International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 3 (Special Issue), 2004, pp 191–208

A Conditional Expectation Method for Improved Residual Estimation and Outlier Identification in Linear Regression

Paul Davies
Institute of Child Health
University of Birmingham, Edgbaston, Birmingham, U.K
Email: pdaviesbch@blueyonder.co.uk

A. H. M. Rahmatullah Imon and M. Masoom Ali
Department of Mathematical Sciences
Ball State University, Muncie, Indiana, USA
Email: rimon@bsu.edu; mali@bsu.edu

[Received May 12, 2004; Accepted September 10, 2004]

Abstract

In linear regression, it is common practice to use the ordinary least squares (OLS) residuals as estimates of the true random disturbances since the latter are unobserved. It is well-known that in many circumstances these OLS residuals are not good estimates of the true disturbances. Residuals based on case deletion have been much studied in recent years particularly in regard to the identification of outliers in linear regression. Masking and swamping can make them unsuccessful in this respect. The least squares method is intended to produce a low sum of squared deviations (SSD) between residuals and true errors. It might be expected that excluding outliers when fitting a model should improve the estimation of errors and hence reduce the SSD. But standard theory tells us that instead of lowering the SSD, arbitrary case deletion usually increases the sum of squared deviations between residuals and true errors. In this paper we propose a new method based on a conditional expectation rule to identify unusual observations in data whose omission from the OLS fitting will tend to reduce SSD between the residuals and errors. We consider some examples and then report a Monte Carlo experiment to see how the newly proposed method can be effective in the identification of outliers in linear regression.

Keywords and Phrases: Residuals, Outliers, masking, deletion residuals, conditional deletion, RSD.

AMS Classification: Primary 62J20; Secondary 62J05.

1 Introduction

Practitioners often use OLS residuals mainly because of tradition and ease of computation. But unfortunate consequences of using the OLS residuals in regression diagnostics are well reported [see Huang and Bolch (1974), Rousseeuw and Leroy (1987)] in the statistical literature. Predictive residual error sum of squares (PRESS) residuals proposed by Allen (1974) were devised to be better than OLS in some respects. But it is now evident [Imon (1999)] that PRESS residuals may have similar disadvantages to the OLS counterparts and may not be very useful for diagnostics. Residuals based on single case deletion or modifications of the OLS residuals are also suggested [Cook and Weisberg (1982), Chatterjee and Hadi (1988), Ryan (1997), Imon (2000) and Imon (2002)] in the quest for a better representation of the true disturbances. Robust regression methods such as least median of squares (LMS) and least trimmed squares (LTS) proposed by Rousseeuw (1984) and reweighted least squares proposed by Rousseeuw and Leroy (1987) should produce residuals which are less affected than least squares by the presence of outliers. But most robust techniques are too prone to declare observations as outliers [see Cook and Hawkins (1990)]. Hampel *et al.* (1986) claimed that a routine data set typically contains 5-10% outliers and even a high quality data cannot be guaranteed free from it. For practical purposes, a technique is needed which works well to handle data sets with roughly 10% of outliers than with 50% outliers. The main objective of this paper is to consider a criterion for the identification of a single or a group of observations whose omission could improve the OLS residuals, in the sense that the resulting residual set has a lower mean squared error as an estimate of the true error set than the set of residuals based on estimation using all data. We set out basic results relating to the OLS fit of a regression model and deletion residuals in section 2. In section 3, we introduce the conditional expectation approach to the analysis of residuals to see whether the deletion residuals have better conditional properties than the OLS residuals. In section 4, we propose a conditional deletion (CD) criterion, which considers whether, conditional on the omission of a single case in estimation, the residual set is close in the mean square error sense to the true error. The criterion is used to define a series of rules for finding a good subset to use in estimation. In section 5, we consider some examples and report simulation results to investigate the usefulness of the proposed method in improving residual estimation and, as a by-product, the identification of outliers.

2 The OLS and Deletion Residuals

Consider a standard linear regression model

$$Y = X\beta + \in$$

where Y is an n-vector of observed responses, X is an $n \times p$ matrix representing p explanatory variables with full column rank, β is a p-vector of unknown finite param-

eters and $\in$ is an $n-$vector of uncorrelated random disturbances with $E(\in) = 0$ and $V(\in) = \sigma^2 I$, where σ^2 is an unknown parameter and I is an identity matrix of order n. The OLS estimator of β is $\hat{\beta} = (X^T X)^{-1} X^T Y$ and the vector of fitted values is $\hat{Y} = X\hat{\beta} = WY$, where $W = X(X^T X)^{-1} X^T$ is known as weight or leverage matrix. The $i-$th diagonal element of W matrix, denoted by w_{ii}, is known as the i-th leverage value. The OLS residual vector is

$$\hat{\in} = Y - \hat{Y} = (I - W) \in \tag{1}$$

Assuming that $Y = (y_1, y_2, \cdots, y_n)^T$, $X = (x_1, x_2, \cdots, x_n)^T$, and $\in = (\in_1, \in_2, \cdots, \in_n)^T$, the j-th OLS residual is defined as

$$\hat{\in}_j = y_j - x_j^T \hat{\beta}, \quad j = 1, 2, \cdots, n \tag{2}$$

The j-th residual can also be expressed [see Weisberg (1980)] in terms of the true errors and the leverage values as

$$\hat{\in}_j = (1 - w_{jj}) \in_j - \sum_{k \neq j}^{n} w_{jk} \in_k \tag{3}$$

Let $\hat{\beta}^{(-i)}$ be the OLS estimate of β with the $i-$th case deleted. Then the j-th deletion residual is defined as

$$\hat{\in}_j^{(-i)} = y_j - x_j^T \hat{\beta}^{(-i)}, \quad j = 1, 2, \cdots, n \tag{4}$$

