# Reweighted Least Trimmed Squares

Cizek, P.

*Publication date:*
2010

No. 2010–91

# REWEIGHTED LEAST TRIMMED SQUARES: AN ALTERNATIVE TO ONE-STEP ESTIMATORS

By Pavel Čížek

August 2010

ISSN 0924-7815

# Reweighted least trimmed squares: an alternative to one-step estimators

Pavel Čížek

*Department of Econometrics & OR, Tilburg University, P.O.Box 90153, 5000 LE Tilburg, The Netherlands*

*E-mail: P.Cizek@uvt.nl*

**Abstract**

A new class of robust regression estimators is proposed that forms an alternative to traditional robust one-step estimators and that achieves the $\sqrt{n}$ rate of convergence irrespective of the initial estimator under a wide range of distributional assumptions. The proposed reweighted least trimmed squares (RLTS) estimator employs data-dependent weights determined from an initial robust fit. Just like many existing one- and two-step robust methods, the RLTS estimator preserves robust properties of the initial robust estimate. However contrary to existing methods, the first-order asymptotic behavior of RLTS is independent of the initial estimate even if errors exhibit heteroscedasticity, asymmetry, or serial correlation. Moreover, we derive the asymptotic distribution of RLTS and show that it is asymptotically efficient for normally distributed errors. A simulation study documents benefits of these theoretical properties in finite samples.

*Keywords:* asymptotic efficiency, breakdown point, least trimmed squares

*JEL codes:* C13, C21

# 1. Introduction

In statistics, techniques robust to atypical observations have recently been studied since such observations can arise for many reasons: heavy-tailed data distributions, miscoding, or heterogeneity not captured or presumed in a model. This is of high importance especially in (non)linear regression models and time series as the least squares (LS) and maximum likelihood (MLE) estimators are heavily influenced by data contamination. For example, Balke and Fomby (1994) document presence of outliers in macroeconomic time series and Sakata and White (1998) evidence data contamination in financial time series and its adverse effects on estimators and tests. The need for estimation procedures insensitive to data contamination and large errors have been recognized by many authors, for example, Hampel et al. (1986), Simpson et al. (1992), Stromberg et al. (2000), and Gervini and Yohai (2002). On the other hand, the use of methods robust to atypical observations is infrequent in many fields and often limited to detection of outliers (e.g., Temple, 1998; Woo, 2003), although exceptions exist (e.g., Preminger and Franck, 2007). The reasons could range from missing some (easily applicable) results regarding robust inference, low relative efficiency of many robust methods, or the necessity to choose auxiliary tuning parameters. In addition, the detection of outliers by a robust method or eye-balling and, after removing outliers, the subsequent application of a standard method such as LS is not a theoretically justified inference method as the usual standard errors (and statistics based on them) will be biased (Welsh and Ronchetti, 2002).

To address these issues, a new class of robust estimation methods is proposed, the reweighted least trimmed squares (RLTS). While the method and its robust properties rely on an initial robust estimator, RLTS possesses an asymptotic distribution independent of the initial estimator, has a known variance for example under heteroscedas-

ticity or asymmetrically distributed errors, and achieves asymptotic efficiency under normality. This facilitates easy and precise robust estimation and inference. At the same time, RLTS inherits the robust properties of the initial robust fit; for example, the breakdown point, which measures the smallest contaminated fraction of a sample that can arbitrarily change the estimates (see Section 4 for a definition and Genton and Lucas, 2003, and Davies and Gather, 2005, for details). We concentrate here on the equivariant estimators that achieve the maximal asymptotic breakdown point $1/2$ (in contrast, this measure equals zero for LS in usual regression settings).

There is of course a number of high breakdown-point methods, which are insensitive to deviations from the regression model. Many of traditional robust methods however pay for their robustness by a low relative efficiency with non-contaminated data, especially with normally distributed data. For example, the least median of squares (LMS; Rousseeuw, 1984) converges only at rate $n^{-1/3}$ and the least trimmed squares (LTS; Rousseeuw, 1985) and S-estimators (Rousseeuw and Yohai, 1984), while achieving the usual $\sqrt{n}$ consistency, exhibit under normality the asymptotic relative efficiency of 8% and 28%, respectively. To improve the quality of estimation of high breakdown-point methods, Rousseeuw and Leroy (1987) initially suggested using weighted least squares (WLS), where observations with (robustly-estimated) standardized residuals beyond some fixed cut-off point are assigned zero weight. Even though this reduces the variability of estimates, this method converges at the same rate as the initial robust estimator (He and Portnoy, 1992) and has the asymptotic distribution dependent on the initial robust fit (Welsh and Ronchetti, 2002). A more general class of such iterated estimators are the one-step M-estimators (e.g., Simpson et al., 1992), which start from an initial robust fit and perform one Newton-Raphson iteration of an M-estimation algorithm, for instance. In general, the convergence rate and asymptotic distribution of the one-step M-estimators also depend on the initial

robust estimate: while they can be often asymptotically equivalent to the non-iterated M-estimators under symmetrically distributed and homoscedastic errors (Welsh and Ronchetti, 2002), this does not hold when errors become heteroscedastic or asymmetrically distributed (Simpson et al., 1992). Further, to combine efficiency under normality and a high breakdown point, Gervini and Yohai (2002) proposed to use the WLS strategy with a data-dependent cut-off point by means of the robust and efficient weighted least squares (REWLS). Apart from the optimal case of Gaussian data, the convergence rate and asymptotic distribution of REWLS again depend on the initial estimator even for homoscedastic symmetrically distributed errors.

While one-step estimators and REWLS represent (efficient) robust estimators suitable for the standard linear regression model with independent and identically distributed errors and continuously-distributed explanatory variables, they are less practical in areas, where regression variables or errors often exhibit dependence, heteroscedasticity, and non-normality (e.g., all these issues can be present in microeconomic and other panel data; see e.g. Baltagi et al., 2010). In such models, statistical inference requires the knowledge of the (asymptotic) distributions of REWLS and an initial robust estimator, for instance. Therefore, even if REWLS were studied in a more general setting than by Gervini and Yohai (2002), inference would be difficult since the asymptotic distributions of many high breakdown-point regression estimators is known only for independent and identically distributed (iid) data.

In this paper, we propose a new efficient high breakdown-point regression estimator, RLTS. Similarly to Gervini and Yohai (2002), we construct data-dependent weights using the empirical distribution of regression residuals. Instead of using WLS, we however employ the weights for the LTS estimator. This approach eliminates the asymptotic first-order dependence of the RLTS estimates on the initial estimator under various distributional assumptions: the asymptotic distribution is derived for

heteroscedastic, asymmetric, and serially correlated errors. This results in the asymptotic efficiency of RLTS in the models with Gaussian errors, extends currently known results for LTS (cf. Čížek, 2006), and facilitates new applications of robust methods (e.g., Aquaro and Čížek, 2010). Altogether, precise and correct inference using RLTS is possible irrespective of the initial estimator. This is important especially for data exhibiting heteroscedasticity, asymmetry, and other departures from the assumption of iid symmetric errors since many highly robust estimators have not been (asymptotically) studied for such data yet. In the case of the standard linear regression with iid data, the independence of the initial estimator leads at least to a better performance of RLTS compared to REWLS in small samples. Finally, even though we concentrate here on linear regression, the principle of RLTS is straightforward to generalize to (robust) nonlinear regression and the maximum (trimmed) likelihood estimation (e.g., using Čížek, 2008).

The paper is organized as follows. The existing LTS and REWLS estimators are introduced in Section 2. Next, RLTS is proposed in Section 3 and its robust and asymptotic properties are studied in Sections 4 and 5, respectively. The finite-sample properties of the proposed and existing methods are evaluated and compared using Monte Carlo experiments in Section 6. Proofs are given in the appendices.

# 2. Least trimmed squares and efficient robust estimation

Let us now introduce the LTS and REWLS estimators of the linear regression model

$$y_i = x_i^\top \beta^0 + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$ denote the response and explanatory variables and $\beta^0 \in \mathbb{R}^p$ is the true value of the $p$ unknown regression parameters; $x_i$ is assumed to contain the intercept. Rousseeuw (1985) proposed to robustly estimate this model by LTS,

$$\hat{\beta}_n^{(LTS)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{h_n} r_{[i]}^2(\beta), \tag{2}$$

where $r_{[i]}^2(\beta)$ represents the $i$th smallest order statistics of squared regression residuals $r_1^2(\beta), \ldots, r_n^2(\beta)$ and $r_i(\beta) = y_i - x_i^\top \beta$. The trimming constant $h_n$, $\frac{n}{2} < h_n \leq n$, is usually defined in such a way that $h_n/n \to \lambda \in \langle 1/2, 1\rangle$. It determines the breakdown point of LTS since definition (2) implies that $n - h_n$ observations with the largest residuals do not directly affect the estimator. The maximal breakdown point equals asymptotically $1/2$ for $h_n = [n/2] + [(p+1)/2]$ (Rousseeuw and Leroy, 1987), whereas it asymptotically equals 0 for $h_n = n$, which corresponds to LS. Note that, using weights $w_i = w(i/n)$ and $w(z) = I(z \leq \lambda)$, LTS can be alternatively defined by

$$\hat{\beta}_n^{(LTS)} \;=\; \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w_i r_{[i]}^2(\beta) = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w\big[G_n\big\{r_i^2(\beta)\big\}\big]\, r_i^2(\beta). \tag{3}$$

This facilitates the introduction of general weights in LTS (Víšek, 2002).

For Gaussian data, the relative asymptotic efficiency of LTS with the maximal breakdown point is only 8%. Therefore, Rousseeuw and Leroy (1987) proposed to combine robust estimators with WLS. Given initial robust estimates $\hat{\beta}_n^0$ and $\hat{\sigma}_n^0$ of regression parameters and residual standard deviation, one can define in the simplest case the hard-rejection weights $w_i(\hat{\beta}_n^0, \hat{\sigma}_n^0) = I\{|r_i(\hat{\beta}_n^0)/\hat{\sigma}_n^0| < c\}$ for some $c > 0$ and $i = 1, \ldots, n$ and then estimate using WLS. The constant $c$, representing a high quantile of the normal distribution, equals frequently $c = 2.5$ in the literature. While this method decreases the variability of the estimates, it converges to $\beta^0$ at the same

rate as the initial estimator and its asymptotic distribution depends on the initial estimator as well (Welsh and Ronchetti, 2002).

As a further improvement, Gervini and Yohai (2002) proposed REWLS, a method to adaptively determine the observations that needs to be trimmed and to apply LS to the rest of data. Specifically, the weights for sample observations are defined by

$$w_i(\hat{\beta}_n^0, \hat{\sigma}_n^0) = I\{|r_i(\hat{\beta}_n^0)/\hat{\sigma}_n^0| < t_n\} \tag{4}$$

for some data-dependent $t_n > 0$ and the estimation is done using WLS. To find $t_n$, one measures the largest discrepancy between the distribution functions $F^+$ and $F_0^+$ of absolute standardized residuals underlying data and assumed in the model (1), respectively, in the tail of $F_0^+$. It is theoretically defined for $c > 0$ (e.g., $c = 2.5$) by

$$d_0 \quad = \quad \sup_{t \geq c} \max\{0, F_0^+(t) - F^+(t)\} \tag{5}$$

and it is estimated using the empirical distribution function $F_n^+$ of $|r_i(\hat{\beta}_n^0)/\hat{\sigma}_n^0|$:

$$d_n \quad = \quad \sup_{t \geq c} \max\{0, F_0^+(t) - F_n^+(t)\}. \tag{6}$$

As $d_n$ measures the fraction of observations too large for model (1) with $\varepsilon_i \sim F_0^+$, the cut-off point $t_n$ is set to the $(1 - d_n)$th quantile of $F_n^+$: $t_n = \min\{t : F_n^+(t) \geq 1 - d_n\}$. Typically, $F_0^+$ is constructed under the assumption $\varepsilon_i \sim N(0, \sigma)$, which guarantees the efficiency of LS and a low probability of outliers. The REWLS estimator preserves the breakdown-point properties of the initial estimator and achieves asymptotic efficiency under the normal model. In general, the convergence rate and asymptotic distribution of REWLS nevertheless depend on the initial robust estimator.

