Detecting outliers and/or leverage points: a robust two-stage procedure with bootstrap cut-off points

Ettore Marubini[1], Annalisa Orenti[1]

ABSTRACT

BACKGROUND: Identification and assessment of outliers, have a key role in Ordinary Least Squares (OLS) regression analysis. This paper presents a robust two-stage procedure to identify outlying observations in regression analysis.

METHODS: The exploratory stage identifies leverage points and vertical outliers through a robust distance estimator based on Minimum Covariance Determinant (MCD). After deletion of these points, the confirmatory stage carries out an OLS analysis on the remaining subset of data and investigates the effect of adding back in the previously deleted observations. Cut-off points pertinent to different diagnostics are generated by bootstrapping and the cases are definitely labelled as good-leverage, bad-leverage, vertical outliers and typical cases.

RESULTS: This procedure is applied to four examples taken from the literature and it is effective in rightly pinpointing outlying observations, even in the presence of substantial masking.

CONCLUSIONS: This procedure is able to identify and correctly classify vertical outliers, good and bad leverage points, through the use of jackknife-after-bootstrap robust cut-off points. Moreover its two-stage structure makes it interactive and this enables the user to reach a deeper understanding of the dataset main features than resorting to an automatic procedure.

Key words: leverage, outlier, Minimum Covariance Determinant, bootstrap, prediction, robust distance

(1) Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy

CORRESPONDING AUTHOR: Annalisa Orenti, Department of Clinical Sciences and Community Health, University of Milan, Via Vanzetti 5, 20133, Milan, Italy. Phone: 02 2390 3282. E-mail: annalisa.orenti@unimi.it

DOI: ***10.2427/9094***

INTRODUCTION

Identification and assessment of outliers, i.e., observations lying far away from the majority of cases in the dataset and probably not following the postulated error model, have a key role in Ordinary Least Squares (OLS) regression analysis. To this end, single-case diagnostic statistics are of common use and their computation is automatically offered to the users by almost all commercial statistical packages. This approach has the advantage of being easy to apply and interpret, though it can fail when clusters of outliers are present and the phenomenon of masking may occur i.e. when outlying cases cannot be detected due to the presence of some extreme outliers in the dataset. The formulae for multiple-case diagnostics are likewise well known, but the relevant diagnostic statistics are little adopted because of the difficulty to identify the cases to be flagged [1]. Alternatively, as advocated by Rousseeuw and Leroy [2], one can resort to robust regression methods, which try to device estimators that are not so strongly affected by outliers as the OLS estimator: it is then by looking at the results from robust regression that outliers may be pinpointed.

An extensive literature on robust fitting methods is available. An exhaustive presentation of these is given by Maronna et al. [3]. Furthermore over the last two decades there has been a great deal of relevant contributions in the important area of outlier identification (see for instance [4-6]).

The two stage procedure, whose main features are going to be presented in this paper, contributes to outlier identification topic. The first stage, exploratory, relies on the robust Minimum Covariance Determinant (MCD) estimator and the second one, confirmatory, relies on the OLS method. The procedure, thus, combines the resistance of the robust estimator, when multiple outliers are present, with the efficiency of the OLS method once the outliers have been detected and deleted. The idea of a two stage procedure is not new (see for instance [7]), nevertheless in our opinion our work qualifies as innovative because of the joint presence of the following items: i) robust distance to measure how far a given case lies from the centroid of all data points in a multivariate space; ii) reliability in defining criteria suitable for labelling cases in more informative terms than outliers; iii) generation of bootstrap cut-off points of single-case diagnostics.

A further important step is needed to decide which cases to delete. The labelled cases should be scrutinised for their validity on the ground of the subject-matter knowledge. In fact, we think that outlier identification is not only important to avoid possible distortions of the statistical model estimates, but it should also be considered as a goal in itself since outliers may be of fundamental interest for the message they are conveying. The procedure is simple to implement and its results are easy to interpret; this makes our approach particularly appealing to practitioners. In this paper methodological aspects of the procedure are outlined and its performance is explored through four examples taken from the literature.

METHODS

Linear model and validity assumptions for OLS method are reported in Table 1. Consider the model [1]: $\mathbf{y}$ is the n response vector; $\mathbf{X}$ is the $(n\times(p+1))$ matrix of regressors; $\boldsymbol{\beta}$ is the $(p+1)$ parameter vector including the intercept, to be estimated, and $\boldsymbol{\varepsilon}$ is the n unknown vector of random errors. The letter G indicates Gaussian distribution, the letter n the number of observations and the letter p the number of explanatory variables (carriers), so that p+1 is the number of regressors, $\mathbf{X}^*$ indicates the matrix of explanatory variables excluding the intercept and $I_{(n)}$ the identity matrix.

Table 1. Linear model and validity assumptions for OLS method

Model	$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ [1]	$y_i=\beta_0+\sum_{j=1}^{p}x_{ij}^*\beta_j+\varepsilon_i$ [2]
Assumptions	In observational studies (with random carriers):	
	$\mathbf{x}_i^*\sim\mathbf{G}_\mathbf{p}(\boldsymbol{\mu}_{\mathbf{x}^*},\boldsymbol{\Sigma}_{\mathbf{x}^*})$ [3]	
	In experimental and observational studies:	
	$\boldsymbol{\varepsilon}\sim\mathbf{G}_\mathbf{n}\left(\mathbf{0},\sigma^2\mathbf{I}_{(\mathbf{n})}\right)$ [4] $\rightarrow$ $\mathbf{y}\sim\mathbf{G}_\mathbf{n}(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_{(\mathbf{n})})$	$\varepsilon_i\sim G(0,\sigma^2)\ \ \forall i=1,2,\dots,n$

Exporatory stage

Coherently with assumptions [3] and [4] two types of outlying observations in regression analysis can be identified. First, the i-th case may present a large difference between the regressors' vector $\mathbf{x_i}$ and the centroid of the x-data; in other words $\mathbf{x}_\mathrm{i}$ may be an outlier in the (p+1)-dimensional space spanned by the columns of the $\mathbf{X}$ matrix; such a case will be referred to as leverage point. Second, the i-th case may present a big difference between the response y_i and the mean $\mathbf{x}_i'\boldsymbol{\beta}$ predicted by the model. Such a point will be referred to as regression outlier.

Combining these two modalities of being an outlying observation enables labelling the cases as in Table 2:

Table 2. Outlying observation labelling

		Leverage	
		no	yes
Regression outlier	no	typical (bulk of data)	good leverage
	yes	vertical outlier	bad leverage

The goal of the procedure is to label cases reliably. Note that: "...if $(\mathbf{x}_i, y_i)$ does fit the linear relation it will be called a good leverage point, because it improves the precision of the regression coefficients" [8]; therefore searching for cases to be deleted concerns identifying both vertical outliers and bad leverage points.