The deletion residual set has a similar definition to the PRESS residuals, but to compute the j-th PRESS residual only the j-th residual is reestimated from the rest after the deletion of that observation. For deletion residuals, not only the deleted residual is reestimated, but also the entire residual set is reestimated from the rest after deleting the i-th observation. Using the result of Miller (1974) for $\hat{\beta}^{(-i)}$, the j-th deletion residual can be expressed in terms of the OLS residuals as

$$\hat{\in}_j^{(-i)} = y_j - x_j^T \hat{\beta} + \frac{x_j^T (X^T X)^{-1} x_i}{1 - w_{ii}} \hat{\in}_i = \hat{\in}_j + \frac{w_{ij}}{1 - w_{ii}} \hat{\in}_i \tag{5}$$

Replacing i with j in (5) we obtain the j-th deletion residual as

$$\hat{\in}_j^{(-j)} = \hat{\in}_j + \frac{w_{jj}}{1 - w_{jj}} \hat{\in}_j = \frac{\hat{\in}_j}{1 - w_{jj}}, \quad j = 1, 2, \cdots, n \tag{6}$$

which is another well-known form of the j-th PRESS residual.

A review of various properties of the deletion residuals is available in Chatterjee and Hadi (1988), Imon (2002), Sengupta and Jammalamadka (2003). The standard theory tells us that for any observation $j = 1, 2, \cdots, n$, we obtain $E(\hat{\in}_j) = 0$ and

$V(\hat{\in}_j) = \sigma^2(1 - w_{jj})$. We observe from (5) that the expected value of the j-th deletion residual after deleting the i-th case is equal to zero, i.e.

$$E\left(\hat{\in}_j^{(-i)}\right) = 0, \quad j = 1, 2, \cdots, n \tag{7}$$

It is also easy to show [see Imon (2002)]

$$\begin{aligned} V\left(\hat{\in}_j^{(-i)}\right) &= V(\hat{\in}_j) - \frac{w_{ij}^2}{1 - w_{ii}}\sigma^2 \quad \text{for } j \neq i \\ &= \frac{\sigma^2}{1 - w_{jj}} \qquad\qquad\qquad \text{for } j = i \end{aligned} \tag{8}$$

We observe from (8) that when the j-th observation is deleted to estimate the entire set of residuals, the variances of all residuals except the deleted one have lower variances than the OLS, but the variance of the deleted residual is increased.

3 Conditional Distribution of OLS and Deletion Residuals

It is customary to estimate σ^2 by using a sample variance estimate of residuals. Imon (2002) pointed out that irrespective of which observation is deleted, the sample variance of deletion residuals is always greater than that of the OLS residuals. That statement might give the impression that deletion of unusual cases from the least squares fitting would not help the estimation of the entire set of true errors. The following lemma will show in what sense the deletion residuals can remove, on average, the deleterious effect of a deleted observation from the residual analysis.

Let $E_{(i)}\{(\in_j)\}$ denote the expectation operator with respect to all errors $\{\in_j\}$ except $\in_i$, i.e., conditioning on the i-th one. Lemma 1 gives the expected (conditional) difference of any individual residual from its corresponding true error term when both the OLS and deletion residuals are considered.

Lemma 1: *If random errors $\{\in_j\}$ are distributed independently and identically with zero mean, then*

$$\begin{aligned} &(i) \quad E_{(i)}\{(\hat{\in}_j - \in_j)\} = -w_{ij} \in_i \\ &(ii) \quad E_{(i)}\left\{\left(\hat{\in}_j^{(-i)} - \in_j\right)\right\} = 0. \end{aligned}$$

Proof: From equation (3), we observe that

$$\hat{\in}_j - \in_j = -w_{jj} \in_j - \sum_{k \neq j}^{n} w_{jk} \in_k = -w_{ij} \in_i - \sum_{k \neq i}^{n} w_{jk} \in_k \tag{9}$$

Taking expectation on both sides of (9) conditioning on i, shows that, $E_{(i)}\{(\hat{\in}_j - \in_j)\} = -w_{ij} \in_i$ which proves part (i) of Lemma 1. From (3) and (5), the deviation of the j-th deletion residual (after deleting the i-th observation) from the j-th error can be obtained from (5) as

$$\begin{aligned}\hat{\in}_j^{(-i)} - \in_j &= -w_{jj} \in_j - \sum_{k\neq j}^{n} w_{jk} \in_k + w_{ij} \in_i - \frac{w_{ij}}{1-w_{ii}} \sum_{k\neq i}^{n} w_{ik} \in_k \\ &= -\sum_{k\neq i}^{n} \left(w_{jk} + \frac{w_{ij}w_{ik}}{1-w_{ii}} \right) \in_k \end{aligned} \tag{10}$$

Noting that (10) does not depend on $\in_i$, the proof of the second part of the lemma is immediate after taking expectation on both sides of (10).

