Very Robust Regression: Frameworks for Comparisons

Anthony C. Atkinson[1], Marco Riani[2] and Domenico Perrotta[3]
[1]The London School of Economics, London WC2A 2AE
[2] Dipartimento di Economia, Università di Parma, Italy
[3]European Commission, Joint Research Centre, Ispra, Italy
[1] Corresponding author: Anthony Atkinson, e-mail: a.c.atkinson@lse.ac.uk

Abstract

We use a smoothly parameterized series of examples that shows, in a systematic way, how the behaviour of algorithms for very robust regression depends on the closeness of the outliers to the main data. An algorithm based on the Forward Search outperforms Least Trimmed Squares and its reweighted version. An empirical measure of the overlap of the two samples structures our investigation of the bias and variance of the robust estimators. We also consider the power of tests for outliers associated with the estimation methods.

Keywords: forward search, least trimmed squares, multiple outliers, parameterized alternatives, structured simulation, overlapping index

1. Introduction

Multiple regression is one of the main tools of applied statistics, especially in engineering and technological applications. It has however long been appreciated that ordinary least squares as a method of fitting regression models is exceptionally susceptible to the presence of outliers. Instead, very robust methods, that asymptotically resist 50% of outliers, are to be preferred.

Very robust regression was introduced by Rousseeuw (1984) who developed suggestions of Hampel (1975) that led to the Least Median of Squares (LMS) and Least Trimmed Squares (LTS) algorithms. For some history of more recent developments see Rousseeuw and Van Driessen (2006). A different approach, based on the Forward Search (FS) was suggested by Atkinson and Riani (2000). More general discussions of robust methods are in Maronna, Martin, and Yohai (2006) and Morgenthaler (2007), with a description of the current FS algorithm in Torti et al. (2012). We compare the performance of publicly available versions of LTS and its reweighted version LTSr, with that of an FS-based algorithm. However, the comparisons have to be carefully designed to reveal the differences between the performance of the various methods.

Although all algorithms asymptotically resist contamination up to 50%, this resistance depends on the remoteness of the outliers. If they are close to the main body of data, the performance of the various estimators can differ markedly. We

use a parameterized family of models for generation of the outliers in a simulation experiment and so reveal the properties of the estimators as the outlier generation process varies in a smooth way. We choose parameter values so that the outliers start remote from the main data, gradually become almost indistinguishable from it and then move away again. This structure provides a powerful framework for our numerical investigation.

It is our intention in this paper to illustrate the usefulness of this systematic framework. Full details of a more general approach are given in the paper of which this is a summary (Riani, Atkinson, and Perrotta 2013) and in the conference presentation.

2. Models, Data and Robustness

Let the data be generated by the mixture model

$$M(\theta) = (1-\epsilon)M_1(\theta_1) + \epsilon M_2(\theta_2) \qquad (0 < \epsilon < 0.5). \tag{1}$$

Robust methods fit $M_1(\theta)$ to the data in ignorance of the form of the outlier generating model $M_2(\theta_2)$, which can be quite general; it can, for example distribute observations randomly over a large space, concentrate them close to a point or, in the case of interest to us, be a second regression model. There is no difficulty in having $M_1(\theta) = M_2(\theta)$ but then we must have $\theta_1 \neq \theta_2$.

The properties of robust estimators depend on the "distance" between the two models. The breakdown point of an estimator $\tilde{\theta}$ is the largest value of ϵ for which $\tilde{\theta}$ provides information about θ, remaining bounded and bounded away from the edge of the parameter space as the observations take any values (see Maronna, Martin, and Yohai 2006, §3.2). In regression theoretical interest has been in behaviour as $y_{M2} \sim M_2(\theta_2) \to \infty$.

As $y_{M2} \to \infty$ the observations y_{M1} and y_{M2} from the two models become increasing well separated. We are also interested in those data configurations when the observations are not so separated, so that both y_{M1} and y_{M2} may be used in estimating θ because of overlap between the two samples. To distinguish between the estimators we look at finite sample properties as the distance between Y_{M2} and Y_{M1} varies in a smooth way as a function of θ_2.