# 3. Reweighted least trimmed squares

We now propose using data-dependent weights within the LTS estimator so that the reweighted LTS estimator can employ information about the distribution function of errors $\varepsilon_i$ by means of its nonparametric estimate. Similarly to REWLS, this procedure should combine the robustness of the initial estimator and a high precision of estimates. Contrary to REWLS, using data-dependent weights within LTS rather than LS will asymptotically eliminate the dependence of the resulting estimates on the initial estimates (see Section 5), simplifying thus further inference.

Let $\hat{\beta}_n^0$ and $\hat{\sigma}_n^0$ be again the initial estimates of regression parameters and residual standard deviation. Similarly to (4), the aim is to construct hard-rejection weights $w_i(\hat{\beta}_n^0, \hat{\sigma}_n^0)$ determining which observations should be trimmed. Since LTS requires only the total number $h_n$ of observations to be included in the objective function, the total number of observations with non-zero weights has to be found: $\hat{h}_n = \sum_{i=1}^n w_i(\hat{\beta}_n^0, \hat{\sigma}_n^0) = \sum_{i=1}^n I\{|r_i(\hat{\beta}_n^0)/\hat{\sigma}_n^0| < t_n\}$ for some $t_n > 0$. The (second-step) reweighted LTS estimator is then simply defined as LTS using the estimated data-dependent trimming constant $\hat{h}_n$.

In particular, the implementation for the hard-rejection weights (4) proposed by Gervini and Yohai (2002) works as follows. Using $\hat{\beta}_n^0$ and $\hat{\sigma}_n^0$, construct absolute standardized residuals $|r_i(\hat{\beta}_n^0)/\hat{\sigma}_n^0|$ and their empirical distribution function $F_n^+$ and compare it with the distribution $F_0^+$ of absolute standardized residuals assumed in the model (1), where $\varepsilon_i \sim N(0, \sigma^2)$, for instance. Analogously to (6), set

$$\hat{\lambda}_n \quad = \quad \max\left\{1 - \sup_{t \geq c} \max\{0, F_0^+(t) - F_n^+(t)\}, 1/2\right\}, \tag{7}$$

that is, $\hat{\lambda}_n = \max\{1 - d_n, 1/2\}$ for $d_n$ defined in (6), and set the corresponding amount of trimming to $\hat{h}_n = [\hat{\lambda}_n n] = \sum_{i=1}^n I(|r_i(\hat{\beta}_n^0)/\hat{\sigma}_n^0| < t_n)$ for $t_n = \min\{t : F_n^+(t) \geq \hat{\lambda}_n\}$

([$z$] represents here the integer part of $z$). The reweighted least trimmed squares (RLTS) estimator is then defined by

$$\hat{\beta}_n^{(RLTS)} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{\hat{h}_n} r^2_{[i]}(\beta) = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{[\hat{\lambda}_n n]} r^2_{[i]}(\beta). \tag{8}$$

(Alternatively, one can use weights $\hat{w}_n(z) = I(t \leq \hat{\lambda}_n)$ in (3) instead of $w(z)$. This indicates it is possible to define other than 0-1 weights by estimating a general weight function $\hat{w}_n$ and using it in the LTS definition (3); see Čížek (2010) for examples.)

The crucial distinction between LTS and RLTS lies in the fact that the trimming sequence $\hat{\lambda}_n$ of RLTS can converge to an unknown constant $\hat{\lambda} \in \langle 1/2, 1\rangle$ (e.g., depending on the distribution function of $\varepsilon_i$), whereas LTS can be applied only if $\lambda \in \langle 1/2, 1\rangle$ and $h_n = [\lambda n]$ are known. Specifically to achieve asymptotically the breakdown point 1/2, we have to use $\lambda = 1/2$ in LTS. On the other hand, we will show in Section 4 that RLTS with the adaptive trimming $\hat{\lambda}_n$ can achieve the same breakdown point despite the fact that $\hat{\lambda}_n \in \langle 1/2, 1\rangle$ and even that $\hat{\lambda}_n \to 1$ if $\varepsilon_i \sim F^+ = F_0^+$ in (1).

Finally, let us note that the choice of trimming sequences $\hat{\lambda}_n$ and $\hat{h}_n = [\hat{\lambda}_n n]$ are not limited to those defined by (7). The number $\hat{h}_n$ of observations included in the RLTS objective function can be determined in practically any way as long as $\lim_{n\to\infty} \hat{h}_n/n$ exists. In this context, it is beneficial to consider a specific example of discretized trimming constants. Suppose that some initial estimates $\hat{\beta}_n^0$ and $\hat{\sigma}_n^0$ and trimming sequences $\hat{\lambda}_n$ and $\hat{h}_n$ are given. Additionally, let $\Lambda = \{\lambda_j\}_{j=0}^{D+1}, D \in \mathbb{N}$, be a discrete set of $\lambda$-values such that $1/2 = \lambda_0 < \lambda_1 < \ldots < \lambda_D < \lambda_{D+1} = 1$; for example, one could impose that trimming constants $\lambda$ are to be estimated only up to one ($\Lambda = \{0.5, 0.6, \ldots, 1.0\}$) or two digits ($\Lambda = \{0.50, 0.51, \ldots, 1.00\}$). To map the estimated $\hat{\lambda}_n$ to the set $\Lambda$, we further need a decreasing sequence $\{\eta_n\}_{n\in\mathbb{N}}$ that satisfies $0 < \eta_n < \min_{j=0,\ldots,D}(\lambda_{j+1} - \lambda_j)/4$ and slowly converges to zero, $\eta_n \downarrow 0$ as

$n \to \infty$. The discretized trimming sequence can be defined for $n \in \mathbb{N}$ as

$$\hat{\lambda}_n^d = \max_{k=0,\ldots,D} \{\lambda_0\} \cup \{\lambda_k : \lambda_k \leq \hat{\lambda}_n - \eta_n\}, \tag{9}$$

and subsequently, $\hat{h}_n^d = [\hat{\lambda}_n^d n]$. While RLTS with the discretized trimming $\hat{h}_n^d$ trims more observations than necessary, $\hat{\lambda}_n^d \leq \hat{\lambda}_n$, we will see that $\hat{\lambda}_n^d$ converges to its limit faster than the original sequence $\hat{\lambda}_n$ and that RLTS based on $\hat{\lambda}_n^d$ and $\hat{h}_n^d$ will be asymptotically normal even if errors in (1) do not possess finite second moments.

# 4. Fundamental properties

In this section, we will study the asymptotic behavior of trimming sequences $\hat{h}_n$ and $\hat{\lambda}_n$ and the robust properties of RLTS under the data-dependent trimming.

One of the reasons motivating REWLS and RLTS was low relative efficiency of many high breakdown-point estimators. To explain how RLTS improves upon this, we show for example that $\hat{\lambda}_n \to 1$ as $n \to \infty$ if $\varepsilon_i \sim F^+ = N(0, \sigma^2) \equiv F_0^+$ in (1). Hence for normal data, the objective function of RLTS becomes asymptotically identical to the LS criterion. (Note that the following theorem also holds under more general Assumption A introduced later in Section 5; one would have to rely on the results of Engler and Nielsen (2009) instead of Gervini and Yohai (2002).)

**Theorem 1.** *Assume that* $\{(y_i, x_i)\}_{i\in\mathbb{N}}$ *is a random sample from model (1), that* $\{\varepsilon_i\}_{i\in\mathbb{N}}, \varepsilon_i \sim F$, *are independent and identically distributed random variables with finite second moments and stochastically independent of* $x_i$, *and that the initial estimators* $\hat{\beta}_n^0$ *and* $\hat{\sigma}_n^0$ *are consistent,* $\hat{\beta}_n^0 \to \beta^0$ *and* $\hat{\sigma}_n^0 \to \sigma^2$ *in probability as* $n \to \infty$. *Then it holds that*

1. *if* $F$ *is continuous,* $\hat{\lambda}_n \to \hat{\lambda} = \max\left\{1 - \sup_{t\geq c} \max\{0, F_0^+(t) - F^+(t)\}, 1/2\right\}$ *in*

*probability as* $n \to \infty$. *If additionally* $F^+ = F_0^+$, *then* $\hat{\lambda} = 1$.

2. *if* $F$ *is continuous and* $\eta_n = o(1)$ *such that* $|\hat{\lambda}_n - \hat{\lambda}| = o_p(\eta_n)$, *then* $\hat{\lambda}_n^d \to \hat{\lambda}^d = \max_{k=0,\dots,D}\{\lambda_0\} \cup \{\lambda_k : \lambda_k < \hat{\lambda}\}$ *in probability as* $n \to \infty$.

3. *if* $\hat{\beta}_n^0$ *and* $\hat{\sigma}_n^0$ *are* $n^\tau$*-consistent,* $\tau \geq 1/4$, $F$ *is symmetric and absolutely continuous with a differentiable density* $f$ *such that* $f'(z)$ *and* $z^2 f'(z)$ *are bounded, and* $x_i$ *possesses finite second moments, then* $|\hat{\lambda}_n - \hat{\lambda}| = \mathcal{O}_p(n^{-\frac{1}{2}})$ *as* $n \to \infty$.

4. *if* $\hat{\beta}_n^0$ *and* $\hat{\sigma}_n^0$ *are* $n^\tau$*-consistent,* $\tau > 0$, $\eta_n = o(1)$ *such that* $n^{-\tau} = o(\eta_n)$, $F$ *is absolutely continuous with a density* $f$ *such that* $\max\{1, z\}f(z)$ *is bounded, and* $x_i$ *is integrable, then* $|\hat{\lambda}_n - \hat{\lambda}| = \mathcal{O}_p(n^{-\tau})$ *and* $|\hat{\lambda}_n^d - \hat{\lambda}^d| = \mathcal{O}_p(n^{-\frac{1}{2}})$ *as* $n \to \infty$.

Theorem 1 shows that $\hat{\lambda}_n$ and $\hat{\lambda}_n^d$ have well defined limits, and if $F^+ = F_0^+$, these limits are 1 and $\max_{\lambda_k<1} \lambda_k$, respectively. The convergence rates of both $\hat{\lambda}_n$ and $\hat{\lambda}_n^d$ equal to $n^{-\frac{1}{2}}$ if the initial estimators are $\sqrt{n}$ consistent. If $\hat{\beta}_n^0$ and $\hat{\sigma}_n^0$ are just $n^\tau$-consistent, $\tau \in (1/4, 1/2)$, $\hat{\lambda}_n$ can converge at the faster rate $n^{-\frac{1}{2}}$ only for symmetrically distributed errors $\varepsilon_i$, whereas discretized $\hat{\lambda}_n^d$ converges at the rate $n^{-\frac{1}{2}}$ irrespective of the error distribution as long as $\eta_n$ converges to zero slower than $n^{-\tau}$ (this is not a limitation because $\eta_n$ can be chosen arbitrarily).

