Pinpointing outliers and influential cases in regression analysis: a robust method at work

Ettore Marubini[1], Annalisa Orenti[2]

[1] Sezione di Statistica Medica e Biometria (G.A. Maccacaro)
del Dipartimento di Medicina del Lavoro (L. Devoto)
Università degli studi di Milano
[2] Unità di Statistica Medica e Biometria
Fondazione IRCCS Istituto Nazionale dei Tumori, Milano

Corresponding Author:
Annalisa Orenti
Unità di Statistica Medica e Biometria
Fondazione IRCCS Istituto Nazionale dei Tumori
Via Venezian 1, 20133 Milano
E-mail: annalisa.orenti@istitutotumori.mi.it

Summary
Least squares regression analysis is greatly influenced by outlying observations, either in y or in x coordinates. To investigate outliers the traditional approach implies computing LS diagnostics. However sometimes this strategy fails, specifically so when several outliers are present. As an alternative it is possible to adopt a robust regression. In this context the iteratively weighted least squares method proposed by Chatterjee and Mächler (C-M) (2) is particularly appealing. This method is used here as the starting point for a diagnostic tool able to pinpoint outliers and influential cases in regression analysis. This tool is applied to four examples taken from statistical literature and the results are discussed in details.

KEY WORDS: *outliers, leverage, influential cases, robust regression diagnostics.*

Introduction

Regression outliers, either in y or in x, pose a serious threat to standard least squares (LS) analysis. In order to pinpoint outliers the "traditional approach" in LS regression makes use of routine tools such as graphic representations of raw data, computation and assessment of regression diagnostics and graphs of residuals. However the analyst should be aware of possible shortcomings of this approach, particularly when several outliers are present (Rouseeuw and Leroy (1), chapter 6). Alternatively one could resort to robust regression which tries to devise estimators that are not so strongly affected by outliers and it is by looking at the results from a robust regression that outliers may be investigated. In this context the robust regression method suggested by Chatterjee and Mächler (2) appears to be particularly appealing. By taking into consideration outliers and leverage points, the authors propose an iteratively weighted least squares (WLS) method which gives robust fits. As a by product the list of relative weights computed in the last iteration and arranged in increasing order can be obtained. This becomes the starting point of the approach to diagnose outliers which will be presented by examples in this note.

Methodological background

Preliminary notation and terminology

In matrix notation the standard linear model is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad [1]$$

where: $\mathbf{Y}$ ($n\times1$), response vector, $\mathbf{X}$ ($n\times p$), design matrix, $\boldsymbol{\beta}$ ($p\times1$) vector of parameters to be estimated, $\mathbf{e}$ ($n\times1$) unknown vector of random errors. Letter n specifies the number of observations and p the number of regressors (including the intercept). Furthermore:

$$\text{var}(\mathbf{e}) = \sigma^2\mathbf{I} \qquad [2]$$

where **I** is the ($n \times n$) identity matrix.
The predicted value $\hat{\mathbf{Y}}$ of **Y** is defined as:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X(X'X)^{-1}X'Y} = \mathbf{HY}$$

where $\hat{\boldsymbol{\beta}}$ is the LS estimator of $\boldsymbol{\beta}$ (the minimising criterion for $\hat{\boldsymbol{\beta}}$ will be given later);

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

is the projection or leverage matrix (Hat matrix).
The i-th term on the principal diagonal of **H** is:

$$h_{ii} = \mathbf{x'}_i\mathbf{(X'X)^{-1}x}_i, \qquad [3]$$

where $\mathbf{x}_i$ is the i-th row of the design matrix **X** and, therefore, corresponds to the i-th observation. The key feature of a leverage h_{ii} is that it describes how far away the individual data point is from the centroid of all data points in the space of regressors.
The vector of estimated residuals $\hat{\mathbf{e}}$ is:

$$\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{(I - H)Y}$$

and

$$\text{var}(\hat{\mathbf{e}}) = \sigma^2\mathbf{(I - H)}$$

It follows that:

$$\text{standard error}(\hat{e}_i) = \sigma\sqrt{1-h_{ii}}$$

The LS estimator vector $\hat{\boldsymbol{\beta}}$ is obtained by minimising, with respect to $\boldsymbol{\beta}$, Residual Sum of Squares (RSS):

$$RSS = \hat{\mathbf{e}}'\hat{\mathbf{e}}$$

The estimator $\hat{\sigma}^2$ of the variance σ^2 is given by:

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\hat{e}_i^2$$

Among the LS diagnostics the following will be considered:

(i) standardized residuals= $\dfrac{\hat{e}_i}{\hat{\sigma}}$

(ii) studentized residuals= $\dfrac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$

Model [1] can be adopted even in weighted least squares regression analysis; however unlike [2]:

$$\text{var}(\mathbf{e}) = \sigma^2\mathbf{W}^{-1}$$

where **W** is a ($n \times n$) diagonal matrix of proper weights.
We will continue to use the symbol $\hat{\boldsymbol{\beta}}$ for the estimator of $\boldsymbol{\beta}$ even though the estimate will be obtained via generalized, not ordinary, least squares. The estimator $\hat{\boldsymbol{\beta}}$ is chosen now to minimise the generalized Residual Sum of Squares (RSS):

$$RSS = \hat{\mathbf{e}}'\mathbf{W}\hat{\mathbf{e}}$$

Similarly the symbols $\hat{\mathbf{e}}$ and $\hat{\sigma}$ will be used for the estimator of **e** and σ respectively.

