BREAKDOWN AND EFFICIENCY IN ROBUST REGRESSION

Marco Riani[1], Andrea Cerioli[1] and Francesca Torti[2]

[1] Dipartimento di Economia
Università degli Studi di Parma
Via Kennedy 6, 43100 Parma, Italy
(e-mail: mriani@unipr.it, andrea.cerioli@unipr.it)
[2] Dipartimento di Scienze Statistiche
Università degli Studi di Milano Bicocca
Via Bicocca degli Arcimboldi 8, Milano, Italy
Università degli Studi di Parma
Via Kennedy 6, 43100 Parma, Italy
(e-mail: francesca.torti@nemo.unipr.it)

ABSTRACT. In this paper we consider regression data possibly contaminated with the presence of multiple outliers. The goal of this work is to compare the information which can be extracted from the use of traditional robust estimators of regression which try to combine robustness and efficiency with those which use a flexible level of trimming and are based on the forward search estimator.

1 INTRODUCTION

Data are an overwhelming feature of modern life. As the amount of data increases so do the challenges facing the statistician in trying to extract information from ever larger data sets. We argue that larger data sets are also more complex and require flexible multiple analyses in order to reveal their structure. Only then all information can be efficiently extracted.

Given that the presence of outliers can have strong effects on the results of statistical models, it is necessary to use the tools of robust statistics to preliminary detect them. However, many asymptotic results available in the robustness literature require regularity conditions that are difficult to verify in practice, or that may not be valid. For example, typical regularity conditions under which asymptotic properties of robust estimators have been studied include: symmetric errors for the good part of the data (Simpson and Yohai, 1998); known error scale (Markatou and Hettmansperger, 1990); or conditions that involve the expected value of the estimating equations under the unknown distribution of the data (e.g. Huber, 1981). Salibian-Barrera and Zamar (2004) found that there is a trade-off between the breakdown point of the S-estimator (to be described in section 3) and the size of the neighbourhood where uniform asymptotics hold.

In this paper we give an example where even in the presence of a core population and some isolated or clustered outliers, the use of traditional robust methods based on a prefixed level of trimming and a predetermined asymptotic level of efficiency can offer a distort view of the data. The goal of this paper is to compare the information which can be extracted from the use of traditional robust estimators with those which come out from the use a flexible level of trimming like that based on the forward search (Riani, Atkinson and Cerioli, 2009; Atkinson, Riani and Cerioli, 2010).

The paper is structured as follows. In section 2 we introduce the notation. In section 3 we recall the main robust estimators in regression and explain the choices which have to be made to obtain robust estimators which asymptotically try to combine robustness and efficiency. In section 4 we briefly recall the basic steps of the forward search algorithm. In section 5 we use a well known dataset which contains 6 masked outliers and compare the information coming from the different approaches.

2 NOTATION

Consider the usual regression model with random carriers where we observe i.i.d. random vectors $(y_i, x_i^T) \in \Re^{p+1}$, $i = 1, \ldots, n$, where $y_i \in \Re$ and $x_i \in \Re^p$ satisfy

$$y_i = x_i^T \beta + u_i \qquad i = 1, \ldots, n. \tag{1}$$

u_i are random errors independent from the covariates (x_i) which have common variance equal to σ^2 and $\beta \in \Re^p$ is the parameter of interest. We are concerned with the case where a certain proportion of the observations may not follow model (1) above. Given an estimator of β, say $\hat{\beta}$, the residuals are defined as

$$r_i(\hat{\beta}) = y_i - x_i^T \hat{\beta}.$$

3 TRADITIONAL ROBUST ESTIMATORS WHICH USE A FIXED LEVEL OF TRIMMING

Traditional robust estimators attempt to limit the influence of outliers by replacing in the estimation of β the square of the residuals with a less rapidly increasing loss function or by a function ρ of the residuals themselves which is bounded. The regression M-estimate of location β (say $\hat{\beta}_M$) is the value that minimizes the following objective function

$$\sum_{i=1}^{n} \rho \left(\frac{r_i(\hat{\beta}_M)}{\sigma} \right) = \min. \tag{2}$$

Perhaps the most popular choice for the ρ function in (2) is Tukey's Biweight function

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if } |x| \leq c \\ \frac{c^2}{6} & \text{if } |x| > c, \end{cases} \tag{3}$$

where $c > 0$ is a tuning constant which is linked to the breakdown point of the estimator of β. In equation (2) it is assumed that σ is known. However, when this condition is not fulfilled, it is necessary to use an auxiliary robust scale estimate $(\hat{\sigma})$ to make $\hat{\beta}_M$ scale equivariant.

