Computational Statistics & Data Analysis 14 (1992) 75–88
North-Holland

Robust versus classical detection of atypical data

Gérard Antille and Gilbert Ritschard
University of Geneva, Geneva, Switzerland

Received October 1990
Revised March 1991

Abstract: Residuals from a robust fit and robust distances are discussed and used as indicators of outliers or leverage points. The performance of these robust diagnostics are compared to the classical ones by means of a simulation study.

Keywords: Regression diagnostics; Outliers; Leverage points; M and high breakdown point estimators; Robust distances.

1. Introduction

This paper deals with methods to detect atypical data in linear regression. It describes local and global robustness, provides an introduction to the robust regression techniques which underlie robust diagnostics and then discusses the scope and limits of the main classical (Cook, 1977; Belsley et al., 1980; Cook and Weisberg, 1982; Chatterjee and Hadi, 1986) and robust (Rousseeuw and Leroy, 1987) indicators of outlyingness.

Atypical data can be classified in two categories: high leverage points and outliers. High leverage points are outlying in the space of explanatory variables, while outliers are data which show an atypical response to the explanatory variables.

This distinction is essential to evaluate correctly the perverse effects that may result from the presence of atypical data. Indeed, high leverage points and outliers affect the regression results differently, i.e. the coefficient estimates but also the usual synthetic adjustment measures (R^2, Student's t, F, Durbin–Watson, etc...). Figure 1 shows for instance how the least squares line changes when we successively move a point from a normal position (N) to a high leverage (L), to an outlier (O) and to a high leverage outlier (LO) position. The leverage point L does only moderately affect the coefficient estimates, but

Correspondence to: G. Antille, Department of Econometrics, University of Geneva, 2 rue Dancet, CH-1211 Geneva 4, Switzerland.

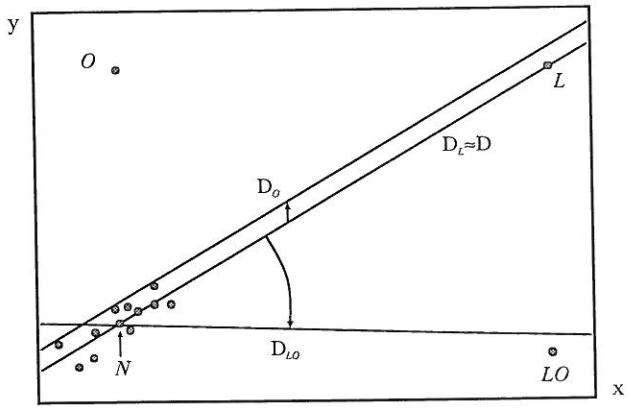

Fig. 1. Outliers and leverage points.

exaggeratedly inflates the R^2, Student's t and F. The outlier O induces mainly a translation of the least squares line, and deteriorates the evaluation statistics. The point LO which is both an outlier and a high leverage point causes a rotation of the regression line. It also affects the R^2, t and F, but more moderately. It can also be shown that outliers (O or LO) will generally give rise to autocorrelation effects. Likewise, leverage points can, as illustrated in Figure 2, mask heteroskedasticity problems.

There are of course diagnostic measures intended to detect influential data regardless of their nature. According to the preceding remarks it is worthwhile, however, to consider also indicators which allow to distinguish between high leverage points and outliers.

Classical indicators are mainly least squares byproducts, while robust measures refer to robust estimators. Before discussing the diagnostics we recall in Section 2 the two main criteria of robustness, i.e. bounded influence and high breakdown point. Section 3 briefly describes the regression robust estimators used as reference. The logic and the conceptual relevance of the outlyingness indicators are discussed in Section 4, and their performance on a simulated data set is commented in Section 5.

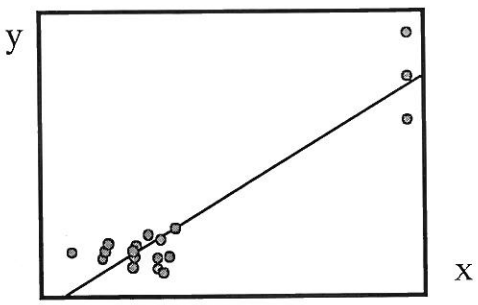

Fig. 2. Leverage points can mask heteroskedasticity.