By virtue of the Gauss-Markov theorem, the OLS method produces minimum sum of squared deviations between observed and fitted responses. Also, in the case of *i.i.d.* random errors, it is readily shown that the expected sum of squared deviations between OLS residuals and true errors is only $p\sigma^2$ [see Freedman (1981)]. That is,

$$E\left[\sum_{i=1}^{n} (\hat{\in}_i - \in_i)^2 \right] = p\sigma^2 \tag{11}$$

Lemma 2 provides the corresponding total performance of deletion residuals in this regard.

Lemma 2: *For i.i.d. random errors* $\{\in_j\}$ *with mean zero and variance* σ^2,

$$E\left[\sum_{j=1}^{n} \left(\hat{\in}_j^{(-i)} - \in_j \right)^2 \right] = \left(p + \frac{w_{ii}}{1-w_{ii}} \right) \sigma^2 \tag{12}$$

Proof: From (10) it follows that

$$\begin{aligned} E\left[\left(\hat{\in}_j^{(-i)} - \in_j \right)^2 \right] &= \sum_{k\neq i}^{n} \left(w_{jk} + \frac{w_{ij}w_{ik}}{1-w_{ii}} \right)^2 \sigma^2 \\ &= \left[\sum_{k\neq i}^{n} w_{jk}^2 + 2\frac{w_{ij}}{1-w_{ii}} \sum_{k\neq i}^{n} w_{ik}w_{jk} + \frac{w_{ij}^2}{(1-w_{ii})^2} \sum_{k\neq i}^{n} w_{ik}^2 \right] \sigma^2 \\ &= \left(w_{jj} + \frac{w_{ij}^2}{1-w_{ii}} \right) \sigma^2 \end{aligned} \tag{13}$$

Taking sums on both sides of (13) extending over n and using the properties of the leverage values that $\sum_{j=1}^{n} w_{jj} = p$ and $\sum_{j=1}^{n} w_{ij}^2 = w_{ii}$, we obtain (12) and thus the proof of Lemma 2 is completed.

Since $0 \leq w_{ii} \leq 1$, comparing the result of Lemma 2 to that obtained from (11) it follows that

$$E\left[\sum_{i=1}^{n} (\hat{\in}_i - \in_i)^2\right] \leq E\left[\sum_{j=1}^{n} \left(\hat{\in}_j^{(-i)} - \in_j\right)^2\right] \tag{14}$$

for any observation $i = 1, 2, \cdots, n$. This might give the impression that deletion of the i-th case from the least squares fitting is likely to worsen the estimation of errors instead of improving it. But the following theorem using expectation conditional on case omission establishes conditions in which case omission is likely to be beneficial.

Theorem 1: *For i.i.d. random errors $\in_j$'s with mean zero and variance σ^2,*

$$E_{(i)}\left\{\sum_{j=1}^{n} \left(\hat{\in}_j^{(-i)} - \in_j\right)^2\right\} \leq E_{(i)}\left\{\sum_{j=1}^{n} \left(\hat{\in}_j^{-} \in_j\right)^2\right\} \tag{15}$$

for any observation i satisfying

$$\frac{\in_i^2}{\sigma^2} > \frac{2 - w_{ii}}{1 - w_{ii}}. \tag{16}$$

Proof: From (10), $E\left[\sum_{j=1}^{n} \left(\hat{\in}_j^{(-i)} - \in_j\right)^2\right]$ does not depend on the i-th observation so the expectation of the sum of squared deviation of deletion residuals from the true errors will not alter by conditioning on i. That is,

$$E_{(i)}\left\{\sum_{j=1}^{n} \left(\hat{\in}_j^{(-i)} - \in_j\right)^2\right\} = E\left[\sum_{j=1}^{n} \left(\hat{\in}_j^{(-i)} - \in_j\right)^2\right] = \left(p + \frac{w_{ii}}{1 - w_{ii}}\right)\sigma^2 \tag{17}$$

Squaring both sides of (9) and summing over j throughout we obtain

$$\sum_{j=1}^{n} (\hat{\in}_j - \in_j)^2 = \in_i^2 \sum_j w_{ij}^2 + 2 \in_i \sum_j w_{ij} \sum_{k \neq i} w_{jk} \in_k + \sum_j \left(\sum_{k \neq i} w_{jk} \in_k\right)^2 \tag{18}$$

Taking expectation on both sides of (18), conditioning on i, we obtain,

$$E_{(i)}\left\{\sum_{j=1}^{n} \left(\hat{\in}_j^{-} \in_j\right)^2\right\} = \in_i^2 w_{ii} + (p - w_{ii})\sigma^2 \tag{19}$$

Comparing (19) with (17), it follows that

$$E_{(i)}\left\{\sum_{j=1}^{n}\left(\hat{\in}_j^{(-i)}-\in_j\right)^2\right\} \leq E_{(i)}\left\{\sum_{j=1}^{n}\left(\hat{\in}_j^{-}\in_j\right)^2\right\}$$

holds when $\frac{\in_i^2}{\sigma^2} > \frac{2-w_{ii}}{1-w_{ii}}$ is satisfied. Conversely reversal of the latter inequality leads to a reversed conclusion.

