A Robust Look at the Use of Regression Diagnostics

Gilbert Ritschard; Gerard Antille

The Statistician, Vol. 41, No. 1. (1992), pp. 41-53.

Stable URL:
http://links.jstor.org/sici?sici=0039-0526%281992%2941%3A1%3C41%3AARLATU%3E2.0.CO%3B2-J

The Statistician is currently published by Royal Statistical Society.

The Statistician (1992) 41, pp. 41–53

A robust look at the use of regression diagnostics

GILBERT RITSCHARD & GÉRARD ANTILLE

Department of Econometrics, University of Geneva, CH-1211 Geneva 4, Switzerland

Abstract. The need to pay special attention to atypical data in regression analysis is generally accepted. First, because of their excessive influence on the regression results. However, also as emphasized by Gray, the unusual data often provide useful information. Thus, even if robust regression techniques offer a remedy to the fitting problem, the need for regression diagnostics remains. Robust techniques lead to powerful remoteness indicators which, unlike the classical measures based on least squares, are themselves insensitive to atypical data. A re-examination of the two examples discussed by Gray shows that these robust indicators advantageously complement the information obtained with classical influence measures.

1 Introduction

In a recent paper, Gray (1989a) provides very interesting hints on the use of regression diagnostics, which he illustrates on two real data sets. He focuses his discussion on classical, i.e. least squares based, diagnostics, arguing that they are easily calculable and, more importantly, available in the major statistical packages. The purpose of this paper is to complete the discussion by showing how robust outlyingness indicators can provide further insight on unusual data. The same two real data sets are used as illustrations.

Indeed, robust diagnostics have not yet been implemented in common statistical packages like SAS, SPSS, etc. Packages exist, however, which offer robust techniques: S-PLUS (which runs under UNIX) from Statistical Sciences (1988), SC the Statistical Calculator (UNIX, DOS) from Dusoir (1989), and ROBETH (a set of FORTRAN routines developed at the Swiss Institute of Technology ETH, see Marazzi, 1985) to mention just a few. Also, there are of course the two nifty and easy to use PROGRESS (Leroy & Rousseeuw, 1984) and PROCOVIEV (Rousseeuw & Van Zomeren, 1987) programs which have been used to compute the numerical results exhibited in this paper.

Robust indicators of outlyingness have been seen to be superior to classical diagnostics to detect atypical data (see, for instance, Rousseeuw & Leroy, 1987, or Antille & Ritschard, 1990, for an introduction to the topic). This superiority lies especially in their ability to cope with multiple outliers, contrarily to the classical measures which suffer from the masking effect in the sense that one atypical point can make all other outliers have small values of the diagnostic measures. Nevertheless, robust indicators do not measure the influence of the atypical data on the least squares regression. Hence, robust remoteness information usefully complements the influence indications obtained from classical diagnostics, but does not replace them.

Section 2 recalls the goal and definition of the main classical and robust diagnostic measures. Section 3, then, reconsiders, from a robust viewpoint, the two examples discussed in Gray (1989a).

2 Atypical data indicators

For our discussion, we partition the atypical data indicators into influence and direct remoteness measures. Influence diagnostics measure the outlyingness of data indirectly through their influence on the regression results. Classical diagnostics are mainly of this type. Robust techniques are more concerned with the direct measure of outlyingness.

In order to introduce some notations, let us consider the regression model

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is the n vector of the dependent variable, $\mathbf{X}$ the $n\times p$ matrix of p independent variables, $\boldsymbol{\beta}$ the p vector of coefficients, and $\boldsymbol{\varepsilon}$ an n vector of independent errors, which we assume to be symmetrically distributed and independent of the $\mathbf{X}$ variables. We denote by $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ the least-squares (LS) estimator of $\boldsymbol{\beta}$. The hat matrix $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, from which we get the LS fitted values $\hat{\mathbf{y}}=\mathbf{H}\mathbf{y}$, is of special interest for diagnostic purposes. Its diagonal elements are denoted by h_i. We use $\mathbf{r}$ to designate the vector of LS residuals $\mathbf{r}=\mathbf{y}-\hat{\mathbf{y}}$ and $\mathbf{r}_{\text{rob}}$ for the residuals relative to a robust fit. Likewise, s denotes the square root of the usual error variance estimator $s^2=\Sigma r_i^2/(n-p)$ and s_{rob} designates a robust scale estimate. Classical diagnostics extensively use LS results obtained with ith case removed. We designate these results with a subscript i in parentheses. For example, $\hat{\boldsymbol{\beta}}_{(i)}$, $\hat{\mathbf{y}}_{(i)}$ and $s_{(i)}$ denote, respectively, the parameter estimate, the vector of fitted values and the standard error computed after deletion of the ith case.

2.1 Direct remoteness measures

There are two kinds of atypical data in regression analysis: factor-outliers and fit-outliers. Factor-outliers (or $\mathbf{X}$-outliers), classically known as high leverage points because of their effect on the least squares regression, are outlying in the space of the independent variables, i.e. in the row space of $\mathbf{X}$. Fit-outliers (or $\mathbf{y}$-outliers) are cases which show an unusual response to the explanatory variables. Both merit special attention. The distinction is however essential from an interpretation viewpoint. As will be shown in the examples, the lack of fit of the $\mathbf{y}$-outliers often provides fruitful indications about missing explanatory factors. Well fitted factor-outlying points, obviously, cannot by used this way. They may be, however, a good indication of the model's persistence over a wide range of explanatory variables values. The two kinds of atypical data also have different statistical implications. Fit-outliers always deteriorate least squares fittings, while factor-outlying points, when not also fit-outliers, may over-reduce the standard deviation of the estimates and, hence, provide the illusion of a good adjustment.