2. Criteria of robustness

The notion of robustness is simple: Statistical techniques are robust if their output can only be moderately affected by outlying observations. Nevertheless, formal criteria for evaluating or measuring robustness are not so evident. The difficulty is conceptual. Indeed, robustness should be established independently of the data, while outlying precisely refers to the data configuration.

Two main approaches are considered in the literature to overcome this difficulty. The first provides a data independent concept of outlier by considering contaminated model distributions. It leads to what is known as the *influence function* approach. The second, instead of trying to measure the specific influence of one observation, focuses on the number of bad data supported by the statistic before it provides erratic results. This is the *breakdown point* approach. We briefly discuss these two criteria hereafter.

2.1. The influence function: local robustness

Consider a distribution F characterized by a parameter θ. Let $T(F_n)$ where F_n stands for the empirical distribution of n data, denote an estimator or some other statistic. The difference $T(F_n) - T(F_{n-1}^i)$ where F_{n-1}^i is the empirical distribution after deletion of an observation z_i, provides an evident empirical measure of the influence of z_i. The theoretical influence function, first introduced by Hampel (1974), is an extension of this measure to the asymptotic case where the empirical distribution is replaced by the model distribution F. It measures for any point z, the asymptotic bias which results from a small contamination of this distribution F at z. Formally, such a contaminated distribution can be written

$$G_\epsilon^z = (1-\epsilon)F + \epsilon\delta_z,$$

where δ_z stands for the distribution which has unity mass at z. The relative influence of z on T at the model F is then

$$\frac{1}{\epsilon}(T[(1-\epsilon)F + \epsilon\delta_z] - T[F]).$$

The *influence function* of the statistic T at the distribution F is defined as the limit of this expression for ϵ going to zero

$$\mathrm{IF}: \mathscr{Z} \to \mathbb{R}^p,$$

$$z \to \mathrm{IF}(z; T, F) = \lim_{\epsilon \to 0} \frac{1}{\epsilon}(T[(1-\epsilon)F + \epsilon\delta_z] - T[F]).$$

This function allows a first formal definition of robustness:

An *estimator* or *statistic* T is *locally robust* if its influence function IF(.; T, F) is bounded on $\mathscr{Z}$.

This is a concept of local robustness since a bounded influence function implies that, in the space of the contaminated distribution G_ϵ, T does not change drastically as long as the model distribution remains in a neighborhood of F. In other words, this means that whatever z, its impact on T will be limited as long as the rest of the sample can be assumed to be drawn from the hypothesized distribution F.

For example, let μ be the mean of the distribution F. The influence function of the sample mean $\bar{Z}$ at F is then

$$\mathrm{IF}(z;\, \bar{Z},\, F) = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\left[(1-\epsilon)\mu + \epsilon z - \mu\right] = z - \mu.$$

This function is clearly unbounded, which shows the well known nonrobustness of the mean. For the sample median $\dot{Z}$, the influence function at F reads

$$\mathrm{IF}(z;\, \dot{Z},\, F) = \frac{\mathrm{sign}\left[z - F^{-1}(1/2)\right]}{2f\left[F^{-1}(1/2)\right]},$$

which of course exists only for distributions with positive density f at the median $F^{-1}(1/2)$. It is bounded on $\mathcal{Z}$ which establishes the robustness of the median.

A byproduct of the influence function is the *sensitivity* of the statistic T at the hypothesized distribution F. This is defined as the maximum length of the IF vector, i.e. by

$$\gamma(T,\, F) = \sup_z \{\|\mathrm{IF}(z;\, T,\, F)\|\},$$

where $\|\cdot\|$ denotes the Euclidean norm. For non-normalized statistics T we shall however prefer the standardized sensitivity:

$$\gamma_s(T,\, F) = \sup\left\{\left[\mathrm{IF}(z;\, T,\, F)'\mathbf{V}^{-1}IF(z;\, T,\, F)\right]^{1/2}\right\},$$

V being the asymptotic variance matrix of T.

2.2. The breakdown point: global robustness

In contrast to the influence function, which produces a local measure of robustness, the concept of breakdown point gives a global robustness criterion (i.e., independent of the model distribution). The breakdown point is the smallest fraction of contaminated data the estimator can support before taking arbitrary values. To quantify that idea, the finite sample breakdown point introduced by Donoho and Huber (1983) is used throughout this paper.