Thus conditional on the omission of a single case with sufficiently large error, the estimated residual may be close to their corresponding true errors than OLS in the sense of possessing lower total mean square of distances.

4 Algorithm for Residual Set Estimation

In this section, we consider a method based on the above result for deciding whether it is best to omit a single or a group of observations for the improved estimation of true errors. The observations thus identified are termed as outliers though there is no unique definition of an outlier which may be multi-attribute. To apply Theorem 1, plausible estimate of the unknown $\in_i$ and σ^2 are needed.

In view of Lemmas 1 and 2, more realistic estimators than those associated with OLS would be the i-th deletion residual $\hat{\in}_i^{(-i)}$ and $\hat{\sigma}^{(-i)^2}$, the respective estimate of $\in_i$ and σ^2 after deleting the i-th data point both in estimating the residuals and in calculating their mean squared deviation in (16). Then the rule suggests the omission of any point i which satisfies

$$\frac{\hat{\in}_i^{(-i)^2}}{\hat{\sigma}^{(-i)^2}} > \frac{2-w_{ii}}{1-w_{ii}}. \tag{20}$$

The left hand side of the above expression is similar to the diagnostic tools like Cook's distance [see Cook (1977)] and DFFITS [see Belsley *et al.* (1980)] that are widely used in measuring influences of observations but differ in variance estimates. The condition (20) is now used to formulate a conditional deletion (CD) rule that involves the following steps.

Step 1

At the first step (20) is used to identify the suspect cases. It is possible that all deletion candidates for suspect outliers will be identified at the first step. But it may not be so simple if masking/swamping occurs due to the presence of a group of outliers. So we propose repetition of (20) once more on the reduced set to find out more potentially omitted cases. We must make the proviso that the above steps would not pick more than 50% observations as suspect. Otherwise it will be simply impossible to distinguish good observations from bad in meaningless fashion.

Step 2

It is possible that in Step 1 some observations may be wrongly detected as omission candidates because of their association with other unusual cases (i.e. due to swamping and masking effects). The deletion of such harmless cases would not help to produce a better set of residuals. Hence it is also desirable to be able to test whether observation(s) should be replaced into the estimation subset when they are wrongly omitted. Obviously the rule (20) may be over-sensitive because it is based on conditional averaging of the deleted cases. It would be more realistic to consider cases to be genuinely harmful when their rule statistic exceeds at least two or three times the rule threshold. Consider the rule examining whether or not each of the observations individually satisfy

$$\frac{\hat{\in}_i^{(-d)^2}}{\hat{\sigma}^{(-d)^2}} > \frac{3\,(2-w_{ii})}{1-w_{ii}} \tag{21}$$

where d is the group of observations identified as omission candidates at Step 1. Rule (21) is analogous to the thrice-the-mean rule suggested by Velleman and Welsch (1981) in the identification of high leverage points in linear regression. Unless all of the members of the d set individually satisfy (21), we suggest that members of the deletion set which do not satisfy the rule (21) are replaced in the estimation set. For the revised deletion set, $\hat{\in}_i^{(-d)}$ and $\hat{\sigma}^{(-d)^2}$ are then recomputed.

For this deletion set d we may define a statistic

$$D_i^2 = \frac{(1-w_{ii})}{(2-w_{ii})}\frac{\hat{\in}_i^{(-d)^2}}{\hat{\sigma}^{(-d)^2}} \quad i = 1, 2, \cdots, n \tag{22}$$

Henceforth D_i^2 will be referred to as conditional deletion D (CD-D) statistic. Observations that satisfy

$$D_i^2 > 3 \qquad \text{for any } \; i = 1, 2, \cdots, n \tag{23}$$

are finally identified as '*best omitted*' or '*outliers*'.

5 Examples and Simulation Results

We now consider several well-known data sets that have often been used in the study outlier identification. The conditional deletion method will be compared with the standard OLS technique and also the robust regression LMS and RLS techniques for these examples. Other more recent methods like Peña and Yohai (1995) are not considered here as they are focused on criteria such as influence measures and not directly on estimation of a set of residuals. Though related they are therefore not directly comparable.

5.1 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 good observations (cases 15-75). It has been reported by many authors [see Rousseeuw and Leroy (1987)] that most of the single case deletion identification methods fail to identify the true outliers though some of them point out high leverage points (cases 11-14) as outliers. Table 1 shows that the well-known Cook's distance identify only one observation (case 14) but all of the genuine outliers are masked. On the other hand robust detection techniques like LMS and RLS identify outliers correctly.

