A model where the Least Trimmed Squares estimator is maximum likelihood

Vanessa Berenguer-Rico,* Søren Johansen†& Bent Nielsen‡

19 September 2021

Abstract

The Least Trimmed Squares (LTS) estimator is a popular robust regression estimator. It finds a sub-sample of h 'good' observations among n observations and applies least squares on that sub-sample. We formulate a model in which this estimator is maximum likelihood. The model has 'outliers' of a new type, where the outlying observations are drawn from a distribution with values outside the realized range of h 'good', normal observations. The LTS estimator is found to be $h^{1/2}$ consistent and asymptotically standard normal in the location-scale case. Consistent estimation of h is discussed. The model differs from the commonly used ϵ-contamination models and opens the door for statistical discussion on contamination schemes, new methodological developments on tests for contamination as well as inferences based on the estimated good data.

Keywords: Chebychev estimator, Contamination, LMS, Least squares estimator, Leverage, LTS, Maxium Likelihood, Outliers, Regression, Robust statistics.

1 Introduction

The Least Trimmed Squares (LTS) estimator suggested by Rousseeuw (1984) is a popular robust regression estimator. It is defined as follows. Consider a sample with n observations, where some are thought to be 'good' and others are 'outliers'. The user specifies that there are h 'good' observations. The LTS estimator finds the h sub-sample with the smallest residual sum of squares. Rousseeuw (1984) developed LTS in the tradition of Huber (1964) and Hampel (1971), who were instrumental in formalizing robust statistics. Huber suggested a framework of i.i.d. errors from an ϵ-contaminated normal distribution. Hampel formalized robustness and breakdown points. The present work, with its focus on maximum likelihood, deviates from this tradition.

Specifically, we propose a model in which the LTS estimator is maximum likelihood, a new approach to robust statistics. In this model, we first draw h 'good' regression errors from a normal distribution. Conditionally on these 'good' errors, we draw $n - h$ 'outlier' errors from a distribution with support outside the range of the drawn 'good' errors. The model is semi-parametric, so we apply a general notion of maximum likelihood. We provide an

---
*Department of Economics, University of Oxford. E-mail: vanessa.berenguer-rico@economics.ox.ac.uk.

†Department of Economics, University of Copenhagen and CREATES, Department of Economics and Business, Aarhus University, DK-8000 Aarhus C. E-mail: soren.johansen@econ.ku.dk.

‡Department of Economics, University of Oxford. Address for correspondence: Nuffield College, Oxford OX1 1NF, UK. E-mail: bent.nielsen@nuffield.ox.ac.uk.

asymptotic theory for a location-scale model and find that the LTS estimator is $h^{1/2}$-consistent and asymptotically standard normal. More than 50% contamination can be allowed under mild regularity conditions on the distribution function for the 'outliers'. Interestingly, the associated scale estimator does not require a consistency factor. In practice, h is typically unknown. We suggest a consistent estimator for the proportion of 'good' observations, h/n.

The approach of asking in which model the LTS estimator is maximum likelihood is similar to that taken by Gauss in 1809 (Hald, 2007, §5.5, 7.2). In the terminology of Fisher (1922), Gauss asked in which continuous i.i.d. location model is the arithmetic mean the maximum likelihood estimator and found the answer to be the normal model. Maximum likelihood often brings a host of attractive features. The model provides interpretation and reveals the circumstances under which an estimator works well in terms of nice distributional results and optimality properties. Further, we have a framework for testing the goodness-of-fit, which leads to the possibility of first refuting and then improving a model.

To take advantage of these attractive features of the likelihood framework, we follow Gauss and suggest a model in which the LTS estimator is maximum likelihood. The model is distinctive in that the errors are not i.i.d. Rather, the h 'good' errors are i.i.d. normal, whereas the $n - h$ 'outlier' errors are i.i.d., conditionally on the 'good' errors, with distributions assigning zero probability to the realized range of the 'good' errors. When $h = n$, we have a standard i.i.d. normal model, just as the LTS estimator reduces to the ordinary least squares (OLS) estimator. The model is semi-parametric, so we use an extension of traditional likelihoods, in which we compare pairs of probability measures (Kiefer and Wolfowitz, 1956) and consider probabilities of small hyper-cubes including the data point (Fisher, 1922; Scholz, 1980).

In practice, it is of considerable interest to develop a theory for inference for LTS. Within a framework of i.i.d. ϵ-contaminated errors, the asymptotic theory depends on the contamination distribution and the scale estimator requires a consistency correction (Butler, 1982; Rousseeuw, 1984; Croux and Rousseeuw, 1992; Víšek, 2006). Since the contamination distribution is unknown in practice, inference is typically done using the asymptotic distribution of the LTS estimator derived as if there is no contamination. This seems fine for an infinitesimal deviation from the central normal model (Huber and Ronchetti, 2009, §12). However, these approaches would lead to invalid inference in case of stronger contamination - see §7.1. Within the present framework, the asymptotic theory is simpler. Specifically, we derive the asymptotic properties of the LTS estimator for a location-scale version of the presented model and find that the LTS estimator has the same asymptotic theory as the infeasible OLS estimator computed from the 'good' data, when it is known which data are 'good'. As the asymptotic distribution does not depend on the contamination distribution, inference is much simpler.

In regression, a major concern is that the OLS estimator may be very sensitive to inclusion or omission of particular data points, referred to as 'bad' leverage points. LTS provides one of the first high-breakdown point estimators suggested for regression that also avoids the problem of 'bad' leverage. The presented model allows 'bad' leverage points.