Chatterjee-Mächler algorithm: an outline

For an exhaustive discussion on the properties of the Chatterjee and Mächler (hereafter referred to as C-M) robust regression method the reader is referred to the original paper (2). Here it is sufficient to say that the approach suggested is a generalized least squares regression method since "...the weights are determined by the residuals and as these change from iteration to iteration the procedure is an iterative one." (2).
The algorithm starts with an initial fit and iterates. Namely:
Step 0: compute weights $w_i^0 = \dfrac{1}{\max(h_{ii}, \bar{h})}$,
where $\bar{h} = \dfrac{p}{n}$ and h_{ii} is given by [3].

Calculate the WLS regression coefficients:

$$\hat{\boldsymbol{\beta}}^0 = \arg\min_{\beta}\left[\mathbf{e'W^0e}\right]$$

Step k: (k=1,2,...)
compute new weights from the residuals of the last fit

$$\hat{\mathbf{e}}^{j-1} = \mathbf{Y} - \hat{\mathbf{Y}}^{j-1} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{j-1},$$

$$w_i^j = \frac{(1-h_{ii})^2}{\max\left(\left|e_i^{j-1}\right|, med_i\left|e_i^{j-1}\right|\right)}, (i = 1,2,...,n)$$

where $med_i z_i = median(z_1, z_2, \ldots, z_n)$.
Compute the WLS regression coefficients:

$$\hat{\boldsymbol{\beta}}^j = \arg\min_{\beta}\left[\mathbf{e'W}^j\mathbf{e}\right]$$

This is iterated till convergence.
Points with high leverage are down-weighted in step 0 of the algorithm. In subsequent steps the weights are a function of residuals as well as of leverage.
The authors state that: "...the algorithm compute a "Mallows-type" M estimator of regression...".

A diagnostic tool

First of all the cases giving the lowest contributions in estimating the vector $\boldsymbol{\beta}$ are identified by dividing the weights w_i (i=1,2,...,n) by the maximum of their values and arranging the so obtained relative weights in increasing order. These observations may be viewed as suspected "patho-

logical" cases, since the algorithm gives low weights to points with (i) large leverage and (ii) large residuals.

Combining this knowledge with that of leverage h_{ii} and of the residual computed in the last iteration, it is then possible to carry out a differential diagnosis for the suspected pathological cases. As a rule of thumb to identify outliers the cut-off point 2.5 for the absolute value of standardized or studentized residuals is adopted, while to determine potentially influential points, particular attention is paid to leverage $h_{ii} > \frac{2p}{n}$ (1).

Results

In this section four data sets are considered. Firstly two simple linear regression examples are used and outliers are graphically shown in the two-dimensional space; secondly two multiple linear regression examples are introduced. As it is instructive to examine how a procedure works when the correct answer is apparent, *ad hoc* generated datasets are used.

Data were processed by means of software R; *lm* function was adopted for LS regression and an *ad hoc* function *SCreg0*, made available to us by S. Chatterjee, was adopted for C-M regression. A code for C-M procedure is also available in a SAS IML version as well as a Minitab macro.

All sets of original data are given in the appendix.

Pilot Plant data from Daniel and Wood (1971)

The Pilot Plant data are given in Table 1 A. "The independent variable (regressor) x, is the acid number of a chemical determined by titration, and the dependent (response) variable y, is its organic acid content determined by extraction and weighing. The aim was to determine how well values obtained by the relatively inexpensive titration method can serve to estimate those obtained by the more expensive extraction and weighing technique" (3).

Original data are drawn in Panel A of Figure 1 together with the two fitted LS and C-M regression lines. As expected the lines are overlapping since the differences between the respective coefficients are practically negligible as it can be seen from Table 1. By assuming typing or recording errors, two of the original data are modified, namely the response variable y of the observation 17 was reduced from 89 to 59 (data contaminated in y) and the regressor of the observation 20 was decreased from 167 to 0.167 (data contaminated in x). The effect of these contaminations on LS regression estimates is clearly evident from Panels B and C of Figure 1. The two LS straight lines tilt towards the two contaminated cases. On the contrary there is no effect of these contaminations on the C-M regression lines (see also the estimated coefficients in Table 1). However, as far as $\hat{\sigma}$ is concerned the contamination exerts its influence even on the C-M estimate, though less evident than that on LS regression estimate.

Table 1. Estimates of regression coefficients (standard errors) obtained by LS as well as by C-M regression for the three sets of Pilot-Plant data: original, contaminated in y, contaminated in x.

	Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$
Original data	LS	35.4583 (0.6350)	0.3216 (0.0056)	1.2300
	C-M	35.6866 (0.7755)	0.3187 (0.0067)	1.2760
Data contaminated in y	LS	38.1164 (3.5501)	0.2813 (0.0311)	6.8759
	C-M	35.6994 (1.2824)	0.3182 (0.0116)	2.0960
Data contaminated in x	LS	46.6746 (5.1313)	0.2315 (0.0475)	11.0489
	C-M	35.8676 (1.3753)	0.3186 (0.0125)	2.2946