Consider all contaminated samples Z^* obtained by replacing any m of the original data set Z by arbitrary values. The maximal bias is defined by $\mathrm{bias}(m,\, T,\, Z) = \sup_{Z^*} \|T(Z) - T(Z^*)\|$ and the breakdown point by

$$\epsilon^*(T,\, Z) = \inf\left\{\frac{m}{n}\,\middle|\,\mathrm{bias}(m,\, T,\, Z) = \infty\right\},$$

n being the size of Z. This notion provides a second definition of robustness:

An *estimator* or *statistic* T is *globally robust* if its limiting breakdown point (for $n \to \infty$) is strictly positive.

For example the arithmetic mean has a breakdown point equal to $1/n$. Hence, it is not a globally robust estimator. It is obvious that no estimator can have a breakdown point larger that $1/2$ and this upper bound is attained for instance by the sample median.

For multivariate location estimators the problem is much more complex. However, the Stahel–Donoho estimator (Stahel, 1981; Donoho, 1982), which downweights any points being many robust standard deviations away from the sample in some projection, and the Rousseeuw MVE-estimator (Rousseeuw, 1984, 1985), whick takes the center of the minimal volume ellipsoid covering (at least) h points where h can be taken equal to $[n/2]+1$, both have maximal breakdown point.

3. Some robust estimators of regression parameters

The choice of the reference hyperplane is crucial for the detection of outliers in regression. This section presents briefly robust estimators which we shall use later as reference. We first introduce locally robust estimators which have bounded sensitivity at the normal distribution, and then discuss a high breakdown point estimator: the *LMS* (Least Median of Squares estimator).

3.1. Robust M-estimators

Consider the regression model

$$y = X\beta + \epsilon$$

where y is the n-vector of the dependent variable, X the $n \times p$ matrix of p independent variables, β the p-vector of coefficients, and ϵ a vector of independent errors, which we assume to be symmetrically distributed and independent of the X variables.

At the normal model, i.e. when $z_i = (y_i, x_i)$ follows a multinormal distribution, the least squares estimator is the most efficient. Using robust estimators leads to a loss of efficiency in that case. This represents the price to pay for robustness.

It is natural to try to minimize this price. For a given sensitivity limit c, we should then retain the most efficient estimator which satisfies this bound. Formally this leads to the following optimality problem

$$\begin{cases} \min_T \operatorname{tr}[V(T)] \\ \text{u.c. } \gamma(T, F) = c, \end{cases} \tag{3.1}$$

where the trace $\operatorname{tr}[V(T)]$ of the asymptotic variance $V(T)$ represents the efficiency criterion.

This problem is globally untractable, but can be solved in the class of M-estimators. Recall that a regression M-estimator T is any estimator defined by a system of the following form

$$\sum_{i=1}^{n} \eta(x_i, y_i - x_i'T)x_i = 0,$$

where $\eta : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ must be regularly smooth except perhaps on some finite subset of the factor space. (See Hampel et al., 1986, p. 315). This system is indeed a generalization of the least squares normal equations, which are obtained by setting $\eta(x, r) = r$.

Let r denote the residual $y - x'T$. It can be shown that the influence function of the OLS estimator is $\mathrm{IF} = r[E(xx')^{-1}]x$. It is unbounded both with respect to r and x. Bounded influence estimators are obtained by downweighting high leverage points and outliers through an adequate choice of η. The three main classes of robust M-estimators are

Huber	given by	$\eta(x, r) = \psi_c(r)$,
Mallows	given by	$\eta(x, r) = \omega(x) \cdot \psi_c(r)$,
Schweppe	given by	$\eta(x, r) = \omega(x) \cdot \psi_c(r/\omega(x))$,

where $\omega(x)$ is a weight function depending on the position of x in the factor space, and $\psi_c(r)$ is the Huber function:

$$\psi_c(r) = \begin{cases} r & \text{for } |r| \le c, \\ c \cdot \mathrm{sign}(r) & \text{otherwise.} \end{cases}$$

Note that Huber's estimators downweight only outliers, and remain thus non-robust against high leverage points. Mallows estimators separate clearly the weight of outliers from the weight of leverage points, contrary to Schweppe's estimators.

