data with high dimensions and heterogeneous sample points: (a) residuals versus fitted plot; (b) histogram of the residuals; (c) deleted residuals vs. deleted fitted plot; (d) index plot of deleted residuals; (e) histogram of the deleted residuals; (f) index plot of deleted fits; (g) index plot of Cook's distance; (h) index plot of DFFITS; (i) index plot of S_i; (j) histogram of S_i; (k) index plot of M_i; and (l) histogram of M_i.

shows no indication of heterogeneity among the observations. But similar plots based on group deletion (see Figs. 3(c), (d), (e), and (f)) show a clear indication of the presence of unusual observations. We consider Cook's distance, DFFITS, and Pena's S_i to identify the influential cases but their index plots (see Figs. 3(g), (h), and (i)) show they totally fail to identify the influential cases. Index plot of our proposed M_i (see Fig. 3(k)) shows that they can successfully identify all influential cases. The histogram of M_i (see Fig. 3(l)) clearly shows the presence of heterogeneity which is not clearly visible from the histogram of Pena's S_i as shown in Fig. 3(j).

Table 3
Simulation results

Percentages of influential cases	Sample size	Diagnostics	Correct identification (in percentages)	
			$p=1$	$p=5$
10%	$n=50$	S_i	74.2	71.9
		M_i	100.0	100.0
	$n=100$	S_i	5.0	4.1
		M_i	100.0	100.0
20%	$n=50$	S_i	49.6	47.7
		M_i	100.0	100.0
	$n=100$	S_i	3.8	3.1
		M_i	100.0	100.0
30%	$n=50$	S_i	11.8	10.4
		M_i	100.0	100.0
	$n=100$	S_i	0.8	0.6
		M_i	100.0	100.0
40%	$n=50$	S_i	2.6	1.9
		M_i	100.0	100.0
	$n=100$	S_i	0.1	0.0
		M_i	100.0	100.0

4. Simulation Results

In this section, we report a Monte Carlo simulation study that is designed to compare the performance of our newly proposed technique with the Pena statistic (S_i) for measuring influence of observations. We have considered two different designs, a simple linear regression model and a multiple linear regression model with five regressors. We consider two different sample sizes, $n=50$ and 100 with four different levels of influential cases (i.e., $\gamma=10\%$, 20%, 30%, and 40%) and our results are based on 10,000 simulations. At first we have considered a two-variable regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{4.1}$$

We generate the first $100(1-\gamma)\%$ of X from Uniform (1, 4) and the corresponding Y values are computed from (4.1), where $\varepsilon_i \sim N$ (0, 0.2) and we choose $\beta_0=2$ and $\beta_1=1$. The remaining $100\gamma\%$ of X are generated from N (7, 0.5) where the corresponding Y values are generated from N (2, 0.5). In our simulation experiment, we have also considered the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{4.2}$$

The first $100(1-\gamma)\%$ of observations of each X are generated from Uniform (1, 4) and the corresponding Y values are computed from (4.2), where $\varepsilon_i \sim N$ (0, 0.2) and we

choose $\beta_0 = 2$ and $\beta_j = 1$ for j = 1, 2, . . . , 5. The remaining $100\gamma\%$ observations for each X are generated from $N(7, 0.5)$ where the corresponding Y values are generated from $N(2, 0.5)$.

Table 3 reports the correct identification rate (i.e., total number of influential observations identified/total number of influential observations) of Pena's S_i and our proposed M_i for each design. We observe from the results presented in Table 3 that Pena's S_i performs very poor to identify the influential cases most of the times. It performs slightly better only when the sample size is small and the proportion of influential cases is low. But its performance is not satisfactory at all. Our proposed measure performs very effectively in the identification of multiple influential observations. It can successfully identify all influential cases irrespective of sample sizes, number of regressors, and proportion of influential cases.

5. Conclusions

In this article, we introduce a new type of group deletion measure for the identification of multiple influential observations in linear regression. Analyses by a number of well-referred data sets and simulation support the merit of our proposed method while the commonly used methods do not show satisfactory performance. Moreover, the proposed technique is quite satisfactory in situations like clusters of high-leverage points in large data sets with high dimensions that are not easy to handle by the existing influence measures.

Acknowledgments

The authors gratefully thank the Editor and anonymous reviewer for their helpful comments and suggestions that substantially improve this present version of the article.

References

Atkinson, A. C. (1986). Masking unmasked. *Biometrika* 73:533–541.

Atkinson, A. C., Riani, M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer.

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

Billor, N., Hadi, A. S., Velleman, F. (2000). BACON: Blocked adaptive computationally efficient outlier nominator. *Computat. Statist. Data Anal.* 34:279–298.

Chatterjee, S., Hadi, A. S. (2006). *Regression Analysis by Examples*. 4th ed. New York: Wiley.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* 19:15–18.

Cook, R. D., Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

Davies, P., Imon, A. H. M. R., Ali, M. M. (2004). A conditional expectation method for improved residual estimation and outlier identification in linear regression. *Int. J. Statist. Sci.* (Special issue in honour of Professor M. S. Huq). 191–208.

Draper, N. R., John, J. A. (1981). Influential observations and outliers in regression. *Technometrics* 23:21–26.

Habshah, M., Norazan, R., Imon, A. H. M. R. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *J. Appl. Statist.* 36:507–520.

Hadi, A. S., Simonoff, J. S. (1993). Procedure for the identification of outliers in linear models. *J. Amer. Statist. Assoc.* 88:1264–1272.

Hawkins, D. M., Bradu, D., Kass, G. V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics* 26:197–208.

Imon, A. H. M. R. (2002). Deletion residuals and deletion weight matrices: Some of their properties and uses. *Pak. J. Statist.* 12:469–484.

Imon, A. H. M. R. (2005). Identifying multiple influential observations in linear regression. *J. Appl. Statist.* 32:929–946.

Montgomery, D. C., Peck, E. A., Vining, G. G. (2001). *Introduction to Linear Regression Analysis.* 3rd ed. New York: Wiley.

Pena, D. (2005). A new statistic for influence in linear regression. *Technometrics* 47:1–12.

Pena, D., Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *J. Roy. Statist. Soc. Series B* 57:145–156.

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. of Statist.* 9:977–986.

Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* 79: 871–880.

Rousseeuw, P. J., Leroy, A. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Simonoff, J. S. (1991). General approaches to stepwise identification of unusual values in data analysis. In: Stahel, W., Weisberg, S., eds. *Robust Statistics and Diagnostics: Part II*. New York: Springer Verlag, pp. 223–242.

Welsch, R. E. (1982). Influence functions and regression diagnostics. In: Launer, R. L., Siegel, A. F. eds. *Modern Data Analysis*. New York: Academic Press.