It is known that $h_{ii} = x_i'(\mathbf{X}'\mathbf{X})^{-1}x_i$ is a widely used measure of leverage [2]. Its role in measuring the distance of the i-th case from the bulk of the data in the space of regressors is made clear by the following equation: $h_{ii} = \frac{1}{n} + \frac{MD_i^2}{n-1}$, where MD_i^2 is the squared Mahalanobis distance: a standardized form of squared distance between $\mathbf{x}_i^*$ ([2], Table 1) and the centroid $\bar{\mathbf{x}}_n^*$. Namely:

$$MD_i = \sqrt{(\mathbf{x}_i^* - \bar{\mathbf{x}}_n^*)'\left(\mathrm{Cov}_n(\mathbf{X}^*)\right)^{-1}(\mathbf{x}_i^* - \bar{\mathbf{x}}_n^*)}$$

where $\bar{\mathbf{x}}_n^*$ is the (p×1) vector of sample means and $\mathrm{Cov}_n(\mathbf{X}^*)$ is the (p×p) sample covariance matrix of the random carriers.

In observational studies it is assumed that $\mathbf{x}_i^*$ is distributed according to a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_{x^*}$ and covariance matrix $\boldsymbol{\Sigma}_{x^*}$; accordingly MD_i^2 is approximately distributed like a χ^2 r.v. with p degrees of freedom (d.f.).

Unfortunately, the two statistics $\bar{\mathbf{x}}_n^*$ and $\mathrm{Cov}_n(\mathbf{X}^*)$ are not robust; in fact the multivariate vector of outlying cases will tend to change $\bar{\mathbf{x}}_n^*$, to deflate correlations among carriers and perhaps to inflate the corresponding variances. These will in general decrease the Mahalanobis distance. Moreover, owing to the phenomenon of masking, using h_{ii} and equivalently MD_i as measures of leverage, implies the risk to fall in "... a situation where the diagnostics used to identify the high leverage points are undermined by the very points they are designed to detect." [9]. To bypass these shortcomings one should resort to robust estimates of mean and scatter as those given, for instance, by the MCD estimator. The reader is referred to the original paper of Rousseeuw and Van Driessen [10] for a detailed presentation of the algorithm for MCD computation. It suffices to say here: "The MCD objective is to find h observations (out of n) whose classical covariance matrix has the lowest determinant. The MCD estimate of location is then the average of these h points, and the MCD estimate of scatter is their covariance matrix." [10].

The robust Mahalanobis distance suggested by Rousseeuw and van Zomeren [8] results to be:

$${}_xRD_i = \sqrt{(\mathbf{x}_i^* - \bar{\mathbf{x}}_{\mathbf{MCD}}^*)'\left(\mathrm{Cov}_{MCD}(\mathbf{X}^*)\right)^{-1}(\mathbf{x}_i^* - \bar{\mathbf{x}}_{\mathbf{MCD}}^*)}$$

where $\bar{\mathbf{x}}_{\mathbf{MCD}}^*$ and $\mathrm{Cov}_{MCD}(\mathbf{X}^*)$ are MCD high breakdown point estimates of $\boldsymbol{\mu}_{x^*}$ and $\boldsymbol{\Sigma}_{x^*}$, the multivariate means vector and covariance matrix respectively. Their breakdown point is: $\frac{\left[\frac{n-p+1}{2}\right]}{n}$, where [w] denotes the integer part of w [11]. The scatter estimate is adjusted for consistency and small sample sizes according to Pison et al. [12]. Moreover such a robust distance gives the best protection against the masking effect.

After showing that ${}_xRD_i^2$ is an appropriate measure of leverage, a critical value is needed so that cases with ${}_xRD_i^2$ greater than it will be identified as leverage cases. This critical values is based on the asymptotic distribution of the Mahalanobis distance in terms of χ^2; the critical value is then $\chi^2_{p;1-\alpha}$ where the number of d.f. equal to p and α is chosen as 0.01 to have a low false positive rate. The number of pinpointed leverage cases is denoted by m_x. This number includes both good and bad leverage cases.

So far attention was focused exclusively on leverage points; now we also wish to take into account observations outlying in the response. Since in observational studies both response and carriers are random, we observe: $\mathbf{z}_i' = (y_i, \mathbf{x}_i^*)'$, $1 \leq i \leq n$.

These (p+1)-dimensional vectors $\mathbf{z}_i'$ are assumed to be i.i.d. multivariate Gaussian variables with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Considerations like those regarding the sample matrix $\mathbf{X}^*$ apply also to the matrix $\mathbf{Z}$ of size $(n\times(p+1))$. Therefore, the MCD estimator can be adopted to identify m outlying observations on $\mathbf{Z}$ which include bad leverage points, good leverage points and vertical outliers. The MCD breakdown point is now $\frac{\left[\frac{n-p}{2}\right]}{n}$ and the critical value is $\chi^2_{(p+1);(1-\alpha)}$.

The two numbers m_x and m, together with the identification number (ID) of flagged cases are the basic results of the exploratory stage of our procedure. Moreover m is used to split the original dataset in two provisional subsets of m outlying observations and n-m possible cases belonging to the bulk of the data.

Functions *covMcd* in *robustbase* package and *mahalanobis* in *stats* package of R software were used to compute the robust distances.

Confirmatory analysis

Checking the results of the exploratory analysis has a twofold aim: looking at possible further cases which might be considered outliers and assessing the effect of adding back in the individual cases suggested as outlying observations in the exploratory analysis.

With regard to the first objective an OLS regression analysis is carried out on the subset of size (n-m). Following the traditional method, pertinent single-case diagnostics are computed: h_{ii} (leverage), studentized deletion residual and modified Cook's distance. The choice of these statistics is justified by Atkinson [7] and the pertinent formulae are given in the first part of the Appendix. This arm of the procedure hereafter will be referred to as "OLS diagnostics".

Concerning the second objective, the subset of size m is thought of as an "external" subset; for each of these cases, predicted values are computed by the estimate **b** of $\boldsymbol{\beta}$ (obtained in the OLS regression analysis on n-m cases) and the analogues of the three previously mentioned diagnostics are computed. They are derived in the second part of the Appendix. This arm of the procedure hereafter will be referred to as "OLS prediction".

Since we believe that it is important not to underestimate the size of the data bulk, the cut-off points of the single-case diagnostics used in the confirmatory stage were chosen at a probability level α=0.05 in the prediction arm and α=0.01 in the diagnostics arm respectively.

Note that cases labelled as good leverage points by the OLS prediction arm should be added back in to the provisional bulk of n-m typical data, whereas the cases labelled vertical outliers and bad leverage points in the OLS diagnostic arm should be added to the provisional subset of m outlying observations. The final labelling of cases is now obtained: the vertical outliers and bad leverage points now identified constitute the final subset of m_D outlying observations to be scrutinised.

An ad hoc function in R software was written by one of the authors (A.O.) to process the data in the confirmatory analysis of our procedure and it is available on request.

