ROBUST AND CLASSICAL OUTLYINGNESS INDICATORS A SIMULATION STUDY

Gérard ANTILLE and Gilbert RITSCHARD

Department of Econometrics, University of Geneva
2, rue Dancet, CH-1211 Geneva 4, Switzerland

Key Words and Phrases: *regression diagnostics; outliers; leverage points; M and high breakdown point estimators; robust distances*

ABSTRACT

Robust outlyingness indicators provide a more reliable alternative to least square diagnostics. This paper explains why and illustrates this superiority with a simulation study.

1. INTRODUCTION

This paper deals with outlying data in linear regression. It provides a short comparative discussion of the scope and limits of the main classical (Cook, 1977, Belsley et al. 1980, Cook-Weisberg, 1982, Chatterjee-Hadi, 1986) and robust (Rousseeuw-Leroy, 1987) indicators of outlyingness. The conclusions of this discussion are then illustrated with the results of a simulation study.

Due to the specific role of the dependent variable, we can distinguish two kinds of atypical data in regression : those which are outlying in the space of the explanatory variables and those which show an atypical response to the explanatory variables. The former are generally called high leverage points and the latter outliers.

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 differently the coefficient estimates, and the usual synthetic adjustment measures (standard deviations, R^2 , Student's t, F, Durbin-Watson, etc...). Thus, though there exist diagnostic measures, like the Cook distance, which are intended to detect atypical data independently of their nature, we shall also focus on specific outlier and specific high leverage indicators. Classical and robust outlyingness indicators are discussed in Section 2, and their performance on a simulated data set is commented upon in Section 3.

2. OUTLYINGNESS INDICATORS

In this section we describe the indicators used in this article. The indicators can be grouped into three categories: those designed for detecting outliers, those designed for detecting high leverage points, and those intended to detect influential data independently of their nature. The regression model considered is:

$$y = X\beta + \varepsilon$$

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.

2.1. Outliers detection

Outliers are observations (y_i, x_i) which significantly deviate from the model governing the bulk of the data. The classical approach to the detection of outliers focuses on standardized forms of the OLS residuals $r = y - \hat{y} = (I - H)y$, where H is the hat matrix $X(X'X)^{-1}X'$. The two main classical outliers indicators are the *standardized residual* $r_i^s = r_i/\hat{\sigma}$, where $\hat{\sigma}$ is the OLS standard error, and the *studentized residual* $r_i^t = r_i/\hat{\sigma}\sqrt{1 - h_{ii}}$, where h_{ii} is the i-th diagonal term of H.

The drawback to this classical approach is that the OLS reference hyperplane is itself affected by the outliers. A natural alternative way is to refer to a robust regression hyperplane. Several robust regression estimators have been proposed in recent literature (see for instance Hampel et al., 1986, ch. 6 and Rousseeuw-Leroy, 1987). M-estimators with a bounded influence function (Mallows and Schweppe's type estimators) limit the impact of any single observation. To represent this class of estimators we shall consider the Mallows optimal estimator. In the case of scaled residuals, a Mallows estimator T_M is implicitly defined by

$$\sum_i \omega(x)\psi_c(y - x'T_M)x = 0$$

where $\psi_c(\cdot)$ is the Huber function and $\omega(x)$ a weight function which limits leverage effects. For a given bound on the sensitivity, the optimal estimator is obtained by choosing $\omega(\cdot)$ so as to maximize the efficiency of T_M under the classical assumptions of normality (see Hampel et al., 1986, ch. 6, for more details).

Another popular robust regression estimator is the Least Median of Squared residuals (LMS) estimator of Rousseeuw (1984). This is a high breakdown point (50%) estimator. It has thus the advantage, from the diagnostic point of view, to protect against a high number of bad data points, and not only against the individual effect of each observation. The LMS estimator is the solution T_{LMS} of the problem:

$$\min_T \{\operatorname{med}_i (y_i - x_i'T)^2\}$$

Robust outliers indicators are provided by the standardized residuals from a robust fit, i.e. by:

$$r^s_{i,rob} = \frac{r_{i,rob}}{\hat{\sigma}_{rob}}$$

where $r_{i,rob}$ is the residual from the robust fit, and $\hat{\sigma}_{rob}$ the corresponding robust scale estimate.

2.2. Leverage effect indicators

High leverage points are outlying observations in the factor space. Classical measures of outlyingness are based on the Mahalanobis distance DM_i between x_i and the mean point $\bar{x}$. The square of the distance DM_i can be written (Belsley et al. (1980, pp 66-67))

$$DM_i^2 = (n-1)(h_{ii} - 1/n)$$

Thus DM_i^2 is simply an increasing linear transformation of h_{ii}, which provides an equivalent leverage indicator. For inference purposes, we shall however prefer the DM_i^2, which can be compared to a Chi-2 with p-1 degrees of freedom (i.e. the distribution of DM_i^2 when the x_i's are drawn from a multinormal distribution.)

