Robust Estimation Using Least Trimmed Squares

Jurgen A. Doornik*

Institute for Economic Modelling, Oxford Martin School, and
Economics Department, University of Oxford, UK

March 14, 2011

Abstract

A robust procedure is proposed, starting from least trimmed squares as the initial estimator. The asymptotic distribution of the two-step and multi-step estimators is derived. This allows the use with a pre-specified efficiency under normality. It is argued that the good performance, together with the simplicity of the procedure, should make this the robust estimator of choice for applied work.

1 Introduction

It is well known that ordinary least squares estimates of a regression model are very sensitive to even a small amount of contamination of the data. Estimation based on least absolute deviations is only marginally less sensitive. Looking for large outliers relative to the estimated model is not robust, because the OLS estimates of the parameters will already be affected by the outliers.

An extensive research programme, starting with the work of J.W. Tukey, P.J. Huber, F.R. Hampel and others, has led to a large literature on robust estimation. Despite this, the impact on empirical modelling has not been keeping up. To some extent, this can be attributed to the presence of a large number of estimators to choose from, computational issues, and lack of guidance for practitioners. The following points of the framework of Višek (2001) remain important: availability in software, accompanied by the usual range of diagnostics, specification tests, etc., and efficient and acceptable heuristics. Ability to replicate results also matters, requiring precise specification of the initial and final estimators that are used.

When the regressors are fixed, so only allowing for outliers in the dependent variable (the response), it is sufficient to use M-estimation. This is discussed, e.g., in Maronna, Martin and Yohai (2006, Ch.4). Robust estimation is much harder when this assumption is relaxed. Maronna et al.

*I owe special thanks to Søren Johansen and Bent Nielsen for providing detailed comments, as well as many helpful discussions.
(2006, Ch.5) lists three procedures as having desirable robust properties in this case (MM, $\tau$, and reweighted LTS). All algorithms share the property that they are computationally hard, requiring random search for any reasonable sample size. As a rule of thumb, estimation will be between one thousand and ten thousand times slower than OLS (see §2.1).

One of the choices an empirical modeller must make is whether to use a soft weighting scheme, i.e. using weights between zero and one (inclusive), or a hard-rejection scheme. In the latter, outlying observations are identified and removed completely (given a zero weight). This is often preferable from an empirical perspective: much can be learned from identifying the outliers. Indeed, most illustrations in textbooks on robust statistics focus on this aspect. Removing outliers amounts to selecting observations. In a linear regression model, the outliers can also be removed by adding impulse indicator variables that have the value one for that observation and zero otherwise. This emphasizes the relation between outlier identification and variable selection, and so the importance of an inferential approach.

The objective in this paper is robust estimation of the regression model with $n$ observations and $k$ regressors $x'_i = (x_{i1}, ..., x_{ik})$:

$$y_i = \beta'x_i + \epsilon_i, \ i = 1, ..., n.$$ 

The null hypothesis is that $\epsilon_i \sim \text{IID}[0, \sigma^2]$, which for simplicity tends to be replaced by $\epsilon_i \sim \text{IIN}[0, \sigma^2]$. The model is assumed to include an intercept, $x_{i1} = 1$. For an estimate $\hat{\beta}$, the residuals are $r_i = r_i(\hat{\beta}) = y_i - \hat{\beta}'x_i$. The notation for the ordered absolute residuals is $r_{(i)}$, so $|r_{(1)}| \leq |r_{(2)}| \leq ... \leq |r_{(n)}|$.

## 2 Least Trimmed Squares (LTS)

Peter Rousseeuw introduced several robust regression estimators, including least median of squares (LMS) and least trimmed squares (LTS), see Rousseeuw (1984) as well as the monograph Rousseeuw and Leroy (1987). LTS converges at rate $n^{1/2}$ with the same asymptotic efficiency under normality as Huber’s skip estimator. The LMS convergence rate is $n^{1/3}$ and its objective function is less smooth than LTS. As a consequence, as argued in Rousseeuw and van Driessen (2006), LTS is now preferred over LMS. Only LTS will be considered in the remainder.

Least trimmed squares is defined as:

$$\hat{\beta}_{\text{LTS},H,N} = \arg\min_{\beta} \sum_{i=1}^{H} r^2_{(i)}(\beta),$$

(1)

based on the ordered absolute residuals $|r_{(1)}| \leq |r_{(2)}| \leq ... \leq |r_{(n)}|$. $N$ denotes the index set of all observations: $N = \{1, 2, ..., n\}$. When $H = \lceil n/2 \rceil$, LTS locates that half of the observations which has the smallest estimated variance. In that case, the breakdown point is 50%. When $H$ is set to the sample size, LTS and OLS coincide: $\hat{\beta}_{\text{LTS},n,N} = \hat{\beta}_{\text{OLS},N}$. Writing $\mathcal{H}(H)$ for the $H$ observations that are selected by LTS: $\hat{\beta}_{\text{LTS},H,N} = \hat{\beta}_{\text{OLS},\mathcal{H}(H)}$. Note that increasing $H$ from $H_1$ to $H_2$ does not amount to adding observations to $\mathcal{H}(H_1)$: $H_2 > H_1$ does not imply that $\mathcal{H}(H_1) \subset \mathcal{H}(H_2)$.
Rousseeuw and Leroy (1987, §4.4) shows \( n^{1/2} \) consistency and asymptotic normality of LTS in the location–scale model. Víšek (1999) extends this to the regression model with random regressors; the proof for fixed regressors is in a later series of papers: Víšek (2006a, 2006b). Writing \( f \) for the density of \( \sigma^{-1} e_i \), \( F \) for its cdf, and assuming symmetry and true coefficients \( \beta \):