Bootstrap cut-off points

As the studentised deletion residual given by (A.2) of the Appendix has a Student's t-distribution, the pertinent cut-offs may correspond to the $t_{0.025}$ and $t_{0.975}$ percentiles from the relevant t-distribution. However the suitability of the critical points depends on the adequacy of the Gaussianity assumption for linear model errors. If these fail to be reasonably Gaussian, the postulated t-distribution may not properly reflect the true behaviour of the studentized residuals. Furthermore, with regard to the measuring influence diagnostics and particularly Cook's distance note that: "based on large sample theory and rough approximations, the traditional cut-offs may not adequately allow for small sample sizes, or cases where model error distributions exhibit significant skewness or heavy tails." [6]. The alternative resampling method may offer a genuine improvement; hence the confirmatory stage relies upon single term diagnostics bootstrapped cut-off points.

Using a "standard" bootstrap approach to generate sampling distributions of diagnostics could be naive, because of the chances that a highly anomalous case has to be included once or more in several resamples. Alternatively one can resort to Efron's [13] jackknife-after-bootstrap resampling scheme. The underlying idea is "...that within each jackknife-after-bootstrap subgroup of resamples (groups indexed via the missing case), delete-1 diagnostics for each resample can be calculated and the bootstrap distribution of the relevant diagnostic approximated using resamples not contaminated by the point under consideration." [6]. To be explicit let's consider, for instance, the i-th case: it is easy to see that the probability this case does not appear in a bootstrap sample of size n is $(1\text{-}n^{-1})^{n}$. As n increases $(1 - n^{-1})^{n} \rightarrow e^{-1} \cong 0.3679$. Thus among, say, B=5000 resamples, 1839 are expected not to include the i-th case. Each of these bootstrap samples has size n; (n-1) cases of the original $(\mathbf{X}, \mathbf{y})$ dataset, except the i-th one, can be sorted by sampling with replacement in each resample.

Refer now to the first of the 1839 bootstrap samples and focus on a particular diagnostic measure, for instance the studentized deletion residual (s.d.r.). We define t_i the s.d.r. computed for the i-th case in the original dataset $(\mathbf{X}, \mathbf{y})$ and t_k^*, (k=1, 2, …, n) the s.d.r. computed on each unit of the bootstrap sample $(\mathbf{X}^*, \mathbf{y}^*)$. Of course all these n t_k^* result in being independent of the i-th case. The same computation can be done for the remaining 1838 bootstrap samples so that a total of $n \cdot 1839$ (in general $n \cdot B \cdot e^{-1}$) values of t_k^* are obtained. It appears sensible to think that these $n \cdot B \cdot e^{-1}$ t_k^* values generate a "null" bootstrap distribution of the s.d.r. t_i under the hypothesis that the i-th case is not influential. By arranging the t_k^* in increasing order, it is straightforward to get bootstrap 2.5% and 97.5% cut-off points for t_i as the 2.5-th and 97.5-th percentiles of this distribution respectively. These percentiles are indicated $C_{0.025}$ and $C_{0.975}$ respectively as column headings of the Tables showing the results of the second stage analysis. Similarly the boostrap cut-off points can be produced for the s.d.r. of the (n-1) remaining cases of the original dataset. All this justifies the use of a jackknife-after-bootstrap approach to generate the cut-off points of the dataset.

By default our procedure generates B=5000 bootstrap samples.

With regard to the prediction cut-off computation via jackknife-after-bootstrap, the procedure takes advantage of the relationship between diagnostics shown in the Appendix. The value of $\frac{n-m}{p+1} > 6$ is adopted as a rule of thumb to compute the bootstrap cut-offs in the confirmatory analysis; otherwise traditional cut-offs, as shown in the Appendix, are used. The complete procedure can be sketched as shown in Table 3.

Table 3. Brief description of the procedure.

Exploratory analysis	
Step 1	Fit the postulated model to the complete dataset by OLS method. Estimate regression coefficients, error variance (s^2) and coefficient of determination (R^2).
Step 2	Matrix **Z**. What about outlying observations? By MCD robust distance obtain: a) m = number of outlying observations; b) list of ID numbers for outlying observations.
Step 3	Matrix $\mathbf{X}^*$. What about leverage points? By MCD robust distance obtain: a) m_x = number of leverage points; b) list of ID numbers for leverage points.
Step 4	Compare results of Steps 2 b) and 3 b) and provisionally label the cases as vertical outliers and good or bad leverage points.
Confirmatory analysis	
Step 5	Among the m cases, what about good leverage points (g) to be added back in to the bulk of data? OLS prediction arm.
Step 6	Among the (n-m) cases, what about any more bad-leverages and/or vertical outliers (v)? OLS diagnostics arm.
Step 7	Definitely label outlying observations as bad leverage points, good leverage points and vertical outliers.
Step 8	Scrutinize the validity of m_D=(m-g+v) cases on the ground of the subject-matter knowledge.

RESULTS

The performance of our procedure will be illustrated by means of four examples. The first one, known as "Belgian phone calls", is based on a real dataset and as far as we know it is one of the few sets for which also the results of the data scrutiny are available in the literature. The following three examples are artificial datasets, so that it is immediate to assess how our procedure works, since the correct answers are known. Simple linear regression models were fitted to the first two datasets and multiple regression models to the remaining ones.

Belgian phone calls data

This example is taken from Rousseeuw and Yohai [14] and it concerns the time series of the total number of international phone calls made in Belgium between 1950 and 1973. The data are plotted in Figure 1.

Exploratory analysis

Matrix Z: $\chi^2_{2;\,0.99}$= 9.21; m=8, years ’63, ‘64, ‘65, ‘66, ‘67, ‘68, ‘69, ‘70.

Matrix X^*: $\chi^2_{1;\,0.99}$ = 6.63; mx = 0.

Provisional labelling: vertical outliers.

Confirmatory analysis

The 8 cases previously flagged are deleted.

OLS prediction arm: the results are reported in Table 4. For all cases the studentized prediction residual and modified Cook’s distance exceed the relative cut-offs, whereas $\tilde{h}_{ii}$ do not. Therefore these 8 cases can be definitively labelled vertical outliers and cannot be added back in to the bulk.

OLS diagnostic arm (data not shown): no further cases are pinpointed.

A scrutiny is required to assess whether any gross error contaminated the time series. Concerning this, Rousseeuw and Leroy [2], state that: “...it turned out that from 1964 to 1969 another recording system was used, giving the total number of *minutes* of these calls. The years 1963 and 1970 are also partially affected because the transitions did not happen exactly on New Year’s Day, so the number of calls of some months were added to the number of minutes registered in the remaining months!”.

On this ground it appears sensible to delete the 8 cases previously mentioned and to consider the remaining ones belonging to the bulk of homogeneous data. This justifies the use of OLS method to estimate the regression line on the 16 homogeneous data (OLS-16 in Figure 1).