An M-estimator of scale $\hat{\sigma}_M$ is defined as the solution to the following equation

$$\frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{r_i}{\hat{\sigma}_M} \right) = K, \tag{4}$$

where $r_i = r_i(\hat{\beta}_M)$, $\hat{\sigma}_M = \hat{\sigma}(r_1(\hat{\beta}_M), \ldots, r_n(\hat{\beta}_M))$ and K is a constant which is linked to the breakdown point of the scale estimator. It is worthwhile to notice that in equation (4) we have used the symbol ρ_1 because the ρ function which is used to obtain the scale estimator is not necessarily the same which is used in (2).

If we take the minimum value of $\hat{\sigma}_M$ which satisfies equation (4), we obtain the so called *S*-estimate of scale ($\hat{\sigma}_S$) and the associated estimate of the vector of regression coefficients ($\hat{\beta}_S$). The word *S* estimator comes from the fact that it is derived from a scale statistic in an implicit way. The computation of S-regression estimators is generally difficult. Recent advances in this direction include Salibian-Barrera and Yohai (2006).

The MM-regression estimator $\hat{\beta}_{MM}$ is defined as any local minimum of $f : R^p \to R_+$, where

$$f(\hat{\beta}_{MM}) = \frac{1}{n}\sum_{i=1}^{n} \rho_2\left(\frac{r_i}{\hat{\sigma}}\right) \tag{5}$$

and ρ_2 is possibly another ρ function. Function f is minimized with respect to β for fixed $\hat{\sigma}$. Among the possible local minima which have been found, we choose the one for which (5) is smallest. In equation (5) $\hat{\sigma}$ is any scale estimator satisfying equation (4). It is common however to use $\hat{\sigma}_S$ (the minimum value).

The consistency and asymptotic distribution of MM-estimates when the observed data follow the central model (1) has been studied by Yohai (1987) for the case of random covariates, and by Salibian-Barrera (2006) for fixed designs. Consistency and asymptotic distribution of S-estimators has been studied by Rousseeuw and Yohai (1984), Davies (1990) and Salibian-Barrera (2006).

4 Robust estimators which use a flexible level of trimming

In the robust estimators, shown in the previous section, the breakdown point and therefore the amount of trimming we are willing to tolerate must be fixed a priori. The forward search, on the other hand, in order to avoid this problem, fits subsets of observations of size m to the data, with $m_0 \leq m \leq n$. Let $S_*^{(m)}$ be the subset of size m, for which the matrix of regressors is $X(m^*)$. Least squares on this subset of observations yields parameter estimates $\hat{\beta}(m^*)$ and $s^2(m^*)$, the mean square estimate of σ^2 on $m-p$ degrees of freedom. Residuals can be calculated for all observations including those not in $S_*^{(m)}$. The n resulting least squares residuals are

$$e_i(m^*) = y_i - x_i^T \hat{\beta}(m^*). \tag{6}$$

The search moves forward with the augmented subset $S_*^{(m+1)}$ consisting of the observations with the $m+1$ smallest absolute values of $e_i(m^*)$. The estimates of the parameters are based on only those observations giving the central m residuals.

To start we take $m_0 = p$ and search over subsets of p observations to find the subset, out of 1,000, that yields the least median of squares (LMS) estimate of β (Rousseeuw, 1984). Although this initial estimator is not $\sqrt{n}$-consistent (Hawkins and Olive, 2002) our results show that the properties of the initial estimator are not relevant, provided masking is broken. Identical inferences are obtained using the least trimmed squares estimator (LTS) except sometimes when m is small and $n/p < 5$. Random starting subsets also yield indistinguishable results

over the last one third of the search. Examples for multivariate data are in Atkinson and Riani (2007). The forward search, adding, and sometimes deleting, observations provides a bridge between the initial estimate and $\sqrt{n}$-consistent parameter estimates for the uncontaminated observations as the sample size goes to infinity.

5 Data analysis using a fixed vs flexible level of trimming

The purpose of this section is to compare the output which comes out from the use of traditional robust estimators in which we have 1) to choose the kind of ρ function both for the location and the scale estimator; 2) to fix a priori the percentage of trimming one is willing to tolerate; 3) to specify the the number of subsamples which one wants to extract; 4) to specify the number of steps to find the minimum.