\[
n^{\frac{1}{2}} \left( \hat{\beta}_{LTS,H,N} - \beta \right) \xrightarrow{D} N_k \left[ 0, \sigma^2 \Sigma^{-1} \frac{\tau^2_\alpha}{\left( 1 - \alpha - 2c_\alpha f(c_\alpha) \right)^2} \right], \tag{2}
\]

where \( \alpha = (n - H)/n \), \( c_\alpha = F^{-1}(1 - \alpha/2) \), \( \Sigma = E[x_i x_i'] \) with \( x_i \) assumed to be IID with absolutely continuous distribution function (or \( \Sigma = \sum_{i=1}^{n} x_i x_i' \) for fixed regressors). Finally:

\[
\tau^2_\alpha = \int_{-c_\alpha}^{c_\alpha} z^2 f(z) \, dz,
\]

and we assume also that \( 1 - \alpha - 2c_\alpha f(c_\alpha) > 0 \). For the estimated variance derived from LTS:

\[
\hat{\sigma}^2_{LTS,H,N} = \frac{1}{H} \sum_{i=1}^{H} \hat{\epsilon}^2_{(i)}(\hat{\beta}_{LTS,H,N}) \xrightarrow{p} \sigma^2 \frac{\tau^2_\alpha}{1 - \alpha}. \tag{3}
\]

With normal disturbances, \( f \) is the standard normal density and \( \tau^2_\alpha = 1 - \alpha - 2c_\alpha f(c_\alpha) \), in which case the asymptotic efficiency is \( 1 - \alpha - 2c_\alpha \phi(c_\alpha) \) and \( c_\alpha = \Phi^{-1}(1 - \alpha/2) \):

\[
n^{\frac{1}{2}} \left( \hat{\beta}_{LTS,H,N} - \beta \right) \xrightarrow{D} N_k \left[ 0, \sigma^2 \Sigma^{-1} \frac{1}{1 - \alpha - 2c_\alpha \phi(c_\alpha)} \right]. \tag{4}
\]

This low efficiency, e.g. 7% when \( H = n/2 \), is considered by many to be the main deficiency of LTS. However, LTS is often suggested as the starting point for more efficient procedures.

### 2.1 Computation of LTS

When \( n \) is very small, it is possible to generate all subsets of size \( H \) to determine which one minimizes the LTS criterion. For \( n = 40 \) and \( H = n/2 \) there are already more than \( 10^{11} \) subsets, showing the limitations of this approach. Rousseeuw and Leroy (1987, §5.1) base computation of LTS and LMS on estimation of subsets of size \( k \). When \( n \) is small enough:

1. Select \( H \).
2. Generate all possible subsets with \( k \) observations, and compute the regression coefficients (an exact fit), say \( \hat{\beta}(1), \ldots, \hat{\beta}(q) \).
3. Compute the residuals using all \( n \) observations, and from this the LTS (or LMS) criterion.
4. The LTS estimate corresponds to the \( \hat{\beta}(j) \) that minimizes the objective function (1).

The LTS criterion for a given set of residuals involves ordering the residuals, which takes \( O(n \log_2 n) \) operations. The number of subsets is \( q = \binom{n}{k} \), which is usually still too large for realistic applications.
Table 1: Time (in seconds) of LTS and OLS estimation, and ratio of LTS timings to OLS (rounded to nearest 100). \( H = \lfloor (n + k + 1)/2 \rfloor \). Outlier\% is the percentage of the sample with contaminated regressors; Robust is whether LTS found the uncontaminated subsample.

<table>
<thead>
<tr>
<th>n</th>
<th>k</th>
<th>Outlier%</th>
<th>Robust</th>
<th>Time</th>
<th>Ratio to OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2</td>
<td>40%</td>
<td>yes</td>
<td>0.029</td>
<td>7300</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>40%</td>
<td>yes</td>
<td>0.044</td>
<td>4800</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>40%</td>
<td>yes</td>
<td>0.063</td>
<td>2800</td>
</tr>
<tr>
<td>1000</td>
<td>2</td>
<td>35%</td>
<td>yes</td>
<td>0.088</td>
<td>3800</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>35%</td>
<td>yes</td>
<td>0.48</td>
<td>2500</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>35%</td>
<td>yes</td>
<td>95</td>
<td>7200</td>
</tr>
<tr>
<td>10000</td>
<td>5</td>
<td>40%</td>
<td>yes</td>
<td>1.7</td>
<td>2300</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>40%</td>
<td>yes</td>
<td>2.6</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>40%</td>
<td>yes</td>
<td>580</td>
<td>2300</td>
</tr>
<tr>
<td>100000</td>
<td>5</td>
<td>40%</td>
<td>yes</td>
<td>20</td>
<td>1100</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>40%</td>
<td>yes</td>
<td>33</td>
<td>600</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>40%</td>
<td>yes</td>
<td>6800</td>
<td>2000</td>
</tr>
</tbody>
</table>

The proposed solution to this is to generate a fixed number of random subsets of size \( k \) instead. To allow replication of the results, it would be useful to always use the same seed for the random subsets.\(^1\)

Rousseeuw and van Driessen (1999) propose a fast algorithm for computing LTS, which was finally published as Rousseeuw and van Driessen (2006). The trick, which is also used in other robust algorithms,\(^2\) is to iterate a few steps on a large number of starting values, and keep the 10 (say) most promising ones. These are then used for full iteration, yielding the final estimate. The resulting algorithm makes LTS estimation faster than LMS, providing an additional argument to prefer it.