Table 1: Conditional deletion diagnostics for Hawkins *et al.* (1984) data

Index	Cook D	CD-D	Index	Cook D	CD-D	Index	Cook D	CD-D
1	0.040	147.77	26	0.000	0.766	51	0.002	0.676
2	0.053	161.83	27	0.004	0.858	52	0.006	0.469
3	0.046	166.56	28	0.000	0.262	53	0.000	1.577
4	0.031	143.82	29	0.000	0.241	54	0.006	0.744
5	0.039	158.29	30	0.004	0.008	55	0.001	0.013
6	0.052	154.60	31	0.000	0.009	56	0.001	0.004
7	0.079	181.09	32	0.001	0.183	57	0.000	0.877
8	0.052	167.89	33	0.000	0.549	58	0.000	0.047
9	0.034	147.22	34	0.001	0.422	59	0.001	0.051
10	0.048	156.91	35	0.000	0.131	60	0.006	0.625
11	0.348	0.006	36	0.002	1.170	61	0.000	0.028
12	0.851	0.061	37	0.000	0.266	62	0.001	0.823
13	0.254	0.589	38	0.002	1.351	63	0.001	0.148
14	2.114	0.045	39	0.003	0.625	64	0.001	0.425
15	0.001	0.396	40	0.000	0.227	65	0.000	0.843
16	0.003	0.322	41	0.004	0.019	66	0.000	0.928
17	0.001	0.008	42	0.004	0.309	67	0.000	0.711
18	0.000	0.002	43	0.010	0.737	68	0.001	1.081
19	0.001	0.053	44	0.007	0.427	69	0.000	0.069
20	0.000	0.147	45	0.001	0.317	70	0.000	1.178
21	0.001	1.226	46	0.004	0.084	71	0.000	0.078
22	0.002	0.275	47	0.008	1.335	72	0.000	0.008
23	0.000	1.100	48	0.002	0.022	73	0.000	0.571
24	0.002	0.798	49	0.001	1.074	74	0.000	0.824
25	0.000	0.052	50	0.000	0.198	75	0.000	0.351

If the proposed algorithm is applied then the initial stage rule (20) identifies 12 observations (cases 1-3, 5-8, and 10-14) to be unusual. Two more observations (cases 4 and 9) as suspects are identified when Step 1 is repeated once more. Thus the initial deletion set contains 14 observations (cases 1-14) as prime suspects. These 14 observations are now omitted to compute the CD-D values for the entire data set. At this stage all the deleted cases except observations 11-14 possess significantly high CD-D values (the results are not shown for brevity). These observations are then put back into the estimation subset sequentially. When observations 1-10 are omitted from the OLS fit, we observe from Table 1 that all of their corresponding CD-D values are over 3 and no other observations possesses high CD-D value. Thus observations 1-10 are finally deemed to be outliers.

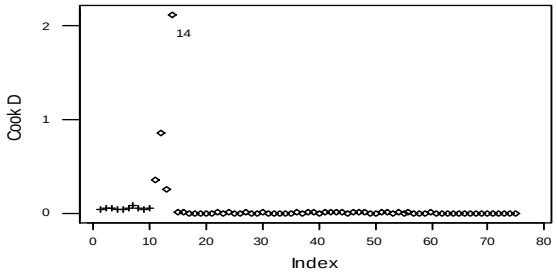

Figure 1.a. Index plot of Cook's distances for Hawkins *et al*. (1984) data

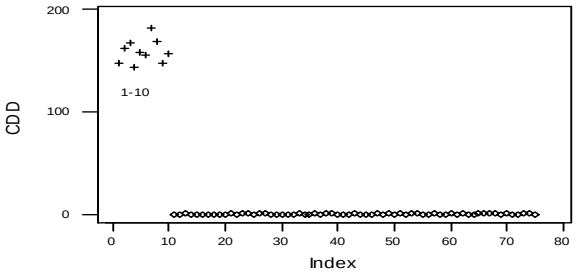

Figure 1.b. Index plot of conditional deletion diagnostics for Hawkins *et al*. (1984) data

Figure 1(a) is an index plot of Cook's distances and CD-D values for this data. Here the outliers are marked by + while the good observations are plotted as o on this graph.

All of the outliers are heavily masked in the index plot of Cook's distances. None of the 10 outliers is identified, but high leverage points appear as outliers. Figure 1(b) is an index plot of CD-D values. All 10 outliers are clearly identified by the conditional method and they are clearly separated from the rest of the data.

5.2 Brownlee's stack loss data

The next example is a real set of data known as the stack loss data, presented by Brownlee (1965) which has 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 observations 1, 3, 4, and 21 generally considered to be outliers [see Atkinson (1985)].

Table 2: Conditional deletion diagnostics for Brownlee's stack loss data

Index	Residuals	Cook D	CD-D	Index	Residuals	Cook D	CD-D
1	1.2095	0.154	10.13	12	0.9667	0.065	0.063
2	-0.7051	0.060	0.342	13	-0.4687	0.011	1.830
3	1.6179	0.126	11.90	14	-0.0170	0.000	0.511
4	2.0518	0.131	19.82	15	0.8006	0.039	0.515
5	-0.5305	0.004	0.140	16	0.2912	0.003	0.006
6	-0.9632	0.020	0.477	17	-0.5996	0.066	0.033
7	-0.8260	0.049	0.050	18	-0.1487	0.001	0.003
8	-0.4737	0.017	0.093	19	-0.1972	0.002	0.099
9	-1.0486	0.045	0.330	20	0.4431	0.005	1.142
10	0.4262	0.012	0.037	21	-3.3305	0.692	19.78
11	0.8783	0.036	0.271				

When the OLS technique is used only one of the outliers (observation 21) is apparent although robust regression techniques can successfully detect all of them [see Rousseeuw and Leroy (1987) and Atkinson (1985)]. The index plot (see Figure 2.a.) of Cook's distances for this data clearly shows how the other outliers are masked. When the CD rule is used the prime suspects are observations 3, 4 and 21. At the next step observations 1 and 13 are marked as suspect. Thus we obtain 5 observations as possible suspects. When all 5 suspects are omitted from fitting, 4 cases (observations 1, 3, 4, and 21) satisfy condition (21). Table 2 shows the conditional deletion D for the entire data set after deleting these 4 observations. The D_i^2 values corresponding to these 4 observations are all over 3 and thus the 4 outliers are correctly identified. The index plot of CD-D (see Figure 2.b.) shows the how clearly the outliers are separated from the rest of the data.