Next, another feature of the RLTS estimator is that, similarly to REWLS, it trims only a (small) adaptively chosen proportion of observations. To show that this feature does not reduce the breakdown properties of RLTS compared to the initial estimator, we first have to define the breakdown point. Given a random sample $Z = (y_i, x_i)_{i=1}^n$, the finite-sample breakdown point of a linear-regression estimator $\hat{\beta}_n = T\{(y_i, x_i)_{i=1}^n\}$ can be defined as (Rousseeuw and Leroy, 1987)

$$\varepsilon_n^*(T, Z) = \max_{m \geq 0}\left\{\frac{m}{n} : \max_{I_m = \{i_1,\dots,i_m\}} \sup_{(\tilde{y}_i, \tilde{x}_i): i \in I_m} \left\| T\left\{(y_i, x_i)_{i \in \{1,\dots,n\} \setminus I_m} \cup (\tilde{y}_i, \tilde{x}_i)_{i \in I_m}\right\}\right\| < \infty\right\}.$$

In other words, it is the maximal number $m$ of observations that can be replaced by arbitrary values $(\tilde{y}_i, \tilde{x}_i), i \in I_m$, without making the estimate infinite and completely uninformative. The asymptotic breakdown point of the estimator $T$ is then the limit $\varepsilon^*(T) = \lim_{n\to\infty} \varepsilon_n^*(T, Z)$, providing it exists.

Now, we show that the breakdown point of RLTS preserves the breakdown point of the initial estimator $\hat{\beta}_n^0$ if $\hat{\sigma}_n^0$ is the standardized median absolute deviation (MAD) estimator, $\hat{\sigma}_n^0 = \mathrm{MAD}_{i=1,\ldots,n} r_i(\hat{\beta}_n^0)/\Phi^{-1}(3/4)$, which has the breakdown point $1/2$ for continuously distributed $\varepsilon_i$ (Davies and Gather, 2005). The claim however holds also for other high breakdown-point M-estimators of scale (Gervini and Yohai, 2002).

**Theorem 2.** *Let $Z = (y_i, x_i)_{i=1}^n$ be a random sample from model in (1), which is almost surely in a general position for $n > p$, that is, any $p + 1$ points do not lie on a hyperplane almost surely. Further, let $\varepsilon_n^{0*}(Z)$ be the finite-sample breakdown point of an initial estimator $\hat{\beta}_n^0$ of regression parameters with limit $\varepsilon^{0*} = \lim_{n\to\infty} \varepsilon_n^{0*}(Z)$. If $\hat{\sigma}_n^0 = \mathrm{MAD}_{i=1,\ldots,n} r_i(\hat{\beta}_n^0)/\Phi^{-1}(3/4)$ and $F_0$ has a finite variance, the finite-sample breakdown point $\varepsilon_n^{1*}(Z)$ of the RLTS estimator using trimming $\hat{h}_n$ or $\hat{h}_n^d$ is larger than $\varepsilon_n^{0*}(Z)$, $\varepsilon_n^{1*}(Z) \geq \min\{\varepsilon_n^{0*}(Z), \{[n/2] - (p+1)\}/n\}$, and tends to $\varepsilon^{0*}$ as $n \to \infty$.*

In Theorem 2, we limit ourselves only to independent observations so that the traditional definition of the breakdown point holds. Under dependence, exact breakdown-point results often depend on a specific model; see Genton and Lucas (2003), who indicate that the breakdown point $\varepsilon_n^*(Z)$ of an estimator in cross-sectional regression reduces to $\varepsilon_n^*(Z)/(1 + L)$ in time-series models with at most the $L$th lagged variable.

There are other characteristics of global robustness than just the breakdown point, for example, the maximum bias of an estimator caused by a given fraction of outliers. Since such a measure is not easy to derive theoretically, we attempt to estimate the maximum bias of RLTS by means of simulations in Section 6.

# 5. Asymptotic properties

In this section, we first introduce the assumptions necessary for proving the main asymptotic results. Later, the asymptotic distribution of LTS and RLTS are derived.

## 5.1 Assumptions

Let us now introduce some notation and definitions. First, the distribution functions of $\varepsilon_i$ and $\varepsilon_i^2$ in model (1) are referred to as $F$ and $G$, respectively, their density functions are denoted $f$ and $g$, provided that they exist, and the corresponding quantile functions are $F^{-1}$ and $G^{-1}$, respectively. More generally, the distribution functions of $r_i(\beta)$ and $r_i^2(\beta)$ are denoted $F_\beta$ and $G_\beta$ and the corresponding quantile functions are $F_\beta^{-1}$ and $G_\beta^{-1}$, respectively (i.e., $F \equiv F_{\beta^0}$ and $G \equiv G_{\beta^0}$). Next, the true parameter value in model (1) is referred to by $\beta^0$, where the first element of vector $\beta^0$ is assumed to represent the intercept. The true parameter value with the intercept being changed by a constant $C$ is denoted $\beta_C^0$, that is, $\beta_C^0 = \beta^0 + (C, 0, \ldots, 0)^\top$.

Further, the concept of $\beta$-mixing is introduced, which is central to the assumptions made here. A sequence of random variables $\{x_i\}_{i\in\mathbb{N}}$ is said to be absolutely regular (or $\beta$-mixing) if $\omega_m = \sup_{i\in\mathbb{N}} \mathrm{E}\{\sup_{B\in\sigma_{i+m}^f} |P(B|\sigma_i^p) - P(B)|\} \to 0$ as $m \to \infty$, where $\sigma$-algebras $\sigma_i^p = \sigma(x_i, x_{i-1}, \ldots)$ and $\sigma_i^f = \sigma(x_i, x_{i+1}, \ldots)$; see Davidson (1994) for details. Numbers $\omega_m, m \in \mathbb{N}$, are called mixing coefficients. For example, a stationary ARMA process with continuously distributed innovations is absolutely regular.

Now, the assumptions necessary to derive the asymptotic distribution of LTS and RLTS are presented. Let us only recall in this context that $\lambda \in \langle 1/2, 1\rangle$ refers to the limits $\lim_{n\to\infty} h_n/n$ or $\lim_{n\to\infty} \hat{h}_n/n$.

**Assumption A**

**A1** Random vectors $\{(y_i, x_i)\}_{i\in\mathbb{N}}$ form a strongly stationary absolutely regular sequence with mixing coefficients $\omega_m$ satisfying $m^{r/(r-2)}(\log m)^{2(r-1)/(r-2)}\omega_m \to 0$ as $m \to \infty$ for some $r > 2$ and have finite $r$th moments.

**A2** Let $\{\varepsilon_i\}_{i\in\mathbb{N}}$ be a sequence of random variables with finite second moments and $\mathrm{E}(\varepsilon_i|x_i) = 0$. The unconditional distribution function $F$ of $\varepsilon_i$ is assumed to be strictly unimodal and absolutely continuous and its density function $f$ has to be bounded and continuously differentiable. Further, $\varepsilon_i$ has to be symmetrically distributed condionally on $x_i$ or to be independent of $x_i$.

**A3** Let $Q_s(\lambda) = \mathrm{E}\{x_i x_i^\top I[|F(\varepsilon_i) - F(-\varepsilon_i - 2C)| \le \lambda]\}$ be a nonsingular matrix for any fixed $C \in \mathbb{R}$.

**A4** Assume that $\sup_{\beta\in\mathbb{R}^p} \sup_{z>\alpha} g_\beta(z) < \infty$ for any $\alpha > 0$, and if $\lambda < 1$, that $\inf_{\beta\in\mathbb{R}^p} \inf_{z\in(-\delta,\delta)} g_\beta\left(G_\beta^{-1}(\lambda) + z\right) > 0$ for some $\delta$.

Assumption A1 formulates standard conditions of the (uniform) central limit theorem. For independent $(y_i, x_i)$, the existence of finite second moments is sufficient, $r = 2$.

Assumption A2 presents standard assumptions on the error term $\varepsilon_i$, although they are more restrictive than necessary for the sake of simplicity. For example, if $\lambda < 1$ is imposed (e.g., by using the discretized trimming sequence $\hat{\lambda}_n^d$), only trimmed moments such as $E\{\varepsilon_i^2 \cdot I(\varepsilon_i^2 \le \varepsilon_{[h_n]}^2)\}$ have to exist (Čížek, 2008). Similarly, random variables $\varepsilon_i$ and $x_i$ are assumed to be independent if $F$ is asymmetric to avoid specifying an adequate kind of dependence between $\varepsilon_i$ and $x_i$. The strict unimodality of $F$ is needed for the identification of an intercept and can be relaxed if only slopes have to be identified. On the other hand, a differentiable density $f$ is necessary when the asymptotic behavior of order statistics is analyzed (cf. Stromberg et al., 2000).

Assumption A3 formulates an analog of the standard full-rank condition, $\mathrm{E}(x_i x_i^\top) >$

0, taking into account that some observations are trimmed from the (R)LTS objective function. If $\varepsilon_i$ is independent of $x_i$, Assumption A3 is equivalent to $\mathrm{E}(x_i x_i^\top) > 0$. Additionally, Assumption A3 has to hold only for $C = 0$ if $\varepsilon_i$ is symmetrically distributed conditionally on $x_i$.

Assumption A4 formalizes the fact that the distribution $G_\beta$ should be absolutely continuous: its density should not approach $\infty$ at any point and any $\beta$, which would correspond to the distribution becoming discontinuous at some point. Assumption A4 is usually implied by $F \equiv F_{\beta^0}$ being absolutely continuous with a density function $f \equiv f_{\beta^0}$ positive, bounded and differentiable around the points of trimming $\pm\sqrt{G^{-1}(\lambda)}$; see Čížek (2006) for the case of $\varepsilon_i$ and $x_i$ being stochastically independent.

## 5.2 Asymptotic normality

Let us derive the asymptotic results for LTS, that is, estimator (2) defined by a deterministic sequence of trimming constants $h_n = [\lambda n], n \in \mathbb{N}$, for some $\lambda \in \langle 1/2, 1\rangle$.