The example in §3 shows the differing effect on the robust regression procedures of lack of separation between the two groups. Insight into this behaviour comes from a measure of the overlap of the observations coming from M_1 and M_2.

There is a sample $\mathcal{S}_1$ of n_1 observations from $M_1(\theta_1)$ the distribution of which is $F_1(y_i; x_i, \theta_1)$ and a sample $\mathcal{S}_2$ of n_2 observations from $M_2(\theta_2)$ with expectation $\mathrm{E}(y; x_i, \theta_2)$. In this paper, with $M_1(\theta_1)$ normal theory regression, we count the total number of observations in $\mathcal{S}_2$ the expectations of which lie in $(1-\gamma)$ symmetrical probability intervals around the expectation of $M_1(.)$. As γ decreases, the strip becomes broader and this number, which we call the overlapping index O,

tends to n_2, the number of observations in $\mathcal{S}_2$.

3. The Numerical Effect of Overlap

We compare and contrast the properties of three methods for very robust regression. The algorithms that we use are all publicly available from the Forward Search Data Analysis (FSDA) Matlab toolbox at `www.riani.it/MATLAB`. A recent discussion of the Forward Search is given by Atkinson, Riani, and Cerioli (2010). In order to calibrate our outlier detection rule to provide the desired size for the samplewise test, we adapt the algorithmic outlier rejection procedure for multivariate data of Riani, Atkinson, and Cerioli (2009). An important factor in our ability to conduct as many simulations as were necessary is the efficient sampling of subsets provided in FSDA.

For Least Trimmed Squares (LTS) we search for the subset of size h for which the LS estimate of β has the minimum residual sum of squares. With $h = [n/2] + [(p+1)/2]$, LTS has an asymptotic breakdown point of 50%. To increase efficiency, reweighted versions of LTS estimators can be computed. These reweighted estimators, denoted LTSr, are computed by giving weight 0 to observations which LTS suggests are outliers. We then obtain a sample of reduced size to which OLS is applied. For comparison of results from LTSr with those from the FS we perform the outlier test at a Bonferronised size $\alpha^* = \alpha/n$, so taking the $1 - \alpha^*$ cutoff value of the reference distribution. In our calculations $\alpha = 0.01$. See Torti et al. (2012) for the important computational details of all algorithms .

In our numerical investigation we generate data from a mixture of regression models, so that both $M_1(\theta_1)$ and $M_2(\theta_2)$ are of the form $\alpha + \beta x$. The parameters for $M_1(.)$ were kept fixed ($\theta_{1,k} = \theta_{1,0}\ \forall\ k$) with $\alpha_1 = 1$ and β_1 = 3. For Model 2 we had $\beta_2 = 2$, so that the slopes of the lines were not the same.

In addition to β_2, the set Θ_2 consisted of a range of 21 values of α_2 from 0.5 to 2.5 in steps of 0.1. To emphasize that only α_2 changes we call this set $\mathcal{A}$. We are therefore looking at properties as the regression line for the outliers moves vertically. The right-hand panel of Figure 1 indicates that, over this part of $\mathcal{A}$, Model 2 goes from being close Model 1 to lying above it.

In our simulations we take $n_1 = 100$ with $x_{M1,i} \sim U(0,1)$ and $n_2 = 30$ with now $x_{M2,i} \sim U(0.21, 0.91)$. These values of x were generated once for the whole set of simulations. The observational errors for both populations were i.i.d. $\mathcal{N}(0, 0.1)$. A single regression model was fitted and the overlapping index calculated for $1 - \gamma = 99\%$. There were 300 hundred simulations for each value of α_2.

