Outlier detection and the distribution of residuals in robust regression

A. Cerioli[1*], S. Salini[2], M. Riani[1], F. Laurini[1] and A. Ghiretti[3]

[1] *Department of Economics, University of Parma, Italy; andrea.cerioli@unipr.it, mriani@unipr.it, fabrizio.laurini@unipr.it.*
[2] *Department of Economics Management and Quantitative Methods, University of Milan, Italy; silvia.salini@unimi.it.*
[3] *Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Italy; a.ghiretti@disia.unifi.it.*
[*] *Presenting author*

Keywords. *Forward Search; LTS; Masking and Swamping; Residuals; S-estimation*

1 Introduction

Recent work in robust statistics has focused on the attempt to reconcile the two enemy brothers of high-breakdown estimation: robustness against a large fraction of masked outliers and good statistical properties, comparable to those of classical estimators, when the normal model holds for all the data. From the point of view of estimation, the goal of this body of work has been the construction of estimators that can achieve both a high breakdown point and high efficiency at the normal distribution (see, e.g., **?**). From a diagnostic perspective, reaching satisfactory statistical properties under the normal model also implies good control of the number of false discoveries in situations of practical interest. There are many application fields, such as high-dimensional genomics, quality control and anti-fraud analysis, where such a property is highly desirable (see, e.g., **??**). However, high-breakdown techniques may produce a potentially large number of spurious outliers. The main target of the present work is to address the diagnostic behaviour of high-breakdown techniques at the normal model from a regression perspective, by considering a wide variety of alternative estimators and different approximations to the null distribution of the resulting robust residuals.

2 Framework and main results

Our basic diagnostic quantities in robust regression are the squared scaled residuals

$$\hat{s}_i^2 = \frac{\hat{\epsilon}_i^2}{\hat{\sigma}^2}, \qquad i = 1, \ldots, n, \tag{1}$$

where $\hat{\epsilon}_i$ is the estimated regression residual for observation i, $\hat{\sigma}^2$ is the model-based estimate of the error variance, and n is the sample size. Precise outlier identification is based on the asymptotic approximation

$$\hat{s}_i^2 \simeq \chi_1^2. \tag{2}$$

Also informal diagnostic methods, such as Q-Q plots of squared scaled residuals, rely on (**??**). However, the reference χ_1^2 distribution holds only in the limit and may provide poor approximations in small or moderate samples when parameters are estimated by high-breakdown techniques. Additional problems may occur due to the effect of alternative tuning choices in the algorithm used to compute the parameter estimates.

One goal of our work is to investigate to what extent the most popular high-breakdown regression methods provide accurate rules for outlier detection using the squared scaled residuals $\hat{s}_i^2$. We compute appropriate corrections when the null performance of the resulting procedure is poor. In particular, we place outlier detection in a testing scenario and we develop robust regression diagnostics that are able to control empirical test sizes at a prescribed level for all the procedures that we analyze. We also evaluate the loss of power that can be expected from our corrections under different contamination schemes and we show that this loss is often not dramatic. See **?** for details. A second goal of our work is to find simple and accurate approximations to the finite sample distribution of (**??**), thus extending the results of **?** to the case of regression. The availability of these approximations will provide more flexible outlier detection rules and more accurate diagnostics tools to be used, e.g., in Q-Q plots of squared scaled residuals.

References

Cerioli A (2010) Multivariate outlier detection with high-breakdown estimators. Journal of the American Statistical Association 105:147–156

Cerioli A, Farcomeni A (2011) Error rates for multivariate outlier detection. Computational Statistics and Data Analysis 55:544–553

Cerioli A, Perrotta D (2014) Robust clustering around regression lines with high density regions. Advances in Data Analysis and Classification 8:5–26

Salini S, Cerioli A, Laurini F, Riani M (2016) Reliable Robust Regression Diagnostics. International Statistical Review, in press, doi:10.1111/insr.12103

Van Aelst S, Willems G, Zamar RH (2013) Robust and efficient estimation of the residual scale in linear regression. Journal of Multivariate Analysis 116:278–296