To check our implementation of Fast-LTS, we repeat the experiment of Rousseeuw and van Driessen (2006). The data generation process is:

\[
\begin{align*}
\epsilon_i & \sim \text{Uniform}[0,1], \quad i = 1, \ldots, n, \\
x_{ij} & \sim \text{Uniform}[0,100], \quad i = 1, \ldots, n, \quad j = 2, \ldots, k.
\end{align*}
\]

The data for estimation consists of \( y_i \) and \( x^*_{ij} \), where \( x^*_{ij} \) coincides with \( x_{ij} \), except for a percentage of the sample where it is replaced by \( 100 + x_{ij} \). This contamination of the regressors is randomly spread over the sample.

\(^1\)An alternative would be to save the seed. However, modern random number generators have a large number of seeds (the default Ox generator has 1024 seeds).

\(^2\)All truly robust methods seem to require a combinatorial search over observations.
Table 1 shows that LTS is at least one thousand times slower than OLS.\footnote{Timings are on a Intel Core 2 Quad Q6600 at stock speed of 2.66Ghz, using only one core. The code is written in Ox 6, see Doornik (2007). Base on a rough comparison of SPEC ratings, our computer is about 150 to 200 times faster than that used in Rousseeuw and van Driessen (2006, Table 2). As a consequence, our Ox implementation of Fast-LTS is about 1.5 to 5 times faster than their Fortran version for $n \leq 1000$. For larger samples we are more than three times slower, probably because we use a more comprehensive blocking approach in the algorithm.} Also note that the robust solution is always found. For sample sizes of 1000 and higher, going from 10 to 100 regressors slows LTS 200 times (from just under a half to almost a hundred seconds).

Some alternative approaches are discussed in Agulló (2001) and Nguyen and Welsch (2010).

3 Improving the efficiency of LTS

Various proposals have been made to avoid the very low efficiency of the LTS estimator. Rousseeuw (1984) introduces reweighted least squares (RLS), where the weights are based on the initial LMS estimation. Rousseeuw and Leroy (1987) develops this further, while Rousseeuw and Hubert (1997) describes the LTS equivalent:

A1. Select $H$ and run LTS to obtain $\hat{\beta}_{\text{LTS}}$ (short for $\hat{\beta}_{\text{LTS},H,N}$). From (3) compute:

$$\hat{\sigma}_{0}^{2} = \frac{1 - \alpha}{1 - \alpha - 2c_{\alpha}\phi(c_{\alpha})} \hat{\sigma}_{\text{LTS},H,N}^{2},$$

(5)

Where $\alpha$ and $c_{\alpha}$ are defined below (2).

Construct t-ratios for all $n$ residuals:

$$u_{0,i} = \left( y_{i} - x_{i}^{\prime}\hat{\beta}_{\text{LTS}} \right) / \hat{\sigma}_{0}.$$

A2. Use the following weights (the suggested default is $c_{2} = 2.5$):

$$w_{i} = I \left[ |u_{0,i}| \leq c_{2} \right] \quad i = 1, \ldots, n$$

for an updated scale estimate:

$$\hat{\sigma}_{1}^{2} = \frac{1}{\sum_{i=1}^{n} w_{i} - k} \sum_{i=1}^{n} w_{i} \left( y_{i} - x_{i}^{\prime}\hat{\beta}_{\text{LTS}} \right)^{2}.$$

and new t-ratios for all observations:

$$u_{1,i} = \left( y_{i} - x_{i}^{\prime}\hat{\beta}_{\text{LTS}} \right) / \hat{\sigma}_{1}.$$

A3. Select observations according to $|u_{1,i}| \leq c_{2}$ and apply OLS to this ‘clean’ sample.$^{4}$

\footnote{OLS estimates of a subset $N^{0}$ of dimension $n^{0}$ refers to the coefficients estimated by OLS from that subset and corresponding scale:

$$\hat{\sigma}_{\text{OLS}}^{2} = \frac{1}{n^{0} - k} \sum_{i \in N^{0}} \left( y_{i} - x_{i}^{\prime}\hat{\beta}_{\text{OLS}} \right)^{2}.$$}
We shall call this procedure LTS-RLS. One of the few published applications in economics of this procedure is in Zaman, Rousseeuw and Orham (2001). It is pointed out there that only high-leverage outliers should be removed, an aspect which is ignored here.

Maronna et al. (2006, p.132) describe this procedure somewhat differently:

B1. Select $H$ and run LTS to obtain $\hat{\beta}_{\text{LTS}}$ and residuals

$$r_{i,\text{LTS}} = y_i - x_i'\hat{\beta}_{\text{LTS}}, \quad i = 1, \ldots, n.$$ 

Compute a robust scale estimate such as:

$$\tilde{\sigma}_0 = \frac{1}{\Phi^{-1}(0.75)} \text{median}(|r_{1,\text{LTS}}|, \ldots, |r_{n,\text{LTS}}| \text{ for } r_{i,\text{LTS}} \neq 0).$$

Construct t-ratios for all $n$ residuals:

$$u_{0,i} = r_{i,\text{LTS}}/\tilde{\sigma}_0.$$

B2. Select observations according to $|u_{0,i}| \leq c_2$ and apply OLS to this ‘clean’ sample.