Optimal M-estimators, i.e. solutions of the program (3.1) are of the Schweppe type. If one chooses to bound the non-standardized sensitivity γ the estimator of Hampel and Krasker (Hampel, 1978; Krasker, 1980) is the solution. For a bound on the standardized sensitivity γ_s, the first order conditions are satisfied by the estimator of Krasker and Welsch (1982). These are very complex estimators whose weights are only implicitly defined. For instance the weight function $\omega(x)$ of the *Krasker–Welsch estimator* is given by

$$\omega(x_i) = \| Ax_i \|^{-1},$$

where A is the solution of an implicit equation of the form (see Hampel et al., 1986, for more details)

$$(A'A)^{-1} = \frac{1}{n} \sum g(c/\| Ax_i \|) x_i x_i'.$$

This estimator requires a bound $c \ge \sqrt{p}$.

Mallows-type estimators have the advantage to allow separate bounds for the sensitivity to outliers and the sensitivity to leverage points. For standardized sensitivity, the weight function $\omega(x)$ of the *optimal Mallows estimator* is given by

$$\omega(x_i) = \omega_b(\| Bx_i \|) = \inf\{1, b/\| Bx_i \|\},$$

where the matrix B is the solution of the implicit equation

$$(B'B)^{-1} = \frac{1}{n} \sum \omega_b^2(\| Bx_i \|) x_i x_i'.$$

The bound b must satisfy $b \geq \sqrt{p}$.

3.2. The least median of squares estimator

Robust M-estimators were developed by minimizing $\Sigma\rho(r_i)$ instead of Σr_i^2, but they did not touch the summation sign. On the contrary the LMS (Least Median of Squares) estimator, introduced by Rousseeuw (1984) according to an idea of Hampel (1975), is the solution of $\min_\beta \text{med}_i r_i^2$.

In the simple regression case the LMS estimator is the midline of the narrowest strip covering at least $[n/2] + 1$ of the points.

For the multidimensional case the extension is straightforward. However, computational problems can arise very quickly. Algorithms are discussed in Rousseeuw and Leroy (1987), and their PROGRESS computer program performs multiple regression by means of the LMS method.

The LMS breakdown point is $([n/2] - p + 2)/n$ and asymptotically $\epsilon^* = 1/2$. In that sense LMS is better than M-estimators whose breakdown points are smaller than $1/p + 1$. Moreover the LMS estimator attains the maximal breakdown point among all regression equivariant estimators. See Rousseeuw and Leroy (1987, p. 112).

Figure 3 illustrates the difference between M and LMS estimators for a data set composed of a subset A containing 60% of the data and a subset B with the remaining part.

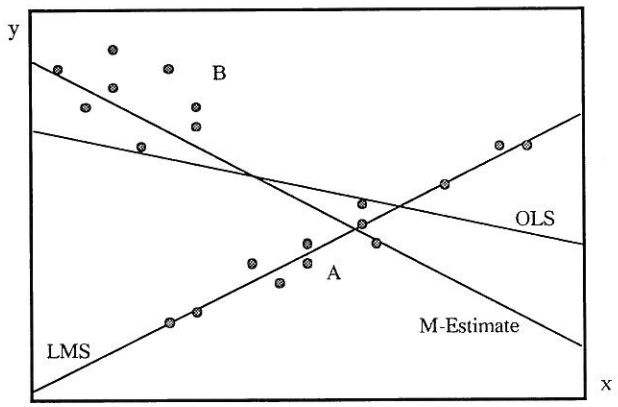

Fig. 3. M and LMS estimators behave differently.

4. Detecting influential observations

Several measures of outlyingness have been designed to detect an observation or a group of observations having a large influence on the LS estimator. These classical measures suffer from the fact that they are based on the LS estimator which is not robust. Obviously robust diagnostics rely on robust estimates. Diagnostic measures with high breakdown point are able to cope with multiple outliers without suffering from the masking effect (i.e., the effect that one outlier can make all other outliers have small values of the diagnostic).

However it is interesting to note that classical diagnostics as well as robust diagnostics have to be used in conjunction with their respective residuals.

4.1. Classical diagnostics

Classical diagnostics are closely related to the matrix $H = X(X'X)^{-1}X'$, called the hat matrix because it transforms the vector of observations y into its least squares estimate $\hat{y} = Hy$. So the elements h_{ij} of H measure the effect of the jth observation on $\hat{y}_i$ and large values of h_{ii} are supposed to indicate high leverage points.