**Theorem 3.** *Let Assumption A hold and let* $\mathfrak{C}$ *solve the equation* $\mathrm{E}\{(\varepsilon_i + \mathfrak{C})I[(\varepsilon_i + \mathfrak{C})^2 \le q^2_{\lambda,\mathfrak{C}}]\} = 0$, *where* $q_{\lambda,\mathfrak{C}} = \sqrt{G^{-1}_{\beta^0_{\mathfrak{C}}}(\lambda)}$. *Next, let* $Q_s(\lambda) = \mathrm{E}[x_i x_i^\top I((\varepsilon_i + \mathfrak{C})^2 \le q^2_{\lambda,\mathfrak{C}})]$, $J_s(\lambda) = -\mathrm{E}[x_i x_i^\top q_{\lambda,\mathfrak{C}}\{f_i(-q_{\lambda,\mathfrak{C}}) + f_i(q_{\lambda,\mathfrak{C}})\}]$, *and* $Q_s(\lambda) + J_s(\lambda)$ *be a non-singular matrix, where* $f_i$ *represents the conditional distribution of* $(\varepsilon_i + \mathfrak{C})|x_i$. *Then the LTS estimator* $\hat\beta_n^{(LTS)}$ *defined by trimming* $h_n = [\lambda n]$ *for* $n \in \mathbb{N}$ *and* $\lambda \in \langle 1/2, 1\rangle$ *is a* $\sqrt{n}$*-consistent and asymptotically normal estimator of the unique* $\beta^0_{\mathfrak{C}}$, $\sqrt{n}(\hat\beta_n^{(LTS)} - \beta^0_{\mathfrak{C}}) \xrightarrow{\mathcal{L}} N(0, V(\lambda))$ *as* $n \to \infty$, *where the asymptotic covariance matrix equals*

$$V(\lambda) \;=\; \{Q_s(\lambda) + J_s(\lambda)\}^{-1}\Sigma(\lambda)\{Q_s(\lambda) + J_s(\lambda)\}^{-1} \tag{10}$$

*and*

$$\Sigma(\lambda) \;=\; \mathrm{E}\left[\sum_{j=-\infty}^{\infty}(\varepsilon_1 + \mathfrak{C})(\varepsilon_{1+|j|} + \mathfrak{C})x_1 x_{1+|j|}^\top I((\varepsilon_1 + \mathfrak{C})^2 \le q^2_{\lambda,\mathfrak{C}})I((\varepsilon_{1+|j|} + \mathfrak{C})^2 \le q^2_{\lambda,\mathfrak{C}})\right].$$

Theorem 3 generalizes the existing asymptotic results concerning LTS (e.g., Čížek, 2006) to the case of heteroscedastic, asymmetrically distributed, or serially correlated errors and facilitates new applications of LTS (e.g., Aquaro and Čížek, 2010). By Theorem 3, LTS identifies $\beta^0_{\mathfrak{C}} = \beta^0 + (\mathfrak{C}, 0, \ldots, 0)^\top$, that is, the slope parameters are consistently estimated and the intercept estimate is asymptotically "shifted" by $\mathfrak{C}$. Since $r_i(\beta^0_{\mathfrak{C}}) = \varepsilon_i + \mathfrak{C}$, $\varepsilon_i + \mathfrak{C}$ can be consistently estimated by the regression residuals $y_i - x_i^\top \hat{\beta}_n^{(LTS)}$. This enables the estimation of functions of $r_i(\beta^0_{\mathfrak{C}}) = \varepsilon_i + \mathfrak{C}$ including the distribution and quantile functions $G_{\beta^0_{\mathfrak{C}}}$ and $G^{-1}_{\beta^0_{\mathfrak{C}}}$, the density function $f_i$ of $(\varepsilon_i + \mathfrak{C})|x_i$, and the asymptotic variance matrix $V(\lambda)$ (see Čížek, 2010). However, if the errors $\varepsilon_i$ are symmetrically distributed, $\mathfrak{C} = 0$ and LTS identifies all regression parameters including intercept. In addition, if $\varepsilon_i$ is independent of $x_i$ and $\varepsilon_j, j \neq i$, and $f_i = f$, the asymptotic variance (10) reduces to the one found for LTS by Čížek (2006).

On the other hand, the proposed RLTS estimator uses data-dependent (random) trimming sequences $\hat{h}_n = [\hat{\lambda}_n n]$ and $\hat{h}^d_n = [\hat{\lambda}^d_n n]$, see Section 3. They nonetheless have, similarly to deterministic weights, well-defined limits $\lim_{n\to\infty} \hat{h}_n/n = \hat{\lambda}$ and $\lim_{n\to\infty} \hat{h}^d_n/n = \hat{\lambda}^d$, where $\hat{\lambda}_n$ and $\hat{\lambda}^d_n$ converge to these limits at rate $n^{-\frac{1}{2}}$ by Theorem 1 (to achieve this in the case of $\hat{\lambda}_n$, the initial estimators have to be $\sqrt{n}$ consistent or the error distribution has to be symmetric). In the following theorem, we can therefore show that the asymptotic distribution of RLTS is the same as the one specified in Theorem 3 for LTS using the sequence of trimming $h_n = [\hat{\lambda} n], n \in \mathbb{N}$.

**Theorem 4.** *Let the assumptions of Theorem 3 hold. Consider the RLTS estimator* $\hat{\beta}_n^{(RLTS)}$ *defined by a trimming sequence* $\{\hat{h}_n\}_{n\in\mathbb{N}}$ *such that* $\hat{h}_n/n = [\hat{\lambda}_n n]/n \to \hat{\lambda} \in \langle 1/2, 1\rangle$ *in probability and* $|\hat{\lambda}_n - \hat{\lambda}| = \mathcal{O}_p(n^{-\frac{1}{2}})$ *as* $n \to \infty$. *Then*

$$\sqrt{n}\left(\hat{\beta}_n^{(RLTS)} - \hat{\beta}_n^{(LTS)}\right) = (K\mathcal{O}_p(1), 0, \ldots, 0)^\top + o_p(1) \tag{11}$$

*as $n \to \infty$, where $K = I(\text{“}\varepsilon_i$ is asymmetrically distributed”).*

This result shows that the RLTS estimator converges at $\sqrt{n}$ rate and follows asymptotically the same normal distribution as LTS with the same (limit) amount of trimming. This result is independent of the initial estimate under very general conditions. On the one hand, the initial first-stage estimator has to be only $n^\tau$-consistent, $\tau > 0$, as long as it guarantees $|\hat{h}_n/n \to \hat{\lambda}| = \mathcal{O}_p(n^{-\frac{1}{2}})$ as $n \to \infty$, see Theorem 1. On the other hand, errors are allowed to be heteroscedastic, asymmetric, or serially correlated. This contrasts with the necessary assumptions, for example, iid symmetric errors, required by the existing one- and two-step robust methods such as the one-step M-estimators (cf. Simpson et al., 1992). Thus, RLTS improves the convergence rate of the initial estimator, and at the same time, is first-order asymptotically independent of the initial estimator. The only limitation is that this equivalence of RLTS and LTS does not hold for the estimate of intercept if the errors $\varepsilon_i$ are asymmetrically distributed. Note though that the majority of robust estimators also does not identify the mean intercept under asymmetry (see e.g. Simpson et al., 1992, and Stromberg et al., 2000).

# 6. Finite-sample properties

In this section, we present a Monte Carlo study done to assess finite-sample behavior of the proposed RLTS estimator both under various error distributions (Section 6.1) and under the worst-case data contamination (Section 6.2). In particular, we study to which extent the first-order asymptotic independence of RLTS on the initial estimator holds also in finite samples and what implications does it have for the performance of RLTS. To this end, only trimming sequences $\hat{h}_n = [\hat{\lambda}_n n]$ defined by (7) are considered, which by definition exhibit more variation with respect to initial

estimates than the discretized sequences $\hat{h}_n^d$. For comparison, we use the REWLS estimator with the hard-rejection weights (4). Both estimators (using $c = 2.5$ in (6) and (7)) are evaluated using three initial high breakdown-point estimators: the LMS, LTS, and S estimators set up for the maximal breakdown point 1/2 (see Rousseeuw and Leroy, 1987, for details). All initial robust estimators are computed using the R-package 'robustbase.' The LS estimates are reported for comparison.

## 6.1 Behavior under various error distributions

We evaluate the performance of all estimators for the regression model

$$y_i = 0.5 + x_{1i} - 2x_{2i} + \varepsilon_i, \tag{12}$$

where $x_{1i}, x_{2i} \sim N(0,1)$. The errors $\varepsilon_i$ are generated from the standard normal $\varepsilon_i \sim N(0,1)$, heteroscedastic normal $\varepsilon_i \sim N(0, \exp(x_{1i} + x_{2i}))$, and asymmetric chi-square $\varepsilon_i \sim \chi^2_4 - 4$ distributions, which are further referred to as NORM, NHET, and CHISQ, respectively. Additionally, data OUT10 contaminated by 10% outliers are studied: while 90% of observations are generated using model (12) and $\varepsilon_i \sim N(0,1)$, 10% of observations are generated as $x_{1i}, x_{2i} \sim N(2,1)$ and $y_i \sim U(-20, 20)$, where $U(a,b)$ denotes the uniform distribution on $\langle a, b\rangle$.

The performance of each estimator $T$ is measured by the mean squared error (MSE). Having an experiment consisting of $S$ simulated samples of size $n$, we obtain $S$ estimates $\hat{\beta}_n^{(T,s)}$, $s = 1, \ldots, S$, and report MSE, $MSE = \sum_{s=1}^{S} \sum_{j=J}^{p} |\hat{\beta}_{j,n}^{(T,s)} - \beta_j^0|^2/S$, either for the whole parameter vector ($J = 1$) for designs with symmetrically distributed errors or only for the slope parameters ($J = 2$) for designs with asymmetrically distributed errors. MSEs are in all cases evaluated for sample sizes from $n = 25$ to 400 and are based on 2500 simulated samples; see results in Table 1.

Table 1: Mean squared errors of the LS, REWLS, and RLTS estimators in the linear regression with normal, heterescedastic, contaminated, and asymmetric data.

| Model | Sample size | LS | REWLS using | | | RLTS using | | |
|---|---|---|---|---|---|---|---|---|
| | | | LMS | LTS | S | LMS | LTS | S |
| NORM | 25 | 0.139 | 0.507 | 0.371 | 0.309 | 0.223 | 0.210 | 0.211 |
| | 50 | 0.064 | 0.144 | 0.118 | 0.105 | 0.087 | 0.084 | 0.084 |
| | 100 | 0.031 | 0.047 | 0.044 | 0.040 | 0.038 | 0.038 | 0.037 |
| | 200 | 0.015 | 0.019 | 0.019 | 0.018 | 0.018 | 0.018 | 0.018 |
| | 400 | 0.008 | 0.009 | 0.009 | 0.009 | 0.009 | 0.009 | 0.009 |
| NHET | 25 | 0.545 | 0.598 | 0.504 | 0.452 | 0.396 | 0.386 | 0.386 |
| | 50 | 0.268 | 0.228 | 0.208 | 0.189 | 0.167 | 0.168 | 0.168 |
| | 100 | 0.135 | 0.108 | 0.099 | 0.089 | 0.085 | 0.085 | 0.084 |
| | 200 | 0.068 | 0.053 | 0.050 | 0.043 | 0.043 | 0.043 | 0.043 |
| | 400 | 0.034 | 0.030 | 0.028 | 0.022 | 0.022 | 0.022 | 0.022 |
| OUT10 | 25 | 2.809 | 0.521 | 0.392 | 0.330 | 0.253 | 0.243 | 0.238 |
| | 50 | 1.543 | 0.157 | 0.139 | 0.119 | 0.099 | 0.099 | 0.099 |
| | 100 | 0.902 | 0.058 | 0.058 | 0.051 | 0.047 | 0.047 | 0.047 |
| | 200 | 0.593 | 0.027 | 0.027 | 0.023 | 0.023 | 0.023 | 0.023 |
| | 400 | 0.445 | 0.014 | 0.014 | 0.012 | 0.012 | 0.012 | 0.012 |
| CHISQ | 25 | 0.734 | 1.518 | 1.171 | 0.974 | 0.755 | 0.711 | 0.715 |
| | 50 | 0.350 | 0.405 | 0.350 | 0.310 | 0.279 | 0.281 | 0.277 |
| | 100 | 0.168 | 0.142 | 0.144 | 0.131 | 0.129 | 0.131 | 0.131 |
| | 200 | 0.078 | 0.064 | 0.066 | 0.060 | 0.060 | 0.061 | 0.061 |
| | 400 | 0.040 | 0.032 | 0.034 | 0.030 | 0.031 | 0.031 | 0.032 |

First, let us discuss the data NORM with the standard normal errors, for which all compared estimators should be asymptotically equivalent to LS. In this case, LS is the optimal estimator, and for $n = 400$, both REWLS and RLTS exhibit the same MSEs, which are practically equal to those of LS. At smaller samples, the precision of REWLS estimates depend on the initial estimates, which result in the 18%, 37%, and 64% differences between the best and worst REWLS estimates at samples with 100, 50, and 25 observations, respectively. On the other hand, the differences in the MSEs of RLTS do not exceed 5% for various initial estimators even at $n = 25$ and

are negligible for $n \geq 100$. Additionally, RLTS provides at smaller samples, $n \leq 100$, much smaller MSEs than REWLS.