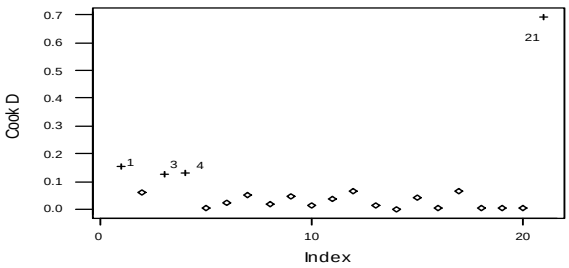

Figure 2.a. Index plot of Cook's distances for stack loss data

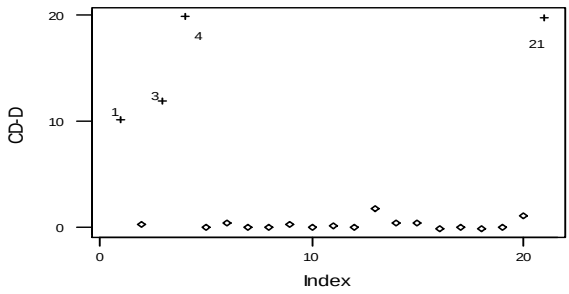

Figure 2.b. Index plot of conditional deletion diagnostics for stack loss data

5.3 Artificial data - Normal errors with 10% of cases as outliers

Table 3 shows an artificial three-predictor data set containing 20 observations. The X's are generated from independent Uniform (0,1) distribution and the Y is computed from equation

$$Y = 20 + 4.5X_1 - 1.5X_2 + 2.8X_3 + \in \tag{24}$$

The errors for the cases 1 and 2 are fixed as 5 and the last 18 $\in$'s are generated from a Normal (0,1) distribution so that this example can be considered as a 10% outlier data set.

The index plot of squares of the standardized errors for this data set in Figure 3.a. clearly indicates observations 1 and 2 as outliers. Table 3 also shows that single case diagnostic like Cook's distance does not unambiguously identify just these two observations as outliers. The index plot of Cook's distances (Figure 3.b) shows that

Table 3: Conditional deletion residuals and diagnostics for artificial 10% outlier data

Index	Y	X_1	X_2	X_3	$\in$	OLS	RLS	CDR	Cook D	CD-D
1	25.90	0.1870	0.9745	0.5415	5.0000	2.0507	0.8451	4.7895	0.324	8.65
2	27.40	0.3745	0.8131	0.6880	5.0000	2.5831	2.4719	4.8730	0.161	9.96
3	24.05	0.8082	0.6417	0.9994	-1.4182	-0.8812	-0.9062	-1.1991	0.052	0.568
4	23.73	0.0203	0.2147	0.9236	1.3781	0.0987	0.0574	0.3573	0.001	0.049
5	22.94	0.5424	0.8337	0.4558	0.4733	-0.4351	-2.1042	0.5927	0.009	0.144
6	25.70	0.7970	0.2137	0.9759	-0.3009	0.3278	1.4821	-0.4074	0.011	0.063
7	24.53	0.6284	0.7165	0.9905	0.0069	-0.3183	-0.4308	0.0690	0.005	0.002
8	24.77	0.5453	0.5622	0.8146	0.8764	0.2027	0.1462	0.7501	0.001	0.242
9	21.71	0.1579	0.2709	0.8303	-0.9145	-1.2660	-2.1421	-1.7178	0.073	1.204
10	24.54	0.7342	0.8591	0.7902	0.3103	-0.1679	-0.7204	0.6305	0.001	0.162
11	24.24	0.7715	0.3856	0.7708	-0.8048	-0.2893	-0.2224	-0.7868	0.003	0.257
12	24.34	0.4322	0.2028	0.3519	1.7152	0.8489	0.4680	1.2528	0.039	0.633
13	24.89	0.0247	0.3527	0.9888	2.5431	0.7465	0.9702	1.6204	0.047	1.007
14	22.72	0.0862	0.0910	0.5949	0.8113	-0.1154	-0.7427	-0.1776	0.001	0.012
15	24.65	0.7798	0.3801	0.3600	0.7051	0.5681	0.2341	0.7850	0.018	0.249
16	20.41	0.4802	0.5094	0.0042	-1.0016	-1.5023	-3.9852	-1.1334	0.292	0.449
17	20.57	0.0226	0.6727	0.5304	-0.0048	-1.5261	-3.8453	-0.6314	0.149	0.156
18	24.24	0.8306	0.0855	0.3008	-0.2089	0.4207	0.3024	-0.2778	0.023	0.028
19	25.30	0.8280	0.6874	0.6058	0.9074	0.4891	0.2915	1.2375	0.010	0.635
20	21.83	0.5038	0.9243	0.7216	-1.0680	-1.6041	-3.3706	-0.9639	0.108	0.380

observation 2 is heavily masked. Also the robust detection technique RLS fails to identify the true outliers and picks the wrong ones (cases 16, 17, and 20) as shown in Figure 3.c. However the conditional deletion method clearly singles out only cases 1 and 2 as the best omitted for the estimation of residuals. The index plot, Figure 3.d. of CD-D values is clearly analogous to Figure 3.a. of the true squared errors.