Figure 1. Scatter plot of Belgian phone calls (tens of millions) together with the regression lines computed on the whole dataset (OLS-24) and on the subset after deleting outlying observations (OLS-16)

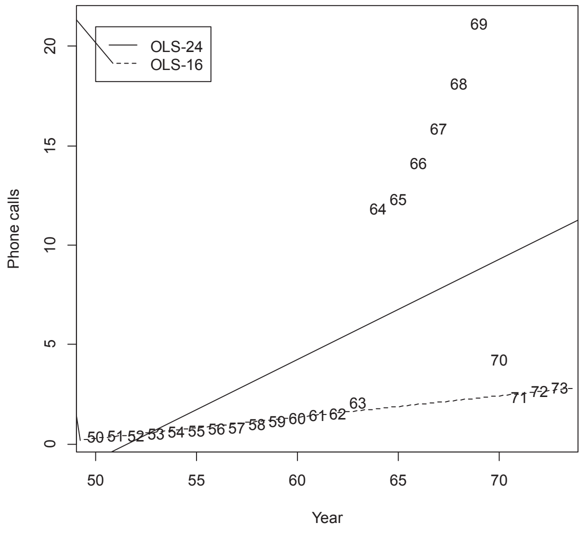

Table 4. Belgian phone calls data; results of the confirmatory analysis: OLS Prediction arm.

	Studentized prediction residual			$\tilde{h}_{ii}$		Modified Cook's distance	
year	$C_{0.025}$	Estimate	$C_{0.975}$	Estimate	$C_{0.95}$	Estimate	$C_{0.95}$
63	-1.894	**4.478**	2.503	0.082	0.416	**3.267**	1.771
64	-1.868	**99.962**	2.502	0.093	0.411	**77.313**	1.753
65	-1.880	**103.188**	2.494	0.107	0.416	**84.898**	1.751
66	-1.898	**118.928**	2.478	0.123	0.415	**104.189**	1.750
67	-1.883	**133.338**	2.505	0.142	0.420	**124.286**	1.768
68	-1.897	**153.112**	2.502	0.163	0.415	**151.556**	1.747
69	-1.893	**179.001**	2.515	0.186	0.420	**187.662**	1.756
70	-1.884	**17.556**	2.510	0.212	0.415	**19.436**	1.758

Simulated dataset

This dataset was created following Rousseeuw [15]. A bulk of 30 "homogeneous" observations was generated according to the linear relationship $y_i = 2 + x_i + \varepsilon_i$ where: x_i is uniformly distributed on (1,4) and $\varepsilon_i \sim G(0, 0.04)$. A cluster of 20 observations was added, having a spherical bivariate Gaussian distribution with mean vector (7; 2)' and variances 0.25. This yielded a high level (40%) of contamination in the pooled sample. Figure 2 reports the simulated dataset together with the line fitted to the whole sample (OLS-50) and to the bulk (OLS-30). According to Figure 2 of Rousseeuw and van Zomeren [8] the 20 contaminating observations are bad leverage points. Table 5 gives the raw data together with the robust distances ${}_{x}RD_i^2$ and RD_i^2 (see section 2.1); the values exceeding the relative χ^2 threshold (column headings) are highlighted in bold.

Figure 2. Simulated dataset; scatter plot together with the regression lines computed on the whole dataset (OLS-50) and on the bulk of homogeneous observations (OLS-30).

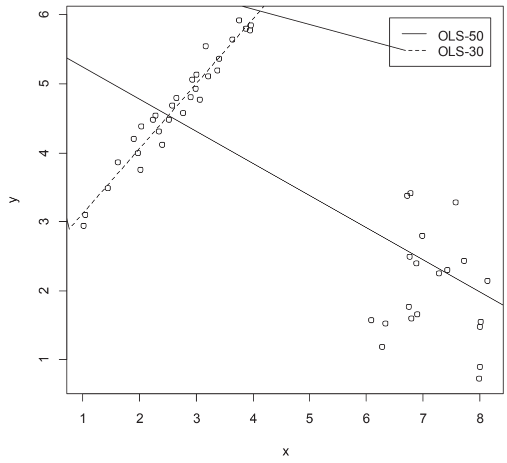

Table 5. Simulated dataset; raw data and robust distances.

ID	x	y	${}_{x}RD_i^2$ (6.63)	RD_i^2 (9.21)
1	1.44	3.486	0.38	0.885
2	1.97	3.998	0.165	0.29
3	3.965	5.85	0.149	1.007
4	2.515	4.476	0.036	0.064
5	2.346	4.31	0.066	0.122
6	1.047	3.095	0.597	1.552
7	2.019	3.755	0.149	1.423
8	2.575	4.688	0.028	0.099
9	1.621	3.865	0.297	0.957
10	3.208	5.111	0.007	0.252
11	1.009	2.935	0.62	1.977
12	3.376	5.194	0.023	0.589
13	3.41	5.369	0.027	0.319
14	2.995	4.921	0	0.123
15	2.24	4.476	0.089	0.543
16	2.034	4.382	0.145	1.241
17	3.062	4.764	0.001	1.096
18	1.904	4.195	0.187	0.951
19	3.171	5.547	0.005	1.882
20	3.635	5.645	0.065	0.56
21	2.775	4.579	0.007	0.51
22	3.753	5.916	0.091	1.163
23	2.647	4.796	0.019	0.196
24	2.998	5.138	0	0.293
25	2.927	5.065	0.001	0.248
26	2.287	4.545	0.078	0.637
27	2.399	4.118	0.056	1.189
28	3.952	5.774	0.145	1.125
29	2.904	4.811	0.001	0.145
30	3.878	5.803	0.124	0.846
31	6.901	1.661	2.409	**584.478**
32	6.766	2.491	2.245	**437.87**
33	7.577	3.286	3.314	**438.031**
34	7.273	2.247	2.889	**548.664**
35	6.876	2.4	2.378	**466.486**
36	8.017	1.552	3.98	**796.479**
37	6.788	1.596	2.271	**577.152**
38	7.426	2.304	3.099	**563.27**
39	8.132	2.148	4.164	**707.086**
40	7.99	0.72	3.937	**957.616**
41	6.27	1.182	1.694	**562.8**
42	7.998	1.48	3.95	**806.72**
43	6.782	3.418	2.264	**318.897**
44	6.079	1.572	1.502	**474.043**
45	6.988	2.8	2.517	**425.157**
46	6.752	1.767	2.229	**544.121**
47	6.337	1.526	1.764	**518.84**
48	7.995	0.893	3.945	**922.915**
49	6.716	3.378	2.186	**315.888**
50	7.722	2.437	3.526	**588.687**

Exploratory analysis

Matrix Z: $\chi^2_{2;\,0.99}$= 9.21; m=20, cases 31-50.

Matrix X*: $\chi^2_{1;\,0.99}$ = 6.63; mx = 0.

Provisional labelling: vertical outliers.

Confirmatory analysis

The 20 cases previously flagged are deleted.

OLS prediction arm (data not shown): for all cases the studentized prediction residual, $\tilde{h}_{ii}$ and the modified Cook's distance exceed the relative cut-offs. Therefore these 20 cases can be definitively labelled bad leverage points, as expected by construction, and they cannot be added back in to the bulk.

