Discussion of 'Asymptotic theory of outlier detection algorithms for linear time series regression models'

Christophe Croux and Ines Wilms

Faculty of Economics and Business, KU Leuven

We would like to congratulate the authors on their great effort in studying the asymptotic theory of outlier detection algorithms. In the first part, they provide an overview of the asymptotic distribution of estimators based on outlier detection. In the second part, they discuss the gauge, the fraction of wrongly detected outliers, for which they derive an asymptotic theory. The results for the forward search procedure are particularly impressive and of practical relevance.

In this discussion note, we (i) investigate the sampling variation of the empirical gauge, and (ii) study the power of the outlier detection test by means of a modest simulation study. An important contribution of the paper by Johansen and Nielsen (2015) is that the asymptotic results are valid for a wide range of regressors. For this reason, we also consider two time series settings in our simulation study.

Sampling variation of the empirical gauge

The authors prove that the empirical gauge of Huber skip estimators (Huber, 1964) converge in probability to the size of the underlying test. This results does not depend on the type of regressors, hence, regressors may have a deterministic or stochastic trend. We consider the 1-step and 2-step Robustified Least Squares, a special case of the m-step Huber skip estimator where the initial estimator is the full sample Least Squares. Besides, we also take the Least Trimmed Squares (Rousseeuw, 1984) with trimming proportion 50% as initial estimator.

Let the number of regressors k be one or five, and the sample size $n = 100$. Bold symbols refer to vectors, and a bold constant refers to a vector of length k where every component equals the constant. The three simulation designs are

1. **Linear regression model**:
$$y_i = \alpha + \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i,$$
for $i = 1, \dots, n$, where $\alpha = 0.5$, $\boldsymbol{\beta} = \mathbf{1}$, $\mathbf{x}_i$ are i.i.d. $N_k(\mathbf{0}, \mathbf{I})$, and ϵ_i are i.i.d. $N(0, 0.1^2)$.

2. **Stationary time series model**:
$$y_i = \alpha + \phi_y y_{i-1} + \boldsymbol{\phi}_x'\mathbf{x}_{i-1} + \epsilon_i,$$
for $i = 1, \dots, n$, where $\alpha = 0.5$, $\phi_y = 0.4$, $\boldsymbol{\phi}_x = \mathbf{0.2}$, $\mathbf{x}_i$ are i.i.d. $N_k(\mathbf{0}, \mathbf{I})$, and ϵ_i are i.i.d. $N(0, 0.1^2)$.

3. **Trending time series model**:
$$y_i = \alpha + \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i,$$

Table 1: Empirical gauges for the three designs (with each $n = 100$), population gauge γ, and number of regressors k with either the Least Squares or the Least Trimmed Squares as initial estimator.

Design	Steps	Initial estimator	$\gamma = 0.05$		$\gamma = 0.005$	
			$k = 1$	$k = 5$	$k = 1$	$k = 5$
Linear regression	1-step	Least Squares	0.0504	0.0479	0.0049	0.0040
	2-step	Least Squares	0.0518	0.0508	0.0050	0.0041
	1-step	Least Trimmed Squares	0.0625	0.0689	0.0092	0.0112
	2-step	Least Trimmed Squares	0.0617	0.0692	0.0086	0.0105
Stationary time series	1-step	Least Squares	0.0500	0.0478	0.0044	0.0039
	2-step	Least Squares	0.0518	0.0505	0.0044	0.0039
	1-step	Least Trimmed Squares	0.0653	0.0707	0.0096	0.0126
	2-step	Least Trimmed Squares	0.0648	0.0714	0.0090	0.0118
Trending time series	1-step	Least Squares	0.0477	0.0478	0.0044	0.0041
	2-step	Least Squares	0.0492	0.0507	0.0044	0.0042
	1-step	Least Trimmed Squares	0.0609	0.0698	0.0085	0.0123
	2-step	Least Trimmed Squares	0.0600	0.0703	0.0079	0.0116

for $i = 1, \ldots, n$, where $\alpha = 0.5$, $\boldsymbol{\beta} = \mathbf{1}$, regressors $\mathbf{x}_i = \mathbf{0.5} + \boldsymbol{i} + \boldsymbol{\nu}_i$, with $\boldsymbol{\nu}_i$ are i.i.d. $N_k(\mathbf{0}, \mathbf{I})$ and autoregressive error terms $\epsilon_i = 0.5\epsilon_{i-1} + u_i$, with u_i are i.i.d. $N(0, 0.1^2)$.

The empirical gauges for the three simulation designs are in Table 1. The empirical gauges of the Huber skip estimators with Least Squares as initial estimator are accurately sized, also for the two time series models.

For $k = 1$, empirical gauges of the 1-step and 2-step estimators are very similar and close to the population gauges. For $k = 5$, the empirical gauges slightly deviate from the population ones. A small-sample size correction might improve the results, but this requires future research.

Using the Least Trimmed Squares, a robust estimator, as initial estimator, the empirical gauges are more distorted. This holds for both the 1-step and 2-step estimator. The Least Trimmed Squares estimator is less efficient than the Least Squares estimator (in absence of outliers, and at normal errors), which creates small size distortions.