In order to assess a technique's effectiveness in estimating the true $\in$ values, define the sum of squared distances (SSD) by

$$SSD(*) = \sum_{i=1}^{n} (\in_i^* - \in_i)^2$$

for any set of residuals $\in_1^*, \in_2^*, \cdots, \in_n^*$. As the OLS residuals satisfy $E\left[\sum_{i=1}^{n} (\hat{\in}_i - \in_i)^2\right] = p\sigma^2$, a criterion for whether a method performs well in estimating true errors is the ratio of squared distances (RSD) proposed by Imon (2003) defined by

$$RSD(*) = \frac{\sum_{i=1}^{n} (\in_i^* - \in_i)^2}{p\sigma^2} \tag{25}$$

The RSD quantities can be expected not far from 1 when the OLS residuals are used. Also note that a high correlation coefficient between the residuals and errors does not necessarily imply a good set of residuals although the converse should be true.

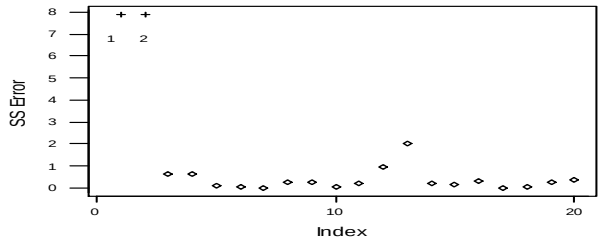

Figure 3.a. Index plot of squared standardized errors for artificial 10% outlier data

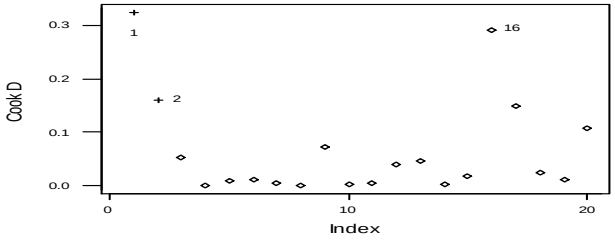

Figure 3.b. Index plot of Cook's distances for artificial 10% outlier data

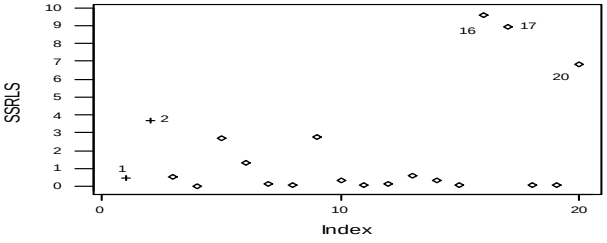

Figure 3.c. Index plot of squared standardized RLS residuals for artificial 10% outlier data

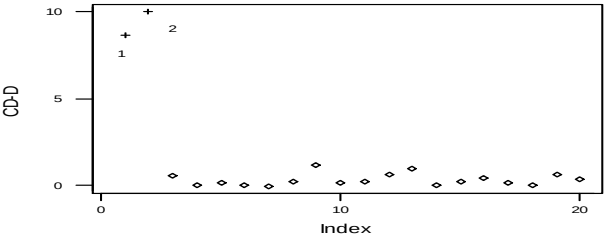

Figure 3.d. Index plot of conditional deletion diagnostics for artificial 10% outlier data

Table 4 shows the RSD values for different estimation techniques used to estimate the true errors for the artificial data. It also contains the indices of the observations indicated as unusual by different methods and correlation coefficients between the residuals and the errors. The LMS and RLS residuals were computed using the PROGRESS program developed by Rousseeuw and Leroy (1987). Table 4 shows that the RLS and LMS methods, which were not specifically devised for the purpose of outlier detection, fail to identify the correct cases as outliers and consequently produce very high residuals for some inliers so that their corresponding RSD quantities are large and much worse than OLS. The CD method is successful in identifying the correct outliers and Table 4 confirms that the estimation of residuals is markedly better than OLS when cases 1 and 2 are omitted from the fit.

Table 4: RSD values for the artificial 10% outlier data

Estimation technique	Observations omitted	RSD	Correlation
OLS	.....	1.73	0.872
LMS	.....	7.14	0.517
RLS	16, 17, 20	5.37	0.630
CD	1, 2	0.32	0.970

6 Simulation results

The merit of the CD-D diagnostic method is now further demonstrated by the results of a Monte Carlo simulation. We re-use the model (24) for artificial three-predictor data sets for sample sizes n =20, 40 and 100. The X variables which are distributed independently of $\in$, are drawn from a Uniform (0,1) distribution and are held constant for experiments based upon a given sample size. In the simulation experiment 90% of the true errors are generated as Normal (0,1) while the first 10% of error values are fixed at 10. Hence the mixed distribution of true errors has mean 1 and standard deviation 3.15 and use of the distribution can be alternatively viewed as equivalent to a biased regression model. Four types of residual estimation, OLS, robust LMS and RLS, and conditional deletion CD are considered.