The left-hand panels of this figure shows boxplots of the values of the three estimators for some values of α_2. The plots in Figure 1 start with high overlapping and show what happens as the outliers move above M_1. In the topmost plot $\alpha_2 = 1.5$ and $O_{1.5} = 29$. With this amount of overlap the three estimators all have means around 2.85 and similar variances, with a few values nearer 3 for FS. When, in the

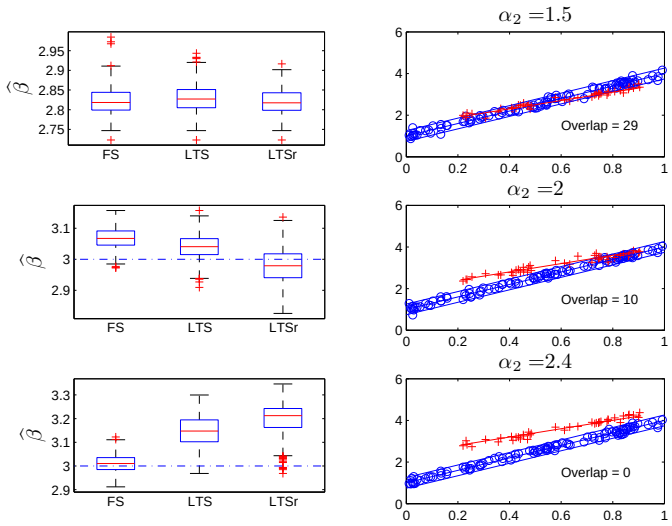

Figure 1: Three simulated data sets with $n_1 = 100$ and $n_2 = 30$ for α_2 = 1.5, 2 and 2.4. Left-hand panels: boxplots of estimates of β (dotted line: $\beta_1 = 3$); FS, Forward search, LTSr, reweighted Least Trimmed Squares. Right-hand panels: typical simulations for these three members of $\mathcal{A}$

middle plot, $\alpha_2 = 1.5$ and the value of O has decreased to 10, the closest outliers are for values of x above 0.5. Now FS shows some positive bias, LTS a little less, while LTSr still has some negative bias. The variance of LTSr is again largest.

The final panels in Figure 1 are for $\alpha_2 = 2.4$. Now $O_{2.4} = 0$ and the FS has very small bias. However the other two estimators have appreciable upward bias due to inclusion of observations from M_2 with larger values of x. This effect is most appreciable for LTSr, which also has the largest variance.

These and the results for the other values of α_2 are summarised in Figure 2 where the upper two panels show the behaviour of the three estimators of α_1. The left-hand panel shows the partial sums of the bias, summed over $\mathcal{A}$. All figures show the same ordering with FS best, followed by LTS with LTSr worst, sometimes clearly so.

For the extreme values of α_2 the horizontal value of the summed squared bias for FS shows that the bias is zero. However even at these edges of $\mathcal{A}$ the other two estimators are producing biased estimates. Their behaviour is similar until around $\alpha_2 = 1.7$. Thereafter the bias of LTSr is larger than that of LTS, as shown, for example, in the central panel of Figure 1, so the two curves separate. By the time the extreme value of $\alpha_2 = 2.5$ is reached, all three lines are virtually horizontal, so that the estimates are unbiased.

The plots of partial sums of variances, on the other hand, increase steadily, since the estimators always are subject to the effect of the random variability in the observations. The variance of LTS is slightly greater than that of FS, both being appreciably less than that of LTSr.

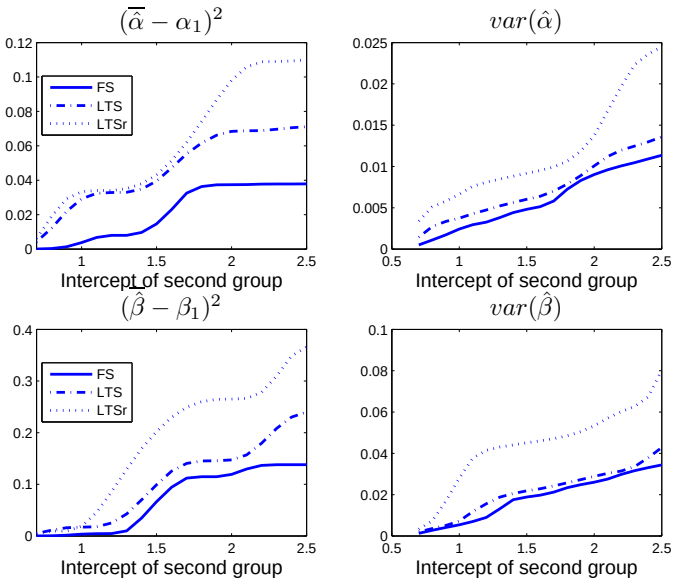

Figure 2: Partial sums over $\mathcal{A}$ of simulated squared bias and variance of the three estimators. Left-hand panels squared bias, right hand panels variance. Top line $\hat{\alpha}$, bottom line $\hat{\beta}$.