Another key feature of the LTS model presented here is a separation of 'good' and 'outlying' errors, where the 'outliers' are placed outside the realized range of the 'good', normal observations. This way, the asymptotic error in estimating the set of 'good' observations is so small that it does not influence the asymptotic distribution of the LTS estimator. This throws light on the discussion regarding estimators' ability to separate overlapping populations of 'good' and 'outlying' observations (Riani et al., 2014; Doornik, 2016).

LTS is widely used in its own right and often as a starting point for algorithms such as the MM estimator (Yohai, 1987) and the Forward Search (Atkinson et al., 2010). Many variants of LTS have been developed: non-linear regression in time series (Čížek, 2005), algorithms for

fraud detection (Rousseeuw et al., 2019) and sparse regression (Alfons et al., 2013).

In all cases, the practitioner must choose h, the number of 'good' observations. In our reading, this remains a major issue in robust statistics. We propose a consistent estimator for the proportion of 'good' observations, h/n, in a location-scale model and study its finite sample properties by simulation. We apply it in the empirical example in §8.

The Least Median of Squares (LMS) estimator, also suggested by Rousseeuw (1984), is closely related to LTS. The LMS estimator seemed numerically more attractive than LTS until the advent of the fast LTS approximation to LTS (Rousseeuw and van Driessen, 2000). Nonetheless, LMS remains of considerable interest. Replacing the normal distribution in the LTS model with a uniform distribution gives a model in which LMS is maximum likelihood. In the supplementary material, we show that the LMS estimator is h-consistent and asymptotically Laplace in the location-scale case. This is at odds with the slow $n^{1/3}$ consistency rate found in the context of i.i.d. models.

We start by presenting the LTS estimator in §2. The LTS regression model is given in §3. The general maximum likelihood concept and its application to LTS are presented in §4 with details in Appendix A. §5 presents an asymptotic theory for LTS in the location-scale case with proofs in Appendix B. Estimation of h is discussed in §6. Monte Carlo simulations are presented in §7. An empirical illustration is given in §8. In a supplement, the LMS estimator is analyzed and detailed derivations of some some identities in the LTS proof are given.

Notation: Vectors are column vectors. The transpose of a vector v is denoted v'.

2 The LTS estimator

The available data are a scalar y_i and a p-vector of regressors x_i for $i = 1, \ldots, n$. We consider a regression equation $y_i = \beta' x_i + \sigma \varepsilon_i$ with regression parameter β and scale σ.

The Least Trimmed Squares (LTS) estimator suggested by Rousseeuw (1984) is defined as follows. Given a value of β, the residuals are $r_i(\beta) = y_i - \beta' x_i$. The ordered squared residuals are denoted $r_{(1)}^2(\beta) \leq \cdots \leq r_{(n)}^2(\beta)$. The user chooses an integer $h \leq n$. Given that choice, the sum of the h smallest residual squares is obtained. Minimizing over β gives the LTS estimator

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{h} r_{(i)}^2(\beta). \tag{2.1}$$

The LTS minimization classifies the observations as 'good' or 'outliers'. The set of indices of the h 'good' observations is an h-subset of $(1, \ldots, n)$. We let $\hat{\zeta}$ denote the set of indices i of the estimated 'good' observations satisfying $r_i^2(\hat{\beta}) \leq r_{(h)}^2(\hat{\beta})$. Rousseeuw and van Driessen (2000) point out that $\hat{\beta}$ is a minimizer over least squares estimators, that is $\hat{\beta} = \hat{\beta}_{\hat{\zeta}}$, where

$$\hat{\zeta} = \arg\min_{\zeta} \sum_{i \in \zeta} (y_i - \hat{\beta}_{\zeta}' x_i)^2 \quad \text{where} \quad \hat{\beta}_{\zeta} = (\sum_{i \in \zeta} x_i x_i')^{-1} \sum_{i \in \zeta} x_i y_i. \tag{2.2}$$

In the model proposed below, the maximum likelihood estimator for the scale is the residual variance of the estimated 'good' observations

$$\hat{\sigma}^2 = \frac{1}{h} \sum_{i \in \hat{\zeta}} (y_i - \hat{\beta}' x_i)^2. \tag{2.3}$$

We show consistency of $\hat{\sigma}^2$ in the location-scale case. The problem of estimating the scale is intricately linked to the choice of model for the innovations ε_i. Previously, Croux and

Rousseeuw (1992) considered $\hat{\sigma}^2$ in the context of a regression model with i.i.d. innovations with distribution function F. For that model, they found that $\hat{\sigma}^2$ should be divided by a consistency factor defined as the conditional variance of ε_i given that $|\mathsf{F}(\varepsilon_i) - 1/2| \leq h/(2n)$. In practice F is unknown, so the normal distribution is often chosen for the consistency factor.

3 The LTS regression model

We formulate the LTS model where h 'good' errors are assumed i.i.d. normal, while $n - h$ 'outliers' are located outside the realized range of the 'good' errors. This differs from the ϵ-contaminated normal model of Huber (1964), where i.i.d. errors are normal with probability $1 - \epsilon$ and otherwise drawn from a contamination distribution. The supports of the normal and the contamination distributions overlap in the ϵ-contamination case, whereas 'good' and 'outlying' observations are separated in the below LTS model.