The classical approach to the detection of fit-outliers focuses on standard forms of the LS residuals. The three main classical indicators are the standardized residual $r_i^s=r_i/s$, the Studentized residual $r_i^t=r_i/s(1-h_i)^{1/2}$ and the jack-knifed residual $r_i^j=r_i/s_{(i)}(1-h_i)^{1/2}$. The terminology used here is taken from Rousseeuw & Leroy (1987). There is, however, no universal agreement about it. For instance, Gray's (1989a) standardized residual corresponds to our r^t, and Belsley *et al.* (1980) use the term Studentized to describe our jack-knifed residual. The latter has the advantage to have a known Student distribution with $n-p$ degrees of freedom when the errors ε are normally distributed. In practice, though the Studentized and jack-knifed residuals score slightly higher than the standardized r^s, a cut-off of 2·5 seems reasonable for all three measures.

The drawback of the LS residuals is that they are themselves influenced by the atypical data. This makes them especially unreliable in the case of multiple unusual data. This problem can be avoided by considering the residuals relative to a robust fit. A robust standardized residual r_{rob}^s is obtained by standardizing the robust residual r_{rob} with the corresponding robust scale estimate s_{rob}:

$$r_{i,\text{rob}}^s=\frac{r_{i,\text{rob}}}{s_{\text{rob}}}$$

Here again, a cut-off of 2·5 is recommended.

Robust regression estimators which can be used to determine the robust fit include R-estimators (based on ranks, cf. Hettmansperger, 1984), influence bounded

M-estimators (see, for instance, ch. 6 in Hampel *et al.*, 1986) and high breakdown point estimators (Rousseeuw & Leroy, 1987). The third of these which protect against a high number of unusual data, are the most suitable for diagnostic purposes. Rousseeuw's (1984) least median of squared residuals (LMS) estimator is the best known in that family. It is, for instance, available in S-PLUS and in the Statistical Calculator SC. Formally, it is defined as the solution $\hat{\boldsymbol{\beta}}_{\text{LMS}}$ of the problem

$$\min_{\beta} \{\underset{i}{\text{med}} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\}$$

and corresponds to the centre of the smallest band which covers at least 50% of the data.

For factor-outliers, the classical indicators are the diagonal elements h_i of the hat matrix. Each h_i is linearly related to the squared Mahalanobis distance MD_i^2 from the ith case to the mean point of the observed explanatory variables (Rousseeuw & Leroy, 1987, p. 225) such that

$$h_i = \frac{1}{n-1} \text{MD}_i^2 + \frac{1}{n}$$

Clearly, large h_i values, or equivalently large MD_i distances, indicate high leverages, i.e. factor-outlyingness. Usual cut-offs are $2p/n$ for h_i, and $\chi^2_{(p-1),0.975}$ for the squared Mahalanobis distance.

Robust factor-outlyingness indicators are obtained by using robust distances instead of the classical Mahalanobis distance. One can, for instance, consider the minimum volume ellipsoid (MVE) squared distance (Rousseeuw & Leroy, 1987, p. 260, Rousseeuw & Van Zomeren, 1990):

$$\text{DMVE}_i^2 = (\mathbf{x}_i - \mathbf{c}_{\text{MVE}})' \mathbf{V}_{\text{MVE}}^{-1} (\mathbf{x}_i - \mathbf{c}_{\text{MVE}})$$

where $\mathbf{c}_{\text{MVE}}$ is the centre of the MVE which covers 50% of the data and $\mathbf{V}_{\text{MVE}}$ is the covariance matrix computed on these covered data. Rousseeuw & Leroy suggest a cut-off equal to $\chi^2_{(p-1),0.975}$ for DMVE_i^2. Let us notice that the DMVE_i indicator requires much computation time. It is, therefore, only applicable in the case of a reasonable number of variables. For instance, we did not get any results for the four variables example discussed in Section 3.2 after a full night on a 386 PC. The results were finally computed in 28 min of CPU on a mainframe IBM 3090 machine.

Finally, let us consider global direct remoteness indicators which attempt to detect atypical data without bothering with the distinction between fit- and factor-outliers. A classical measure is the Andrews & Pregibon (1978) determinantal R_i ratio:

$$R_i = \frac{\det(\mathbf{Z}_{(i)}'\mathbf{Z}_{(i)})}{\det(\mathbf{Z}'\mathbf{Z})} = 1 - h_i - \frac{r_i^2}{(n-p)s^2}$$

where the matrix $\mathbf{Z}$ is the $\mathbf{X}$ matrix augmented by the vector $\mathbf{y}$ and $\mathbf{Z}_{(i)}$ is the $\mathbf{Z}$ matrix with the ith case deleted. Small values of R_i identify extreme cases in the case space. Another classical measure, which obeys the logic of the h_i values, is given by the diagonal elements of the matrix $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ (see, for instance, Cook & Weisberg, 1982). Indeed, the ith diagonal term is equal to $1 - R_i$. This last measure has the advantage to have, like other outlyingness measures, large values associated with atypical data. The inverse $1/R_i$ of R_i would, however, better emphasize the extreme cases.

Likewise, the DMVE_i distance computed on the rows of matrix $\mathbf{Z}$ provides a robust global remoteness indicator. Rousseeuw's resistant diagnostic (cf. Rousseeuw & Leroy, 1987, pp. 238–240) is another robust measure. It is obtained as a byproduct of the LMS estimates. It presents, therefore, obvious computational advantages. Broadly, the resistant diagnostic RD_i considers, for each case i, its maximal relative residual to an hyperplane.