OLS diagnostic arm (data not shown): studentized deletion residual of case 19 slightly exceeds the upper cut-off. Coherently with the 99% level of the cut-offs, the case can be considered false positive.

Note that in example Belgian phone call data the x-values were homogeneous on the whole dataset, while the y-values were obtained by two different recording systems. Here on the contrary the contaminated set differs from the bulk for both x and y values. This justifies the different labelling in the two examples, even though the results of the exploratory stages appeared to be similar.

Hawkins-Bradu-Kass data

These data are taken from Hawkins et al. [16]. “This is a much referenced artificial dataset consisting of 75 cases and 3 carriers; it is known to be troublesome in terms of masking and swamping and new methods for detecting outliers have been tested on the dataset” [17]. The data were generated so that the first ten cases were bad leverage points, and the next four cases were good leverage points; the remaining 61 constitute the bulk of homogeneous data.

Exploratory analysis

Matrix Z: $\chi^2_{4;\,0.99}$= 13.28; m=14, cases 1-14.

Matrix X*: $\chi^2_{3;\,0.99}$= 11.34; mx=14, cases 1-14.

Provisional labelling: bad leverage points.

Confirmatory analysis

The 14 cases previously flagged are deleted.

OLS prediction arm: the results are reported in Table 6. All the three diagnostics of case 1-10 exceed the respective cut-offs. Therefore these cases are labelled bad leverage points and cannot be added back in to the bulk of typical cases. As far as cases 11-14 are concerned, the studentized prediction residuals result to be within the cut-off intervals, whereas both $\tilde{h}_{ii}$ and the modified Cook’s distances are greater than their respective cut-offs, hence these 4 cases are labelled good leverage points and may be added back in to the bulk.

OLS diagnostic arm (data not shown): no further cases are pinpointed.

Note that the definite labelling by our procedure matches Hawkins Bradu Kass [16] specifications.

Table 6. Hawkins-Bradu-Kass data; results of the confirmatory analysis: OLS Prediction arm.

	Studentized prediction residual			$\tilde{h}_{ii}$		Modified Cook’s distance	
ID	$C_{0.025}$	Estimate	$C_{0.975}$	Estimate	$C_{0.95}$	Estimate	$C_{0.95}$
1	-1.675	**5.353**	1.811	**14.464**	0.118	**19.541**	1.367
2	-1.669	**5.442**	1.813	**15.223**	0.117	**19.900**	1.370
3	-1.677	**5.319**	1.811	**16.967**	0.118	**19.511**	1.367
4	-1.678	**4.889**	1.815	**18.015**	0.118	**17.965**	1.368
5	-1.675	**5.145**	1.812	**17.381**	0.118	**18.886**	1.370
6	-1.676	**5.314**	1.812	**15.611**	0.117	**19.445**	1.369
7	-1.676	**5.647**	1.815	**15.705**	0.117	**20.667**	1.367
8	-1.680	**5.589**	1.814	**14.817**	0.118	**20.421**	1.370
9	-1.673	**5.040**	1.812	**17.034**	0.117	**18.492**	1.366
10	-1.673	**5.308**	1.813	**15.974**	0.118	**19.438**	1.368
11	-1.669	0.946	1.816	**22.389**	0.117	**3.496**	1.370
12	-1.677	0.902	1.822	**24.026**	0.117	**3.336**	1.367
13	-1.673	1.197	1.815	**22.732**	0.117	**4.422**	1.370
14	-1.680	0.872	1.814	**28.158**	0.118	**3.234**	1.370

Hawkins two-system data

This dataset is taken from Billor et al. [18]. The authors state: “Doug Hawkins (personal communication) has constructed a dataset in which there are 32 observations with 4 predictor variables x_1, x_2, x_3, x_4 and y. The data come from 2 regression systems, a subset of the observations (2,6,10,14,18,22,26,30) is related to x_2 only, while the remaining observations are related to the other 3 variables. The multiple correlation for the two regression systems is very high, approximately 0.97.”

Exploratory analysis

Matrix Z: $\chi^2_{5;\,0.99}$= 15.09; m=8: cases 2, 6, 10, 14, 18, 22, 26, 30.

Matrix X*: $\chi^2_{4;\,0.99}$= 13.28, mx=8: cases 2, 6, 10, 14, 18, 22, 26, 30.

Provisional labelling: bad leverage points.

Confirmatory analysis

The 8 cases previously flagged are deleted.

OLS prediction arm: the results are reported in Table 7. For all cases the studentized prediction residual, $\tilde{h}_{ii}$ and the modified Cook’s distance exceed the relative cut-offs. Therefore these 8 cases can be definitively labelled bad leverage points, as expected by construction, and they cannot be added back in to the bulk.

OLS diagnostic arm (data not shown): no further cases are pinpointed.

The analysis suggests the presence of two different populations, indeed, by construction, cases 2, 6, 10, 14, 18, 22, 26, 30 are generated by a regression system different from the one generating the bulk of homogeneous data.

Table 7. Hawkins two-system data; results of the confirmatory analysis: OLS Prediction arm.

	Studentized prediction residual			$\tilde{h}_{ii}$		Modified Cook’s distance	
ID	$C_{0.025}$	Estimate	$C_{0.975}$	Estimate	$C_{0.95}$	Estimate	$C_{0.95}$
2	-2.0173	**-14.2379**	1.8914	**1103.07**	0.6404	**27.7422**	1.2996
6	-2.018	**14.2062**	1.8742	**760.708**	0.6445	**27.6749**	1.2992
10	-2.0283	**-13.9295**	1.8704	**162.0509**	0.641	**27.0703**	1.2924
14	-2.0172	**-14.2991**	1.8809	**280.1902**	0.6358	**27.8246**	1.296
18	-2.0284	**14.1681**	1.8923	**171.48**	0.6405	**27.5386**	1.3016
22	-2.0163	**12.7828**	1.8749	**4.4487**	0.6441	**22.5159**	1.2983
26	-2.0375	**-13.8631**	1.8809	**18.0377**	0.6389	**26.3049**	1.2978
30	-2.0218	**14.2915**	1.8812	**582.7356**	0.6363	**27.8354**	1.2989

Performance comparison between our procedure and the lmrob R function

“Leverage plus outliers” automatic labelling is available in the default plot of *lmrob* function for linear robust regression estimator in the *robustbase* package in R software (2.15.2 version). This function estimates the robust MM-estimator for linear regression models [19] and it also gives a plot of scaled residuals versus robust distances computed through MCD estimates of location and scatter.

Comparison of the two procedures is made in terms of number and type of outlying observations. Namely:

Belgian phone calls. *lmrob* function tends to be slightly more conservative, i.e., to pinpoint a smaller number of cases (64-70) to be scrutinized than the number of cases our procedure pinpoints (63-70).