The plots for the estimate of β in the lower panels of the plot tell a similar story. The bias of FS is zero at the extremes of $\mathcal{A}$ whereas the cumulative sums of those of LTS and, especially, LTSr, are already increasing. Both LTS and LTSr have appreciable biases for larger values of α_2, whereas that of FS is already zero. The cumulative plot of variances of $\hat{\beta}$ is similar to that for $\hat{\alpha}$; FS is slightly better than LTS and both are much better than LTSr.

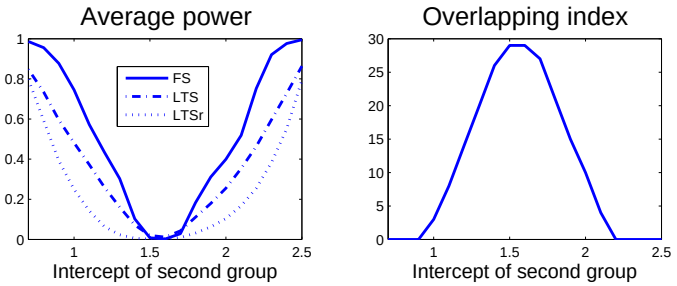

Figure 3: Left-hand panel: simulated average power of the three procedures over $\mathcal{A}$. Right-hand panel: the overlapping index O over $\mathcal{A}$

These plots illustrate the superior performance of the FS estimator. In addition to good parameter estimates we would also like our estimate to signal the presence of outliers if the model fitted to the data is incorrect. For all of the values in $\mathcal{A}$ we calculated the average power, that is the proportion of outliers correctly detected

by the three estimators. The results are in the left-hand panel of Figure 3, with the values of the overlapping index O in the right-hand plot. In the centre of $\mathcal{A}$ the index is at its highest and virtually no outliers are detected. However as overlapping decreases, that is as α_2 increases or decreases, the number of outliers detected by the three methods increases. FS detects the most outliers, with LTS second and LTSr the least powerful procedure for outlier detection.

4. Discussion

Our conclusion, from this and other studies, is that forward search regression combined with a calibrated outlier detection rule provides highly robust and efficient estimates of the regression function. Our comparisons with other very robust methods show that our method has superior performance especially, but not only, when the outliers are close to observations from the main model. The simulation structure that we propose enables these numerical results to be obtained in a systematic, informative and efficient manner

5. References

Atkinson, A. C. and M. Riani (2000). *Robust Diagnostic Regression Analysis*. New York: Springer–Verlag.

Atkinson, A. C., M. Riani, and A. Cerioli (2010). The forward search: theory and data analysis (with discussion). *Journal of the Korean Statistical Society 39*, 117–134. doi:10.1016/j.jkss.2010.02.007.

Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods. *Bulletin of the International Statistical Institute 46*, 375–382.

Maronna, R. A., R. D. Martin, and V. J. Yohai (2006). *Robust Statistics: Theory and Methods*. Chichester: Wiley.

Morgenthaler, S. (2007). A survey of robust statistics. *Statistical Methods and Applications 15*, 271–293. Erratum: 16, 171–172.

Riani, M., A. C. Atkinson, and A. Cerioli (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B 71*, 447–466.

Riani, M., A. C. Atkinson, and D. Perrotta (2013). Very robust regression. (Submitted).

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

Rousseeuw, P. J. and K. Van Driessen (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery 12*, 29–45.

Torti, F., D. Perrotta, A. C. Atkinson, and M. Riani (2012). Benchmark testing of algorithms for very robust regression: FS, LMS and LTS. *Computational Statistics and Data Analysis 56*, 2501–2512. doi:10.1016/j.csda.2012.02.003.