According to Maronna et al. (2006, p.133) these procedures will give more efficient estimates, ’but its asymptotic distribution is complicated’. They note that the asymptotic efficiency is zero because $c_2$ is fixed. However, this is not so relevant in practice, because $c_2$ can be taken as an increasing function of the sample size (e.g. using $\alpha = n^{-1/2}$).

Gervini and Yohai (2002) proposes to choose the weights in a data-dependent manner, based on the empirical distribution function. The resulting method, called REWLS, achieves full asymptotic efficiency under normality (also see Maronna et al., 2006, p.133–134):

C1. As B1. Write the order statistics of the absolute t-ratios as $|u_0|(i)$. Remember that, when $U \sim N[0, 1]$ and $A = |U|$, then the cdf of $A$ is $F_A(a) = 2\Phi(a) - 1$.

C2. Choose $\alpha$ (at 1%, e.g.) and corresponding critical value $c_2 = \Phi^{-1}(1 - \alpha/2)$. Compute the difference between theoretical and empirical cdf:

$$d_i = \left[2\Phi\left(|u_0|(i)\right) - 1\right] - \frac{i - 1}{n}.$$

The number of observations that are classified as outliers is:

$$n^* = \left[n \left(\max_{\{i:|u_0|(i) > c_2, d_i \geq 0\}} d_i\right)\right].$$

In other words, we pick the critical value $c_3$ as any value in the interval (setting $|u_0|(n+1) = \infty$):

$$(|u_0|(j-1), |u_0|(j)) \quad \text{where } j = n - n^* + 1.$$

Select observations according to $|u_{0,i}| \leq c_3$ and apply OLS to this ‘clean’ sample.

The main motivation of this procedure is that it improves the efficiency when LMS is the initial estimator (or another estimator with lower than $\sqrt{n}$ convergence). This is also achieved through estimation based on weighted order statistics, see Čížek (2007). Here we use LTS as the initial estimator.
3.1 A modified two-step procedure

A small modification of the reweighting procedure of Rousseeuw (1984) involves re-estimation by OLS, using a bias correction at each stage:

D1. As A1, except that we use \( \frac{1 - \alpha_1}{\tau_2^2} \frac{H}{H - k} \hat{\sigma}_{\text{LTS}, H, N}^2 \) in (3):

\[
\hat{s}_0^2 = \frac{1 - \alpha_1}{\tau_2^2} \frac{H}{H - k} \hat{\sigma}_{\text{LTS}, H, N}^2,
\]

and corresponding t-ratios for testing:

\[ u_{0,i} = \left( y_i - x_i' \hat{\beta}_{\text{LTS}} \right) / \hat{s}_0. \]

D2. Choose \( \alpha_2 \) with corresponding critical value \( c_2 = \Phi^{-1}(1 - \alpha_2/2) \). Select observations according to \( I[|u_{0,i}| \leq c_2] \), and use OLS on this subsample to compute \( \hat{\beta}_1, \hat{\sigma}_1^2 \). Adjust \( \hat{\sigma}_1^2 \) with the bias correction based on \( \tau_2^2 = 1 - \alpha_2 - 2c_2\phi(c_2) \):

\[
\hat{s}_1^2 = \frac{1 - \alpha_2}{\tau_2^2} \hat{\sigma}_1^2.
\]

D3. Construct t-ratios for testing:

\[ \hat{u}_{1,i} = (y_i - x_i' \hat{\beta}_1) / \hat{s}_1, \]

Select observations according to \( I[|\hat{u}_{1,i}| \leq c_2] \), and use OLS on this subsample to compute \( \hat{\beta}_2, \hat{\sigma}_2^2 \). Adjust \( \hat{\sigma}_2^2 \) with the same bias correction as in the previous step. So \( \hat{\beta}_2 \) and \( \hat{\sigma}_2^2 \) are the final estimates.

This two-step estimator is called ELTS(2). Stopping at the end of step D2 gives the one-step estimator ELTS(1). The results in Johansen and Nielsen (2009) may be used to derive the asymptotic distribution of these procedures, see the Appendix. The efficiency gain arises again from switching from a large \( \alpha_1 \) to a small \( \alpha_2 \) (e.g. from 0.5 to 0.975).

Figure 1 shows the efficiency of the one and two-step continuations of the suggested procedure from an initial breakdown point of \( \alpha_1 = 0.2, 0.35, 0.5 \). The lines are bounded by the efficiency of \( \hat{\beta}_{\text{LTS}} \) based on \( \alpha_2 \). The right panel of Figure 1 zooms in on the interesting part of the left panel. For example, a breakdown of \( \alpha_1 = 50\% \) followed by the one-step estimator at \( \alpha_2 = 2\% \) has an efficiency of just over 70\%. The two-step estimator has an efficiency of over 85\% in this case.

3.2 A fully iterated version

Step D3 can be iterated until the estimates do not change anymore. This procedure may have the following distribution under normality:

\[
n^{-\frac{1}{2}} \left( \hat{\beta}_s - \beta \right) \overset{D}{\rightarrow} N_k \left[ 0, \sigma^2 \Sigma^{-1} \frac{1}{1 - \alpha_2 - 2c_2\phi(c_2)} \right], \tag{7}
\]
Figure 1: Efficiency of $\hat{\beta}_2$ from ELTS(2) and $\hat{\beta}_1$ from ELTS(1) for three choices of $\alpha_1$; $\alpha_2$ is on the horizontal axis. The right panel zooms in on the top left corner of the first panel.

writing $c_i = \Phi^{-1}(1 - \alpha_i/2)$. This is the conjectured extension of the $m$ step procedure discussed in the Appendix. Now the breakdown point is that corresponding to $\alpha_1$, but the efficiency is based on $\alpha_2$, irrespective of the value of $\alpha_1$ that was chosen. The fully iterated procedure is labelled ELTS.