Model 3.1 (LTS regression model)**.** *Consider the regression model $y_i = \beta' x_i + \sigma \varepsilon_i$ for data y_i, x_i with $i = 1, \ldots, n$. Let $h \leq n$ be given. Let $x_1, \ldots, x_n$ be deterministic. Let ζ be a set with h elements from $1, \ldots, n$.*
For $i \in \zeta$, let ε_i be i.i.d. $\mathsf{N}(0,1)$ distributed.
For $j \notin \zeta$, let ξ_j be independent with distribution functions $\mathsf{G}_j(z)$ for $z \in \mathbb{R}$, where G_j are continuous at 0. The 'outlier' errors are defined by

$$\varepsilon_j = (\max_{i \in \zeta} \varepsilon_i + \xi_j) 1_{(\xi_j > 0)} + (\min_{i \in \zeta} \varepsilon_i + \xi_j) 1_{(\xi_j < 0)}. \tag{3.1}$$

The parameters are $\beta \in \mathbb{R}^{\dim x}$, $\sigma > 0$, ζ which is any h-subset of $1, \ldots, n$ and G_j which are $n - h$ arbitrary distributions on $\mathbb{R}$, that are continuous at 0.
Finally, suppose $\sum_{i \in \zeta} x_i x_i'$ is invertible for all ζ.

The 'outlier' distributions G_j are parameters in the LTS model. They are constrained to be continuous at zero to avoid overlap of 'good' and 'outlier' errors. Although the regressors are deterministic in Model 3.1, randomness of $x_1, \ldots, x_n$ can be easily accommodated by a conditional model where the set of distributions $\mathsf{G}_j(z)$ are replaced by the set of conditional distributions $\mathsf{G}_j(z|x_j)$. The likelihood below then becomes a conditional likelihood.

The LTS model allows leverage points, since G_j can vary with j. Informally, an observation is a 'bad' leverage point for the OLS estimator, if that estimator is very sensitive to inclusion or omission of that observation (Rousseeuw and Leroy, 1987, §1.1). An example is the star data shown in Figure 8.1 and used in the empirical illustration in §8.

A consequence of the model is that the 'good' errors must have consecutive order statistics. Randomly, according to the choice of G_j, some 'outlier' errors are to the left of the 'good' errors. The count of left 'outlier' errors is the random variable

$$\delta = \sum_{j \notin \zeta} 1_{(\xi_j < 0)} = \sum_{j \notin \zeta} 1_{(\varepsilon_j < \min_{i \in \zeta} \varepsilon_i)}. \tag{3.2}$$

Thus, the ordered errors satisfy

$$\underbrace{\varepsilon_{(1)} \leq \cdots \leq \varepsilon_{(\delta)}}_{\delta \text{ left 'outliers'}} < \underbrace{\varepsilon_{(\delta+1)} < \cdots < \varepsilon_{(\delta+h)}}_{h \text{ 'good'}} < \underbrace{\varepsilon_{(\delta+h+1)} \leq \cdots \leq \varepsilon_{(n)}}_{\bar{n} = n - h - \delta \text{ right 'outliers'}}. \tag{3.3}$$

The set ζ collects the indices of the observations corresponding to $\varepsilon_{(\delta+1)}, \ldots, \varepsilon_{(\delta+h)}$. The difficulty in robust regression is of course that the errors are unknown. We note, that the extreme 'good' errors, $\varepsilon_{(\delta+1)}$ and $\varepsilon_{(\delta+h)}$, are finite in any realization. As the 'good' errors are normal, the extremes grow at a $(2 \log h)^{1/2}$ rate for large h, see Example 5.1 below. Likewise, the 'outlier' errors are also finite, but located outside the realized range from $\varepsilon_{(\delta+1)}$ to $\varepsilon_{(\delta+h)}$.

4 Maximum likelihood estimation in the LTS model

The LTS model is semi-parametric. We therefore start with a general maximum likelihood concept before proceeding to analysis of the LTS model.

4.1 A general definition of maximum likelihood

Traditional parametric maximum likelihood is defined in terms of densities, which are not well-defined here. Thus, we follow the generalization proposed by Scholz (1980), which has two ingredients. First, it uses pairwise comparison of measures, as suggested by Kiefer and Wolfowitz (1956), see Johansen (1978) and Gissibl et al. (2020) for applications. This way, a dominating measure is avoided. Second, it compares probabilities of small sets that include the data point, following the informality of Fisher (1922). This way, densities are not needed. Scholz' approach is suited to the present situation, where the candidate maximum likelihood estimator is known and we are searching for a model.

We consider data in $\mathbb{R}^n$ and can therefore simplify the approach of Scholz. Let $\mathcal{P}$ be a family of probability measures on the Borel sets of $\mathbb{R}^n$. Given a (data) point $z \in \mathbb{R}^n$ and a distance $\epsilon > 0$ define the hypercube $C_z^\epsilon = (z_1 - \epsilon, z_1] \times \cdots \times (z_n - \epsilon, z_n]$, which is a Borel set.

Definition 4.1. *For* $\mathsf{P}, \mathsf{Q} \in \mathcal{P}$ *write* $\mathsf{P} <_z \mathsf{Q}$ *if* $\limsup_{\epsilon \to 0} \{\mathsf{P}(C_z^\epsilon)/\mathsf{Q}(C_z^\epsilon)\} < 1$ *and* $\mathsf{P} \leq_z \mathsf{Q}$ *if* $\limsup_{\epsilon \to 0} \{\mathsf{P}(C_z^\epsilon)/\mathsf{Q}(C_z^\epsilon)\} \leq 1$, *where by convention* $0/0 = 1$.