The relative residual is $rr_i(\boldsymbol{\beta}) = |r_i(\boldsymbol{\beta})| / \text{med}_j |r_j(\boldsymbol{\beta})|$, where $r_i(\boldsymbol{\beta})$ is the residual to the hyperplane defined by $\boldsymbol{\beta}$. Let $u_i = \max_{\beta} rr_i(\boldsymbol{\beta})$ denote its maximum over all hyperplanes. The resistant diagnostic is then defined as the following normalized u_i

$$\text{RD}_i = \frac{u_i}{\underset{i}{\text{med}}\, u_j}$$

In practice, the RD_i values are computed by considering only the hyperplanes passing through p of the n data points. These are, indeed, the hyperplanes which are successively checked to determine the LMS estimate. Rousseeuw & Leroy again suggest a 2·5 cut-off.

2.2 *Influence measures*

The goal of the influence indicators is not to measure remoteness, but to quantify the individual statistical effects of the data on the regression results. Incidentally, they also provide indirect remoteness information. The quantified influence information is obviously interesting *per se*. It remains however only valid for a given regression estimator. Furthermore, it makes sense for least squares techniques which are very sensitive to atypical data. For robust estimators, the influence information is much less important since robust estimators limit themselves the influence of atypical data. The following measures refer, therefore, only to the influence on LS regression results. We successively consider the influence on the estimate of $\boldsymbol{\beta}$, on the fitted values and on the overall fit. All measures use the same deletion principle: the LS results are compared with those obtained with case i removed.

The influence of case i on the estimate $\hat{\boldsymbol{\beta}}$ of the regression parameters, is measured, for instance, by the distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$. Of course, different metrics can be used. Cook (1977) proposes the metric $\mathbf{X}'\mathbf{X}/ps^2$, i.e. a Mahalanobis-like metric based on the classical estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$. Cook's squared distance is thus

$$\text{CD}_i^2 = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{ps^2} = \frac{h_i r_i^2}{ps^2(1-h_i)^2}$$

A high value of CD_i^2 indicates a great influence of the ith observation on $\hat{\boldsymbol{\beta}}$. As for the cut-off, Cook (1977) suggests comparing the CD_i^2 with a central F distribution with p and $n-p$ degrees of freedom. Cook & Weisberg (1982) simply retained a cut-off of 1·0. We propose a cut-off equal to 6·25 ($=2{\cdot}5^2$) times the median of the squared Cook's distances which, as will be shown in Section 3, proves to be more helpful.

Measures of the influence of case i on the fitted values are given by distances between the vectors $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$. The Euclidean distance can be shown to be equal to the Cook distance CD_i between parameter estimates. By standardizing this Euclidean distance by $s_{(i)}\sqrt{h_i}$ we get the DFFITS_i measure introduced by Belsley *et al.* (1980). These authors, indeed, introduced DFFITS_i as the scaled change in the fit for the deleted case:

$$\text{DFFITS}_i = \frac{|\hat{y}_i - \hat{y}_{i(i)}}{s_{(i)}\sqrt{h_i}} = \left\{ \frac{(n-p-1)h_i r_i^2}{(1-h_i)[s^2(n-p)(1-h_i) - r_i^2]} \right\}^{1/2}$$

(Note, as compared with the formula in Gray, (1989a, b), the correct place of s^2.) Again, we suggest for DFFITS_i a cut-off equal to 2·5 times its median.

Finally, for measuring the influence on the overall fit, one can consider the change or ratio of overall fit indicators. One usual measure, considered for instance by Gray (1989b), is the ratio MSRATIO_i between the mean square errors s^2 and $s_{(i)}^2$:

$$\text{MSRATIO}_i = \frac{s^2}{s_{(i)}^2} = \frac{(n-p-1)(1-h_i)}{(n-p)(1-h_i) - r_i^2/s^2}$$

In the above formulae, the classical indicators have been expressed in terms of the LS residual r_i and the classical leverage indicator h_i. We can thus notice that influence indicators like the squared distance CD_i^2 of Cook or the $DFFITS_i$, by depending upon the product $h_i r_i^2$, take large values only for cases with both important leverage effect and residual. This makes such measures unsuitable to detect very well fitted **X**-outliers with very low leverage. Generalizations of influence measures to multiple-case diagnostics have also been considered. See for instance Cook & Weisberg (1982) for an extension of Cook's distance and Belsley *et al.* (1980) for a generalized $DFFITS_i$. These generalizations have, however, little practical scope because they require consideration of an intractable number of subsets of data.

As explained before, influence measures do not make much sense for robust estimators. Nevertheless, such estimators can provide alternative influence information. Typically, the inverse of the automatic weights affected to the data by robust M-estimators have such a flavour of influence measure. It has been shown (Antille & Ritschard, 1990), however, that these weights exhibit, except perhaps for the residual effect, results very similar to those of the classical measures.

2.3 Graphical summaries

A diagnostic analysis requires dealing with a very large volume of information: each diagnostic provides one value for each case. Graphical summaries should obviously help to capture the essential of this information. Several plots have been introduced in the literature. Indicators can simply be plotted versus the case number. Recently, Doreian & Hummon (1990) introduced a plot of single parameter change with error bars versus case number. Among more sophisticated plots, we can mention Atkinson's (1981, 1985) half-normal plots of the jack-knifed residual and of a transformed Cook's distance. Also, of course, there is Gray's (1989b) four-measure influence plot which has the advantage of visualizing at once the CD_i^2, $DFFITS_i$, $MSRATIO_i$ and $1 - R_i$. To detect remoteness, the most useful seems to be leverage–residual (L–R) plots (the terminology was introduced by Gray, 1986) which usually display the squared or absolute value of a standardized residual against a factor-outlyingness (leverage) measure. These plots exist in different variants, depending on the kind of standardized residual and the factor-outlyingness indicator considered. Hoaglin & Kempthorne (1986) also added contours of a cut-off for $DFFITS_i$ on L–R plots. The main interest of L–R plots lies in their ability to pinpoint simultaneously fit-outliers and factor-outlying points. Fit-outliers lie in the upper left-hand corner, well fitted factor-outlying points in the lower right one, and **y**-outliers with high leverage effect in the upper right-hand part of the plot. Superimposing the partition defined by the cut-off values on the plot clearly points out the atypical data and their nature.