Next, the data NHET with heteroscedastic normal errors represent data, where trimming always takes place and REWLS thus depends on the initial estimator even asymptotically. The results in Table 1 indicate that, at least for this type of heteroscedasticity, LS is not the optimal estimator anymore and it actually exhibits the worst MSE for $n \geq 50$. Although REWLS performs better than LS, its performance considerably depends on the initial estimator: the MSE of REWLS estimates using LMS or LTS as the initial estimator are always 20–30% or 10–20% larger than those obtained with the initial S estimator. On the contrary, RLTS behaves practically independently of the initial estimator for $n \geq 50$ and provides the smallest MSEs.

Now, we analyze data OUT10 with normal errors, but contaminated by 10% outliers this time. The behavior of REWLS and RLTS is similar to the case with data NORM, whereas LS is extremely biased. The performance of RELWS depends on the initial estimator, strongly at small samples and less so at larger samples (for $n \geq 100$, the differences in MSEs due to the initial estimator used with REWLS are around 15%). The MSEs of RLTS are independent of the initial estimator at samples with $n \geq 50$ and significantly smaller than those of REWLS at small samples.

Finally, data CHISQ with asymmetrically distributed errors are studied. In Table 1, the MSEs for the slope parameters are reported. Again, LS is not optimal and exhibits the largest MSEs of all the estimators for $n \geq 100$. As in the previous cases, the RLTS results depend much less on the initial estimator than the REWLS results, and additionally, RLTS performs much better than REWLS in small samples with $n \leq 50$. The only surprising result is that both two-step estimators perform generally better using LMS rather than LTS as the initial estimator for $n \geq 100$, although this might be attributed to the fact that only slope parameters are considered.

Altogether, RLTS preserves its asymptotic independence of the initial estimator also in finite samples to a large extent. Typically, the differences in MSEs due to the choice of initial estimator are negligible for $n \geq 50$. Additionally, RLTS exhibits much smaller MSEs than REWLS in very small samples with $n \leq 50$ and outperforms LS in all models except data NORM (this is true also for errors from the Student or double exponential distribution, for instance, as unreported simulations show).

## 6.2 Behavior under point contamination

In Section 6.1, RLTS and REWLS were studied under various distributional schemes. To estimate the worst effect of outliers on an estimator, we consider again normal data with two explanatory variables, which are now contaminated by several identical outliers. For a given sample size $n$ and a contamination fraction $\alpha \in (0, 0.5)$, the point-contamination model can be defined as follows: $n - [\alpha n]$ observations follow the normal model $y_i = 0 + 0x_{1i} + 0x_{2i} + \varepsilon_i$, where $x_{1i}, x_{2i} \sim N(0,1)$ and $\varepsilon_i \sim N(0,1)$; the remaining $[\alpha n]$ observations are identical outliers fixed at, without loss of generality due to the sphericity of the normal model, $x_{1i} = x_1 \in \mathbb{R}$ and $y_i = K_o \geq 0$. We consider $n \in \{25, 50, 100\}$, $\alpha \in \{0.05, 0.10, 0.20\}$, and $x_1 \in \{1, 8\}$, where $x_1 = 1$ corresponds to low-leverage contamination and $x_1 = 8$ to high-leverage contamination. The values of $K_o$ vary on a grid from 0 to 50 (higher values of $K_o$ do not affect the results) and, for each estimator, the values of $K_o$ leading to the worst MSE are determined. The maximum MSEs of REWLS and 2S-LWS estimators based on the minimax-bias LMS estimator for $x_1 = 1$ and $x_1 = 8$ evaluated using 1500 simulations are reported in Table 2. Larger sample sizes are not reported since the results for REWLS and 2S-LWS are rather similar, especially for $n = 400$.

The maximum MSEs are smaller for $x_1 = 1$ than for $x_1 = 8$, but the overall pattern

Table 2: The maximum mean squared errors of REWLS and RLTS under the point contamination.

| Contamination [%] | Sample size | $x_1 = 1$ | | $x_1 = 8$ | |
|---|---|---|---|---|---|
| | | REWLS | RLTS | REWLS | RLTS |
| 5 | 25 | 0.578 | 0.460 | 0.588 | 0.486 |
| | 50 | 0.165 | 0.135 | 0.228 | 0.196 |
| | 100 | 0.069 | 0.066 | 0.156 | 0.157 |
| 10 | 25 | 0.731 | 0.604 | 0.830 | 0.688 |
| | 50 | 0.351 | 0.287 | 0.591 | 0.504 |
| | 100 | 0.160 | 0.152 | 0.452 | 0.423 |
| 20 | 25 | 4.877 | 2.885 | 4.854 | 3.257 |
| | 50 | 2.128 | 1.454 | 2.902 | 1.950 |
| | 100 | 1.029 | 0.912 | 2.016 | 1.901 |

is similar. The maximum MSEs increase with an increasing level of contamination and with a decreasing sample size. The MSEs of RLTS are generally smaller than those of REWLS, but the differences are not very large for sample sizes $n \geq 100$. On the other hand, RLTS exhibits much smaller maximum MSEs at small sample sizes similarly to simulations in Section 6.1. The differences are more pronounced at higher levels of contamination, especially at 20%, where the maximum bias of REWLS exceed that of RLTS by 45%–70%.

# 7. Conclusion

In this paper, the two-step robust estimation method RLTS is introduced, which combines a high breakdown point and the asymptotic efficiency for Gaussian data. The main feature of RLTS is its first-order asymptotic independence of the initial estimator for a general underlying error distribution including heteroscedastic, asymmetric, and serially correlated errors. This property permits an initial estimator to be selected only with respect to its robust properties, allows easy and correct inference for

robust RLTS estimates, and additionally, renders stable and precise estimates even in very small samples. Although this method is proposed and discussed in the context of linear regression, many extensions are straightforward. In particular, an extension of the RLTS concept to nonlinear regression and maximum-likelihood based estimation are feasible using the results of Čížek (2008).

# Appendix

## A. Proofs of the fundamental properties

*Proof of Theorem 1:* 1. Under the assumptions of the theorem, Gervini and Yohai (2002, Lemma 4.1) state for $d_n$ and $d_0$ defined in (6) and (5) that $d_n \to d_0$ in probability as $n \to \infty$ since $|d_n - d_0| \leq \sup_{z\in\mathbb{R}} \|F_n^+(z) - F^+(z)\| = o_p(1)$. The claim 1 follows from definitions (6) and (5) implying $\hat{\lambda}_n = \max\{1-d_n, 1/2\}$ and $\hat{\lambda} = \max\{1-d_0, 1/2\}$ and the fact that $d_0 = \sup_{t\geq c} \max\{0, F_0^+(t) - F^+(t)\} = 0$ if $F_0^+ = F^+$.

2. Suppose now that $\hat{\lambda} \in (\lambda_k, \lambda_{k+1}\rangle$ for some $k \in \{0, \ldots, D\}$ (the case $\hat{\lambda} = \lambda_0 = 1/2$ follows from point 1). For any $\varepsilon > 0$, there is some $n_0 \in \mathbb{N}$ such that $|\hat{\lambda}_n - \hat{\lambda}| < \eta_n < (\hat{\lambda} - \lambda_k)/2$ for $n > n_0$ with probability higher than $1 - \varepsilon$. Hence, $\hat{\lambda} \in (\lambda_k, \lambda_{k+1}\rangle$ implies $\hat{\lambda}_n - 2\eta_n \in \langle\lambda_k, \lambda_{k+1})$, and by definition (9), $\hat{\lambda}_n^d = \lambda_k$ for all $n > n_0$ with probability at least $1-\varepsilon$. Thus, $\hat{\lambda}_n^d \to \lambda_k = \hat{\lambda}^d$ in probability as $n \to \infty$.

3. Gervini and Yohai (2002, Lemma 4.2) proved under the assumptions of the theorem that $\sup_{z\in\mathbb{R}} |F_n^+(z) - F^+(z)| = \mathcal{O}_p(n^{-1/2})$ for $F$ being symmetric. The claim 3 follows from the inequality $|\hat{\lambda}_n - \hat{\lambda}| \leq |d_n - d_0| \leq \sup_{z\in\mathbb{R}} |F_n^+(z) - F^+(z)|$.

4. The proof of Gervini and Yohai (2002, Lemma 4.2) shows that $\sup_{z\in\mathbb{R}} |F_n^+(z) - F^+(z)| = \mathcal{O}_p(n^{-\tau})$ if $\hat{\beta}_n^0$ and $\hat{\sigma}_n^0$ are $n^\tau$-consistent (only the density $f = F'$ and one moment of $|x_i|$ have to exist because the first-order Taylor expansion is used instead of

the second-order Taylor expansion in point 3). Next, $|\hat{\lambda}_n - \hat{\lambda}| \leq \sup_{z\in\mathbb{R}} |F_n^+(z) - F^+(z)|$ implies $|\hat{\lambda}_n - \hat{\lambda}| = \mathcal{O}_p(n^{-\tau})$ as $n \to \infty$. Since $n^{-\tau} = o(\eta_n)$, the assumption $|\hat{\lambda}_n - \hat{\lambda}| = o_p(\eta_n)$ of point 2 holds. In point 2, we showed that, for $\hat{\lambda} \in (\lambda_k, \lambda_{k+1}\rangle$ and any $\varepsilon > 0$, $P(\hat{\lambda}_n^d = \lambda_k = \hat{\lambda}^d) > 1 - \varepsilon$ for all $n > n_0$ and some $n_0 \in \mathbb{N}$. Hence, $|\hat{\lambda}_n^d - \hat{\lambda}^d| = \mathcal{O}_p(n^{-\alpha})$ for any $\alpha > 0$ (e.g., $\alpha = 1/2$). □

*Proof of Theorem 2:* The breakdown point is derived only for RLTS with the trimming sequence $\hat{h}_n = [\hat{\lambda}_n n]$ defined by (7) because $\hat{h}_n^d \leq \hat{h}_n$. In the linear model (1), this guarantees that the breakdown point of RLTS with $\hat{h}_n^d$ is higher than with $\hat{h}_n$ (at least if there are at most $[n/2] - (p+1)$ contaminated observations).