Power of the outlier detection test

We evaluate the power of the outlier detection tests related to the 1-step and 2-step Huber skip estimators. Consider the three simulation designs from the previous section with $k = 1$. We take a contaminated setting where $\epsilon \times 100\%$ of the observations are bad leverage points. We consider two bad leverage point locations[1]: one close to the bulk of the data, one distant from the bulk of the data, and two contamination levels $\epsilon = 0.01, 0.05$. The power of the outlier detection test is assessed by the fraction of actual outliers detected, averaged over $M = 1000$ simulation runs.

[1] For the linear regression design: point concentration at $(x_i, y_i) = (10, 0)$ and at (50,0); for the stationary time series design: point concentration at $(x_i, y_{i-1}, y_i) = (5, 5, 0)$ and at $(100, 100, 0)$; for the trending time series design: point concentration at $(x_i, y_i) = (400, 0)$ and at (500,0).

Table 2: Power of outlier detection test when either the Least Squares or the Least Trimmed Squares is used as initial estimator, for the three designs, different contaminated levels ϵ and two outlier locations.

Design	Steps	Initial estimator	Outliers		Outliers	
			close		distant	
			$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.01$	$\epsilon = 0.05$
Linear regression	1-step	Least Squares	1.00	0.02	0.36	0.00
	2-step	Least Squares	1.00	0.02	0.36	0.00
	1-step	Least Trimmed Squares	1.00	1.00	1.00	1.00
	1-step	Least Trimmed Squares	1.00	1.00	1.00	1.00
Stationary time series	1-step	Least Squares	1.00	0.00	0.00	0.00
	2-step	Least Squares	1.00	0.00	0.00	0.00
	1-step	Least Trimmed Squares	1.00	0.99	1.00	0.99
	2-step	Least Trimmed Squares	1.00	0.99	1.00	0.99
Trending time series	1-step	Least Squares	1.00	0.00	0.00	0.00
	2-step	Least Squares	1.00	0.00	0.00	0.00
	1-step	Least Trimmed Squares	1.00	1.00	1.00	1.00
	2-step	Least Trimmed Squares	1.00	1.00	1.00	1.00

When there is only one outlier ($\epsilon = 0.01$) close to the bulk of the data, all outlier detection algorithms succeed in detecting the actual outlier: the power of all algorithms equals one, as can be seen from Table 2. The power of the outlier detection algorithms based on the Least Squares as initial estimator is, however, affected in all other designs. When the outliers are located far away from the bulk of the data, the power already deteriorates when there is only one outlier ($\epsilon = 0.01$). When the outliers are located close to the bulk of the data, the power goes to zero only for higher amounts of contamination ($\epsilon = 0.05$). The 1-step and 2-step algorithms perform similar, as can be seen from Table 2.

This effect occurs since the Least Squares, a non-robust estimator, is used as initial estimator. As soon as the algorithm includes one non-robust estimator, the power of the corresponding test is affected. Increasing the number of steps in the algorithm will not improve the power. This problem can be circumvented by using a robust estimator as initial estimator, such as the Least Trimmed Squares. The power of the outlier detection algorithms that use the Least Trimmed Squares as initial estimator is much less affected, as can be seen from Table 2.

Conclusion and further comments

We investigate the performance of the empirical gauge, the fraction of wrongly detected outliers. An important novelty of the developed asymptotic theory is that the results are valid for trending time series regressors. Our simulation study on the gauge confirms this. When no contamination is present, using a robust initial estimator results in a small efficiency loss, and hence, small size distortion compared to using a non-robust initial estimator. We study the power of the outlier detection test based on the Robustified Least Squares, and show that it breaks down when bad leverage points are present. This occurs since the outlier detection algorithm uses a non-robust initial estimator. The non-robust Least Squares estimator is heavily influenced by the bad leverage points. This leads to large absolute residual values for the clean observations and small absolute residual values for the outliers. As a result, the actual outliers are not detected as outlying. This effect is known as the "masking effect" (Rousseeuw and Leroy, 1987). This masking effect can be circumvented by using an initial estimator that is robust, such as the Least Trimmed Squares.

The paper assumes that the density of the errors, f, is known up to a scale parameter. In particular, the gauge will only be estimated consistently if f is correctly specified. For the consistency of the associated estimators it is sufficient that f is symmetric. It would be interesting to know whether symmetry is really necessary, since for M-estimators, for instance, the slope parameter can still be estimated consistently for asymmetric errors.

It would also be of interest to see how the results change when the cut-off value c is allowed to depend on the sample size. To address multiple testing issues, it makes sense to let c increase with the sample size. Finally, the results on the Poisson approximation of the gauge are very elegant, and we wonder whether they are of use in cases where the number of outliers is very small, but still increase with the sample size.

References

Huber, P. J. (1964), "Robust estimation of a location parameter," *Ann. Math. Statist.*, 35, 73–101.

Johansen, S. and Nielsen, B. (2015), "Asymptotic theory of outlier detection algorithms for linear time series regression models," *Scandinavian Journal of Statistics.*

Rousseeuw, P. J. (1984), "Least median of squares regression," *J. Amer. Statist. Assoc*, 79, 871–880.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust regression and outlier detection*, New York: Wiley-Interscience.

Christophe Croux, Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven
E-mail: christophe.croux@kuleuven.be