Since robust indicators are mainly robust residuals and robust leverage measures, the L–R plot is obviously the most appropriate to represent robust diagnostic information. Unlike least-squares residuals, robust residuals, like LMS residuals for instance, are not constrained to sum to zero. In order to exhibit the possible asymmetry, we suggest, then, plotting the residual and not its square or absolute value.

3 A robust look at Gray's examples

We illustrate, here, with two examples how robust remoteness indicators can complement classical diagnostics. The examples are those discussed in Gray (1989a). The first concerns a simple regression and the second a multiple regression.

3.1 The homes prices example

In his first example, Gray conducts a single regression analysis on data taken from Mendenhall *et al.* (1986, pp. 544–545). These data concern the sales of 50 single-family residential homes in Eugene, Oregon in 1980. The dependent variable is the sales price (Price) and, among seven explanatory variables, Gray retained only the most significant, i.e. the square footage (Sq ft).

A scatter plot of the data is given in Fig. 1, together with the LS and LMS regression lines. The data can be found in either Mendenhall *et al.* (1986) or Gray (1986a), and are not reproduced here. Table 1, recalls the value of some classical indicators. In addition, it gives the standardized LMS residuals (fit-outlyingness indicators), the MVE distances (factor-outlyingness indicators) and the resistant diagnostics RD_i (global remoteness indicators).

In order to facilitate the comparison between the global diagnostics, the Cook CD_i^2, the $DFFITS_i$ and the Andrews & Pregibon $1-R_i$ have been standardized with respect to their median. Unusual values are marked with asterisks. The cut-offs retained are 6·25 for the standardized CD_i^2, 2·5 for the resistant diagnostic RD_i, as well as for the standardized $DFFITS_i$ and $1-R_i$. For the factor-outlyingness measures MD_i^2 and $DMVE_i^2$, the cut-off is $\chi^2_{1,0.975}=5{\cdot}03$. For the standardized residuals we also retained a cut-off of 2·5.

As shown by Gray, the classical diagnostics designate points 49 and 50 as unusual data. The distance MD_i^2, a linear transformation of h_i, indicates high leverage for observation 50 and low leverage for 49. From the LS residuals, 49 is a strong fit-outlier, while 50 seems to behave like the bulk of the data.

From the robust indicators we get a slightly different view. The robust L–R plot given in Fig. 2, confirms that 50 is outlying in the factor space, but indicates that it is also a fit-outlier. Furthermore, while confirming that 49 is a strong fit-outlier, it pinpoints six additional y-outliers: 8, 25, 44, 45, 46 and 47.

The signs of the LMS residuals provide here interesting information. Indeed, all eight are positive. Thus, all fit-outliers lie on the same side of the LMS line as can be shown in Fig. 1. This suggests that the LMS line determines some starting sales price for a given square footage. The difference between the basic and actual prices can here be explained in terms of some hidden variables giving added value such as neighbourhood, location or high quality. The fact that there are more fit-outliers among the homes with large square footage may, for instance, simply reflect the higher proportion of luxurious or pleasantly

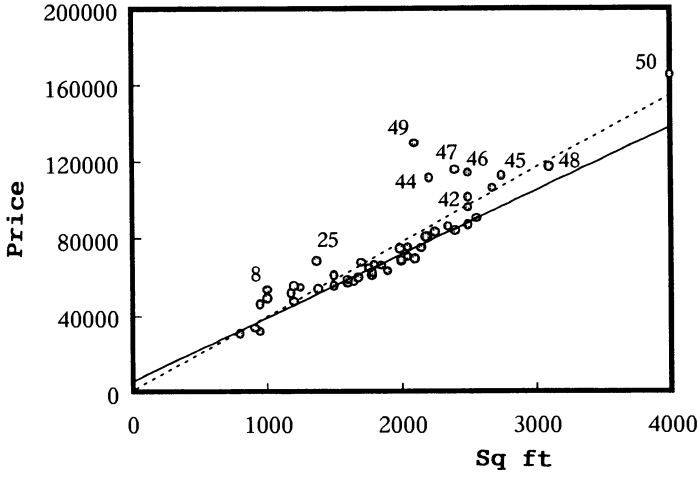

Fig. 1. LS (– – –) and LMS (——) on the residential homes data.