Following Scholz, define P, Q to be equivalent at z and write $\mathsf{P} =_z \mathsf{Q}$ if $\mathsf{P} \leq_z \mathsf{Q}$ and $\mathsf{Q} \leq_z \mathsf{P}$. We get that (i) $\mathsf{P} =_z \mathsf{Q}$ if and only if $\lim_{\epsilon \to 0} \{\mathsf{P}(C_z^\epsilon)/\mathsf{Q}(C_z^\epsilon)\}$ exists and equals 1; (ii) $\mathsf{P} <_z \mathsf{Q}$ and $\mathsf{Q} <_z \mathsf{R}$ imply $\mathsf{P} <_z \mathsf{R}$ (transitivity); and (iii) $\mathsf{P} =_z \mathsf{P}$ for all $\mathsf{P} \in \mathcal{P}$ (reflexivity).

Definition 4.2. *Let* $\mathcal{P} = \{\mathsf{P}_\theta | \theta \in \Theta\}$ *be a parametrized family of probability measures. Then* $\mathsf{L}^\epsilon(\theta) = \mathsf{P}_\theta(C_y^\epsilon)$ *is the* ϵ-*likelihood at the data point* y. *Further,* $\hat{\theta}$ *is a maximum likelihood estimator if* $\mathsf{P}_\theta \leq_y \mathsf{P}_{\hat{\theta}}$ *for all* $\theta \in \Theta$.

The maximum likelihood estimator is unique if $\mathsf{P} <_z \hat{\mathsf{P}}$ for all $\mathsf{P} \neq \hat{\mathsf{P}}$. As for Scholz, traditional maximum likelihood is a special case. Further, the empirical distribution function is maximum likelihood estimator in a model of i.i.d. draws from an unknown distribution.

4.2 The LTS likelihood

To obtain the ϵ-likelihood for the LTS regression Model 3.1, we find the probability that the random n-vector of observations y_i belongs to an ϵ-cube C_z^ϵ. Throughout the argument the regressors x_i are kept fixed. Conditioning 'outliers' on 'good' observations, we get

$$\mathsf{P}(y \in C_z^\epsilon) = \prod_{i \in \zeta} \mathsf{P}(z_i - \epsilon < y_i \leq z_i) \prod_{j \notin \zeta} \mathsf{P}(z_j - \epsilon < y_j \leq z_j \mid y_i \text{ for } i \in \zeta). \tag{4.1}$$

For the first product for 'good' observations, we note that $\varepsilon_i = (y_i - \beta' x_i)/\sigma$ is standard normal and define $z_i^{\beta\sigma} = (z_i - \beta' x_i)/\sigma$ and $z_i^{\beta\sigma\epsilon} = z_i^{\beta\sigma} - \epsilon/\sigma$. Then, the factors of the first product in (4.1) are $\Phi(z_i^{\beta\sigma}) - \Phi(z_i^{\beta\sigma\epsilon})$, which we denote $\Delta^\epsilon \Phi(z_i^{\beta\sigma})$.

For the second product, let $y_i^{\beta\sigma} = (y_i - \beta' x_i)/\sigma = \varepsilon_i$ and introduce

$$\tilde{z}_j^{\beta\sigma} = (z_j^{\beta\sigma} - \min_{i \in \zeta} y_i^{\beta\sigma}) 1_{(z_j^{\beta\sigma} < \min_{i \in \zeta} y_i^{\beta\sigma})} + (z_j^{\beta\sigma} - \max_{i \in \zeta} y_i^{\beta\sigma}) 1_{(z_j^{\beta\sigma} > \max_{i \in \zeta} y_i^{\beta\sigma})}$$

and a similar expression $\tilde{z}_j^{\beta\sigma\epsilon}$ for $z_j^{\beta\sigma\epsilon}$, noting that $\tilde{z}_j^{\beta\sigma\epsilon} = 0$ if $z_j^{\beta\sigma\epsilon}$ falls within the range from $\min_{i \in \zeta} \varepsilon_i$ to $\max_{i \in \zeta} \varepsilon_i$. A derivation in Appendix A shows that the factors of the second product

in (4.1) are $\Delta^\epsilon \mathsf{G}_j(\tilde{z}_j^{\beta\sigma}) = \mathsf{G}_j(\tilde{z}_j^{\beta\sigma}) - \mathsf{G}_j(\tilde{z}_j^{\beta\sigma\epsilon})$. With this notation, the probability (4.1) of the ϵ-cube is

$$\mathsf{P}(y \in C_z^\epsilon) = \prod_{i \in \zeta} \Delta^\epsilon \Phi(z_i^{\beta\sigma}) \prod_{j \notin \zeta} \Delta^\epsilon \mathsf{G}_j(\tilde{z}_j^{\beta\sigma}). \qquad (4.2)$$

The ϵ-likelihood arises from the probability (4.2) evaluated in the data point. That is, we replace the vector z by the data vector y. We define $\tilde{y}_j^{\beta\sigma}$ and $\Delta^\epsilon \mathsf{G}_j(\tilde{y}_j^{\beta\sigma})$ in a similar fashion as before. Thus, we get the ϵ-likelihood

$$\mathsf{L}^\epsilon(\beta, \sigma, \zeta, \mathsf{G}_j \text{ for } j \notin \zeta) = \prod_{i \in \zeta} \Delta^\epsilon \Phi(y_i^{\beta\sigma}) \prod_{j \notin \zeta} \Delta^\epsilon \mathsf{G}_j(\tilde{y}_j^{\beta\sigma}). \qquad (4.3)$$

We note that the 'good' errors cannot be repeated due to the assumptions that the 'good' observations are continuous and that the centered 'outliers' are continuous at zero. Thus, parameter values resulting in repetitions of the 'good' errors are to be ignored. As an example, suppose we have $n = 10$ observations and we have chosen a β so that the residual $y_i - \beta' x_i$ takes the ordered values: 1,1,1,2,3,6,6,7,8,9. The values 1 and 6 are repetitions and cannot be 'good'. Thus, for $h = 3$, we can only select ζ as the index triplet 7,8,9. All other choices are asigned a zero likelihood. The issue is essentially the same as in ordinary least squares theory, where the least squares estimator may be useful when there are repetitions, but it cannot be maximum likelihood, as repetitions happen with probability zero under normality.