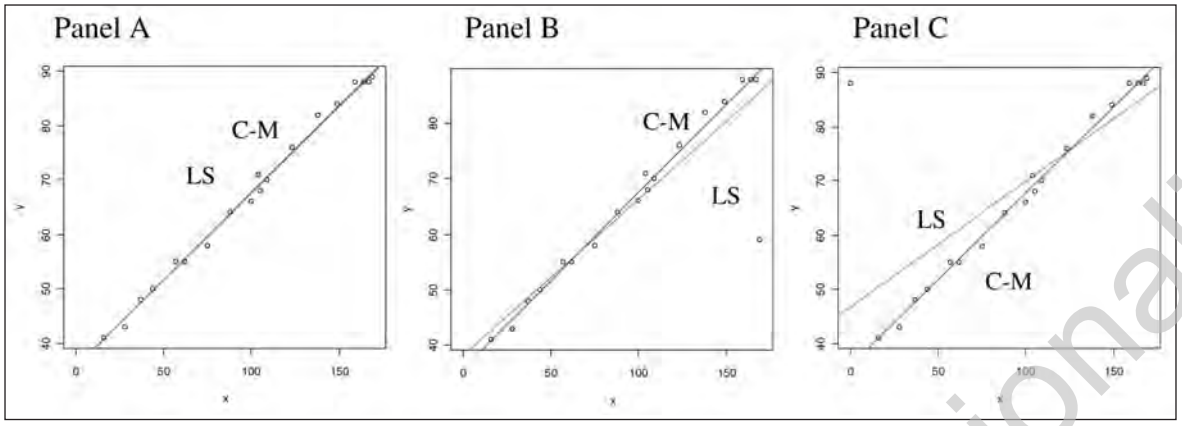

Figure 1. Graphical display of: original Pilot-Plant data (Panel A); data contaminated in y (Panel B); data contaminated in x (Panel C) with the pertinent LS and C-M regression lines.

The C-M regression diagnostics are reported in Table 2. Considering the original homogenous data, the absolute values of C-M residuals are expected not to be over the 2.5 cut-off point. Moreover the relative C-M weights range from 0.35 to 1.0 (see also Panel A of Figure 2) indicating that no case is strongly down-weighted. The leverages h_{ii} (i=1,2,...,n) are all less than the cut-off point (0.2) with the exception of case 9, which could be considered a pathological case. However this hypothesis is rejected by the procedure, which does not down-weight it.

In the dataset contaminated in y, case 17 has rank 1, with the small value 0.027 in the column of relative C-M weights (see also Panel B of Figure 2) informing the reader that the case contaminated in y has been correctly down-weighted. As the absolute value of the corresponding residual is much greater than the 2.5 cut-off point and the corresponding h_{ii} is clearly under its cut-off, it seems sensible to diagnose case 17 as an outlier.

Similarly in the third dataset, the case contaminated in x has rank 1 in the column of relative C-M weights (see also Panel C of Figure 2). Note that now both the C-M residual and the leverage h_{ii} are over the respective cut-off points, suggesting the presence of an outlying influential (very bad) point.

Ad hoc data for simple linear regression

Data for this example are reported in Table 2.A; they correspond to the data in table 11.3 (pg 372) by Ryan (4) and were used by the author to illustrate the Least Trimmed Squares (LTS) method with sequential trimming. Figure 3 gives the scatter plots of these data.

One can recognize a bulk of twelve observations on the left side of the panels, three supposed good leverage points (13, 14, 15) and three bad leverage

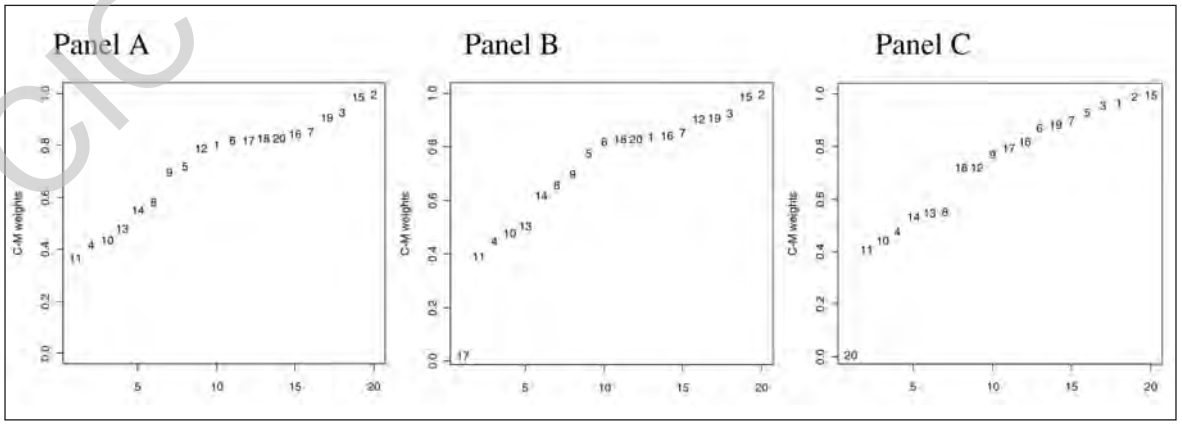

Figure 2. C-M weights on the ordinate with the corresponding increasing order number (1-20) in abscissa: original Pilot-Plant data (Panel A); data contaminated in y (Panel B); data contaminated in x (Panel C).

Table 2. C-M weights, C-M standardized residuals and leverages for the three sets of Pilot-Plant data: Original, contaminated in y, contaminated in x.