Table 1. Classical and robust diagnostics for the homes prices example

	Median standardized global diagnostics				Factor-outlyingness indicators		Standardized residuals	
i	CD_i^2	$DFFITS_i$	$(1-R_i)$	RD_i	DM_i^2	$DMVE_i^2$	LS	LMS
1	0·20	0·40	2·08	1·67	3·08	3·45	−0·13	−0·33
2	2·40	1·60	1·85	1·68	2·30	2·71	−0·56	−1·10
3	0·60	0·70	1·80	1·52	2·50	2·90	−0·26	−0·51
4	4·20	2·00	1·95	2·15	2·30	2·71	0·73	1·65
5	0·00	0·00	1·13	1·25	1·25	1·69	0·00	0·36
6	4·80	2·20	1·90	2·20	2·06	2·49	0·83	1·91
7	1·00	1·00	1·28	1·68	1·32	1·76	0·45	1·29
8	10·40*	3·20*	2·33	2·71*	2·06	2·49	1·21	2·71*
9	0·00	0·10	0·85	1·09	0·69	1·10	−0·03	0·48
10	1·00	1·00	1·15	1·63	1·08	1·51	0·45	1·36
11	0·40	0·60	0·78	0·84	0·41	0·77	−0·33	−0·04
12	2·40	1·60	1·40	1·95	1·25	1·69	0·71	1·85
13	0·80	0·90	0·78	0·69	0·23	0·55	−0·55	−0·39
14	1·00	0·90	0·78	0·64	0·16	0·45	−0·61	−0·47
15	0·40	0·60	0·70	0·72	0·23	0·55	−0·38	−0·04
16	0·80	0·80	0·73	0·60	0·12	0·39	−0·55	−0·32
17	0·00	0·30	0·73	1·19	0·41	0·77	0·18	1·05
18	1·40	1·20	0·85	0·70	0·04	0·24	−0·79	−0·72
19	1·00	1·00	0·78	0·63	0·03	0·23	−0·69	−0·50
20	2·20	1·50	1·05	0·86	0·00	0·10	−1·02	−1·07
21	0·40	0·60	0·60	0·54	0·05	0·27	−0·40	0·09
22	0·60	0·80	0·68	0·56	0·01	0·15	−0·57	−0·19
23	0·20	0·50	0·58	0·50	0·03	0·21	−0·37	0·19
24	0·00	0·10	0·55	1·00	0·10	0·36	0·08	1·05
25	6·40*	2·50*	1·75	2·87*	0·72	1·13	1·31	3·30*
26	1·60	1·30	0·93	0·77	0·03	0·03	−0·88	−0·67
27	1·60	1·20	0·90	0·74	0·03	0·03	−0·85	−0·62
28	2·80	1·70	1·20	0·97	0·10	0·00	−1·12	−1·07
29	1·60	1·30	0·93	0·76	0·06	0·01	−0·86	−0·59
30	0·20	0·40	0·55	0·62	0·02	0·04	−0·27	0·59
31	1·60	1·20	0·93	0·73	0·16	0·00	−0·80	−0·36
32	0·40	0·60	0·63	0·63	0·06	0·01	−0·43	0·33
33	0·80	0·90	0·78	0·77	0·23	0·01	−0·54	0·24
34	0·60	0·70	0·73	0·80	0·23	0·01	−0·46	0·42
35	0·40	0·60	0·68	0·82	0·20	0·01	−0·39	0·54
36	0·40	0·60	0·75	0·93	0·31	0·03	−0·39	0·62
37	2·60	1·60	1·20	0·91	0·64	0·14	−0·86	−0·21
38	0·80	0·90	0·90	1·00	0·52	0·09	−0·49	0·50
39	3·60	1·90	1·43	1·00	0·92	0·27	−0·93	−0·27
40	3·40	1·80	1·43	1·10	1·11	0·36	−0·84	−0·02
41	0·00	0·20	0·98	1·61	0·92	0·27	−0·12	1·45
42	0·60	0·70	1·03	2·10	0·92	0·27	0·36	2·48
43	0·20	0·30	1·30	2·18	1·55	0·59	0·15	2·20
44	13·40*	3·80*	3·30*	4·53*	0·24	0·01	2·26	6·20*
45	1·60	1·30	1·58	2·68*	1·84	0·75	0·50	3·02*
46	9·40*	3·10*	2·15	3·58*	0·92	0·27	1·50	4·89*
47	13·40*	3·80*	2·83*	4·20*	0·64	0·14	1·96	5·75*
48	1·20	1·00	2·43	2·43	3·66	1·83	−0·31	1·68
49	41·40*	8·10*	9·93*	7·49*	0·10	0·00	4·24*	10·28*
50	35·80*	6·00*	6·65*	5·30*	11·22*	6·81*	0·90	5·17*
Cut-off	6·25	2·50	2·50	2·50	5·03	5·03	2·50	2·50

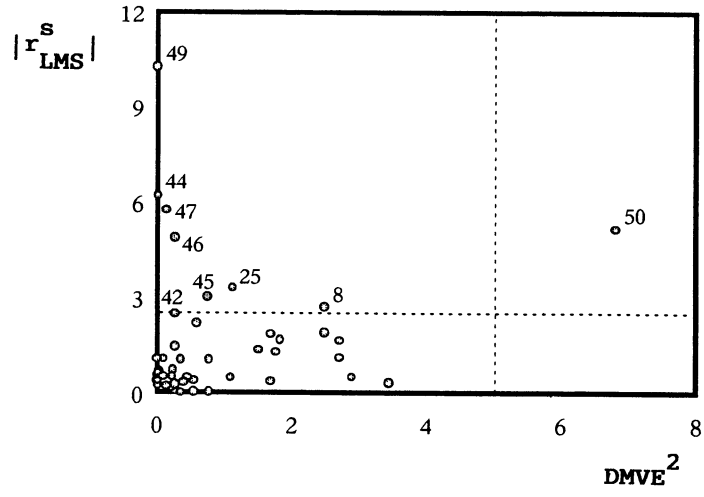

Fig. 2. Robust L–R plot for the residential homes data.

located houses among large homes. The price difference may also suggest an information asymmetry between the buyer and the seller, i.e. a buyer's lack of knowledge may have led to over-payment for a house.

To summarize, the use of robust fit-outlyingness indicators leads, in this example, to a more refined interpretation of the underlying structure of the data set. This interpretation comes up from the unbalanced (i.e. all high residuals of the same sign) configuration of the robust residuals. Indeed, such a configuration is possible only because the LMS residuals are not, contrarily to the LS residuals, constrained to add up to zero.