Table 2 gives some more detailed information on the corresponding critical values that are required to achieve a certain asymptotic efficiency. ELTS(2) already shows very little dependence on $\alpha_2$, and is very close to ELTS.

4 Some simulation results

4.1 Efficiency

First we investigate the efficiency in smaller samples, expressed in relation to the OLS estimates. The experimental design is the same as Čížek (2007, §6.1):

$$y_i = 0.5 + x_{i1} - 2x_{i2} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1), \quad i = 1, \ldots, n,$$
$$x_{i1}, x_{i2} \sim \mathcal{N}(0, 1), \quad i = 1, \ldots, n.$$

The regressors are drawn anew in every replication, and the number of replications is denoted $M$. The relative efficiencies are expressed as:

$$\frac{1}{M} \sum_{m=1}^{M} \left\| \hat{\beta}_{\text{OLS},N}(m) - \beta \right\|_2^2,$$
$$\frac{1}{M} \sum_{m=1}^{M} \left\| \hat{\beta}_{T,N}(m) - \beta \right\|_2^2. \quad (8)$$

So we take the sum of squared deviations from the true coefficients, and average that over all replications. $T$ represents the estimators considered, and the results are normalized with respect to OLS. The
Table 2: Critical values corresponding to efficiencies ranging from 0.6 to 0.99. Breakdown set to 50% and 30%.

<table>
<thead>
<tr>
<th></th>
<th>Asymptotic efficiency under normality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td>100(c_1)</td>
<td></td>
</tr>
<tr>
<td>ELTS(1)</td>
<td>50%</td>
</tr>
<tr>
<td>ELTS(2)</td>
<td>50%</td>
</tr>
<tr>
<td>ELTS</td>
<td>50%</td>
</tr>
<tr>
<td>ELTS(1)</td>
<td>30%</td>
</tr>
<tr>
<td>ELTS(2)</td>
<td>30%</td>
</tr>
<tr>
<td>ELTS</td>
<td>30%</td>
</tr>
</tbody>
</table>

The number of observations selected by LTS is set as \(H = \lceil n/2 \rceil\) and \(H = \lceil 0.7n \rceil\). The corresponding breakdown is \(\alpha_1 = (n - H)/n\), so 50% in the former, and 30% in the latter. For LTS, the asymptotic efficiencies are 7% and 22% at these settings. For ELTS(1), ELTS(2) and ELTS, the critical values are chosen to set the asymptotic efficiency to 90%, see Table 2. LTS-RLS (i.e. steps A1–A3) is run at the critical value of ELTS(1), which it resembles (2.69 and 2.58 respectively). REWLS (procedure C1–C2) has initial critical value \(c_2 = 2.5\). Finally, for comparison we also include MM estimation.

The results are in Table 3. It is not suprising that the 90% efficiency of ELTS in these experiments does not hold at smaller samples (less than 200 in this experiment). However, where we can make a comparison (ELTS(*) and MM), ELTS has the better efficiency in smaller samples.

4.2 Estimation with contaminated data

Let \(\mathcal{N}\) denote the full sample as before, \(\mathcal{N}^0\) the uncontaminated subset of this and \(\mathcal{N}^* = \mathcal{N} - \mathcal{N}^0\) the contaminated data. Table 4 presents two simple data generation processes with contamination. In the first, labelled ‘\(x\)-contamination’, the regressor values (excluding the intercept) are shifted by \(\gamma\) for 30% of the sample. In the second, labelled ‘\(y\)-contamination’, the intercept is shifted by \(\gamma\) (also for 30% of the data). The contamination is randomly spread across the sample, which is why the set notation is used. In time-series modelling, on the other hand, it may be of interest to have the contaminated data in a cluster, because such ‘structural breaks’ frequently happen in practice.

The root mean-squared errors (RMSE) of the estimated coefficients are presented in graphical...
Table 3: Simulated efficiencies under normality relative to OLS, see (8). Breakdown set to $\alpha_1 = 50\%$ and $\alpha_1 = 30\%$; $M = 10\,000$.

<table>
<thead>
<tr>
<th></th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>LTS</td>
<td>0.19</td>
<td>0.16</td>
<td>0.14</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td>LTS-RLS</td>
<td>0.36</td>
<td>0.63</td>
<td>0.81</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>ELTS(1)</td>
<td>0.36</td>
<td>0.55</td>
<td>0.71</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td>ELTS(2)</td>
<td>0.39</td>
<td>0.68</td>
<td>0.82</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td>ELTS</td>
<td>0.43</td>
<td>0.78</td>
<td>0.87</td>
<td>0.90</td>
<td>0.89</td>
</tr>
<tr>
<td>REWLS</td>
<td>0.52</td>
<td>0.76</td>
<td>0.86</td>
<td>0.89</td>
<td>0.93</td>
</tr>
<tr>
<td>MM</td>
<td>0.46</td>
<td>0.79</td>
<td>0.85</td>
<td>0.90</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Table 4: Two DGPs with contaminated data