Original data				Data contaminated in y				Data contaminated in x			
ID	C-M weights	C-M Standardized residuals	h_{ii} (0.2)	ID	C-M weights	C-M Standardized residuals	h_{ii} (0.2)	ID	C-M weights	C-M Standardized residuals	h_{ii} (0.2)
11	0.3719	1.8303	0.0749	17	0.0270	-14.5379	0.1387	20	0.0126	22.6959	0.2151
4	0.4216	1.7011	0.0500	11	0.3962	1.1410	0.0749	11	0.4116	0.9457	0.0846
10	0.4400	-1.2615	0.1649	4	0.4521	1.0542	0.0500	10	0.4484	-0.7790	0.1322
13	0.4859	1.2878	0.1138	10	0.4829	-0.7674	0.1649	4	0.4792	0.8722	0.0516
14	0.5567	-1.2443	0.0660	13	0.5080	0.8157	0.1138	14	0.5377	-0.7671	0.0572
8	0.5872	-1.2184	0.0502	14	0.6216	-0.7457	0.0660	13	0.5497	0.6450	0.1264
9	0.7022	0.1681	0.2045	8	0.6623	-0.7240	0.0502	8	0.5567	-0.7515	0.0505
5	0.7241	0.9002	0.0932	9	0.7022	0.1000	0.2045	18	0.7262	-0.4656	0.1465
12	0.7949	-0.8998	0.0501	5	0.7800	0.5554	0.0932	12	0.7264	-0.5741	0.0520
1	0.8079	0.8744	0.0581	6	0.8227	0.2518	0.1390	9	0.7750	0.0154	0.1644
6	0.8227	0.4093	0.1390	18	0.8334	-0.3987	0.1334	17	0.7983	-0.3075	0.1520
17	0.8232	-0.4261	0.1387	20	0.8334	-0.3987	0.1334	16	0.8234	-0.0491	0.1387
18	0.8334	-0.7103	0.1334	1	0.8414	0.5555	0.0581	6	0.8762	0.1505	0.1115
20	0.8334	-0.7103	0.1334	16	0.8481	0.0567	0.1258	19	0.8902	0.2901	0.1044
16	0.8481	0.0390	0.1258	7	0.8572	0.1434	0.1211	7	0.9041	0.0503	0.0975
7	0.8572	0.2284	0.1211	12	0.9082	-0.5288	0.0501	5	0.9351	0.4245	0.0763
19	0.9128	0.6505	0.0931	19	0.9128	0.4254	0.0931	3	0.9605	-0.2697	0.0698
3	0.9304	-0.3486	0.0844	3	0.9304	-0.2036	0.0844	1	0.9708	0.4134	0.0648
15	0.9918	0.2112	0.0546	15	0.9918	0.1434	0.0546	2	0.9938	-0.2578	0.0538
2	1.0000	-0.3314	0.0507	2	1.0000	-0.1819	0.0507	15	1.0000	0.0429	0.0508

points (16, 17, 18) on the right side of the panels. As stated by Ryan (4): "...the three points that are (somewhat) outlying on x (only) help to define the regression line and thus increase the precision of the parameters estimates, whereas we would want the three points that are also outlying on y to be identified as regression outliers".

In the three panels of figure 3 the LS regression estimated lines are reported together with all the original data. The line drawn in panel A is estimated on the ground of the twelve points on the left of the panel; the line in panel B is estimated by adding observations 13, 14, 15 to the first twelve; the line in panel C is estimated by adding observations 16, 17, 18 to the first twelve.

In moving from panel A to panel B one can see that the three good data points influence the regression line only marginally, but they increase the precision of the estimates; for instance the standard error of $\hat{\beta}_1$ reduces from 0.1918 to 0.0975

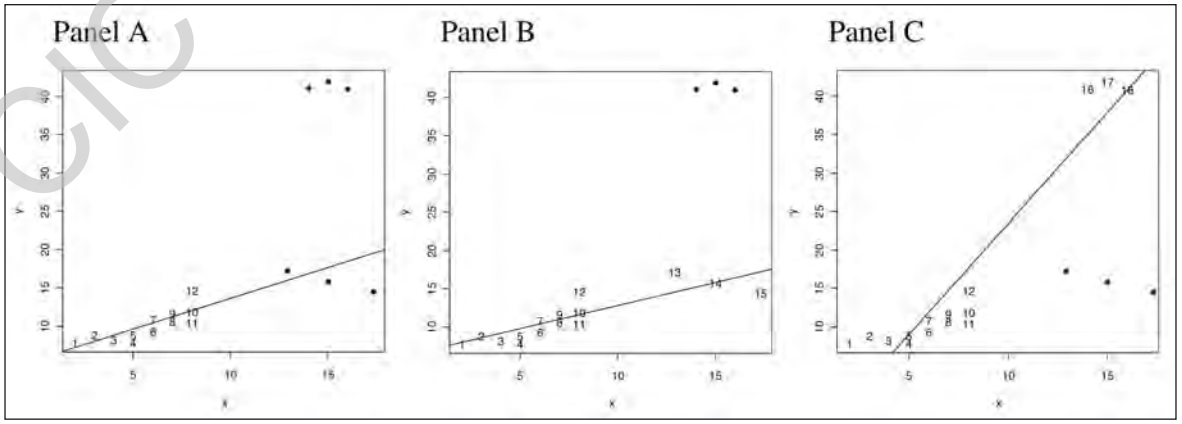

Figure 3. Graphical display of Ryan's data with the LS regression lines estimated on different sets of observations (see text). The numbers indicate data used in estimating the regression line in each panel.