The diagonal element h_{ii} of the hat matrix is related to the Mahalanobis distance MD_i of the ith case to the center of the observed explanatory variables by the relation (see for instance Rousseeuw and Leroy, 1987, p. 225)

$$h_{ii} = \frac{1}{n-1}\mathrm{MD}_i^2 + \frac{1}{n}. \tag{4.1}$$

Thus, it appears that the h_{ii}'s point out observations far from the bulk of data in the factor space.

Obviously one also has to take y_i into account to detect outliers. Least squares residuals $r_i = y_i - \hat{y}_i$, standardized residuals $(n-p)r_i/\Sigma r_i^2$, and studentized residuals $r_i^* = (n-p)r_i/(\Sigma r_i^2\sqrt{1-h_{ii}})$ provide complementary information.

The influence of an observation can also be measured by considering the regression with and without that observation, their respective response being denoted by $\hat{y}$ and $\hat{y}_{(i)}$. The squared distance

$$\mathrm{CD}_i^2 = (n-p)(\hat{y} - \hat{y}(i))'(\hat{y} - \hat{y}(i))/p\sum r_i^2,$$

introduced by Cook (1977) measures the change in the regression. A large value of CD_i^2 indicates that the ith observation has an important effect on the determination of $\hat{\beta}$. Cook suggests that each CD_i^2 be compared to the percentiles of the central F-distribution with p and $n-p$ degrees of freedom. Usually the cut-off value may be taken equal to one (Cook and Weisberg, 1982).

As $\mathrm{CD}_i^2 = (1/p)r_i^{*2}[h_{ii}/(1-h_{ii})]$, it follows that CD_i^2 depends on the number of parameters in the model, on the goodness of fit and on the remoteness of x_i in the space of explanatory variables.

Similar diagnostics' measures are proposed in the literature, see for instance Belsley et al. (1980), but unfortunately these measures, like CD_i^2, provide only single case diagnostics.

Extension of CD_i^2 to multiple-case diagnostics are proposed by Cook and Weisberg (1982), but these methods are not very popular because they require consideration of many or all subsets of the data.

4.2. Robust diagnostics

Classical methods rely on r_i and h_{ii} and by relation (4.1) they depend on the Mahalanobis distance. They all are based on estimates which are not globally robust. Thus they cannot detect multiple outliers.

As an alternative to the Mahalanobis distance one can consider robust distances. Rousseeuw and van Zomeren (1990) proposed to use

$$\mathrm{DMVE}_i^2 = (x_i - T(X))' C_{\mathrm{MVE}}^{-1}(x_i - T(X))$$

based on the minimum volume ellipsoid estimate, where $T(X)$ is the center of that ellipsoid and C_{MVE} the MVE estimator of the covariance matrix. The cut-off value might be taken equal to $\chi^2_{p,0.975}$.

The weight function used in M-estimation depends on the position of x in the factor space. Hence it provides an indicator of leverage. In fact, the inverse of the weight function is a measure of leverage similarly to h_{ii}.

Residuals from a robust fit are simply defined by

$$r_i^{\mathrm{rob}} = y_i - x_i'\hat{\beta}^{\mathrm{rob}},$$

where $\hat{\beta}^{\mathrm{rob}}$ is any robust estimate defined in Section 3. Standardized robust residuals are naturally defined by $r_i^{\mathrm{rob}*} = r_i^{\mathrm{rob}}/\hat{\sigma}^{\mathrm{rob}}$, $\hat{\sigma}^{\mathrm{rob}}$ being a robust scale estimate corresponding to $\hat{\beta}^{\mathrm{rob}}$.

Robust distances and residuals from robust fit should be used in conjunction to determine a weight for the ith observation or to provide an outlier diagnostic.

5. A simulation study

As mentioned in the previous sections, classical measures of outlyingness suffer from the non-robustness of the LS estimator; so robust diagnostics are conceptually superior. A simulated multiple regression problem with three explanatory variables exhibits this superiority by comparing the values of those indicators.

A main model with three explanatory variables and independent normal errors was used to generate 80% of the 50 values of the dependent variable y. The remaining observations were generated with a second model to introduce outliers in our data set.

The values of the three predictors x were generated with autoregressive processes. A simulated structural change in the processes was introduced to increase the leverage effect of the last five points.