Data generation process, $i = 1, \ldots, n$

$$y_i = 1 + \sum_{j=2}^k x_{ij} + \epsilon_i, \quad \epsilon_i \sim \text{IIN}[0, 1],$$
$$x_{ij} \sim \text{IIN}(0, 1), \quad j = 2, \ldots, k.$$  
$$x_{ij}^* \sim x_{ij} + \gamma, \quad j = 2, \ldots, k.$$  
$$y_i^* \sim y_i + \gamma.$$  

Data for estimation, $i = 1, \ldots, n$

$x$-contamination: $y_i, x_{ij}$ for $i \in \mathcal{N}^0$; $y_i^*, x_{ij}^*$ for $i \in \mathcal{N}^*$,
$y$-contamination: $y_i, x_{ij}$ for $i \in \mathcal{N}^0$; $y_i^*, x_{ij}^*$ for $i \in \mathcal{N}^*$.

Table 4: Two DGPs with contaminated data

form. We separately consider the intercept, the slope coefficients and the scale:

\[
\begin{align*}
\text{RMSE intercept} & \quad M^{-1} \sum_{m=1}^M \left| \widehat{\beta}_{1,T,N}(m) - 1 \right|_2 \\
\text{RMSE slope} & \quad M^{-1} (k-1)^{-1/2} \sum_{m=1}^M \left| \widehat{\beta}_{2,k,T,N}(m) - 1 \right|_2 \\
\text{RMSE scale} & \quad M^{-1} \sum_{m=1}^M \left| \widehat{\sigma}_{T,N}(m) - 1 \right|_2
\end{align*}
\]

Figure 2 shows the root mean-squared errors for the intercept in the left panel, and for the slope in the middle panel using the data with contaminated regressors. The robust estimators are all quite similar: the efficiency is close to that of OLS ($\gamma = 0$). Then, from $\gamma = 3$ onwards, the robust methods discover the contamination. The exceptions are that MM takes longer, until $\gamma = 4$, and has considerably higher RMSE for the scale.

Figure 3 shows the results for the contaminated mean. Now the bias is concentrated in the intercept, with the largest RMSE shown by ELTS for $\gamma = 3, 4$ followed by MM. For large magnitudes REWLS is worse than the others (but better with small $\gamma$). Again, the scale estimate that comes out
of MM does not work so well. Neither does the REWLS scale. For large $\gamma$, LTS-RLS and the ELTS estimators are almost identical.

ELTS(2) emerges as the preferred robust estimator from these experiments. Moreover, for the relevant efficiencies, we may as well use the asymptotic distribution of ELTS, as shown in §3.2.

5 A comparison of ELTS and IIS

Johansen and Nielsen (2009) study the so called impulse-indicator saturation estimator (IIS). As the name suggests, this involves adding all possible impulse dummies to the model. Estimation is via a split-sample approach: start by adding the second half of the dummies (i.e. use the first $n/2$ observations), then use the remainder. The half-way split is not necessary, but adopted here. Johansen and Nielsen (2009) derive the asymptotic distributions allowing for stationary and trend-stationary autoregressive processes, so have a much more general setup than IID observations. They show the link between IIS and robust estimators, and provide a powerful framework for theoretical analysis that allows us to derive the asymptotic distribution of ELTS.

The IIS estimator can also be iterated (see §A.2). Figure 4 compares the efficiency of one and two-step ELTS (using $\alpha_1 = 0.5$) to the 1, 2 and 10 step IIS estimator. Both approach the LTS line, but from different sides.

Finally, IIS is not robust in the way that ELTS (or MM is): for the experiments with contaminated
For the data, the RMSE of the two-step IIS estimator (at \( \alpha = 0.01 \), say) is no better than OLS.\(^6\) For the contamination in mean this can be improved by starting from a much higher initial \( \alpha_1 \) (and again the impact of this on efficiency will start to disappear as the number of iterations increases). IIS will also work better when the contamination is clustered (more likely in time-series data). In general, it can be rectified by using LTS as the initial high-breakdown estimator, resulting in ELTS.

\(^6\)This is analogous to Rousseeuw and van Driessen (2006, Fig.2): the probability that each half (or subsample of that size) contains an outlier is very close to one. And one outlier is enough to cause OLS to break down.
6 Conclusions

We have proposed efficient LTS, a modified robust estimator which uses LTS as the initial estimator. The user is required to select the breakdown of the initial estimator, and a critical value (or p-value $\alpha_2$) for subsequent selection. Results from Víšek (1999) and Johansen and Nielsen (2009) allowed us to derive the asymptotic distribution.

The one step version of ELTS is almost the same as the RLS proposal of Rousseeuw and Hubert (1997) (called LTS-RLS here). We prefer the two-step version, ELTS(2), because the effect of the initial breakdown has largely disappeared: for relevant efficiencies, the distribution is essentially indistinguishable from that of LTS based on $\alpha_2$. ELTS(2) also performed well in the contaminated experiments, outperforming MM on root mean-squared errors of the estimated coefficients.

Finally, ELTS(2) fits well in the program of Víšek (2001). The heuristics, in particular, are simple: first find those $H$ observations that have the smallest residual variance. Then use OLS to select observations using simple statistical criteria. The final model is estimated by OLS, and can be presented as full sample estimates with impulse indicator variables added to the model. This may be preferred by practitioners over a soft weighting scheme which is much less transparent. Computation is acceptably slow, provided that the Fast-LTS of Rousseeuw and van Driessen (2006) is used. The fact that only two choices need to be made will facilitate replication of results.