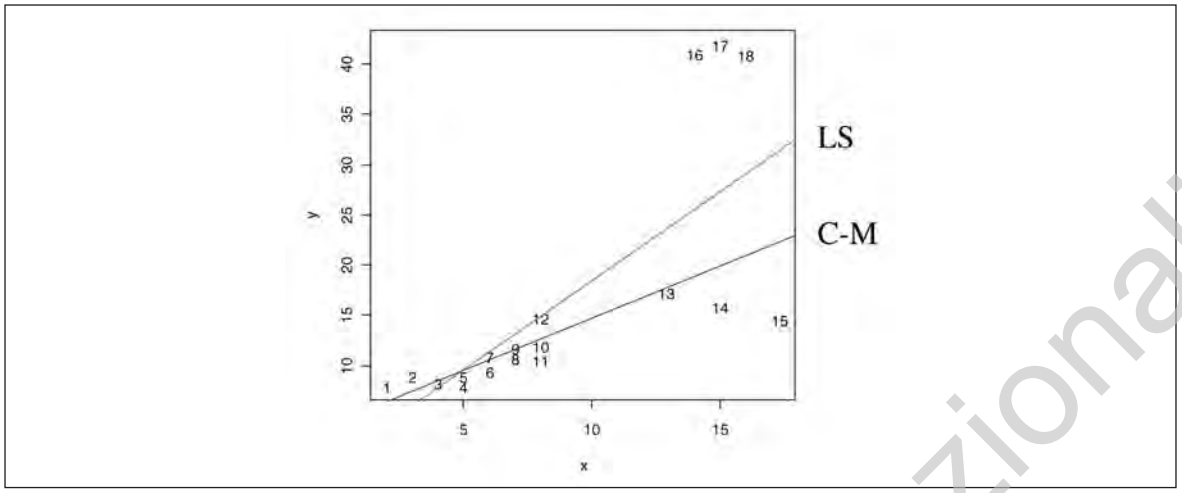

Figure 4. Graphical display of Ryan's data with the pertinent LS and C-M regression lines.

even though the estimate of σ slightly increases from 1.2762 to 1.5783.

On the contrary, in moving from panel A to panel C, one can appreciate the dramatic influence of the bad data points. The results obtained by LS and C-M regressions carried out on the whole set of eighteen datapoints are shown in Figure 4. The large influence of bad observations on the LS estimated line is immediately apparent.

In the LS regression neither the standardized nor the studentized residuals show absolute values over the 2.5 cut-off point; i.e. no outliers are present: any bad observation causes the LS regression line to tilt and inflate the residual variance so that these diagnostics become inefficient in labelling outliers. On the contrary the C-M procedure pinpoints clearly the three bad points as those having the lowest weights; furthermore the corresponding standardized residuals are so much greater than the 2.5 cut-off point to leave no doubt in assessing that these observations are outliers (Table 3).

Furthermore the C-M procedure is also able to identify a further suspected "pathological case" among the three supposed "good leverage" points; observation 15 may be considered an influential point because its leverage $h_{ii} = 0.2340$ is over the pertinent cut-off point $\frac{2p}{n} = 0.2222$ and its C-M weight is low.

On the other hand, observations 13 and 14 have leverage h_{ii} well under the cut-off 0.2222, in line with the h_{ii} of the twelve points forming the bulk of the data.

Ad hoc data set for multiple regression

Data for this example are reported in Table 3.A; they correspond to the data in Table 11.6 (pg 375) by Ryan (4).

"…similar to the previous example, most (24) of the 30 data points constitute a homogeneous set of good data, in addition three good data points that are somewhat removed from the first set in terms of the regressor value and three bad data points." Later it is stated that cases 28, 29 and 30 are the three bad data points.

As regards LS diagnostics, neither standardized, nor studentized residuals have absolute values above the 2.5 cut-off point. On the contrary the three bad points have rank 1-3 in the list of ordered C-M weights and all present C-M residuals over the 2.5 cut-off point thus allowing them to be labelled as outliers (Table 4). Such a result coincides with the one obtained by Ryan adopting the LTS approach with sequential trimming. Furthermore among the three cases (25, 26, 27) supposed by Ryan to be good data points, only case 25 shows leverage h_{ii} over the cut-off point 0.2667, suggesting it may be considered an influential point.

In the generation of data Ryan uses a random error component normally distributed with mean 0 and variance 64. The two estimates of this variance are respectively $\hat{\sigma}^2_{LS} = 106.08$ and $\hat{\sigma}^2_{C-M} = 52.79$. It appears that the three bad points inflate the LS estimate of variance and thus render LS diagnostics inefficient in pinpointing outliers.

Table 3. LS Regression diagnostics, C-M weights, C-M standardized residuals and leverage for Ryan's data at pg. 372.