Table 1
Simulated data

obs	Simulated data				Residuals			
	x_1	x_2	x_3	y	r_{true}	r_{OLS}	r_{MAL}	r_{LMS}
1	15.00	100.00	−10.00	−2.39	0.11	−10.61	−0.65	−0.21
2	15.77	100.38	−8.92	−1.63	0.13	−9.27	−0.51	−0.07
3	15.88	95.91	−7.27	−1.25	−0.63	−8.03	−1.22	−0.95
4	16.58	98.64	−6.53	0.48	0.44	−5.95	−0.04	0.25
5	15.01	95.27	−5.07	7.54	5.59	1.60	4.99	4.90
6	17.12	94.70	−3.71	1.44	−0.75	−3.57	−1.09	−1.02
7	18.20	94.33	−2.89	2.71	0.27	−1.76	0.05	0.20
8	18.12	93.79	−1.79	3.95	0.43	−0.04	0.24	0.26
9	18.53	88.38	−1.13	3.58	0.14	0.10	0.00	−0.01
10	18.95	92.44	−0.40	4.20	−0.17	0.98	−0.22	−0.23
11	19.65	93.39	−0.09	3.88	−0.55	0.87	−0.52	−0.45
12	19.87	82.69	0.80	3.83	−0.30	1.56	−0.25	−0.30
13	18.74	79.09	1.37	5.29	0.38	3.20	0.32	0.03
14	19.21	81.59	1.73	5.79	0.51	3.85	0.51	0.27
15	18.98	83.59	2.33	−0.31	−6.50	−2.08	−6.48	−6.81
16	20.74	79.55	3.06	5.40	−0.24	4.34	−0.04	−0.20
17	19.84	74.07	3.75	6.19	−0.04	5.45	0.07	−0.31
18	22.10	71.07	4.21	5.52	0.26	5.42	0.60	0.47
19	22.91	62.42	4.53	4.41	0.09	4.82	0.48	0.40
20	22.38	58.57	4.12	3.91	0.13	4.17	0.44	0.31
21	22.55	50.74	4.13	2.24	−0.69	2.76	−0.38	−0.52
22	22.45	50.22	3.98	2.01	−0.77	2.47	−0.47	−0.62
23	24.33	47.05	3.82	1.67	0.31	2.45	0.76	0.89
24	24.10	46.83	3.78	1.74	0.32	2.47	0.75	0.85
25	23.79	51.38	3.52	−5.27	−7.03	−4.84	−6.62	−6.52
26	24.35	54.13	3.79	2.56	0.54	3.12	1.01	1.18
27	23.70	49.35	3.96	1.19	−0.85	1.86	−0.44	−0.40
28	23.90	48.11	3.92	2.32	0.54	3.04	0.96	1.02
29	23.96	43.51	4.02	1.32	−0.07	2.23	0.34	0.38
30	24.32	40.25	3.66	−0.05	−0.58	0.85	−0.15	−0.30
31	23.49	40.01	3.52	0.62	−0.15	1.34	0.18	0.19
32	23.57	40.63	3.98	0.60	−0.66	1.52	−0.28	−0.31
33	23.12	43.38	3.81	1.10	−0.49	1.79	−0.16	−0.22
34	22.37	39.94	4.29	2.62	0.53	3.50	0.79	0.55
35	23.16	44.31	3.92	−5.18	−6.95	−4.46	−6.61	−6.68
36	23.03	43.96	4.24	1.82	−0.29	2.66	0.04	−0.07
37	22.68	35.66	4.05	1.13	−0.15	2.08	0.11	−0.07
38	22.82	35.21	4.34	1.55	0.11	2.66	0.39	0.19
39	20.65	31.00	4.01	2.16	0.37	2.92	0.43	−0.07
40	19.77	22.30	3.87	1.54	0.32	2.36	0.26	−0.39

Table 1 (continued)

obs	Simulated data				Residuals			
	x_1	x_2	x_3	y	r_{true}	r_{OLS}	r_{MAL}	r_{LMS}
41	19.96	22.40	4.19	1.42	−0.03	2.40	−0.06	−0.72
42	20.31	18.94	4.17	0.54	−0.37	1.67	−0.37	−1.00
43	20.66	21.68	4.41	2.02	0.77	3.23	0.81	0.22
44	20.09	19.76	4.94	2.28	0.41	3.68	0.41	−0.32
45	21.11	16.34	4.48	−4.12	−4.68	−2.65	−4.61	−5.16
46	24.31	12.22	7.58	−8.01	−9.66	−4.58	−9.18	−9.61
47	26.20	4.29	10.84	−11.45	−14.62	−6.08	−13.87	−14.39
48	28.01	7.60	14.40	−14.40	−20.55	−7.30	−19.49	−20.09
49	28.90	4.69	18.32	−17.55	−26.88	−8.53	−25.59	−26.48
50	28.61	3.03	21.46	−20.11	−32.56	−9.72	−31.20	−32.46