Here again, the drawback to this classical measure is the sensitivity to high leverage of the center $\bar{x}$ and the covariance estimate $(1/(n-1))\tilde{X}'\tilde{X}$, where $\tilde{X}$ is the centered data matrix.

One alternative is to use robust estimators. For instance, Rousseeuw-Leroy (1987, pp. 258-265) consider the Minimum Volume Ellipsoid (MVE) which covers 50% of the data. Obviously, its center $c(X)$, is a robust center estimate and the covariance matrix C_{MVE} computed on the 50% covered data points is a robust estimate of the covariance matrix. A robust leverage indicator is thus provided by the MVE distance, defined by:

$$DMVE_i^2 = (x_i - c(X))'C_{MVE}^{-1}(x_i - c(X))$$

According to Rousseeuw-Leroy (1987, p. 260), the cutoff of this indicator might be taken equal to $\chi_{p,.975}$.

The weight function $\omega(x)$ used in M-estimator to downweigh high leverage points can also be used as a leverage indicator. Practically, it seems however (Ritschard et al., 1988) that the $1/\omega(x_i)$ behave very similarly to the h_{ii}, and suffer thus from the same drawback.

2.3. Global Indicators

Classical global indicators focus on the influence of each observation rather than on their outlyingness.

The Cook distance CD_i (Cook, 1977), measures the distance between the prediction vector obtained with ($\hat{y}$) and without ($\hat{y}_{(i)}$) the i-th observation:

$$CD_i^2 = \frac{1}{p\hat{\sigma}^2}(\hat{y} - \hat{y}_{(i)})'(\hat{y} - \hat{y}_{(i)})$$

It can also be viewed as the distance between the estimates of the regression coefficients $\hat{\beta}$ and $\hat{\beta}_{(i)}$ for the metric $(1/p\hat{\sigma}^2)X'X$. Cook (1977) suggests that the CD_i^2 be compared to a central F distribution with p and $n-p$ degrees of freedom. This produces however exaggeratedly high cutoff values. Practically, a cutoff of one seems more reasonable (Cook-Weisberg, 1982).

The CD_i^2 can be expressed as follows in terms of the studentized residual r_i' and the classical leverage indicator h_{ii} :

$$CD_i^2 = \frac{1}{p}(r_i')^2 \frac{h_{ii}}{1-h_{ii}}$$

High CD_i^2 requires large values of both $(r_i')^2$ and h_{ii}. Thus, for instance, the Cook distance can not detect high leverage points standing on the regression hyperplane. Furthermore, like other classical diagnostic measures, it becomes unreliable in the case of multiple atypical data.

A robust alternative is provided by the resistant diagnostic (RD_i) of Rousseeuw (cf. Rousseeuw-Leroy, 1987, pp. 238-240). Unlike the classical measures, the RD_i focuses directly on the relative position of each observation. It is based on a concept of relative residual to an hyperplane, i.e. $rd_i(\beta) = |r_i(\beta)| / \text{med}_j |r_j(\beta)|$, where $r_i(\beta)$ is the residual to the hyperplane. The resistant diagnostic RD_i is then simply the following normalized form of this maximal relative residual $u_i = \max_\beta \ rd_i(\beta)$

$$RD_i = \frac{u_i}{\underset{j}{\text{med}} \ u_j}$$

Practically, the RD_i is computed by taking the maximum of the rd_i on the set of hyperplanes passing through p of the n data points. Rousseeuw-Leroy (1987) propose a cutoff value of 2.5.

3. A SIMULATION STUDY

From the previous discussion, robust outlyingness indicators appear to be conceptually superior to classical measures. This superiority is here illustrated by showing the values of classical and robust indicators for a simulated multiple regression problem with three explanatory variables.

The data set contains 50 generated observations with both multiple high leverage points and multiple outliers. The values of the three predictors x were obtained by means of autoregressive processes. They look thus as typical time series data. By construction we can expect the first observation as well as the last ones to have high leverage. According to a simulated structural change in the autoregressive processes for the five latest data, the leverage effect of the last points should even be somewhat higher.

A main model with independent normal errors was used to generate 80% of the values of the dependent variable y. Outliers were introduced by generating the remaining 20%, i.e. data 5, 15, 25, 35, 45, 46, 47, 48, 49 and 50 with a second model which differs from the first by the values of the β coefficients.

The values of a choice of indicators computed on these data are given in Table I. These results have been obtained with the following softwares: ROBETH (Marazzi, 1985) for the Mallows estimates and weights; PROGRESS (Leroy-Rousseeuw, 1984) for the LMS estimates and the resistant diagnostic *RD* ; PROCOVIEV (Rousseeuw-Van Zomeren, 1987) for the MVE distance. Stars designate values exceeding the chosen cutoff.