For a given sample $Z = \{y_i, x_i\}_{i=1}^n$, let $\varepsilon_n^*(Z) = \min\{\varepsilon_n^{0*}(Z),\ \{[(n+1)/2] - (p+1)\}/n\}$. The breakdown point of RLTS is larger than or equal to $\varepsilon_n^*(Z)$ if the RLTS estimates $\hat{\beta}_n^s$ obtained for samples $Z^0 = Z$ and $Z_m^s = \{y_i, x_i\}_{i\in\{1,\ldots,n\}\setminus I_m} \cup \{\tilde{y}_i^s, \tilde{x}_i^s\}_{i\in I_m}$ are uniformly bounded in $s \in \mathbb{N}$ for any $m \leq n\varepsilon_n^*(Z)$, an index set $I_m$ of size $m$, and sequences of points $\{\tilde{y}_i^s, \tilde{x}_i^s\}_{s\in\mathbb{N}}, i \in I_m$ ($Z_m^s$ represents a sample with $m$ contaminated observations). Note that $m \leq n\varepsilon_n^*(Z) \leq [n/2] - (p+1)$ and $\lceil n/2 \rceil \leq \hat{h}_n$ by (7), and thus, $p+1 \leq \hat{h}_n - m$. Hence, the objective function of RLTS at any sample $Z_m^s$ always includes at least $p+1$ original non-contaminated points.

Now, the assumptions of the theorem correspond to those of Gervini and Yohai (2002, Theorem 3.3). This is also true for the MAD scale estimator as shown in Gervini and Yohai (2002, p. 18). We can thus apply Gervini and Yohai (2002, Theorem 3.3) for the hard-rejection weights $w_i(\hat{\beta}_n^0, \hat{\sigma}_n^0) = I(|r_i(\hat{\beta}_n^0)/\hat{\sigma}_n^0| < t_n)$ defining the trimming constant $\hat{h}_n = \sum_{i=1}^n w_i(\hat{\beta}_n^0, \hat{\sigma}_n^0)$. In the proof of that theorem (equations (32) and (35)), it is shown for $w_{[i]}(\hat{\beta}_n^0, \hat{\sigma}_n^0) = I(|r_{[i]}(\hat{\beta}_n^0)/\hat{\sigma}_n^0| < t_n)$ that

$$\frac{1}{n}\sum_{i=1}^{\hat{h}_n} r_{[i]}^2(\hat{\beta}_n^0) = \frac{1}{n}\sum_{i=1}^{n} w_{[i]}(\hat{\beta}_n^0, \hat{\sigma}_n^0) r_{[i]}^2(\hat{\beta}_n^0) \leq c^2 + \int_{-\infty}^{+\infty} u^2 dF(u) < +\infty, \qquad (13)$$

and for any $p+1$ indices $1 \leq i_1 < \ldots < i_{p+1} \leq n$ and $\beta \in \mathbb{R}$, that

$$\sum_{j=1}^{p+1}(y_{i_j} - x_{i_j}^\top \beta)^2 \geq \frac{p+1}{2}\delta^2(Z)\|\beta\|^2 - \sum_{i=1}^{n} y_i^2, \tag{14}$$

where $\delta(Z) = \min_{\|v\|=1} \min\{|x_{i_j}^\top v| : 1 \leq i_1 < \ldots < i_{p+1} \leq n\} > 0$ as the points are in a general position. Consequently, the RLTS estimates $\hat{\beta}_n^s$ have to be uniformly bounded in $s \in \mathbb{N}$: on the one hand, the RLTS objective function is uniformly bounded at $\hat{\beta}_n^0$ by (13), and on the other hand, the RLTS objective function at $\hat{\beta}_n^s$ would become unbounded and larger than at $\hat{\beta}_n^0$ if $\limsup_{n\to\infty} \|\hat{\beta}_n^s\| \to +\infty$ by (14) because it always contains at least $p+1$ non-contaminated points. Thus, RLTS does not breakdown if $m \leq n\varepsilon_n^*(Z)$ points are contaminated, which closes the proof. □

# B. Proofs of the asymptotic properties

**Lemma 5.** *Let Assumption A hold and* $\{\lambda_n\}_{n\in\mathbb{N}}$ *satisfy* $|\lambda_n - \lambda| = \mathcal{O}_p(n^{-\frac{1}{2}})$ *for some* $\lambda \in \langle 1/2, 1\rangle$ *as* $n \to \infty$. *Further, let* $\mu_n$ *and* $\mu_0$ *denote the solutions of equations* $\mathcal{E}_n(C;\lambda_n) = \mathrm{E}\{(\varepsilon_i + C)I((\varepsilon_i + C)^2 \leq (\varepsilon_i + C)^2_{[\lambda_n n]})\} = 0$ *and* $\mathcal{E}(C;\lambda) = \mathrm{E}\{(\varepsilon_i + C)I((\varepsilon_i + C)^2 \leq G_C^{-1}(\lambda))\} = 0$, *respectively, where* $G_C$ *is the distribution function of* $(\varepsilon_i + C)^2$. *Then* $\mu_n$ *and* $\mu_0$ *exist, are unique, and* $\mathrm{E}|\mu_n - \mu_0| = \mathcal{O}(n^{-\frac{1}{2}})$ *as* $n \to \infty$.

*Proof:* The solutions $\mu_n$ and $\mu_0$ exist since $\mathcal{E}_n(C;\lambda)$ and $\mathcal{E}(C;\lambda)$ converge to $\pm\infty$ for $C \to \pm\infty$ and they are continuous in $C$. Consequently, solutions $\mu_n$ are uniformly bounded: if $\sup_{n\in\mathbb{N}} |\mu_n| = +\infty$, $\sup_{n\in\mathbb{N}} |\mathcal{E}_n(\mu_n;\lambda_n)| = +\infty$, which would contradict the definition of $\mu_n$. The uniqueness of $\mu_n$ and $\mu_0$ follows from the strict unimodality of $F$. Now, for $\varepsilon > 0$ and $K > 0$, $P(|\lambda_n - \lambda| < Kn^{-\frac{1}{2}}) > 1 - \varepsilon$ for any sufficiently large $n$. Thus with probability $1-\varepsilon$, $\mathcal{E}_n(C;\lambda_n) - \mathcal{E}(C;\lambda) \to 0$ uniformly in $C$ as $n \to \infty$ by Čížek (2008, Corollary A.5), and consequently, $\mu_n \to \mu_0$ in probability.

Next, since $\partial\mathcal{E}(\mu_0;\lambda)/\partial C > 0$ by Assumptions A3 and A4, there are some $\delta > 0$ and $K > 0$ such that $|\mathcal{E}(C;\lambda)| \geq K|C - \mu_0|$ for $C \in U(\mu_0, \delta)$. The consistency of $\mu_n$ thus implies for some $n_0 \in \mathbb{N}$ that $\mu_n \in U(\mu_0, \delta)$ and $|\mathcal{E}(\mu_n;\lambda)| \geq K|\mu_n - \mu_0|$ with probability $1 - \varepsilon$ for $n > n_0$. As $\mathcal{E}(\mu_n;\lambda) = \mathcal{E}(\mu_n;\lambda) - \mathcal{E}_n(\mu_n;\lambda_n) = \mathcal{O}_p(n^{-\frac{1}{2}})$ by Čížek (2008, Corollary A.5), it follows that $|\mu_n - \mu_0| = \mathcal{O}_p(n^{-\frac{1}{2}})$ as $n \to \infty$. Hence, $\mathrm{E}|\mu_n - \mu_0| = \mathcal{O}(n^{-\frac{1}{2}})$ because $\mu_n - \mu_0$ is uniformly bounded. □

*Proof of Theorem 3:* First of all, the objective function of LTS equals

$$S_n(\beta;\lambda) = \sum_{i=1}^{n} r_i^2(\beta) I(r_i^2(\beta) \leq r_{[\lambda n]}^2(\beta))$$

Therefore, we can now employ the existing asymptotic results for general trimmed estimators introduced by Čížek (2008). In this context, note that Assumption A covers all the assumptions relevant for the linear regression model used by Čížek (2008) except for the identification assumptions. Hence, we can now employ the results of Čížek (2008) once we verify the identification assumptions, which state that the limit of $S_n(\beta;\lambda)/n$ has a unique global minimum.

To do so, note that minimizing the objective function $S_n(\beta;\lambda)$ leads to the normal equations $\partial S_n(\beta;\lambda)/\partial\beta = 0$. Čížek (2006, Lemma 1) implies that the normal equations and their derivative wrt. $\beta$ can almost surely be expressed as $(k = 1, 2)$

$$S_n^{(k)}(\beta;\lambda) = \sum_{i=1}^{n} [r_i^2(\beta)]^{(k)} I(r_i^2(\beta) \leq r_{[\lambda n]}^2(\beta)) = 0. \tag{15}$$

Moreover, Assumption A allows us to use the uniform-convergence result of Čížek (2008, Lemma A.1), which implies uniformly in $\beta$ (over any compact subset of $\mathbb{R}^p$) that $S_n(\beta;\lambda)/n \to \mathrm{E}\{r_i^2(\beta) I(r_i^2(\beta) \leq G_\beta^{-1}(\lambda))\} = S(\beta;\lambda)$, $S_n'(\beta;\lambda)/n \to \mathrm{E}\{-2r_i(\beta)x_i \cdot I(r_i^2(\beta) \leq G_\beta^{-1}(\lambda))\} = D(\beta;\lambda)$, and $2Q_n(\beta;\lambda) = S_n''(\beta;\lambda) \to \mathrm{E}\{2x_i x_i^\top I(r_i^2(\beta) \leq G_\beta^{-1}(\lambda))\} =$

$2Q(\beta;\lambda)$ in probability for $n \to \infty$, where $r_i^2(\beta) \sim G_\beta$. The matrix $Q(\beta;\lambda)$ is a positive semidefinite matrix, and at $\beta_C^0$, $Q(\beta_C^0;\lambda)$ is a positive definite matrix for any $C \in \mathbb{R}$ (Assumption A3) because $r_i^2(\beta_C^0) = (\varepsilon_i + C)^2$. This guarantees that the asymptotic objective function $S(\beta;\lambda)$ has a unique global minimum if the minimum of $S(\beta;\lambda)$ is attained at some $\beta_C^0$. If $\varepsilon_i|x_i$ is symmetric, $D(\beta^0;\lambda) = 0$ since $r_i(\beta^0) = \varepsilon_i$ and $S(\beta;\lambda)$ has the unique global minimum at $\beta^0 = \beta_{\mathfrak{C}}^0$ for $\mathfrak{C} = 0$. If $\varepsilon_i|x_i$ is asymmetric, let us first denote $\mathfrak{C}$ the solution of the equation $\mathcal{E}(C;\lambda) = \mathrm{E}\{r_i(\beta_C^0) I(r_i^2(\beta_C^0) \leq G^{-1}_{\beta_C^0}(\lambda))\} = 0$, where $r_i(\beta_C^0) = \varepsilon_i + C$ (it exists and is unique by Lemma 5). Hence, the independence of $\varepsilon_i$ and $x_i$ (Assumption A2) implies that $D(\beta_{\mathfrak{C}}^0;\lambda) = 0$ and the unique global minimum is at $\beta_{\mathfrak{C}}^0 \neq \beta^0$. In both cases, the asymptotic objective function uniquely identifies a parameter vector $\beta_{\mathfrak{C}}^0$, which differs from $\beta^0$ only by the value of intercept if $\mathfrak{C} \neq 0$. Thus, the identification assumptions are satisfied at $\beta_{\mathfrak{C}}^0$.