Finally, let us notice that the classical Mahalanobis distance MD_i and the robust distance $DMVE_i$ provide quite similar indications about factor-outlyingness. Since there is only one independent variable, the two indicators differ only in their choice of centre. MD_i measures the distance to the mean square footage (1900·4), and $DMVE_i$ to the centre (2130) of the smallest interval which covers 50% of the data. Because of the global leverage of the small homes, point 50 is somewhat paradoxically, less distant from the robust centre than from the mean. The next multiple regression example exhibits more significant differences between the two measures.

3.2 The fuel consumption example

The second example in Gray (1989a) is a multiple regression from Weisberg (1980) on data collected from the 48 contiguous United States. Per capita fuel consumption (FUEL) is regressed on the state gasoline tax (TAX), the percentage of licensed drivers in the state (LICENSE), the state average personal income (INCOME), and the total length of Federal-aid primary highways in the state (ROAD). The LS parameters with their standard errors between brackets are given in Table 2. The table also shows the LMS estimates and reweighted LS (RWLS) estimates obtained by giving zero weight to the LMS outliers. The data themselves can be found in either Weisberg (1980) or Gray (1989a).

Table 3 gives the values of classical and robust diagnostics. It includes factor-outlyingness indicators (MD_i^2 and $DMVE_i^2$), LS and LMS standardized residuals, the resistant diagnostic RD_i, and the four classical indicators displayed in Gray's (1989a) four-measure plot. For purposes of comparison, the values of the these last four measures, i.e. the squared Cook distance CD_i^2, the $DFFITS_i$, the $MSRATIO_i$ and the Andrews & Pregibon $1-R_i$, have been standardized with respect to their median. Stars designate

unusual values. The cut-offs retained are similar to those used in example 1, except for the leverage indicators. This cut-off is here $\chi^2_{4,0.975} = 11.14$ since there are four explanatory variables. Note that our cut-off values for the classical diagnostics clearly detect the six influencing cases identified by Gray, i.e. Rhode Island (RI), North Dakota (ND), South Dakota (SD), Texas (TX), Wyoming (WY) and Nevada (NV). They indicate, indeed, that two more cases merit special attention: New York (NY) and Idaho (ID). The resistant diagnostic RD_i, which behaves quite similarly to the Andrews & Pregibon $1-R_i$ remoteness indicator, seems less efficient. It pinpoints only Wyoming (WY) and Nevada (NV).

In order to understand better the role of the influencing cases, let us consider the residuals and factor-outlyingness measures. The classical L–R plot (Fig. 3) indicates that Wyoming (WY) is a fit-outlier, and that Texas (TX), New York (NY) and Nevada (NV) are high leverage points, i.e. factor-outliers. The robust L–R plot (Fig. 4) exhibits broadly, but more clearly, the same results. An important difference, however, concerns the role of Nevada (NV) which appears to be more a fit-outlier than a factor-outlier. This, together with the identification of North Dakota (ND) and South Dakota (SD) as obvious fit-outliers, reinforces Gray's interpretation. Indeed, Gray argues that the bad LS fit results from a missing home isolation factor. This argument gains evidence here with four fit-outliers corresponding to states with high degree of isolation. Note that it is not correct to use Texas (TX), which is relatively well adjusted by the LS and LMS regressions, to argue in favor of this isolation factor. Texas' large influence results only from its high leverage effect and does not reflect a departure from the model. Rhode Island (RI), on the other hand, which was discarded by Gray, provides some further arguments. Its LMS residual is

Table 2. LS, LMS and reweighted LS estimates

	Constant	Tax	License	Income	Road	R^2
LS	377·3	−34·8	1336·4	−0·067	−0·0024	0·68
	(185·5)	(12·9)	(192·3)	(0·017)	(0·0034)	
LMS	319·4	−6·9	1025·4	−0·069	−0·0005	0·80
RWLS	453·2	−15·7	951·4	−0·077	0·0020	0·74
	(128·2)	(9·2)	(141·7)	(0·012)	(0·0024)	

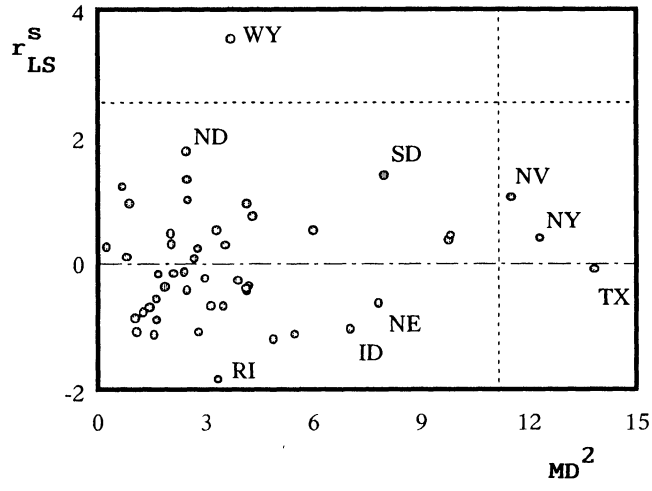

Fig. 3. Classical L–R plot for the fuel consumption data.