For the OLS and Studentized residuals, the cutoff is set at 1.3, which is here justified since we know that there are 20% outliers. For the Mallows and LMS residuals, which are standardized with a robust scale estimate, we choose the cutoff of a 1% significance normal test, i.e. 2.4. The cutoff retained for the classical distance *DM* and the robust distance *DMVE* is respectively $(\chi^2_{2,0.9})^{1/2} = 2.15$ and $(\chi^2_{3,0.99})^{1/2} = 3.37$. The "Mallows" leverage indicator is a standardized inverse of the weight $\omega(x)$, i.e. $\text{med}(\omega(x))/\omega(x)$. We choose the cutoff 2 which points to data with weight less than half the median weight. For the global diagnostics we retained respectively 1 and 2.5 for the Cook distance and the resistant diagnostic *RD*.

A quick glance at Table I shows that:

- robust residuals detect correctly all outliers while least squares residuals not only miss true outliers, but also designate good data as outliers;
- the classical distance and the M-estimator weights $\omega(x)$ behave analogously: both seem to miss some leverages. The robust distance *DMVE* points them out much more clearly;
- global diagnostics provide only limited information. The Cook distance detects nothing with the retained cutoff value and the resistant diagnostic points out only some high leverages.

TABLE I

	Standardized Residuals				Leverage			Global Diagnostics	
	OLS	Studentized	Mallows	LMS	MD	Mallows	DMVE	Cook	RD
1	-2.35 *	-2.53 *	-0.82	-0.35	2.42 *	2.06 *	42.89 *	0.51	4.64 *
2	-2.06 *	-2.18 *	-0.65	-0.11	2.18 *	1.85	39.48 *	0.40	4.28 *
3	-1.78 *	-1.86 *	-1.54	-1.76	1.85	1.56	34.98 *	0.29	3.78 *
4	-1.32 *	-1.37 *	-0.05	0.49	1.71	1.40	32.61 *	0.20	3.53 *
5	0.35	0.37	6.32 *	9.11 *	2.11	1.38	29.73 *	0.07	3.20 *
6	-0.79	-0.81	-1.38	-1.88	1.45	1.01	24.73 *	0.10	2.67 *
7	-0.39	-0.40	0.06	0.36	1.34	1.00	21.82 *	0.05	2.36
8	-0.01	-0.01	0.30	0.50	1.45	1.00	18.98 *	0.00	2.04
9	0.02	0.02	0.00	-0.02	1.26	1.00	16.90 *	0.00	1.82
10	0.21	0.22	-0.28	-0.45	1.53	1.00	14.76 *	0.03	1.59
11	0.19	0.20	-0.66	-0.87	1.60	1.00	13.46 *	0.03	1.45
12	0.35	0.35	-0.32	-0.56	1.18	1.00	10.86 *	0.04	1.17
13	0.71	0.73	0.41	0.09	1.58	1.00	10.13 *	0.10	1.06
14	0.85	0.88	0.65	0.53	1.59	1.00	8.93 *	0.12	0.94
15	-0.46	-0.48	-8.20 *	-12.60 *	2.00	1.19	7.62 *	0.08	1.89
16	0.96	0.99	-0.05	-0.38	1.53	1.00	4.48 *	0.13	0.56
17	1.21	1.26	0.09	-0.57	1.82	1.17	3.46 *	0.20	0.64
18	1.20	1.23	0.76	0.89	1.33	1.00	1.17	0.15	0.52
19	1.07	1.09	0.61	0.76	1.08	1.00	1.54	0.12	0.40
20	0.93	0.93	0.56	0.61	0.65	1.00	0.62	0.08	0.32
21	0.61	0.61	-0.48	-0.95	0.42	1.00	0.42	0.05	0.27
22	0.55	0.55	-0.59	-1.14	0.37	1.00	0.53	0.04	0.29
23	0.54	0.56	0.96	1.67	1.68	1.00	1.34	0.08	0.45
24	0.55	0.56	0.95	1.59	1.54	1.00	1.17	0.08	0.43
25	-1.07	-1.10	-8.38 *	-12.08 *	1.43	1.00	1.24	0.14	1.99

TABLE I (continued)