4.3   Maximum likelihood estimation

The two products in the ϵ-likelihood (4.3) resemble a standard normal likelihood and a product of $n - h$ likelihoods, each with one observations from an arbitrary distribution. We will exploit those examples using profile likelihood arguments.

First, suppose β, σ, ζ are given. Then, the first product in the LTS ϵ-likelihood (4.3) is constant. In the second product, the j-th factor only depends on G_j. Since the G_j functions vary in a product space, we maximize each factor separately. The maximum value for $\Delta^\epsilon \mathsf{G}_j(y_j^{\beta\sigma})$ is unity for any $\epsilon > 0$ and regardless of the value of β, σ, ζ. The maximum is attained for any distribution function that is continuous at zero and that rises from zero to unity over the interval from $y_j^{\beta\sigma\epsilon}$ to $y_j^{\beta\sigma}$. This set of distribution functions includes the distribution with a point mass at $\tilde{y}_j^{\beta\sigma}$, which satisfies the requirement of continuity at zero, because the 'outliers' are separated from the 'good' observations. Moreover, the point mass distribution is unique in the limit for vanishing ϵ. Thus, maximizing (4.3) over G_j gives the profile ϵ-likelihood

$$\mathsf{L}_\mathsf{G}^\epsilon(\beta, \sigma, \zeta) = \mathsf{L}^\epsilon(\beta, \sigma, \zeta, \hat{\mathsf{G}}_j \text{ for } j \notin \zeta) = \prod_{i \in \zeta} \Delta^\epsilon \Phi(y_i^{\beta\sigma}). \qquad (4.4)$$

Second, suppose ζ is given. We argue that the unique maximum likelihood estimator in the sense of Definition 4.2 is given by

$$\hat{\beta}_\zeta = \left(\sum_{i \in \zeta} x_i x_i'\right)^{-1} \sum_{i \in \zeta} x_i y_i \qquad \text{and} \qquad \hat{\sigma}_\zeta^2 = h^{-1} \sum_{i \in \zeta} (y_i - \hat{\beta}_\zeta x_i)^2. \qquad (4.5)$$

We need to show that $\limsup_{\epsilon \to 0} \{\mathsf{L}_\mathsf{G}^\epsilon(\beta, \sigma, \zeta)/\mathsf{L}_\mathsf{G}^\epsilon(\hat{\beta}_\zeta, \hat{\sigma}_\zeta, \zeta)\} < 1$ for all $(\beta, \sigma) \neq (\hat{\beta}_\zeta, \hat{\sigma}_\zeta)$. Note that $\epsilon^{-h} \mathsf{L}_\mathsf{G}^\epsilon(\beta, \sigma, \zeta)$ converges to a standard likelihood for a regression with normal errors. Therefore, the condition is satisfied as long as $\sum_{i \in \zeta} x_i x_i'$ is invertible. Thus, maximizing (4.4) over (β, σ) gives a profile ϵ-likelihood for ζ satisfying, for $\epsilon \to 0$,

$$\mathsf{L}_{\beta,\sigma,\mathsf{G}}^\epsilon(\zeta) = \mathsf{L}_\mathsf{G}^\epsilon(\hat{\beta}_\zeta, \hat{\sigma}_\zeta, \zeta) = \epsilon^h \prod_{i \in \zeta} \{\Delta^\epsilon \Phi(y^{\hat{\beta}_\zeta, \hat{\sigma}_\zeta})/\epsilon\} = \epsilon^h \{(2\pi e \hat{\sigma}_\zeta^2)^{-h/2} + \mathrm{o}(1)\}. \qquad (4.6)$$

6

Third, the profile likelihood is maximized by choosing ζ so that $\hat{\sigma}_\zeta$ is as small as possible. We note that the maximization is subject to the constraint that none of the 'good' residuals are repeated. Apart from the latter constraint, this is the Least Trimmed Squares estimator described in (2.2). We summarize.