Table 3. Classical and robust diagnostics for the fuel consumption example

		Median standardized global diagnositics					Factor-outlyingness indicators		Standardized residuals	
	i	CD_i^2	$DFFITS_i$	$MSRATIO_i$	$1-R_i$	RD_i	MD_i^2	$DMVE_i^2$	LS	LMS
ME	1	0·20	0·44	0·99	1·00	1·01	3·55	1·44	0·27	−0·11
NH	2	0·39	0·62	0·99	0·80	1·04	2·50	1·69	−0·44	−0·74
VT	3	0·16	0·39	0·99	0·88	1·10	3·00	2·29	−0·26	−0·45
MA	4	5·46	2·37	1·03	1·61	1·26	4·90	8·03	−1·20	−1·20
RI	5	8·85*	3·10*	1·08	1·74	1·26	3·36	4·60	−1·84	−2·27
CT	6	1·12	1·06	0·99	2·35	1·59	9·78	5·86	0·35	−0·14
NY	7	1·87	1·36	0·99	2·91*	1·79	12·33*	17·29*	0·38	−0·14
NJ	8	0·45	0·67	0·99	1·16	1·15	4·22	5·84	−0·37	−0·14
PA	9	0·23	0·48	0·99	0·65	0·89	1·88	1·81	−0·38	−0·62
OH	10	0·77	0·88	1·01	0·63	0·84	1·27	1·18	−0·79	−0·44
IN	11	1·29	1·15	1·03	0·70	0·80	0·70	1·13	1·20	1·75
IL	12	1·71	1·30	1·00	2·39	1·80	9·85	8·69	0·43	0·63
MI	13	0·49	0·70	1·00	0·64	0·74	1·64	1·21	−0·58	0·09
WI	14	0·84	0·92	1·01	0·62	0·67	1·06	0·95	−0·88	−0·55
MN	15	1·33	1·16	1·02	0·73	0·73	1·09	1·47	−1·09	−0·49
IA	16	0·37	0·60	1·00	0·70	0·93	2·04	2·20	0·46	1·48
MO	17	0·01	0·08	0·99	0·39	0·63	0·81	0·69	0·08	0·92
ND	18	6·27*	2·60*	1·07	1·49	1·45	2·47	1·31	1·77	3·24*
SD	19	13·05*	3·70*	1·05	2·39	1·92	7·97	11·92*	1·39	3·94*
NE	20	2·75	1·66	1·00	2·00	1·38	7·83	12·24*	−0·64	−0·21
KS	21	1·31	1·15	1·00	1·08	0·91	3·50	5·00	−0·69	0·47
DE	22	0·67	0·82	1·00	1·16	1·23	4·16	3·86	−0·46	0·12
MD	23	1·82	1·35	1·00	1·27	1·13	4·32	6·32	0·74	0·54
VA	24	3·43	1·89	1·03	1·15	1·00	2·49	1·75	1·31	1·26
WV	25	0·68	0·82	1·00	0·64	0·78	1·45	1·73	−0·71	−0·99

Table 3. (*continued*)

NC	26	0·68	0·82	1·00	0·99	0·99	3·31	1·73	0·51	0·27
SC	27	0·05	0·23	0·99	0·68	0·70	2·11	1·09	−0·18	−0·14
GA	28	0·14	0·37	0·99	0·68	0·64	2·06	2·87	0·29	0·91
FL	29	0·04	0·19	0·99	0·28	0·43	0·27	0·17	0·24	0·58
KY	30	2·80	1·69	1·01	1·31	0·93	4·15	1·87	0·94	0·52
TN	31	0·01	0·09	0·99	0·79	0·81	2·69	1·44	0·06	0·55
AL	32	0·55	0·74	0·99	1·15	1·00	4·14	1·73	−0·41	−0·17
MS	33	5·48	2·37	1·02	1·70	1·09	5·50	6·22	−1·13	−1·36
AR	34	0·10	0·32	0·99	0·83	0·87	2·80	1·12	0·21	0·76
LA	35	0·26	0·51	0·99	1·08	1·02	3·92	2·81	−0·29	−0·63
OK	36	2·70	1·66	1·02	1·11	0·89	2·83	4·75	−1·10	−0·14
TX	37	0·19	0·43	0·99	3·20*	2·03	13·83*	11·68*	−0·11	1·41
MT	38	0·88	0·94	1·01	0·61	0·98	0·88	0·86	0·94	2·30
ID	39	6·40*	2·56*	1·02	2·00	1·37	7·05	12·82*	−1·05	−0·81
WY	40	36·02*	7·27*	1·47	3·98*	2·84*	3·71	5·54	3·54*	6·73*
CO	41	1·86	1·38	1·02	0·86	0·93	1·58	1·50	−1·14	−0·30
NM	42	2·01	1·43	1·02	0·99	1·26	2·52	1·30	0·99	2·31
AZ	43	0·05	0·23	0·99	0·59	0·93	1·70	1·35	−0·19	0·94
UT	44	1·26	1·12	1·00	1·57	1·29	6·02	5·51	0·51	1·32
NV	45	12·18*	3·53*	1·02	2·95*	2·57*	11·52*	7·23	1·03	3·86*
WA	46	0·04	0·20	0·99	0·74	0·91	2·42	1·27	−0·15	−0·41
OR	47	1·22	1·11	1·01	0·77	0·85	1·67	1·73	−0·91	0·03
CA	48	1·16	1·08	1·00	1·00	1·04	3·15	1·73	−0·69	−0·04
Cut-off		6·25	2·50	2·50	2·50	2·50	11·14	11·14	2·50	2·50

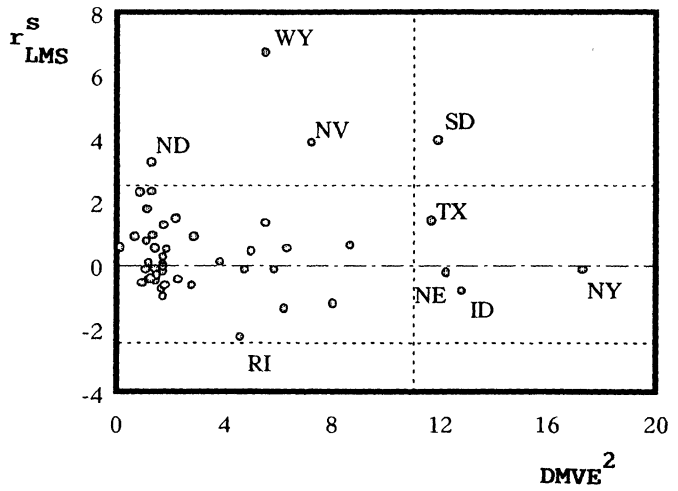

Fig. 4. Robust L–R plot for the fuel consumption data.

very near the 2·5 cut-off and is of the opposite sign of the four other high residuals. This is exactly what we would expect from a state with low degree of isolation.