Simulated dataset. A different classification emerges: it deserves to be commented upon. Our procedure, in the confirmatory stage, compares the couple of coordinates (y, x) of each observation from 31 to 50 to the linear regression line estimated on the bulk of 30 data (1-30) and recognizes that each point is far from such a line in terms of both y and x values; hence every observation (31-50) is labelled bad leverage point. On the contrary, *lmrob* function separately assesses the scaled distance of the y-coordinate from the regression line and the scaled distance of the x-coordinate from the centre of x-values of the whole dataset. This reduces the sensitivity of the x distances, hence *lmrob* function classifies cases from 31 to 50 as vertical outliers only.

Hawkins-Bradu-Kass and Hawkins two-system data: the results of the two procedure are overlapping.

DISCUSSION

The two stage procedures suggested by Atkinson [7], Rousseeuw and van Zomeren [8] and Simpson et al. [1] resorted to Least Median of Squares (LMS) estimator in the first stage. Being aware of the shortcomings of LMS [17,9], an affine equivariant estimator of location and scatter [11], like MCD was chosen for the first stage of our procedure.

The regression examples given in the section 3 show the effectiveness of our procedure in pinpointing vertical outliers, good and bad leverage points, even in the presence of substantial masking. In addition to these four examples, we processed several (> 20) real datasets which are reported by Rousseeuw and Leroy [2] and considered difficult in the statistical literature [20]. The identification of outlying observations was similar to that given by different authors, with a trend of our procedure to be slightly more sensitive and to pinpoint a slightly larger number of cases to be scrutinised. Though one might object that certain data configurations could imply the risk of failure for any procedure aiming at identifying multiple outliers, we are confident in the performance of the one we propose. As a matter of fact, in the first stage, i.e., dealing with the original dataset, the high breakdown point MCD estimator is required to identify leverage points and regression outliers by processing the matrices X^* and Z respectively. In the second stage, dealing with the subset of cases obtained after eliminating the extreme and possibly masked cases, the BLU Ordinary Least Squares Estimator is required to accomplish the two tasks of distinguishing between good and bad leverage points [7], as well as indicating the size of the bulk of typical data.

Our procedure is interactive and this enables the practitioner to reach a deeper understanding of the dataset main features than resorting to an automatic procedure; this is helpful also to make decisions during the scrutiny of the data.

Traditional diagnostics of influence measurements (like leverage, DFFITS, DFBETAS, Cook's distance, modified Cook's distance) do not have obvious theoretical threshold values, so that different practical cut-offs have been proposed in the literature. Alternatively our procedure uses bootstrap cut-offs. These compare favourably to the traditional ones particularly when the original data have heavy tails or skewness in their error distribution as shown by Martin and Roberts [6].

Detection of outliers has been carried out under the assumption that the model is correct. Model identification and outlier detection are fundamentally interconnected. After scrutinizing the dataset validity one could question the appropriateness of the model fitted, looking for an alternative one [21], but such a topic is beyond the scope of this note.

Appendix

To make the appendix of practical impact, the computation detail are here shown with reference to the Hawkins two-system example.
n (# cases) = 32, p (# explanatory variables) = 4, p+1 (# regressors) = 5; from section 2.1 **X**=matrix of regressors of size (32×5), $\mathbf{X}^*$ = matrix of explanatory variables (without the intercept) of size (32×4), $\mathbf{Z} = (\mathbf{y}, \mathbf{X}^*)$, still of size (32×5).
Process matrices $\mathbf{X}^*$ and **Z** is processed to obtain MCD estimates of location and scatter and compute ${}_x\mathrm{RD}_i^2$ and RD_i^2 respectively.
Flag m_x = 8 cases (2, 6, 10, 14, 18, 22, 26, 30), as ${}_x\mathrm{RD}_i^2$ of each of them is greater than the cut-off point: $\chi^2_{p+1,\ 1-\alpha} = \chi^2_{4,\ 0.99} = 13.28$.
Flag m = 8 cases (2, 6, 10, 14, 18, 22, 26, 30), as ${}_x\mathrm{RD}_i^2$ of each of them is greater than the cut-off point: $\chi^2_{p+1,\ 1-\alpha} = \chi^2_{5,\ 0.99} = 15.09$.
Delete from the **X** matrix the rows corresponding to the 8 cases flagged in terms of RD_i^2 and call **U** the resulting matrix of size $((n-m)\times(p+1)) = (24\times5)$.

OLS diagnostics arm

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}'_1 = \mathbf{x}'_1 \\ \mathbf{u}'_2 = \mathbf{x}'_3 \\ \mathbf{u}'_3 = \mathbf{x}'_4 \\ \mathbf{u}'_4 = \mathbf{x}'_5 \\ \mathbf{u}'_5 = \mathbf{x}'_7 \\ \mathbf{u}'_6 = \mathbf{x}'_8 \\ \mathbf{u}'_7 = \mathbf{x}'_9 \\ \mathbf{u}'_8 = \mathbf{x}'_{11} \\ \mathbf{u}'_9 = \mathbf{x}'_{12} \\ \mathbf{u}'_{10} = \mathbf{x}'_{13} \\ \mathbf{u}'_{11} = \mathbf{x}'_{15} \\ \mathbf{u}'_{12} = \mathbf{x}'_{16} \\ \mathbf{u}'_{13} = \mathbf{x}'_{17} \\ \mathbf{u}'_{14} = \mathbf{x}'_{19} \\ \mathbf{u}'_{15} = \mathbf{x}'_{20} \\ \mathbf{u}'_{16} = \mathbf{x}'_{21} \\ \mathbf{u}'_{17} = \mathbf{x}'_{23} \\ \mathbf{u}'_{18} = \mathbf{x}'_{24} \\ \mathbf{u}'_{19} = \mathbf{x}'_{25} \\ \mathbf{u}'_{20} = \mathbf{x}'_{27} \\ \mathbf{u}'_{21} = \mathbf{x}'_{28} \\ \mathbf{u}'_{22} = \mathbf{x}'_{29} \\ \mathbf{u}'_{23} = \mathbf{x}'_{31} \\ \mathbf{u}'_{24} = \mathbf{x}'_{32} \end{bmatrix} \qquad {}_{\mathrm{U}}\mathbf{y} = \begin{bmatrix} y_1 \\ y_3 \\ y_4 \\ y_5 \\ y_7 \\ y_8 \\ y_9 \\ y_{11} \\ y_{12} \\ y_{13} \\ y_{15} \\ y_{16} \\ y_{17} \\ y_{19} \\ y_{20} \\ y_{21} \\ y_{23} \\ y_{24} \\ y_{25} \\ y_{27} \\ y_{28} \\ y_{29} \\ y_{31} \\ y_{32} \end{bmatrix}$$