Table 5 shows the effectiveness of the CD technique in the estimation of residuals in the presence of outliers. The simulation experiment consisted of generating 10000 samples of size n for each of which the CD procedure was used and the mean RSD values were calculated for the OLS, LMS, RLS and CD methods. This table also shows the mean correlation coefficients between estimated residuals $\hat{\in}^*$ and true errors $\in$. Unsurprisingly, the performance of the OLS residuals is not at all satisfactory. Throughout the simulation they produce high RSD values and the method's performance tends to deteriorate with the increase in sample size. The performances

of the robust LMS and RLS are not very satisfactory though they perform better than the OLS. Here the RLS performs marginally better than the LMS. The CD method is very successful in improving the estimation of $\in$ when the data set contains 10% outliers. The RSD values clearly show that CD improves the estimation of residuals. It is also interesting to note that the mean RSD values for OLS, LMS and RLS residuals increase with sample size while they remain almost the same for CD residuals. This suggests that irrespective of sample sizes, the CD method produces more robust residuals in the presence of outliers. The CD method also give a higher correlation with true errors than other methods.

Table 5: Simulated RSD and correlation values for 10% outlier data

Measures	Estimation technique	$n = 20$	$n = 40$	$n = 100$
RSD	OLS	1.88	2.14	3.38
	LMS	0.47	0.51	0.74
	RLS	0.38	0.41	0.58
	CD	0.12	0.11	0.10
Correlation	OLS	0.854	0.943	0.984
	LMS	0.943	0.948	0.989
	RLS	0.954	0.961	0.996
	CD	0.991	0.996	0.999

7 Conclusions

In this paper the main objective was to propose a criterion and method for obtaining a better estimated set of residuals than provided by the least squares regression. Theory established conditions under which, conditional on the omission of a single case in estimation, it is possible to obtain a residual set that is close in the mean square error sense to the true errors. The criterion was used to define a stepwise conditional deletion rule for finding a good subset to use in estimation. The examples considered, both real and artificially constructed, clearly showed how this method can be effective not only to produce a more accurate set of residuals but consequently to identify outliers. The simulation results of the CD procedures also support its effectiveness as a tool in regression diagnostics. This paper has confined itself to 'one case at a time' deletion rule. However, as shown in Imon (1996), the theory and method can be readily extended to considering simultaneous deletion of subsets of cases.

References

[1] Allen, D.M. (1974). The relationship between variable selection and augmentation and a method of prediction, *Technometrics*, 16, 125-127.

[2] Atkinson, A.C. (1985). *Plots, Transformations, and Regression*, Clarendon Press, Oxford.

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

[4] Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression*, Wiley, New York.

[5] Cook, R.D. (1977). Detection of influential observations in linear regression, *Technometrics*, 19, 15-18.

[6] Cook, R.D. and Hawkins, S. (1990). Comment on paper by Rousseeuw, P.J. and van Zomeren, B.C., *Journal of the American Statistical Association*, 85, 640-644.

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

[8] Freedman, D.A. (1981). Bootstrapping regression models, *The Annals of Statistics*, 9, 1218-1228.

[9] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions,* Wiley, New York.

[10] Huang, C.J. and Bolch, B.W. (1974). On the testing of regression disturbances for normality, *Journal of the American Statistical Association*, 69, 330-335.

[11] Imon, A.H.M.R. (1996). *Subsample Methods in Regression Residual Prediction and Diagnostics.* Unpublished Ph.D. thesis, School of Mathematics and Statistics, University of Birmingham, U.K.

[12] Imon, A.H.M.R. (1999). On PRESS residuals, *Journal of Statistical Research*, 33, 57-65.

[13] Imon, A.H.M.R. (2000). Modified residuals in tests for normality, *Journal of Statistical Studies*, 20, 21-26.

[14] Imon, A.H.M.R. (2002). A note on deletion residuals, *Calcutta Statistical Association Bulletin*, 52, 65-79.

[15] Imon, A.H.M.R. (2003). Residuals from deletion in added variable plots, *Journal of Applied Statistics*, 30, 841-855.

[16] Miller, R.G. (1974). An unbalanced jackknife, *The Annals of Statistics*, 2, 880-891.

[17] Peña, D. and Yohai, V.J. (1995). The detection of influential subsets in linear regression by using an influence matrix, *Journal of the Royal Statistical Society Series*-B, 57, 145-156.

[18] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, 79, 871-880.

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

[20] Ryan, T.P. (1997). *Modern Regression Methods*, Wiley, New York.

[21] Sengupta, D. and Jammalamadaka, S.R. (2003). *Linear Models: An Integrated Approach*, World Scientific, New Jersey.

[22] Vellman, P.F. and Welsch, R.E. (1981). Efficient computing of regression diagnostics, *The American Statistician*, 35, 234-242.

[23] Weisberg, S. (1990). Comment on 'Some large-sample tests for nonnormality in the linear regression model' by White, H. and MacDonald, G.M. *Journal of the American Statistical Association*, 75, 28-31.