Theorem 4.1. *The LTS regression Model 3.1 has ϵ-likelihood $\mathsf{L}^\epsilon(\beta, \sigma, \zeta, \mathsf{G}_j$ for $j \notin \zeta)$ defined in (4.3). The maximum likelihood estimator is found as follows. For any h-subsample ζ, let $\hat{\beta}_\zeta$ and $\hat{\sigma}_\zeta$ be the least squares estimators in (4.5). Let $\hat{\zeta} = \arg\min_\zeta \hat{\sigma}_\zeta^2$ subject to the constraint that $\hat{\varepsilon}_i \neq \hat{\varepsilon}_\ell$ for $i \in \zeta$, $1 \leq \ell \leq n$ and $\ell \neq i$ and where $\hat{\varepsilon}_i = y_i - \hat{\beta}_\zeta' x_i$. Then $\hat{\beta} = \hat{\beta}_{\hat{\zeta}}$ and $\hat{\sigma} = \hat{\sigma}_{\hat{\zeta}}$.*

5 Asymptotic theory for the location-scale model

We present an asymptotic theory of the LTS estimator in the special case of a location-scale model $y_i = \mu + \sigma\varepsilon_i$. In this model, the observations y_i and the unknown errors ε_i have the same ranks. Thus, the ordering in (3.3) is observable and we only need to consider values of ζ corresponding to indices of observations of the form $y_{(\delta+1)}, \ldots, y_{(\delta+h)}$ for some $\delta = 0, \ldots, n-h$. Following Rousseeuw and Leroy (1987, p. 171), the LTS estimators then reduce to $\hat{\mu} = \hat{\mu}_{\hat{\delta}}$ and $\hat{\sigma} = \hat{\sigma}_{\hat{\delta}}$, where

$$\hat{\mu}_\delta = \frac{1}{h}\sum_{i=1}^h y_{(\delta+i)}, \qquad \hat{\sigma}_\delta^2 = \frac{1}{h}\sum_{i=1}^h \{y_{(\delta+i)} - \hat{\mu}_\delta\}^2 \qquad \text{and} \qquad \hat{\delta} = \arg\min_{0 \leq \delta \leq n-h} \hat{\sigma}_\delta^2. \tag{5.1}$$

5.1 Sequence of data generating processes

We consider a sequence of LTS models indexed by n, so that $h_n \to \infty$ as $n \to \infty$. In this sequence, the 'outliers' have a common distribution G, which is continuous at zero. If $h_n = n$ the LTS estimator reduces to the full sample least squares estimator with standard asymptotic theory. Here, we choose

$$h_n/n \to \lambda \quad \text{for} \quad 0 < \lambda < 1, \tag{5.2}$$

where λ is the asymptotic proportion of 'good' observations. The parameters ζ_n, δ_n vary with n, while μ, σ, G are constant in n.

We reparametrize G in terms of separate distributions for left and right 'outliers'. Let

$$\rho = \mathsf{G}(0), \quad \overline{\mathsf{G}}(x) = (1-\rho)^{-1}\{\mathsf{G}(x) - \rho\}1_{(x>0)}, \quad \underline{\mathsf{G}}(x) = 1 - \rho^{-1}\lim_{\epsilon\downarrow 0}\mathsf{G}(-x-\epsilon)1_{(x>0)}. \tag{5.3}$$

Thus, as ξ_j is G-distributed, we have that $\underline{\varepsilon}_j = -\xi_j$ is $\underline{\mathsf{G}}$-distributed when $\xi_j < 0$ and $\overline{\varepsilon}_j = \xi_j$ is $\overline{\mathsf{G}}$-distributed when $\xi_j > 0$. This leads to

$$\mathsf{G}(x) = \{\rho + (1-\rho)\overline{\mathsf{G}}(x)\}1_{(x>0)} + \rho\{1 - \lim_{\epsilon\downarrow 0}\underline{\mathsf{G}}(-x+\epsilon)\}1_{(x\leq 0)}.$$

This way, the 'outliers', ε_j for $j \notin \zeta_n$, can be constructed through a binomial experiment as follows. Draw $n - h$ independent $\mathsf{Bernoulli}(\rho)$ variables. If the jth variable is unity then $\varepsilon_j = \min_{i\in\zeta}\varepsilon_i - \underline{\varepsilon}_j$. If it is zero then $\varepsilon_j = \max_{i\in\zeta}\varepsilon_i + \overline{\varepsilon}_j$. Hence, the number of left 'outliers' is

$$\delta_n = \sum_{j\in\zeta_n} 1_{(\varepsilon_j < \min_{i\in\zeta_n}\varepsilon_i)}, \quad \text{so that} \quad \delta_n/(n-h_n) \overset{a.s.}{\to} \rho, \tag{5.4}$$

by the Law of Large Numbers. In summary, the sequence of data generating processes is defined by $\mu, \sigma, \rho, \underline{\mathsf{G}}, \overline{\mathsf{G}}$ and h_n, δ_n.

5.2 The main result

We show that the asymptotic distribution of the LTS estimator is the same as it would have been, if it had been known which observations were 'good'. Here, we consider the case where the 'good' observations are normal and there are more 'good' observations than 'outliers'. The result does not require any regularity conditions for the common 'outlier' distribution G.

Theorem 5.1. *Consider the sequence of LTS location-scale models, where $h_n/n \to \lambda$ with $1/2 < \lambda < 1$. Suppose that ε_i for $i \in \zeta_n$ are i.i.d. $\mathsf{N}(0,1)$ distributed. Then, for any $\eta > 0$,*

$$\hat{\delta} = \delta_n + \mathsf{o}_\mathsf{P}(h_n^\eta), \quad h_n^{1/2}(\hat{\mu} - \mu)/\sigma \xrightarrow{\mathsf{D}} \mathsf{N}(0,1), \quad \hat{\sigma}^2 \xrightarrow{\mathsf{P}} \sigma^2.$$

We expect that Theorem 5.1 would generalize to the LTS regression Model 3.1. That is, the LTS regression estimator will have the same asymptotic theory as if we knew which observations were 'good'. The proof of such a result is, however, an open problem.