A Appendix

A.1 Asymptotic distribution of ELTS(2)

The derivations here build upon the results in Johansen and Nielsen (2009), referred to as JN below. Theorem JN-1.1 shows under what conditions $I[|\{y_i - x_i \hat{\beta}\}/\hat{\omega}_i| < c]$ can be approximated by $I[|\varepsilon_i| < c\sigma]$. These conditions are satisfied by the initial LTS estimator and the assumptions around (2). Further generalization is explored by JN, and will be applicable provided that the asymptotic expansion of LTS remains valid.

Assume symmetry, so that $\xi_1^c = 2cf(c)$, and follow the proof of Theorem JN-1.7. Using equation (JN-1.20) with $N_n = n^{-1/2}I_n$ and $\xi_2^c = 0$:

$$ (1 - \alpha_2) \Sigma n^{\frac{1}{2}} (\hat{\beta}_1 - \beta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \varepsilon_i u_i + \xi_1^{\varepsilon_2} \Sigma n^{\frac{1}{2}} (\hat{\beta}_{\text{LTS}} - \beta) + o_p(1), \quad (10) $$

where $u_i = I[|\varepsilon_i| < c_2\sigma]$. Insert the asymptotic expansion of $\hat{\beta}_{\text{LTS}}$ as given by Víšek (2006b, Theorem 1):

$$ \Sigma n^{\frac{1}{2}} (\hat{\beta}_{\text{LTS}} - \beta) = n^{-\frac{1}{2}} \frac{1}{1 - \alpha_1 - \xi_1^{\varepsilon_1}} \sum_{i=1}^{n} x_i \varepsilon_i w_i + o_p(1), \quad (11) $$

where $w_i = I[|\varepsilon_i| < c_1\sigma]$, into (10):

$$ (1 - \alpha_2) \Sigma n^{\frac{1}{2}} (\hat{\beta}_1 - \beta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \varepsilon_i \left\{ u_i + w_i \frac{\xi_1^{\varepsilon_2}}{1 - \alpha_1 - \xi_1^{\varepsilon_1}} \right\} + o_p(1). $$
When $\alpha_1 > \alpha_2$, then $c_1 < c_2$ and $u_i w_i = w_i$, and the sum of the conditional variances converges in probability to:

$$\sigma^2 \Sigma \left[ \tau_2^{c_2} + 2 \tau_2^{c_1} \frac{\xi_2^{c_2}}{1 - \alpha_1 - \xi_1^{c_1}} + \tau_2^{c_1} \left( \frac{\xi_1^{c_2}}{1 - \alpha_1 - \xi_1^{c_1}} \right)^2 \right].$$

Writing

$$q_1 = \frac{\xi_2^{c_2}}{1 - \alpha_1 - \xi_1^{c_1}}$$

we find:

$$n^\frac{1}{2} \left( \hat{\beta}_1 - \beta \right) \xrightarrow{D} N_k \left[ 0, \sigma^2 \Sigma^{-1} \left( \frac{\tau_2^{c_2} + (2q_1 + q_1^2)\tau_2^{c_1}}{(1 - \alpha_2)^2} \right) \right].$$

Iterate one additional step:

$$(1 - \alpha_2) \Sigma n^\frac{1}{2} \left( \hat{\beta}_2 - \beta \right)
= n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \xi_i u_i + \frac{\xi_2^{c_2}}{1 - \alpha_2} (1 - \alpha_2) \Sigma n^\frac{1}{2} \left( \hat{\beta}_1 - \beta \right) + o_p(1)
= n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \xi_i \left\{ u_i \left( 1 + \frac{\xi_1^{c_2}}{1 - \alpha_2} \right) + w_i \frac{\xi_1^{c_2}}{1 - \alpha_1 - \xi_1^{c_2}} \right\} + o_p(1).$$

Then, writing

$$q_2 = \frac{\xi_1^{c_2}}{1 - \alpha_2},$$

we find:

$$n^\frac{1}{2} \left( \hat{\beta}_2 - \beta \right) \xrightarrow{D} N_k \left[ 0, \sigma^2 \Sigma^{-1} \left( 1 + q_2 \right)^2 + 2q_1 q_2 (1 + q_2) \tau_2^{c_1} + q_1^2 q_2^2 \tau_2^{c_1} \right].$$

For the $(m + 1)$-step estimator we find:

$$(1 - \alpha) \Sigma n^\frac{1}{2} \left( \hat{\beta}_{m+1} - \beta \right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \xi_i \left\{ u_i \left[ \sum_{j=0}^{m} \left( \frac{\xi_1^{c_2}}{1 - \alpha_2} \right)^j \right] + \left( \frac{\xi_1^{c_2}}{1 - \alpha_2} \right)^m q_1 w_i \right\} + o_p(1).$$

If we let $m \to \infty$, because $\xi_1^{c_2}/(1 - \alpha) < 1$, and provided that the iteration of the $o_p(1)$ term does not blow up:

$$(1 - \alpha_2) \Sigma n^\frac{1}{2} \left( \hat{\beta}_2 - \beta \right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \xi_i u_i \left( \frac{1 - \alpha_2}{1 - \alpha_2 - \xi_1^{c_2}} \right) + o_p(1).$$

This is (11) but in terms of $u_i, \alpha_2, c_2$, so the same asymptotic distribution as if we would have run LTS at $\alpha_2$. This is also the same asymptotic distribution as the 2S-LWS estimator of Čižek (2007, Corollary 5.2).