Differences between robust and classical factor-outlyingness indicators are less spectacular. However, the $DMVE_i^2$ complements the information provided by the classical Mahalanobis distance. It indicates, for instance, that, despite their lower individual classical leverages, South Dakota (SD), Nebraska (NE) and Idaho (ID) are more outlying than Nevada (NV) in the factor space. This should warn us about some masked global leverage effect of these three states.

4 Conclusions

Atypical data in regression analysis obviously merit special attention. They are sources of fitting problems which can be solved through robust regression. However, as claimed by Gray (1989a), they also often provide useful unexpected information about the process under study. Many remoteness and influence indicators are now available to detect these unusual data. Classical diagnostics are mainly concerned with measuring the individual influence of the data on the LS results. On the other hand, robust techniques do not measure influence, but provide powerful direct remoteness indicators. Because of their insensitivity to masking effects, robust indicators usefully complement the classical influence measures. They excel, for instance, at distinguishing fit-outliers from factor-outlying points which is essential from the interpretation viewpoint.

Classical diagnostics, which are simply LS byproducts, have been incorporated into the major statistical packages. Robust indicators are somewhat more complicated to compute. They are, nevertheless, available in some more specialized packages such as S-PLUS, ROBETH and the Statistical Calculator SC. The LMS estimates and residuals require more computation time than LS results. We get, however the results within a few seconds on a PC for our multiple regression example. The $DMVE_i$ is much more time consuming: 28 min of CPU was necessary on an IBM 3090 for our second example. This obviously limits its practical scope.

Acknowledgement

The authors are very grateful to an anonymous referee for his helpful comments.

References

ANDREWS, D. F. & PREGIBON, D. (1978) Finding the outliers that matter, *Journal of the Royal Statistical Society, Series B*, 40, pp, 85–93.

ANTILLE, G. & RITSCHARD, G. (1990) Robust and classical outlyingness indicators: a simulation study, *Communications in Statistics: Simulation and Computation*, 19 (2), pp. 505–512.

ATKINSON, A. C. (1981) Two graphical displays for outlying and influential observations in regression, *Biometrika*, 68, pp. 13–20.

ATKINSON, A. C. (1985) *Plots, Transformations and Regression* (Oxford, Oxford University Press).

BELSLEY, D. A., KUH, E. & WELSH, R. E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (New York, Wiley).

CHATTERJEE, S. & HADI, A. S. (1986) Influential observations, high leverage points, and outliers in linear regression, *Statistical Science*, 1 (3), pp. 379–416.

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

COOK, R. D. & WEISBERG, S. (1982) *Residuals and Influence in Regression* (New York, Chapman & Hall).

DOREIAN, P. & HUMMON, N. P. (1990) Regoo plots as a regression diagnostic tool, *Quality and Quantity*, 24, pp. 213–229.

DUSOIR, T. (1989) *SC Statistical Calculator* (Hove, Lawrence Erlbaum).

GRAY, J. B. (1986) A simple graphic for assessing influence in regression, *Journal of Statistical Computation and Simulation*, 24, pp. 121–134.

GRAY, J. B. (1989a) On the use of regression diagnostics, *The Statistician*, 38, pp. 97–105.

GRAY, J. B. (1989b) The four-measure influence plot, *Computational Statistics and Data Analysis*, 8, pp. 179–188.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. & STAHEL, W. A. (1986) *Robust Statistics: The Approach Based on Influence Functions* (New York, Wiley).

HETTMANSPERGER, T. P. (1984) *Statistical Inference Based on Ranks* (New York, Wiley).

HOAGLIN, D. C. & KEMPTHORNE, P. J. (1986) Comment, *Statistical Science*, 1 (3), pp. 408–412.

LEROY, A. & ROUSSEEUW, P. J. (1984) PROGRESS: a program for robust regression, *Technical Report 201* (Brussels, Belgium, Free University).

MARAZZI, A. (1985) ROBETH, robust linear programs, *Documents 1 to 6* (Lausanne, Division de statistique et informatique, Institut Universitaire de Médecine Sociale et Préventive).

MENDENHALL, W., REINMUTH, J. E., BEAVER, R. & DUHAN, D. (1986) *Statistics for Management and Economics*, 5th edn (Boston, MA, Duxbury Press).

ROUSSEEUW, P. J. (1984) Least median of squares regression, *Journal of the American Statistical Association*, 79, pp. 871–880.

ROUSSEEUW, P. J. & LEROY, A. (1987) *Robust Regression and Outlier Detection* (New York, Wiley).

ROUSSEEUW, P. J. & VAN ZOMEREN, B. C. (1987) Identification of multivariate outliers and leverage points by means of robust covariance matrices, *Technical Report* (Delft University of Technology, The Netherlands, Faculty of Mathematics and Informatics.

ROUSSEEUW, P. J. & VAN ZOMEREN, B. C. (1990) Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, pp. 633–639.

STATISTICAL SCIENCES, INC. (1988) Modern regression module for S-PLUS, ch. 5 in: *S-PLUS User's Manual* (Seattle, WA, SSI).

WEISBERG, S. (1990) *Applied Linear Regression* (New York, Wiley).