Theorem 5.1 for the LTS location-scale model differs from the known results for i.i.d. models. Butler (1982) proved a general result for the model with i.i.d. errors, showing that $n^{1/2}(\hat{\mu} - \mu)$ is asymptotically normal when the errors are symmetric with a strongly unimodal distribution. Further, the asymptotic variance involves an efficiency factor that differs from unity even in the normal case. He also gives an asymptotic theory for non-symmetric errors. Further, Víšek (2006) analyzed the regression estimator for i.i.d. errors. Johansen and Nielsen (2016b, Theorem 5) provide an asymptotic expansion of the scale estimator under normality.

A corresponding result for the LMS estimator is given in the supplement. The LMS estimator is maximum likelihood in a model where the 'good' observations are uniform. When there are no outliers, $n = h$, the LMS estimator is a Chebychev estimator. Knight (2017) provides asymptotic analysis of Chebychev estimators in regression models. Coinciding with that theory, Theorem C.2 shows that the LMS estimator is super-consistent at an n^{-1} rate in that model, which constrasts with the well-known $n^{-1/3}$ rate for i.i.d. models. However, Theorem C.2 does require regularity conditions for the 'outliers' to give sufficient separation from the 'good' observations.

5.3 Extentions of the main result

5.3.1 Non-normal errors

We start by relaxing the normality assumption. In the spirit of OLS asymptotics, we present sufficient conditions for the 'good' errors: 2+ moments, symmetry, and exponential tails.

Assumption 5.1. *Suppose ε_i for $i \in \zeta_n$ are i.i.d. with distribution F satisfying*
(i) $\mathsf{E}\varepsilon_i = 0$ and $\mathsf{E}\varepsilon_i^2 = 1$ for $i \in \zeta_n$;
(ii) $\mathsf{E}|\varepsilon_i|^{2+\omega} < \infty$ for $i \in \zeta_n$ and some $\omega > 0$;
(iii) F has infinite support: $\inf\{x : \mathsf{F}(x) > 0\} = -\infty$ and $\sup\{x : \mathsf{F}(x) < 1\} = \infty$;
(iv) $\varepsilon_{(\delta_n+1)}/\varepsilon_{(\delta_n+h_n)} + 1 = \mathsf{o}_\mathsf{P}(1)$;
(v) $\forall 0 < \eta < 1, \exists C_\eta < 1 : \varepsilon_{(\delta_n+h_n^{1-\eta})}/\varepsilon_{(\delta_n+1)}, \varepsilon_{(\delta_n+h_n-h_n^{1-\eta})}/\varepsilon_{(\delta_n+h_n)} \le C_\eta + \mathsf{o}_\mathsf{P}(1)$.

Example 5.1. *The normal distribution satisfies Assumption 5.1.*
For (iv) use that $a_n\{\varepsilon_{(\delta_n+h_n)} - b_n\}$ converges to a type I extreme value distribution when $a_n \sim b_n \sim (2\log h_n)^{1/2}$, see Leadbetter et al. (1982, Theorem 1.5.3). Here, $\sim$ denotes asymptotic equivalence. In particular, $\varepsilon_{(\delta_n+h_n)}/b_n \to 1$ in probability.
For (v) use Mill's ratio $\Phi(x) \sim -\varphi(x)/x$ for $x \to -\infty$. Take log to see that $\Phi^{-1}(1/s_n) \sim -(2\log s_n)^{1/2}$ for $s_n \to \infty$. We find $C_\eta \sim (2\log h_n^{\eta-1})^{1/2}/(2\log h_n^{-1})^{1/2} = (1-\eta)^{1/2} < 1$.

Example 5.2. *Assumption 5.1 is satisfied by the Laplace distribution and the double geometric distribution with probabilities $(1-p)^{|x|}p/(2-p)$ for $x \in \mathbb{Z}$. The latter is not in the domain of attraction of an extreme value distribution (Leadbetter et al., 1982, Theorem 1.7.13).*

Assumption 5.1(iv) is not satisfied by t_d distributions with $d \geq 1$ degrees of freedom. These are in the domain of attraction of extreme value distributions of type II. Gumbel and Keeney (1950) show that $\varepsilon_{(\delta_n+1)}/\varepsilon_{(\delta_n+h_n)}$ converges to a non-degenerate distribution with median 1.

Theorem 5.2. *Consider the sequence of LTS location-scale models. Let $1/2 < \lambda < 1$ and suppose Assumption 5.1. Then, the limiting results in Theorem 5.1 apply.*

5.3.2 More 'outliers' than 'good' observations

In this section, we allow for more 'outliers' than 'good' observations. Although in the traditional breakdown point analysis there has to be more 'good' observations than 'outliers' (Rousseeuw and Leroy, 1987, §3.4), in practice, LTS estimators are sometimes used with more 'outliers' than 'good' observations, as in the Forward Search algorithm (Atkinson et al., 2010) for instance. We show that, within the LTS model framework, it is possible to allow for more 'outliers' than 'good' observations, but in this case we need some regularity conditions on the 'outliers' to make sure the 'good' observations can be found.