LS			C-M			
ID	Standardized residuals	Studentized residuals	ID	C-M weights	Standardized residulas	h_{ii} (0.222)
1	0,4516	0,4964	17	0.0566	6.1366	0.1501
2	0,3589	0,3872	18	0.0576	5.5705	0.1833
3	0,0625	0,0664	16	0.0601	6.1756	0.1219
4	-0,1859	-0,1952	15	0.1325	-2.1301	0.2340
5	-0,0661	-0,0694	14	0.3142	-1.1054	0.1501
6	-0,2187	-0,2275	12	0.7007	0.6101	0.0573
7	-0,0391	-0,0406	1	0.7233	0.4553	0.1725
8	-0,2755	-0,2848	11	0.7697	-0.5553	0.0573
9	-0,1557	-0,1610	2	0.7994	0.4441	0.1408
10	-0,3443	-0,3546	3	0.8831	-0.0388	0.1141
11	-0,5120	-0,5273	13	0.9184	-0.1107	0.0966
12	-0,0089	-0,0091	4	0.9269	-0.4107	0.0924
13	-0,7507	-0,7898	5	0.9269	-0.1332	0.0924
14	-1,3647	-1,4803	6	0.9613	-0.2832	0.0757
15	-2,0092	-2,2956	7	0.9613	0.1330	0.0757
16	1,8665	1,9919	8	0.9858	-0.2111	0.0640
17	1,7618	1,9111	9	0.9858	0.0663	0.0640
18	1,4295	1,5819	10	1.0000	-0.1668	0.0573

Table 4. Regression LS diagnostics, C-M weights, C-M standardized residuals and leverage for Ryan's data at page 375.

LS			C-M			
ID	Standardized residuals	Studentized residuals	ID	C-M weights	Standardized residuals	h_{ii} (0.2667)
1	0,8708	1,0016	30	0,1071	3,8708	0,2224
2	0,4001	0,4199	28	0,1105	4,1579	0,1815
3	0,9846	1,0668	29	0,1963	2,5579	0,1441
4	-0,0355	-0,0364	23	0,2471	-1,9143	0,1694
5	0,0561	0,0586	20	0,2796	-1,3487	0,2583
6	0,0273	0,0278	27	0,2925	-1,6548	0,1597
7	0,6632	0,7464	26	0,3334	-1,5618	0,1285
8	-0,2844	-0,2894	3	0,3504	1,4200	0,1481
9	0,4451	0,4654	15	0,3536	-1,3163	0,1760
10	0,1420	0,1482	1	0,3745	1,0457	0,2442
11	1,0844	1,1711	7	0,4226	1,0112	0,2105
12	-0,7371	-0,7854	11	0,4399	1,1455	0,1427
13	0,5870	0,6761	16	0,4797	-1,2648	0,0592
14	0,5067	0,5266	25	0,5687	-0,2098	0,2718
15	-0,9738	-1,0727	13	0,6094	0,4803	0,2462
16	-1,0257	-1,0575	22	0,7564	-0,7099	0,1151
17	-0,4839	-0,5088	12	0,7933	-0,6703	0,1192
18	0,1377	0,1462	19	0,7991	-0,5655	0,1368
19	-0,6304	-0,6785	18	0,8456	0,3321	0,1121
20	-1,1534	-1,3393	17	0,8778	-0,6080	0,0953
21	0,1878	0,1951	2	0,8845	0,1238	0,0919
22	-0,1291	-0,1373	9	0,8969	0,5644	0,0855
23	-1,1453	-1,2567	5	0,9005	0,0584	0,0837
24	-0,2128	-0,2182	10	0,9035	-0,0177	0,0822
25	-0,3803	-0,4456	14	0,9192	0,4109	0,0743
26	-1,8131	-1,9422	21	0,9212	0,1593	0,0733
27	-1,9520	-2,1294	4	0,9713	-0,1331	0,0484
28	2,0588	2,2757	24	0,9719	0,2157	0,0481
29	1,0548	1,1402	6	0,9944	0,2866	0,0371
30	1,7504	1,9849	8	1,0000	-0,2114	0,0344

Interstitial Lung Disease data (Narula et al. (5))

The data reported in table 4.A are the same as those used in Narula *et al.* (5); they regard the results of a study performed to investigate the association between objective indicators of lung damage and severity of functional impairment in patients affected by interstitial lung disease (ILD)

The objective of Narula et al. (5) was to introduce the minimum sum of absolute errors regression as an alternative to the LS regression that is sensitive to outliers. In their paper fourteen regressors were initially considered; here we chose to focus on the four regressors kept in their final model. The variables are:

Response: y = FVC (Forced Vital Capacity).

Regressors: x_2 = AGE (in years);

x_4 = EPIT (epithelial cells): area fraction of epithelial cells/10000 μm^2 of alveolar tissue;

x_8 = CELL (cellular infiltration): total cellularity/10000 μm^2 of alveolar tissue;

x_{13} = HONEY (honeycombing): score of honeycombing (zero to four).

The LS and C-M diagnostics are reported in table 5. As regards LS diagnostics no absolute value of residuals are over the cut-off point with the exception of case 11. Instead, along with case 11, also case 15 can be considered as an outlier from C-M standardized residuals. In addition C-M procedure leads us to suspect two further influential points; these are case 23 and 3 because they have the smallest C-M weights. Since the absolute value of their C-M standardized residuals are clearly under the 2.5 cut-off point, their low C-M weights are likely to be a result of their leverage h_{ii} being much greater than the pertinent cut-off point.

In trying to study the behaviour of the latter two cases with respect to each independent variable,

Table 5. Regression LS diagnostics, C-M weights, C-M standardized residuals and leverage for Narula et al. data (5).