Atkinson and Riani (2000), pp. 5-9, give an example of a regression dataset with 60 observations on three explanatory variables where there are 6 masked outliers that cannot be detected using standard analyses. The scatter plot of the response y against the three explanatory variables of and the traditional plot of residuals against fitted values and the qqplot of OLS residuals (not given here for lack of space) do not reveal particular observations far from the bulk of the data.

Figure 1 shows the index plots of the scaled MM residuals. In the left panel we have used a preliminary S estimate of σ based on a 50% breakdown point and in the MM step an efficiency of 90%. In the right panel we have used the same preliminary estimate of σ as before but an efficiency of 95%. As the reader can see, these two figures produce a very

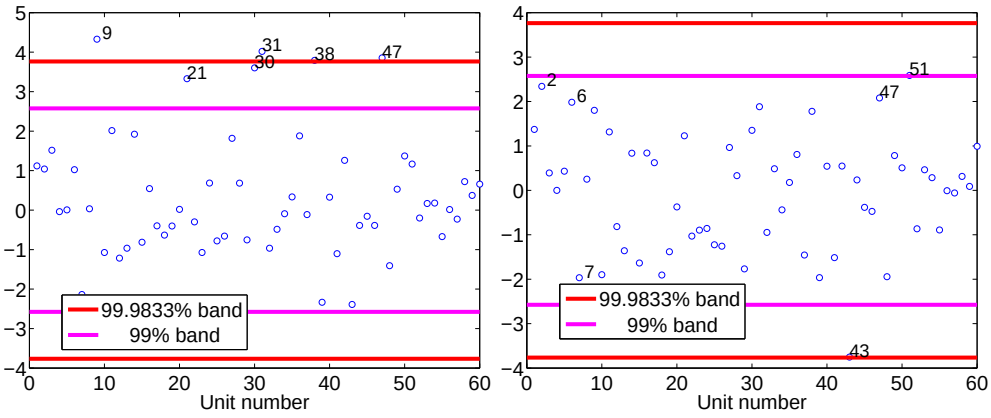

Figure 1. Index plots of robust scale residuals obtained using $\hat{\beta}_{MM}$ using a preliminary S-estimate of scale based on a 50% breakdown point. Left panel: 90% nominal efficiency; right panel: 95% nominal efficiency. The horizontal lines correspond to the 99% individual and simultaneous bands using the standard normal.

different output. While the plot of the right (which is similar to the masked index plot of OLS residuals) highlights the presence of a unit (number 43) which is on the boundary of a

simultaneous confidence band, the plot on the left (based on a smaller efficiency) suggests that there may be 6 atypical units (9, 21 30, 31, 38 47). Many questions, however, still remain unanswered such as whether these 6 units form a group or what is the connection of these units with unit 43 and what is the effect that these units exert on the fitted model. Because of the way in which models are fitted (either with LS, LTS or LMS, M, S or MM approach) we lose information about the effect of individual observations on inferences about both the form and the parameters of the model. In order to understand the effect that each unit, outlier or not, exerts on the fitted model, it is necessary to start with a subset of data and monitor the required diagnostics. In the example above, if we start with a least squares fit to 4 observations, robustly chosen, we can calculate the residuals for all 60 observations and next fit to the five observations with smallest squared residuals. In general, given a fit to a subset of size m, we can order the residuals and take, as the next subset, the $m + 1$ cases with smallest squared residuals. This gives a forward search through the data (Atkinson and Riani 2000; Riani, Atkinson and Cerioli, 2009), ordered by closeness to the model. We expect that the last observations to enter the search will be those which are furthest from the model and so may cause changes once they are included in the subset used for fitting. Figure 2 shows the monitoring of the scaled squared residuals for the 60 units of the dataset. In this case