Hence by construction the data set contains three groups of atypical data:

- observations 5, 15, 25, 35 and 45 are outliers which are not leverage points;
- observations 1, 2, 3, ... are leverage points but not outliers;
- observations 50, 49, 48, 47, 46 are outliers and leverage points.

The simulated data, the true residuals (i.e., with respect to the main model) and the residuals for the OLS, for Mallows with $b = 2.25$, $c = \sigma = 0.5$ and for the LMS are given in Table 1. The values of classical and robust outlyingness indicators displayed in Figures 4, 5 and 6 can be found in Antille and Ritschard (1990).

A convenient way to detect simultaneously outliers and leverage points is given by a scatterplot of standardized residuals versus leverage indicators. Outliers lie outside of the central band, leverage points in the right part of the scatterplot. Data which are both outliers and leverage points are located on the

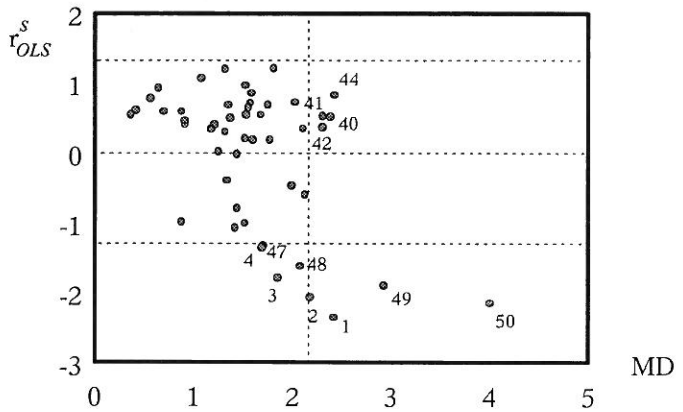

Fig. 4. Standardized OLS residuals versus Mahalanobis distances.

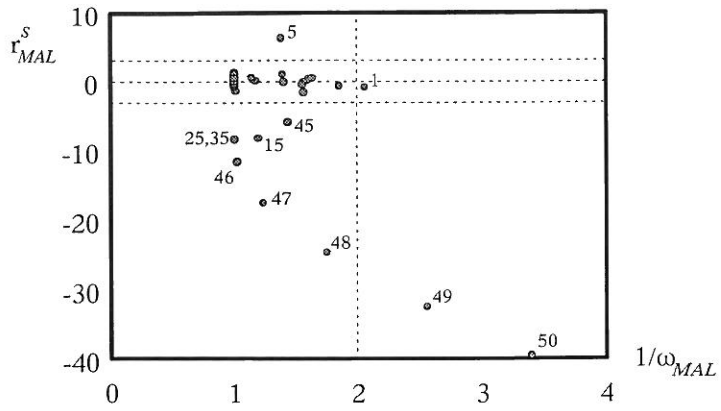

Fig. 5. Standardized Mallows residuals versus inverse of Mallows leverages.

right and outside of the central band. The partition of the graphic is defined by the cut-off values of the two measures plotted.

For the OLS and studentized residuals, the cutoff is set at 1.3 which is justified by the fact that there are 20% of outliers. For the Mallows and LMS residuals, standardized with a robust scale estimate, the cutoff value is set at 2.4, the 1% critical value for a t distribution.

For the leverage measures the cut-off chosen for the classical Mahalanobis' distance MD, and the robust distance RD based on the minimum ellipsoid estimate, is respectively

$$\sqrt{\chi^2_{2,0.9}} = 2.15 \quad \text{and} \quad \sqrt{\chi^2_{3,0.99}} = 3.37.$$

For the "Mallows" leverage indicator, $\text{med}[\omega(\boldsymbol{x})]/\omega(\boldsymbol{x})$, we choose the cut-off 2 which points to data with weight less than half the median weight.