LS			C-M			
ID	Standardized residuals	Studentized residuals	ID	C-M weights	Standardized residuals	h_{ii} (0.4167)
1	-0,0355	-0,0410	23	0,0799	-1,2592	0,6220
2	0,1500	0,1596	3	0,1299	0,0832	0,6719
3	-0,1559	-0,2722	15	0,1494	2,9294	0,2116
4	-0,2821	-0,3541	11	0,1789	-3,2591	0,0899
5	1,1924	1,2579	7	0,2629	-1,8900	0,1600
6	1,1133	1,2161	6	0,3065	1,6139	0,1618
7	-1,1725	-1,2793	5	0,3154	1,8025	0,1015
8	-0,8189	-0,8657	18	0,3575	1,3744	0,1646
9	-0,6429	-0,6777	14	0,4318	1,1804	0,1491
10	0,4211	0,4567	4	0,4861	0,2687	0,3653
11	-2,5234	-2,6451	16	0,5757	-0,7494	0,2173
12	-0,1684	-0,1765	22	0,5795	-0,4294	0,3070
13	-0,5787	-0,6328	8	0,5991	-0,9409	0,1053
14	1,0666	1,1563	1	0,6748	0,0954	0,2522
15	1,8704	2,1065	20	0,7182	0,2357	0,2285
16	-0,3005	-0,3397	19	0,8416	-0,6104	0,1458
17	0,2175	0,2337	13	0,8438	-0,4709	0,1638
18	1,0953	1,1983	21	0,8522	0,4373	0,1597
19	-0,6026	-0,6520	10	0,8726	0,5564	0,1496
20	0,1712	0,1950	17	0,9059	0,2872	0,1336
21	0,3669	0,4003	24	0,9072	0,2950	0,1329
22	-0,2234	-0,2684	9	0,9295	-0,6135	0,1000
23	-0,4551	-0,7402	2	0,9412	-0,0570	0,1168
24	0,2950	0,3168	12	1,0000	-0,3033	0,0897

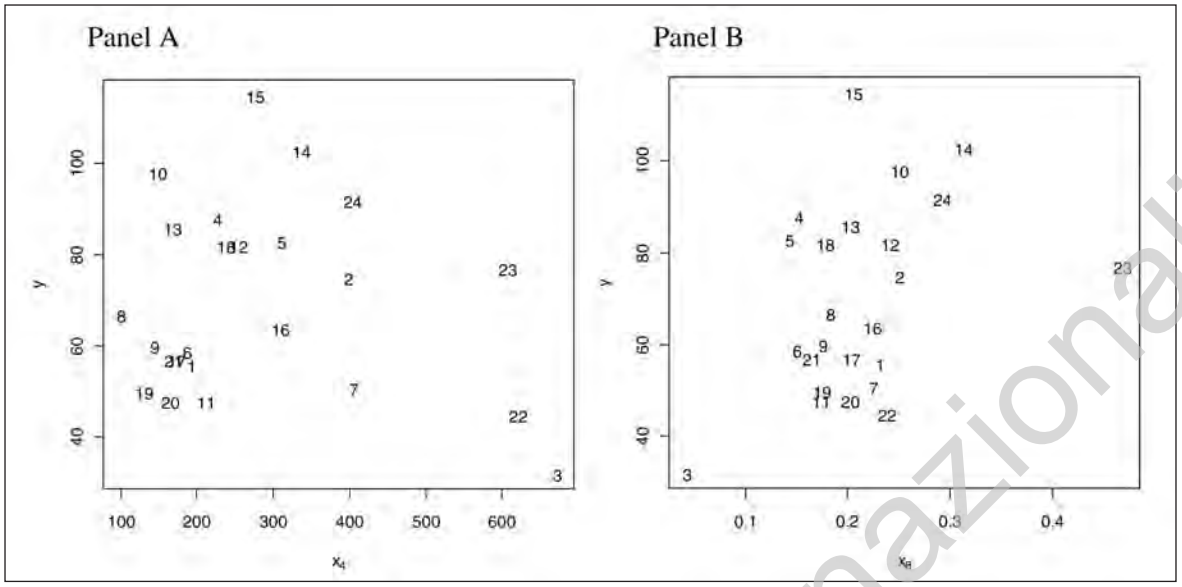

Figure 5. Graphical display of Narula et al. data: response variable versus x_4 (EPIT) (Panel A) and response variable versus x_8 (CELL) (Panel B).

graphs of response versus each of the four regressors were drawn. The plot of y versus x_4 (Panel A of Figure 5) shows that cases 3, 23 (together with 22) lie on the right side of the figure, rather far from the bulk of the remaining data. The plot of y versus x_8 (Panel B of Figure 5) shows that case 3 is the most extreme on the left, whereas case 23 is the most extreme on the right side.

The other two plots, y vs. x_2 and y vs. x_{13}, seem uninformative with regard to cases 3 and 23. This suggests that the validity of values x_4 and x_8 should be investigated for cases 3 and 23.

In a medical context, looking for a therapy is the next step after a diagnosis; likewise in the present context it is necessary to scrutinise the original dataset and, on the ground of the subject-matters knowledge, assess whether outlying cases are valid or not valid. If it is believed that the model is correct, in the first case it seems sensible to down-weight outliers and leverage points to accommodate the model (this is the typical strategy of C-M or of any other robust regression method), in the second case one can delete the non valid observations and refit the model by means of LS regression. On the other hand, when the correctness of the model is questionable or validity assumptions of LS regression are lacking a great deal of remedies are available as shown in many books on regression analysis or on related topics (see for instance Draper and Smith (6), Weisberg (7) and Atkinson and Riani (8)).

It is now interesting to see how Narula et al. face the problem: in their initial analysis a model by LS regression was fitted and the authors state that "…an analysis of the residuals identified two outliers (cases 11, 15) and a leverage point.…Further investigation confirmed that these two outliers were valid observations. Therefore we decided not to discard them and to use a more robust procedure than least squares to estimate the parameters of the model. The minimum sum of absolute errors (MSAE) regression is one such alternative." However they do not take into consideration the influential points any longer; this is surprising as it is well known that the MSAE regression does not protect against outlying x ((1), pg 11). This shortcoming could have been avoided through the application of the robust C-M regression procedure.