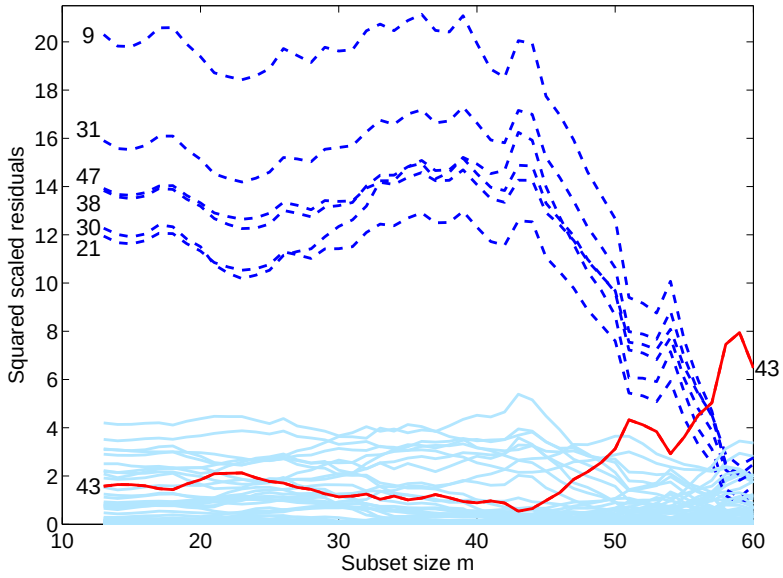

Figure 2. Monitoring of squared scaled residuals. The outliers have been drawn with dotted lines, while the trajectory of the case which in the final step shows the largest residual has been drawn with a solid line. All the other unimportant trajectories have been shown in faint grey.

we have initialized the search with LTS, investigating all possible 60 choose 4 subsets and taking the one with the smallest sum of the 50% smallest squared residuals, although virtually

identical results are obtained starting from different starting points. This fascinating plot not only reveals the presence of six masked outliers but also that

1. the six outliers form a cluster, because their trajectories are very similar. In other words, they respond in a similar way to the introduction of units into the subset;

2. the residuals of the six outliers at the end of the search are completely mixed with those of the other units, therefore traditional methods based on single deletion diagnostics cannot detect them;

3. the entry of the six outliers causes a big increase in the trajectory of the residual for unit 43. Indeed this is the unit which in the final step has the largest residual and may be wrongly considered as an outlier from the traditional plot of residuals against fitted values or an index plot of robust residuals which uses a typical 95% asymptotic efficiency (see right panel of Figure 1).

REFERENCES

ATKINSON, A.C., RIANI, M. (2000): *Robust Diagnostic Regression Analysis*, Springer-Verlag, New-York.

ATKINSON A.C., RIANI M. (2007): Exploratory Tools for Clustering Multivariate Data. *Computational Statistics and Data Analysis. vol. 52*, 272–285

ATKINSON, A.C., RIANI, M. CERIOLI A. (2010): The forward search: theory and data analysis (with discussion), *Journal of the Korean Statistical Society, vol. 39*, 117–134.

HAWKINS, D. M., OLIVE, D. J. (2002): Inconsistency of Resampling Algorithms for High-Breakdown Regression Estimators and a New Algorithm, *Journal of the American Statistical Association, Vol. 97, No. 457*, 156–159.

DAVIES, L. (1990): The asymptotics of S-estimators in the linear regression model. *The Annals of Statistics, 18*, 1651–1675.

HUBER, P. J. (1981): *Robust Statistics*. Wiley, New York.

MARKATOU, M., HETTMANSPERGER, T. P. (1990): Robust bounded-influence tests in linear models. *Journal of the American Statistical Association, 85*, 187–190.

RIANI, M. ATKINSON, A. C., CERIOLI, A. (2009): Finding an unknown number of multivariate outliers, *Journal of the Royal Statistical Society B, 71*, 1–21.

ROUSSEEUW, P. J. (1984): Least Median of Squares Regression, *Journal of the American Statistical Association, Vol. 79, No. 388*, 871–880.

ROUSSEEUW, P. J., YOHAI, V. J. (1984): Robust regression by means of S-estimators. In J. Franke, W. Hardle, D. Martin (Eds.), *Robust and Nonlinear Time Series. Lecture Notes in Statistics, vol. 26*, 256–272. Berlin: Springer.

SALIBIAN-BARRERA, M. (2006): The asymptotics of MM-estimators for linear regression with fixed designs. *Metrika, 63*, 283–294.

SALIBIAN-BARRERA, M., YOHAI, V. J. (2006): A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics, 15*, 414–427.

SALIBIAN-BARRERA, M., ZAMAR, R. H. (2004): Uniform asymptotics for robust location estimates when the scale is unknown. *The Annals of Statistics, 32*, 1434–1447.

SIMPSON, D. G., YOHAI, V. J. (1998): Functional stability of one-step GM-estimators in approximately linear regression. *The Annals of Statistics, 26*, 1147–1169.

YOHAI, V. J. (1987): High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics, Vol. 15, No. 2*, 642–656.