The conditional variance converges in probability towards:

$$\sigma^2 \Sigma \left[ \frac{\tau_2^{c_2}}{(1 - \alpha_2 - \xi_1^{c_2})^2} \right].$$
Then
\[ n^{\frac{1}{2}} \left( \hat{\beta}_s - \beta \right) \xrightarrow{D} N_k \left( 0, \sigma^2 \Sigma^{-1} \frac{\tau_{\hat{\beta}_s}^2}{(1 - \alpha_2 - \xi_1^2)^2} \right). \quad (16) \]
Under normality \( \tau_{\hat{\beta}_s}^2 = 1 - \alpha_2 - \xi_1^2 \), in which case the efficiency is \( 1 - \alpha_2 - \xi_1^2 \). Iterating one more time from \( \hat{\beta}_s \) does not change the asymptotic distribution. The distribution of \( \hat{\beta}_s \) is the same as that of LTS based on \( \alpha_2 \).

The efficiency of the variance could be derived in a similar fashion. The results can also be derived for a more general distribution than the normal.

A.2 The iterated impulse-indicator saturation estimator

Again assume symmetry and follow the proof of Theorem JN-1.7. Using equation (JN-1.20) with \( N_n = n^{-1/2} I_n \):
\[ (1 - \alpha) \Sigma n^{\frac{1}{2}} \left( \hat{\beta}_2 - \beta \right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \varepsilon_i w_i + \frac{\xi_1^c}{1 - \alpha} (1 - \alpha) \Sigma n^{\frac{1}{2}} \left( \hat{\beta}_1 - \beta \right) + o_p(1). \quad (17) \]
Insert (JN-1.24), which is:
\[ (1 - \alpha) \Sigma n^{\frac{1}{2}} \left( \hat{\beta}_1 - \beta \right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \varepsilon_i w_i + \xi_1^c n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \varepsilon_i + o_p(1), \quad (18) \]
into (17):
\[ (1 - \alpha) \Sigma n^{\frac{1}{2}} \left( \hat{\beta}_2 - \beta \right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \varepsilon_i w_i \left[ 1 + \frac{\xi_1^c}{1 - \alpha} \right] + \frac{\xi_1^c}{1 - \alpha} \xi_1^c n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \varepsilon_i + o_p(1). \]
So for the \((m + 1)\)-step estimator:
\[ (1 - \alpha) \Sigma n^{\frac{1}{2}} \left( \hat{\beta}_{m+1} - \beta \right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} x_i \varepsilon_i \left\{ w_i \left[ \sum_{j=0}^{m} \left( \frac{\xi_1^c}{1 - \alpha} \right)^j \right] + \left( \frac{\xi_1^c}{1 - \alpha} \right)^m \xi_1^c \right\} + o_p(1). \quad (19) \]

A.3 Some simulation results for the \( \tau \)-estimator

The \( \tau \)-estimator, Yohai and Zamar (1988), did not perform as well as expected in our simulations, which is why it is excluded from the main results. The algorithm we use is based on Salibian-Barrera, Willems and Zamar (2008). In the \( x \)-contaminated experiments above, the \( \tau \)-estimator does no better than OLS, which is puzzling, because it is supposed to have the ‘exact-fit’ property. To investigate this more closely, we consider a slightly modified experiment, see Table 5. Now the magnitude of the contamination is always the same, at \( \gamma = 10 \), but the variance of the regressor (just a single one here) is varied. The amount of contamination is kept at 30%.

Table 6 reports some results. The RMSE of OLS gets smaller as \( \sigma_x \) increases, because the contamination gets relatively smaller (i.e. \( \gamma/\sigma_x \) falls). Nonetheless, the robust estimaters’ performance relative to OLS improves as \( \sigma_x \) increases. The puzzling result is that \( \tau \)-estimation still performs like OLS up to \( \sigma_x = 2 \), while the others are already doing much better.
Data generation process, $i = 1, \ldots, n$

\[ y_i = 1 + \sum_{j=2}^{k} x_{ij} + \epsilon_i, \quad \epsilon_i \sim \text{IIN}[0, 1], \]
\[ x_{ij} \sim \sigma_i \text{IIN}[0, 1], \quad j = 2, \ldots, k, \]
\[ x_{ij}^* \sim x_{ij} + 10, \quad j = 2, \ldots, k. \]

Data for estimation, $i = 1, \ldots, n$

$x$-contamination: $y_i, x_{ij}$ for $i \in \mathbb{N}^0; y_i, x_{ij}^*$ for $i \in \mathbb{N}^*$

Table 5: Modified DGP with contaminated regressors

<table>
<thead>
<tr>
<th>$\sigma_x$</th>
<th>RMSE slope OLS</th>
<th>ELTS</th>
<th>MM</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.96</td>
<td>0.79</td>
<td>0.79</td>
<td>0.96</td>
</tr>
<tr>
<td>$\sqrt{2}$</td>
<td>0.91</td>
<td>0.38</td>
<td>0.38</td>
<td>0.92</td>
</tr>
<tr>
<td>$\sqrt{3}$</td>
<td>0.88</td>
<td>0.15</td>
<td>0.15</td>
<td>0.87</td>
</tr>
<tr>
<td>2</td>
<td>0.84</td>
<td>0.08</td>
<td>0.08</td>
<td>0.76</td>
</tr>
<tr>
<td>5</td>
<td>0.46</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>10</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 6: RMSE of the slope parameters using the DGP of Table 5. $M = 1000$.

References


16