Final comments

The presence of several outlying observations (in y co-ordinate and/or in x co-ordinates) is a challenge for any robust method. Optimal robust methods, which are theoretically complex and computer intensive, aim at having a 50% breakdown point, i.e. to be resistant for a contamination of the 50% of the dataset.

Chatterjee and Mächler (2) argue that the majority of practicing statisticians would not fit a model to data which are contaminated at the level of 50% without pre-screening procedure. Therefore they pursue a humbler objective for their procedure with a 20-25% breakdown point, because they are convinced that "most statisticians would be satisfied with a procedure which would be robust against 20 to 25 percent contamination." They test their method on a wide series of datasets previously used to study the performance of other robust regression methods. At the end of the assessment process they conclude that the C-M algorithm could be adopted by all practicing statisticians. Implicitly the limit of 20-25 percent contamination applies also to the diagnostic tool suggested in section "A diagnostic tool" of this paper.
When "pathological cases" are scrutinised and found to be valid, the diagnostic tool presented here offers a further advantage; in fact the subsequent steps of the statistical analysis flow into the mainstream of statistics as the C-M robust regression approach is a weighted linear regression method.

Acknowledgments
We are very grateful to Professor Chatterjee for sending us his program SCReg0, which made possible the realization of the present paper.

References

1. Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. New York: John Wiley and Sons, 1987.
2. Chatterjee S, Mächler M. Robust regression: a weighted least squares approach. Communication in statistics. Theoretical Method 26(6), 1381-1394, 1997
3. Daniel C, Wood FS. Fitting equations to data. New York: John Wiley and Sons, 1971
4. Ryan T. Modern regression methods. New York: John Wiley and Sons, 1997.
5. Narula SC, Saldiva PHN, Andre CDS, Elian SN, Favero Ferreira A, Capellozzi V. The minimum sum of absolute errors regression: a robust alternative to the least squares regression. Statistics in medicine 1999; 18: 1401-1417.
6. Draper NR, Smith H. Applied regression analysis. New York: John Wiley and Sons, 1981.
7. Weisberg S. Applied linear regression. New York: John Wiley and Sons, 1980.
8. Atkinson A, Riani M. Robust diagnostic regression analysis. New York: Springer-Verlag, 2000

APPENDICE

Table 1 A. Pilot Plant Data Set from Daniel and Wood (1971, pg 45).

ID	*y*	*x*
1	76	123
2	70	109
3	55	62
4	71	104
5	55	57
6	48	37
7	50	44
8	66	100
9	41	16
10	43	28
11	82	138
12	68	105
13	88	159
14	58	75
15	64	88
16	88	164
17	89	169
18	88	167
19	84	149
20	88	167

Table 2 A. Dataset from Ryan (pg 372)

ID	*y*	*x*
1	8	2
2	9	3
3	8.3	4
4	8	5
5	9	5
6	9.5	6
7	11	6
8	10.8	7
9	11.8	7
10	12	8
11	10.6	8
12	14.8	8
13	17.3	12.9
14	15.9	15
15	14.6	17.3
16	41.1	14
17	42	15
18	41	16

Table 3 A. Dataset from Ryan (pg 375).

ID	y	x_1	x_2	x_3
1	192,435	11	32	68
2	140,125	14	28	50
3	131,522	16	39	49
4	167,656	18	35	61
5	181,396	23	36	63
6	181,191	22	42	66
7	131,443	21	40	48
8	169,829	20	41	64
9	182,263	16	35	65
10	137,330	12	33	53
11	153,592	19	29	50
12	137,975	17	46	59
13	200,989	15	28	68
14	142,088	13	31	52
15	163,871	12	38	67
16	162,010	24	41	62
17	182,666	25	38	65
18	139,104	20	41	53
19	137,298	15	45	59
20	131,868	10	44	61
21	152,650	13	35	58
22	138,864	15	26	50
23	139,007	22	31	52
24	180,265	24	48	68
25	232,376	31	42	79
26	194,843	32	58	78
27	203,432	34	60	81
28	239,322	31	63	82
29	230,985	33	59	80
30	247,940	35	66	85

Table 4.A Dataset from Narula et al.

ID	y	x_2	x_4	x_8	x_{13}
1	56	64	192,405	0,231	4
2	75	39	398,588	0,251	0
3	32	39	671,674	0,043	0
4	88	69	227,424	0,153	0
5	83	41	310,136	0,143	0
6	59	42	187,597	0,150	3
7	51	32	405,836	0,225	0
8	67	45	100,237	0,183	1
9	60	53	144,290	0,176	2
10	98	46	149,187	0,251	0
11	48	44	211,614	0,174	0
12	82	44	254,398	0,242	0
13	86	57	167,728	0,203	0
14	103	49	337,145	0,313	0
15	115	65	276,864	0,206	0
16	64	26	309,206	0,224	0
17	57	46	173,373	0,204	3
18	82	28	238,277	0,178	0
19	50	52	130,308	0,175	3
20	48	49	165,546	0,203	4
21	57	32	168,547	0,165	2
22	45	57	621,861	0,238	2
23	77	72	607,268	0,468	2
24	92	57	404,735	0,293	0