For a sufficiently large $n$, we will now show that there is a solution to the normal equations (15) in an arbitrarily small neighborhood of $\beta_{\mathfrak{C}}^0$. If such a solution exists, it has to be unique (with an arbitrarily high probability) and equal to the LTS estimate minimizing $S_n(\beta;\lambda)$ because $D(\beta_{\mathfrak{C}}^0;\lambda) = 0$, $Q_n(\beta;\lambda)$ is positive definite around $\beta_{\mathfrak{C}}^0$, and positive semidefinite elsewhere: due to the continuity of $Q(\beta;\lambda)$ at $\beta_{\mathfrak{C}}^0$ and the uniform convergence of $Q_n(\beta;\lambda)$ to $Q(\beta;\lambda)$, it is possible for $\varepsilon > 0$ to find some $n_0 \in \mathbb{N}$ such that the matrix $Q_n(\beta;\lambda)$ is positive definite in a neighborhood of $\beta_{\mathfrak{C}}^0$ with a probability greater than $1 - \varepsilon$ for any $n > n_0$.

To find the solution of (15) for $k = 1$, the asymptotic linearity of LTS is employed in a neighborhood $U(\beta_{\mathfrak{C}}^0, n^{-\frac{1}{2}}M)$ of $\beta_{\mathfrak{C}}^0$, where $M > 0$. To characterize $\beta \in U(\beta_{\mathfrak{C}}^0, n^{-\frac{1}{2}}M)$, one can express it as $\beta = \beta_{\mathfrak{C}}^0 - n^{-\frac{1}{2}}t$ for $t \in T_M = \{t : \|t\| \leq M\}$. Thus, using the asymptotic linearity theorem (Čížek, 2008; Lemma A.7) for LTS,

$$\frac{\partial S_n(\beta_{\mathfrak{C}}^0 - n^{-\frac{1}{2}}t;\lambda)}{\partial \beta} = \frac{\partial S_n(\beta_{\mathfrak{C}}^0;\lambda)}{\partial \beta} - 2\{Q_s(\lambda) + J_s(\lambda)\} \cdot n^{\frac{1}{2}}t + o_p(1) \qquad (16)$$

uniformly for all $t \in T_M$ and $M > 0$, where $Q_s(\lambda) = \mathrm{E}\{x_i x_i^\top I(r_i^2(\beta_{\mathfrak{C}}^0) \leq G_{\beta_{\mathfrak{C}}^0}^{-1}(\lambda))\}$ and $J_s(\lambda) = \frac{\partial}{\partial \beta^\top} \mathrm{E}\{-x_i r_i(\beta_{\mathfrak{C}}^0) I(r_i^2(\beta) \leq G_\beta^{-1}(\lambda))\}\big|_{\beta=\beta_{\mathfrak{C}}^0}$. An analytic form of $J_s(\lambda)$ for $\lambda < 1$ was derived by Čížek (2009, Lemma 3): $J_s(\lambda) = \mathrm{E}\left\{-x_i x_i^\top q_\lambda \left[f_i(-q_\lambda) + f_i(q_\lambda)\right]\right\}$, where $f_i$ denotes the conditional probability density function of $\varepsilon_i | x_i$ and $q_\lambda = \sqrt{G_{\beta_{\mathfrak{C}}^0}^{-1}(\lambda)}$ (for $\lambda = 1$, it trivially holds $J_s(\lambda) = 0$).

Thus, we have to show that, with an arbitrarily high probability, there is a $t_n^* \in T_M$ such that $\beta_{\mathfrak{C}}^0 - n^{-\frac{1}{2}} t_n^*$ solves the normal equations $S_n'(\beta_{\mathfrak{C}}^0 - n^{-\frac{1}{2}} t_n^*; \lambda) = 0$. At such a solution $t_n^*$, equation (16) implies $S_n'(\beta_{\mathfrak{C}}^0; \lambda) = 2\{Q_s(\lambda) + J_s(\lambda)\} \cdot n^{\frac{1}{2}} t_n^* + o_p(1)$ and, recalling that $Q_s(\lambda) + J_s(\lambda)$ is assumed to be a nonsingular matrix,

$$t_n^* = \{Q_s(\lambda) + J_s(\lambda)\}^{-1} \cdot \frac{1}{2\sqrt{n}} S_n'(\beta_{\mathfrak{C}}^0; \lambda) + o_p(n^{-\frac{1}{2}}) \tag{17}$$

as $n \to \infty$. To prove that $t_n^*$ is bounded in probability, we have to show that

$$\begin{aligned} \frac{-1}{2\sqrt{n}} S_n'(\beta_{\mathfrak{C}}^0; \lambda) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n r_i(\beta_{\mathfrak{C}}^0) x_i I(r_i^2(\beta_{\mathfrak{C}}^0) \leq r_{[\lambda n]}^2(\beta_{\mathfrak{C}}^0)) & (18) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n r_i(\beta_{\mathfrak{C}}^0) x_i I(r_i^2(\beta_{\mathfrak{C}}^0) \leq G_{\beta_{\mathfrak{C}}^0}^{-1}(\lambda)) + o_p(1) & (19) \end{aligned}$$

is bounded in probability (the equality was derived in Čížek, 2008, Theorem 3.2, and in particular, equations (B.3) and (B.4)). Since $r_i(\beta_{\mathfrak{C}}^0) = \varepsilon_i + \mathfrak{C}$, the right-hand side of (19) is a sum of identically distributed random variables with zero mean and finite second moments (see also next paragraph), and as such, it is asymptotically normally distributed (e.g., Arcones and Yu, 1994). Hence, (19) and $t_n^*$ in (17) are bounded in probability, and for some $n_0 \in \mathbb{N}$ and $\varepsilon > 0$, the right-hand side of (16) equals zero for some $t_n^* \in T_M, n > n_0$, with probability higher than $1 - \varepsilon$. Then $\beta^0 - n^{-\frac{1}{2}} t_n^*$ is the unique solution of (15), and consequently, the LTS estimate itself,

$\hat{\beta}_n^{(LTS,\lambda)} = \beta^0 - n^{-\frac{1}{2}} t_n^*$. Apparently, it holds that $\sqrt{n}\left(\hat{\beta}_n^{(LTS,\lambda)} - \beta^0\right) = t_n^* = \mathcal{O}_p(1)$, which implies the $\sqrt{n}$ consistency of LTS.

Finally, we have to prove the asymptotic normality of LTS, that is, to find the asymptotic distribution of $t_n^*$. Because $Q_s(\lambda)$ and $J_s(\lambda)$ in (17) are constants, we just have to derive the asymptotic distribution of (19). The summands of (19), $r_i(\beta_{\mathfrak{C}}^0)x_i^\top I(r_i^2(\beta_{\mathfrak{C}}^0) \leq G_{\beta_{\mathfrak{C}}^0}^{-1}(\lambda))$, form by the construction of $\beta_{\mathfrak{C}}^0$ ($r_i(\beta_{\mathfrak{C}}^0) = \varepsilon_i + \mathfrak{C}$) a sequence of identically distributed random variables with zero mean and finite variances because the expectation of $r_i^2(\beta_{\mathfrak{C}}^0)x_i x_i^\top I(r_i^2(\beta_{\mathfrak{C}}^0) \leq G_{\beta_{\mathfrak{C}}^0}^{-1}(\lambda))$ is finite by Assumption A3. Hence, Assumption A1 allows us to employ the central limit theorem for $\beta$-mixing sequences by Arcones and Yu (1994) for (19), proving that (19) is asymptotically normal with finite variance

$$\Sigma(\lambda) = \mathrm{E} \sum_{j=-\infty}^{\infty} (\varepsilon_1+\mathfrak{C})(\varepsilon_{1+|j|}+\mathfrak{C})x_1 x_{1+|j|}^\top I[(\varepsilon_1+\mathfrak{C})^2 \leq G_{\beta_{\mathfrak{C}}^0}^{-1}(\lambda)] I[(\varepsilon_{1+|j|}+\mathfrak{C})^2 \leq G_{\beta_{\mathfrak{C}}^0}^{-1}(\lambda)].$$

By (17) and $\sqrt{n}(\hat{\beta}_n^{(LTS,\lambda)} - \beta_{\mathfrak{C}}^0) = t_n^*$, it follows that $\sqrt{n}(\hat{\beta}_n^{(LTS,\lambda)} - \beta_{\mathfrak{C}}^0) \xrightarrow{\mathcal{L}} N(0, V(\lambda))$, where $V(\lambda) = \{Q_s(\lambda) + J_s(\lambda)\}^{-1}\Sigma(\lambda)\{Q_s(\lambda) + J_s(\lambda)\}^{-1}$ due to $r_i(\beta_{\mathfrak{C}}^0) = \varepsilon_i + \mathfrak{C}$. □

*Proof of Theorem 4:* Assumption A and the identification assumptions verified in the proof of Theorem 3 allow us again to employ the results of Čížek (2008). Since $|\hat{\lambda}_n - \hat{\lambda}| = \mathcal{O}_p(n^{-\frac{1}{2}})$, the asymptotic linearity stated by Čížek (2008; Lemma A.7) applies also for the stochastic sequence of trimming constants $\hat{\lambda}_n$ with an arbitrarily high probability. Analogously to (16), we can thus write $S_n'(\beta_{\mathfrak{C}}^0 - n^{-\frac{1}{2}}t; \hat{\lambda}_n) = S_n(\beta_{\mathfrak{C}}^0; \hat{\lambda}_n) - 2\{Q_s(\hat{\lambda}) + J_s(\hat{\lambda})\} \cdot n^{\frac{1}{2}}t + o_p(1)$. Following the steps (16)–(17) of the proof of Theorem 3 to solve for $t$, we find that the solution $\hat{t}_n^* = \sqrt{n}(\hat{\beta}_n^{(LTS,\hat{\lambda}_n)} - \beta_{\mathfrak{C}}^0)$ equals

$$\sqrt{n}(\hat{\beta}_n^{(LTS,\hat{\lambda}_n)} - \beta_{\mathfrak{C}}^0) = \{Q_s(\hat{\lambda}) + J_s(\hat{\lambda})\}^{-1} \cdot n^{-\frac{1}{2}} S_n'(\beta_{\mathfrak{C}}^0; \hat{\lambda}_n) + o_p(1). \tag{20}$$

Naturally, equation (17) formulated for the (fixed) limit value $\hat{\lambda}$ holds as well:

$$\sqrt{n}(\hat{\beta}_n^{(LTS,\hat{\lambda})} - \beta_{\mathfrak{C}}^0) = \{Q_s(\hat{\lambda}) + J_s(\hat{\lambda})\}^{-1} \cdot n^{-\frac{1}{2}} S_n'(\beta_{\mathfrak{C}}^0; \hat{\lambda}) + o_p(1). \tag{21}$$

Since LTS with a fixed $\hat{\lambda}$ is defined by (21) and RLTS with a data-dependent $\hat{\lambda}_n$ by (20), the claim of the theorem is equivalent to showing for $n \to \infty$ that

$$\{Q_s(\hat{\lambda}) + J_s(\hat{\lambda})\}^{-1}[n^{-\frac{1}{2}} S_n'(\beta_{\mathfrak{C}}^0; \hat{\lambda}_n) - n^{-\frac{1}{2}} S_n'(\beta_{\mathfrak{C}}^0; \hat{\lambda})] = (\mathcal{O}_p(1), 0, \ldots, 0)^\top + o_p(1). \tag{22}$$