$h_l = \mathbf{u}'_l(\mathbf{U}'\mathbf{U})^{-1}\mathbf{u}_l$
$\mathbf{b} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\,{}_{\mathrm{U}}\mathbf{y}$
$e_l = y_l - \mathbf{u}'_l\mathbf{b}$, l=1, 2,…, 24
$\mathrm{RSS} = \mathbf{e}'\mathbf{e}$
d.f.=(n-m) - (p+1)

$$s^2 = \frac{\mathrm{RSS}}{\mathrm{d.f.}} = \frac{\mathrm{RSS}}{n-m-p-1} \qquad (\mathrm{A}.1)$$

$$\text{Studentized residual: } r_l = \frac{e_l}{\sqrt{s^2(1-h_l)}}$$

$$\text{Studentized deletion residual: } t_l = \frac{e_l}{\sqrt{s^2_{(l)}(1-h_l)}} \qquad (\mathrm{A}.2)$$

where $s^2_{(l)}$ is the estimate of σ^2 when the entire regression is run again on the n-m sample without the l-th case.

$$\text{Modified Cook's distance: } c_l = \sqrt{\frac{(n-m)-(p+1)}{p+1}\,\frac{h_l}{1-h_l}\,t_l^2}$$

OLS prediction arm

Let us now focus on case 6:

$\mathbf{x}_6'$ (6th row of $\mathbf{X}$ matrix) = $(1, x_{61}, x_{62}, x_{63}, x_{64}) = (1, \mathbf{x}_6^{*\prime})$

$\tilde{h}_6 = \mathbf{x}_6'(\mathbf{U}'\mathbf{U})^{-1}\mathbf{x}_6$

$\tilde{e}_6 = y_6 - \mathbf{x}_6'\mathbf{b}$

It can be shown [22], that the residual $\tilde{e}_6$ has $E(\tilde{e}_6) = 0$ and estimated variance $\widehat{var(\tilde{e}_6)} = s^2(1 + \tilde{h}_6)$, where s^2 is given by (A.1). Therefore:

Studentized prediction residual: $\tilde{t}_6 = \dfrac{\tilde{e}_6}{s\sqrt{1 + \tilde{h}_6}}$

Since $\tilde{e}_6$ is independent of s^2, $\tilde{t}_6$ is distributed according to the Student's t-distribution with (n-m-p-1)=19 d.f.

As shown by Atkinson [7], the analogue of the modified Cook's Distance is:

$$\tilde{c}_6 = \sqrt{\frac{(n-m)-(p+1)}{p+1}\frac{\tilde{h}_6}{1+\tilde{h}_6}\tilde{t}_6^2}$$

Alternatively one could think of adding the case in question to the reduced subset generating an augmented set of size (n-m+1) and computing the pertinent single case diagnostics on the latter.

Add $\mathbf{x}_6$ back in the matrix $\mathbf{U}$ and call $\mathbf{U}^+$ this augmented matrix so that:

$$\mathbf{U}^+ = \begin{bmatrix} \mathbf{u}_1^{+\prime} = \mathbf{x}_1' \\ \mathbf{u}_2^{+\prime} = \mathbf{x}_3' \\ \mathbf{u}_3^{+\prime} = \mathbf{x}_4' \\ \mathbf{u}_4^{+\prime} = \mathbf{x}_5' \\ \mathbf{u}_5^{+\prime} = \mathbf{x}_6' \\ \mathbf{u}_6^{+\prime} = \mathbf{x}_7' \\ \mathbf{u}_7^{+\prime} = \mathbf{x}_8' \\ \mathbf{u}_8^{+\prime} = \mathbf{x}_9' \\ \mathbf{u}_9^{+\prime} = \mathbf{x}_{11}' \\ \mathbf{u}_{10}^{+\prime} = \mathbf{x}_{12}' \\ \mathbf{u}_{11}^{+\prime} = \mathbf{x}_{13}' \\ \mathbf{u}_{12}^{+\prime} = \mathbf{x}_{15}' \\ \mathbf{u}_{13}^{+\prime} = \mathbf{x}_{16}' \\ \mathbf{u}_{14}^{+\prime} = \mathbf{x}_{17}' \\ \mathbf{u}_{15}^{+\prime} = \mathbf{x}_{19}' \\ \mathbf{u}_{16}^{+\prime} = \mathbf{x}_{20}' \\ \mathbf{u}_{17}^{+\prime} = \mathbf{x}_{21}' \\ \mathbf{u}_{18}^{+\prime} = \mathbf{x}_{23}' \\ \mathbf{u}_{19}^{+\prime} = \mathbf{x}_{24}' \\ \mathbf{u}_{20}^{+\prime} = \mathbf{x}_{25}' \\ \mathbf{u}_{21}^{+\prime} = \mathbf{x}_{27}' \\ \mathbf{u}_{22}^{+\prime} = \mathbf{x}_{28}' \\ \mathbf{u}_{23}^{+\prime} = \mathbf{x}_{29}' \\ \mathbf{u}_{24}^{+\prime} = \mathbf{x}_{31}' \\ \mathbf{u}_{25}^{+\prime} = \mathbf{x}_{32}' \end{bmatrix} \qquad {}_{\mathbf{U}^+}\mathbf{y} = \begin{bmatrix} y_1 \\ y_3 \\ y_4 \\ y_5 \\ y_7 \\ y_8 \\ y_9 \\ y_{11} \\ y_{12} \\ y_{13} \\ y_{15} \\ y_{16} \\ y_{17} \\ y_{19} \\ y_{20} \\ y_{21} \\ y_{23} \\ y_{24} \\ y_{25} \\ y_{27} \\ y_{28} \\ y_{29} \\ y_{31} \\ y_{32} \end{bmatrix}$$

$h_6^+ = \mathbf{x}_6'(\mathbf{U}^{+\prime}\mathbf{U}^+)^{-1}\mathbf{x}_6$

$\mathbf{b}^+ = (\mathbf{U}^{+\prime}\mathbf{U}^+)^{-1}\mathbf{U}^{+\prime}{}_{\mathbf{U}^+}\mathbf{y}$

$e_6^+ = y_6 - \mathbf{x}_{6\,\mathbf{U}^+}'\mathbf{b}$

$RSS^+ = \mathbf{e}^{+\prime}\mathbf{e}^+$

$d.f.^+ = (n\text{-}m+1)\text{-}(p+1)$

$s^{+2} = \dfrac{RSS^+}{d.f.^+} = \dfrac{RSS^+}{n-m-p}$

$s_{(6)}^{+2} = \dfrac{RSS}{d.f.^+ - 1} = \dfrac{RSS}{n-m-p-1}$

Studentized deletion residual: t_6^+

$$= \frac{e_6^+}{s_{(6)}^+\sqrt{1-h_6^+}}$$

$$c_6^+ = \sqrt{\frac{(n-m+1)-(p+1)}{p+1}\frac{h_6^+}{1-h_6^+}t_6^{+2}}$$

The relationship between matrices $(\mathbf{U}'\mathbf{U})^{-1}$ and $\left(\mathbf{U}^{+\prime}\mathbf{U}^{+}\right)^{-1}$ is preliminary used to explore the relationships between $\tilde{h}_6$ and h_6^+, $\tilde{t}_6$ and t_6^+, $\tilde{c}_6$ and c_6^+. It can be shown [23], that:

$$(\mathbf{U}'\mathbf{U})^{-1} = \left(\mathbf{U}^{+\prime}\mathbf{U}^{+}\right)^{-1} + \frac{\left(\mathbf{U}^{+\prime}\mathbf{U}^{+}\right)^{-1}\mathbf{x}_6\mathbf{x}_6'\left(\mathbf{U}^{+\prime}\mathbf{U}^{+}\right)^{-1}}{1 - h_6^+}$$