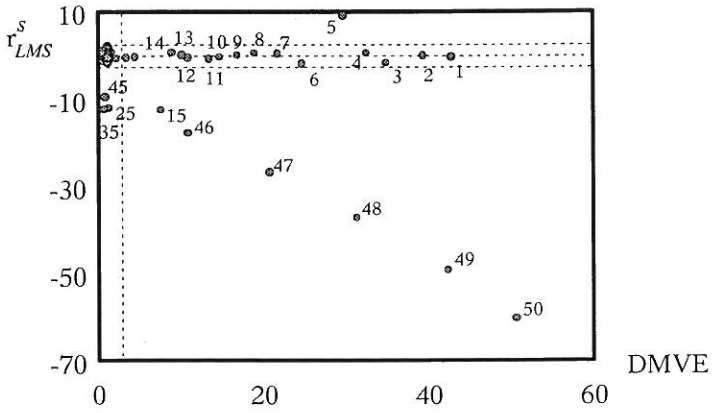

Fig. 6. Standardized LMS residuals versus MVE distances.

Figure 4 shows such a plot for OLS residuals versus classical distances MD. Figure 5 exhibits a plot of Mallows residuals versus "Mallows" indicator of leverage, and Figure 6 presents a plot of LMS residuals versus DMVE. A quick glance at these different plots shows the superiority of the high breakdown point estimators.

Results and computation have been obtained with the following software: ROBETH (Marazzi, 1985), PROGRESS (Rousseeuw and Leroy, 1987) and PROCOVIEV (Rousseeuw and Van Zomeren, 1990).

6. Conclusion

Classical measures of detection of atypical data are mainly least squares byproducts, while robust indicators refer to robust estimators and by construction appear to be more reliable. The numerical results presented in Section 5 exhibit this superiority.

References

Antille, G., Ritschard, G. (1990). Robust and Classical Outlyingness Indicators A Simulation Study, *Communications in Statistics, Simulation and Computation*, **19**(2), 505–512.

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

Chatterjee, S. and Hadi, A.S. (1986). Influential Observations, High Leverage Points and Outliers in Linear Regression, *Statistical Science*, 1 No 3, 379–416.

Cook, R.D. (1977). Detection of Influential Observations in Linear Regression, *Technometrics*, **19**, 15–18.

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

Donoho, D.L. (1982). *Breakdown Properties of Multivariate Location Estimators*, qualifying paper, Harvard University.

Donoho, D.L. and Huber, P.J. (1983). The Notion of Breakdown Point, in: *A Festschrift for Erich L. Lehmann*, P.J. Bickel, K.A. Doksum and J.L. Hodges Jr, Eds. (Belmont, CA: Wadsworth) 157–184.

Hampel, F.R. (1974). The Influence Curve and its Role in Robust Estimation, *Journal of the American Statistical Association*, **69**, 383–393.

Hampel, F.R. (1975). Beyond Location Parameters: Robust Concepts and Methods, *Bulletin of the International Statistical Association*, **46**, 375–382.

Hampel, F.R. (1978). Optimally Bounding the Gross-Error-Sensitivity and the Influence of Position in Factor Space, *Proceedings of the Statistical Computing Section, American Statistical Association,* 59–64.

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).

Krasker, W.S. (1980). Estimation in Linear Regression Models with Disparate Data Points, *Econometrica*, **48**, 1333–1346.

Krasker, W.S. and Welsh, R.E. (1982). Efficient bounded influence regression estimation. *Journal of the American Statistical Association*, **77**, 595–604.

Marazzi, A. (1985). "ROBETH, Robust Linear Programs," Documents 1 to 6, Division de statistique et informatique, Institut Universitaire de Médecine Sociale et Préventive, Lausanne.

Rousseeuw, P.J. (1984). Least Median of Squares Regression, *Journal of the American Statistical Association*, **79**, 871–880.

Rousseeuw, P.J. (1985), Multivariate Estimation with High Breakdown Point, *Mathematical Statistics and Applications*, Vol. B, W. Grossmann, G. Pflug, I. Vincze and W. Wertz, Eds. (Dordrecht: Reidel Publishing Company) 283–297.

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

Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking Multivariate Outliers and Leverage Points, *Journal of the American Statistical Association*, **85**, 633–639.

Stahel, W.A. (1981). "Robust Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen", Ph.D. Thesis, ETH Zürich.