	Standardized Residuals				Leverage			Global Diagnostics	
	OLS	Studentized	Mallows	LMS	MD	Mallows	DMVE	Cook	RD
26	0.69	0.72	1.28	2.19	1.76	1.00	1.22	0.11	0.44
27	0.41	0.42	-0.56	-0.76	1.21	1.00	0.98	0.05	0.35
28	0.68	0.69	1.22	1.91	1.36	1.00	1.11	0.09	0.40
29	0.50	0.51	0.43	0.72	1.37	1.00	1.36	0.06	0.40
30	0.19	0.19	-0.19	-0.05	1.78	1.00	1.44	0.03	0.45
31	0.30	0.30	0.23	0.37	1.32	1.00	1.37	0.04	0.38
32	0.34	0.34	-0.35	-0.57	1.18	1.00	1.09	0.04	0.35
33	0.40	0.40	-0.20	-0.40	0.92	1.00	0.80	0.04	0.30
34	0.78	0.78	1.00	1.06	0.57	1.00	0.60	0.06	0.44
35	-0.99	-1.00	-8.37 *	-12.37 *	0.88	1.00	0.72	0.10	1.96
36	0.59	0.60	0.05	-0.11	0.70	1.00	0.96	0.05	0.28
37	0.46	0.47	0.14	-0.11	0.91	1.00	0.67	0.05	0.41
38	0.59	0.60	0.49	0.40	0.89	1.00	1.17	0.06	0.42
39	0.65	0.67	0.54	-0.11	1.56	1.14	1.32	0.09	0.78
40	0.52	0.56	0.33	-0.69	2.39 *	1.60	2.25	0.11	1.03
41	0.53	0.57	-0.08	-1.29	2.31 *	1.56	1.37	0.11	1.00
42	0.37	0.39	-0.47	-1.80	2.31 *	1.55	1.17	0.07	1.00
43	0.72	0.75	1.03	0.46	2.03	1.39	0.57	0.13	0.97
44	0.82	0.88	0.52	-0.52	2.43 *	1.63	1.17	0.18	1.06
45	-0.59	-0.62	-5.84 *	-9.52 *	2.13	1.43	0.85	0.11	1.43
46	-1.01	-1.05	-11.62 *	-17.80 *	1.53	1.02	10.85 *	0.14	2.76 *
47	-1.35 *	-1.40 *	-17.56 *	-26.67 *	1.69	1.23	20.79 *	0.20	4.14 *
48	-1.62 *	-1.71 *	-24.67 *	-37.28 *	2.08	1.75	31.40 *	0.30	5.89 *
49	-1.89 *	-2.10 *	-32.39 *	-49.16 *	2.93 *	2.56 *	42.46 *	0.52	7.76 *
50	-2.15 *	-2.66 *	-39.49 *	-60.25 *	4.01 *	3.40 *	50.66 *	0.97	9.44 *

4. CONCLUSION

By construction robust outlyingness indicators appear to be superior to classical ones. Our simulation study exhibits this superiority. It also shows that usual cutoff values for classical indicators should be lowered when multiple atypical data can be expected.

BIBLIOGRAPHY

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

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*. New York and London, Chapman and Hall.

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

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.

Leroy, A. and Rousseeuw, P.J. (1984). "PROGRESS : A program for Robust Regression," Technical Report 201, Free University, Brussels, Belgium.

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

Ritschard, G., Antille, G., Alfaro, W. and Parmeggiani, L. (1988). "Régression robuste et données atypiques," Cahiers du Département d'économétrie, 88-02, Université de Genève.

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

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

Rousseeuw, P.J. and Van Zomeren, B.C. (1987). "Identification of Multivariate Outliers and Leverage Points by Means of Robust Covariance Matrices," Technical Report, Faculty of Mathematics and Informatics, Delft University of Technology, The Netherlands.

Received August 1989; Revised March 1990.

Recommended Anonymously.

COMMUN. STATIST. -SIMULA., 19(2), 513-526 (1990)

COMPARISON OF TEST STATISTICS FOR THE CORRELATION COEFFICIENT IN BIVARIATE NORMAL SAMPLES WITH TYPE II CENSORING

D.C. Vaughan

Department of Mathematics
Wilfrid Laurier University
Waterloo, Ontario, Canada N2L 3C5

Key Words and Phrases: *bivariate life testing; censored samples; correlation coefficient; MML estimators; robustness; Fisher z-transformation; concomitant.*

ABSTRACT

A comparison of five possible statistics for testing the null hypothesis $Ho:\rho = \rho_0$, with ρ_0 not necessarily 0, is undertaken, using the modified maximum likelihood estimators of Tiku and Gill (1989). Symmetric and asymmetric Type II censored samples are considered, as is the robustness of these statistics to departures from normality.

1. INTRODUCTION

Suppose that the random variable (X, Y) has a bivariate normal distribution $BN(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2; \rho)$. In certain experimental situations, a number of the smallest and largest observations on one of the two variables, say Y, may not be available. For example Harrell and Sen (1979), in an experiment designed to study the relationship between arteriosclerosis and length of life, a number N of rhesus monkeys are fed an atherogenic diet. An autopsy is performed on the $K \leq N$ monkeys which die during the term of the experiment in order to measure X: the amount of fatty plaque in the aorta.