By simple algebra the relationships between $\tilde{h}_6$ and h_6^+, $\tilde{t}_6$ and t_6^+, $\tilde{c}_6$ and c_6^+ result to be:

$$\tilde{h}_6 = \mathbf{x}_6'(\mathbf{U}'\mathbf{U})^{-1}\mathbf{x}_6 = \mathbf{x}_6'\left(\mathbf{U}^{+\prime}\mathbf{U}^{+}\right)^{-1}\mathbf{x}_6 + \frac{\mathbf{x}_6'\left(\mathbf{U}^{+\prime}\mathbf{U}^{+}\right)^{-1}\mathbf{x}_6\mathbf{x}_6'\left(\mathbf{U}^{+\prime}\mathbf{U}^{+}\right)^{-1}\mathbf{x}_6}{1 - h_6^+} = \frac{h_6^+}{1 - h_6^+} \qquad (A.3)$$

$$\tilde{t}_6 = \frac{\tilde{e}_6}{s\sqrt{1 + \tilde{h}_6}} = \frac{e_6^+}{s_{(6)}^+\sqrt{1 - h_6^+}} = t_6^+$$

$$\tilde{c}_6 = c_6^+\sqrt{(1 - h_6^+)\frac{(n - m) - (p + 1)}{(n - m + 1) - (p + 1)}} \qquad (A.4)$$

The same computations apply to the remaining 7 outliers.

Provided that the cases are numerically identified as in the original dataset, the relationships between the diagnostics for the two mentioned options can be generalized in terms of i-th case.

Confirmatory stage cut-off points

As a rule of thumb the threshold T for h_i^+ is: $h_i^+(T) = \frac{2(p+1)}{n-m+1}$ [2]. With reference to the Hawkins two-system data it is easy to see that $h_i^+(T) = \frac{2\cdot5}{25} = 0.4$. By means of (A.3), this translates in the following $\tilde{h}_i$ threshold: $\tilde{h}_i(T) = \frac{0.4}{1-0.4} = 0.667$.

With regard to c_i^+ the threshold is: $c_i^+(T) = 2\sqrt{\frac{(n-m+1)-(p+1)}{(n-m+1)}}$. See Table 3 in [24]. With reference to the Hawkins two-system data $c_i^+(T) = 2\sqrt{\frac{25-5}{25}} = 1.789$. Using (A.4) this translates in the following $\tilde{c}_i$ threshold: $\tilde{c}_i(T) = 1.789\sqrt{\left(1 - \frac{2\cdot5}{25}\right)\frac{24-5}{25-5}} = 1.351$.

Moreover, traditional thresholds for the OLS diagnostics arm are: $h_i(T) = \frac{3(p+1)}{n-m} = \frac{3\cdot5}{24} = 0.625$, $|t_i|(T) = t_{0.995;\, n-m-p-2} = t_{0.995;\, 18} = 2.878$, $c_i(T) = 2\sqrt{\frac{n-m-p-1}{n-m}} = 2\sqrt{\frac{19}{24}} = 1.78$.

References

[1] Simpson DG, Ruppert D, Carroll RJ. On One-Step GM Estimates and Stability of Inferences in Linear Regression. Journal of the American Statistical Association 1992; 87(418): 439-50.

[2] Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. New York: John Wiley and Sons, 1987.

[3] Maronna RA, Martin DR, Yohai VJ. Robust Statistics: Theory and Methods. Chichester: John Wiley and Sons, 2006.

[4] Peña D, Yohai V. A fast procedure for outlier diagnostics in large regression problems. Journal of the American Statistical Association 1999; 94(446): 434-45.

[5] Becker C, Gather U. The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. Computational Statistics and Data Analysis 2001; 36: 119–27.

[6] Martin MA, Roberts S. Jackknife-after-bootstrap regression influence diagnostics. Journal of Nonparametric Statistics 2010; 22: 257-69.

[7] Atkinson AC. Masking unmasked. Biometrika 1986; 73: 533-41.

[8] Rousseeuw PJ, van Zomeren BC. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 1990; 85(411): 633-9.

[9] Seber GAF, Lee AJ. Linear regression analysis, second edition. Hoboken, New Jersey: John Wiley and Sons, 2003.

[10] Rousseeuw PJ, Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. Technometrics 1999; 41(3): 212-23.

[11] Becker C, Gather U. The masking breakdown point of multivariate outlier identification rules. Journal of the American Statistical Association 1999; 94(447): 947-55.

[12] Pison G, Van Aelst S, Willems G. Small sample corrections for LTS and MCD. Metrika 2002; 55: 111–23.

[13] Efron B. Jackknife-after-bootstrap standard errors and influence functions. Journal of the Royal Statistical Society 1992; 54: 83-127.

[14] Rousseeuw PJ, Yohai VJ. Robust regression by means of S-estimators. In Robust and nonlinear time series (Franke J, Härdle W, Martin D eds.). Lecture notes in statistics 26: 256-272. New York: Springer, 1984.

[15] Rousseeuw PJ. Least Median of Squares Regression. Journal of the American Statistical Association 1984; 79(388): 871-80.

[16] Hawkins DM, Bradu D, Kass GV. Location of several outliers in multiple regression using elemental sets. Technometrics 1984; 26: 197-208.

[17] Ryan T. Modern regression methods. New York: John Wiley and Sons, 1997.

[18] Billor N, Chatterjee S, Hadi AS. A re-weighted least squares method for robust regression estimation. American journal of mathematical and management sciences 2006; 26(3-4): 229-52.

[19] Yohai VJ. High Breakdown-Point and High Efficiency Robust Estimates for Regression. The Annals of Statistics 1987; 15 (2): 642-56.

[20] Chatterjee S, Mächler M. Robust regression: a weighted least squares approach. Communication in statistics. Theory and Methods 1997; 26: 381-1394.

[21] Atkinson AC, Riani M. Robust diagnostic regression analysis. New York: Springer, 2000.

[22] Sen A, Srivastava M. Regression analysis: theory, methods and applications. New-York: Springer-Verlag, 1990.

[23] Hoaglin DC, Welsch RE. The Hat Matrix in Regression and ANOVA. The American Statistician 1978; 32(1): 17-22.

[24] Chatterjee S, Hadi AS. Influential observations, high leverage points, and outliers in linear regression. Statistical Science 1986; 1: 379-93.

*