Let us now define $\mathcal{E}_n(C; \lambda_n) = \mathrm{E}\{r_i(\beta_C^0) I(r_i^2(\beta_C^0) \leq r_{[\lambda_n n]}^2(\beta_{\mathfrak{C}}^0))\}$ and let $\beta_{\mathfrak{C}1}^0$ and $\beta_{\mathfrak{C}2}^0$ denote the solutions of $\mathcal{E}_n(C; \lambda_{n1}) = 0$ for $\lambda_{n1} = \hat{\lambda}_n$ and of $\mathcal{E}_n(C; \lambda_{n2}) = 0$ for $\lambda_{n2} = \hat{\lambda}$. Since $\|\beta_{\mathfrak{C}j}^0 - \beta_{\mathfrak{C}}^0\| = \mathcal{O}_p(n^{-\frac{1}{2}})$ for $j = 1, 2$ by Lemma 5, the asymptotic linearity of Čížek (2008; Lemma A.7) also applies to $\beta_{\mathfrak{C}j}^0$: $S_n'(\beta_{\mathfrak{C}j}^0; \lambda_{nj}) = S_n'(\beta_{\mathfrak{C}}^0; \lambda_{nj}) - 2\{Q_s(\hat{\lambda}) + J_s(\hat{\lambda})\} \cdot n^{\frac{1}{2}}\{\beta_{\mathfrak{C}j}^0 - \beta_{\mathfrak{C}}^0\} + o_p(1)$, $j = 1, 2$. After differencing these two equations for $j = 1, 2$, the left-hand side of (22) can be expressed as

$$\{Q_s(\hat{\lambda}) + J_s(\hat{\lambda})\}^{-1}[n^{-\frac{1}{2}} S_n'(\beta_{\mathfrak{C}1}^0; \hat{\lambda}_n) - n^{-\frac{1}{2}} S_n'(\beta_{\mathfrak{C}2}^0; \hat{\lambda})] + 2n^{\frac{1}{2}}\{\beta_{\mathfrak{C}1}^0 - \beta_{\mathfrak{C}2}^0\} + o_p(1).$$

As $n^{\frac{1}{2}}\{\beta_{\mathfrak{C}1}^0 - \beta_{\mathfrak{C}2}^0\} = \mathcal{O}_p(1)$ by Lemma 5, we can prove (22) by showing that

$$\{Q_s(\hat{\lambda}) + J_s(\hat{\lambda})\}^{-1}[n^{-\frac{1}{2}} S_n'(\beta_{\mathfrak{C}1}^0; \hat{\lambda}_n) - n^{-\frac{1}{2}} S_n'(\beta_{\mathfrak{C}2}^0; \hat{\lambda})] = o_p(1). \tag{23}$$

Next, we can rewrite the difference $[S_n'(\beta_{\mathfrak{C}1}^0; \hat{\lambda}_n) - S_n'(\beta_{\mathfrak{C}2}^0; \hat{\lambda})]/\sqrt{n}$ as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ r_i(\beta_{\mathfrak{C}1}^0) x_i I(r_i^2(\beta_{\mathfrak{C}1}^0) \leq r_{[\hat{\lambda}_n n]}^2(\beta_{\mathfrak{C}1}^0)) - r_i(\beta_{\mathfrak{C}2}^0) x_i I(r_i^2(\beta_{\mathfrak{C}2}^0) \leq r_{[\hat{\lambda} n]}^2(\beta_{\mathfrak{C}2}^0)) \right]. \tag{24}$$

Denoting $\nu_i$ the summands in the sum (24), $\nu_i = r_i(\beta_{\mathfrak{C}1}^0) x_i I(r_i^2(\beta_{\mathfrak{C}1}^0) \leq r_{[\hat{\lambda}_n n]}^2(\beta_{\mathfrak{C}1}^0)) -$

$r_i(\beta^0_{\mathfrak{C}2})x_i I(r_i^2(\beta^0_{\mathfrak{C}2}) \leq r^2_{[\hat{\lambda}n]}(\beta^0_{\mathfrak{C}2}))$, we can observe that the random vectors $\nu_i$ have zero means: if $\varepsilon_i$ is asymmetrically distributed, this follows from the definition of $\beta^0_{\mathfrak{C}j}$ and the independence of $x_i$; if $\varepsilon_i|x_i$ is symmetrically distributed, this follows from $\beta^0_{\mathfrak{C}j} = \beta^0$ and $r_i(\beta^0_{\mathfrak{C}j}) = \varepsilon_i$, $j = 1, 2$. We will now show that the variance of $n^\alpha \nu_i$ is bounded for $0 < \alpha < 1/4$. Specifically,

$$\begin{aligned} n^\alpha \nu_i &= n^\alpha \left\{ r_i(\beta^0_{\mathfrak{C}1}) - r_i(\beta^0_{\mathfrak{C}2}) \right\} x_i I(r_i^2\big(\beta^0_{\mathfrak{C}1}\big) \leq r^2_{[\hat{\lambda}_n n]}\big(\beta^0_{\mathfrak{C}1}\big)) \\ &+ n^\alpha r_i(\beta^0_{\mathfrak{C}2}) x_i \left[ I(r_i^2\big(\beta^0_{\mathfrak{C}1}\big) \leq r^2_{[\hat{\lambda}_n n]}\big(\beta^0_{\mathfrak{C}1}\big)) - I(r_i^2\big(\beta^0_{\mathfrak{C}2}\big) \leq r^2_{[\hat{\lambda}n]}\big(\beta^0_{\mathfrak{C}2}\big)) \right], \end{aligned}$$

and using the Minkowski inequality, $\mathrm{E}[n^\alpha \nu_i]^2$ can be thus bounded by

$$\mathrm{E}[n\nu_i^\alpha]^2 \leq \mathrm{E} \left\| n^\alpha \left\{ r_i(\beta^0_{\mathfrak{C}1}) - r_i(\beta^0_{\mathfrak{C}2}) \right\} x_i \right\|^2 \tag{25}$$

$$+ \mathrm{E} \left\| n^\alpha r_i(\beta^0_{\mathfrak{C}2}) x_i \left[ I(r_i^2\big(\beta^0_{\mathfrak{C}1}\big) \leq r^2_{[\hat{\lambda}_n n]}\big(\beta^0_{\mathfrak{C}1}\big)) - I(r_i^2\big(\beta^0_{\mathfrak{C}2}\big) \leq r^2_{[\hat{\lambda}n]}\big(\beta^0_{\mathfrak{C}2}\big)) \right] \right\|^2 \tag{26}$$

Since $r_i(\beta^0_{\mathfrak{C}j}) = \varepsilon_i + \mathfrak{C}_j$, the first term (25) behaves as $o(1)$ for $n \to \infty$ due to Lemma 5. Because both $r^2_{[\hat{\lambda}_n n]}(\beta^0_{\mathfrak{C}1}) \to G^{-1}_{\beta^0_{\mathfrak{C}}}(\lambda)$ and $r^2_{[\hat{\lambda}n]}(\beta^0_{\mathfrak{C}2}) \to G^{-1}_{\beta^0_{\mathfrak{C}}}(\lambda)$ as $n \to \infty$, the triangle inequality and Čížek (2008, Corollary A.5) imply that (26) is also asymptotically negligible. Hence, we can apply the law of large numbers for $L^2$-mixingales (Davidson and de Jong, 1997, Corollary 2.1) to (24) written as $n^{-1/2-\alpha} \sum_{i=1}^n n^\alpha \nu_i(K) \to 0$ for some $0 < \alpha < 1/4$ in probability as $n \to \infty$. Hence, (24) is negligible in probability and (23) and (22) are valid, which concludes the proof. □

# References


[1] Andrews, D. W. K. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory* **4**, 458–467.

[2] Aquaro, M. and Čížek, P. (2010). Two-step robust estimation of fixed-effects panel data models. CentER Discussion Paper, Tilburg University.

[3] Arcones, M. A. and Yu, B. (1994). Central limit theorems for empirical and U-processes of stationary mixing sequences. *Journal of Theoretical Probability* **7**, 47–71.

[4] Balke, N. S., and Fomby, T. B. (1994). Large shocks, small shocks, and economic fluctuations: outliers in macroeconomic time series. *Journal of Applied Econometrics* **9**, 181–200.

[5] Baltagi, B. H., Jung, B. C., and Song, S. H. (2010). Testing for heteroskedasticity and serial correlation in a random effects panel data model. *Journal of Econometrics* **154**, 122–124.

[6] Čížek, P. (2006). Least trimmed squares under dependence. *Journal of Statistical Planning and Inference* **136**, 3967–3988.

[7] Čížek, P. (2008). General trimmed estimation: robust approach to nonlinear and limited dependent variable models. *Econometric Theory* **24**, 1500–1529.

[8] Čížek, P. (2009). Generalized method of trimmed moments. CentER discussion paper 2009/25, Tilburg University, The Netherlands.

[9] Čížek, P. (2010). Efficient robust estimation of regression models. *Computational Statistics and Data Analysis*, in press.

[10] Davidson, J. (1994). *Stochastic Limit Theory*. New York: Oxford University Press.

[11] Davies, P. L. and Gather, U. (2005). Breakdown and groups. *The Annals of Statistics* **33**, 977–1035.

[12] Engler, E. and Nielsen, B. (2009). The empirical process of autoregressive residuals. *Econometrics Journal* **12(2)**, 367–381.

[13] Genton, M. G. and Lucas, A. (2003). Comprehensive definitions of breakdown points for independent and dependent observations. *Journal of the Royal Statistical Society, Series B* **65**, 81–94.

[14] Gervini, D. and Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *The Annals of Statistics* **30**, 583–616.

[15] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust statistics: The approach based on influence function*. New York: Wiley.

[16] He, X. and Portnoy, S. (1992). Reweighted LS estimators converge at the same rate as the initial estimator. *The Annals of Statistics* **20**, 2161–2167.

[17] Preminger, A., and Franck, R. (2007). Foreign exchange rates: a robust regression approach. *International Journal of Forecasting* **23**, 71–84.

[18] Ronchetti, E. and Trojani, F. (2001). Robust inference with GMM estimators. *Journal of Econometrics* **101**, 37–69.

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

[20] Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In *Mathematical statistics and applications, Vol. B*, W. Grossman, G. Pflug, I. Vincze and W. Wertz (eds.). Dordrecht: Reidel, pp. 283–297.

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

[22] Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis, Lecture notes in statistics, Vol. 26*, J. Franke, W. Härdle, and R. D. Martin (eds.). New York: Springer, pp. 256–272.

[23] Sakata, S., and White, H. (1998). High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. *Econometrica* **66**, 529–567.

[24] Simpson, D. G., Ruppert, D. and Carroll, R. J. (1992). On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association* **87**, 439–450.

[25] Stromberg, A. J., Hössjer, O. and Hawkins, D. M. (2000). The least trimmed difference regression estimator and alternatives. *Journal of the American Statistical Association* **95**, 853–864.

[26] Temple, J. R. W. (1998). Robustness tests of the augmented Solow model. *Journal of Applied Econometrics* **13**, 361–375.

[27] Víšek, J. Á. (2002). The least weighted squares I. The asymptotic linearity of normal equations. *Bulletin of the Czech Econometric Society* **9(15)**, 31–58.

[28] Welsh, A. H. and Ronchetti, E. (2002). A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference* **103**, 287–310.

[29] Woo, J. (2003). Economic, political, and institutional determinants of public deficits. *Journal of Public Economics* **87**, 387–426.