Let the proportion of 'good' observations satisfy $0 < \lambda < 1$. Recall, that $\delta_n/(n - h_n) \to \rho = \mathsf{G}(0)$ *a.s.* is the proportion of the 'outliers' that are to the left. The number of 'outliers' to the right is $\overline{n} = n - h_n - \delta_n$. Define also the proportion of left and right 'outliers' relative to the number of 'good' observations through

$$\delta_n/h_n \overset{a.s.}{\to} \underline{\omega} = \rho(1-\lambda)/\lambda, \qquad \overline{n}/h_n \overset{a.s.}{\to} \overline{\omega} = (1-\rho)(1-\lambda)/\lambda.$$

Regularity conditions are needed for the 'outlier' distribution when $\underline{\omega} \geq 1$ or $\overline{\omega} \geq 1$. Note that $\underline{\omega} = \overline{\omega} < 1$ when the proportion of 'good' observations is $\lambda > 1/3$ and proportions of left and right 'outliers' are the same, so that $\rho = 1/2$. The following definition is convenient for analyzing the empirical distribution function of the 'outliers', evaluated at a random quantile.

Definition 5.1. *A distribution function H is said to be regular, if it is twice differentiable on an open interval $\mathcal{S} =]\underline{s}, \overline{s}[$ with $-\infty \leq \underline{s} < \overline{s} \leq \infty$ so that $\mathsf{H}(\underline{s}) = 0$ and $\mathsf{H}(\overline{s}) = 1$ and the density h and its derivative $\dot{\mathsf{h}}$ satisfy*
(a) $\sup_{x \in \mathcal{S}} \mathsf{h}(x) < \infty$ and $\sup_{x \in \mathcal{S}} \mathsf{H}(x)\{1 - \mathsf{H}(x)\}|\dot{\mathsf{h}}(x)|/\{\mathsf{h}(x)\}^2 < \infty$;
(b) If $\lim_{x \downarrow \underline{s}} \mathsf{h}(x) = 0$ then h is non-decreasing on an interval to the right of $\underline{s}$.
(c) If $\lim_{x \uparrow \overline{s}} \mathsf{h}(x) = 0$ then h is non-increasing on an interval to the left of $\overline{s}$.

Example 5.3. *The normal distribution is regular. Apply Mill's ratio to see this. The exponential distribution with $\mathsf{H}(x) = 1 - \exp(-x)$ is also regular with $\mathcal{S} = \mathbb{R}_+$ and with $\mathsf{H}(x)\{1 - \mathsf{H}(x)\}|\dot{\mathsf{h}}(x)|/\{\mathsf{h}(x)\}^2 = \mathsf{H}(x) \leq 1$ while $\mathsf{h}(x)$ is decreasing as $x \to \infty$.*

The following assumption requires that the 'outlier' distribution is regular and that the 'outliers' are more spread out than the 'good' observations.

Assumption 5.2. *Let $q > 4$. Let $\underline{\varepsilon}_1$ be $\underline{\mathsf{G}}$-distributed and $\overline{\varepsilon}_1$ be $\overline{\mathsf{G}}$-distributed, see (5.3).*
(i) If $\overline{\omega} \geq 1$ suppose $\overline{\mathsf{G}}$ is regular with $\int_0^\infty x^q d\overline{\mathsf{G}}(x) < \infty$ and
$$\overline{v} = \min_{\overline{\omega}^{-1} \leq \varsigma \leq 1} \mathsf{Var}\{\overline{\varepsilon}_1|\varsigma - \overline{\omega}^{-1} \leq \overline{\mathsf{G}}(\overline{\varepsilon}_1) \leq \varsigma\} > 1.$$
(ii) If $\underline{\omega} \geq 1$ suppose $\underline{\mathsf{G}}$ is regular with $\int_0^\infty x^q d\underline{\mathsf{G}}(x) < \infty$ and
$$\underline{v} = \min_{\underline{\omega}^{-1} \leq \varsigma \leq 1} \mathsf{Var}\{\underline{\varepsilon}_1|\varsigma - \underline{\omega}^{-1} \leq \underline{\mathsf{G}}(\underline{\varepsilon}_1) \leq \varsigma\} > 1.$$

Theorem 5.3. *Consider the sequence of LTS location-scale models. Let $0 < \lambda < 1$ and suppose Assumptions 5.1, 5.2. Then, the limiting results in Theorem 5.1 apply.*

Given the novelty (and perhaps at first surprising content) of Theorem 5.3, a discussion of this result and its relationship with Theorem 5.1 seems in order. Recall that Theorem 5.1 has no regularity conditions on the 'outliers'. The result exploits that there are more than 50% 'good' observations and the thin normal tails of these 'good' observations separate them sufficiently from the 'outliers'. In Theorem 5.3 we allow less than 50% 'good' observations but require regularity conditions on the outliers to ensure sufficient separation. More specifically, the proof of Theorem 5.3 covers two situations. First, if the proportions of left and right 'outliers' are equal and more than $1/3$ of the observations are 'good', then $\overline{\omega} = \underline{\omega} < 1$, so that Assumptions 5.2 is non-binding. Second, in general, Assumption 5.2 (i, ii) implies that the 'outliers' are spread out more than the 'good' observations.

This has a potential parallel with the breakdown point analysis of (Rousseeuw and Leroy, 1987, §3.4), which shows that for a given dataset the distribution of the LTS estimator is bounded with less than 50% contamination of an arbitrary type. It may still be possible to obtain the same result allowing for more than 50% contamination that satisfies regularity conditions of the type considered in Theorem 5.3, but we leave this as an open question.

5.4   The OLS estimator in the LTS model

We show that the least squares estimator $\bar{\mu} = n^{-1} \sum_{i=1}^{n} y_i$ can diverge in the sequence of LTS models. This implies that the least squares estimator is not robust within the LTS model in the sense of Hampel (1971). We assume all 'outliers' are to the right, so that $\rho = 0$.

Theorem 5.4. *Consider the sequence of LTS location-scale models. Let $0 < \lambda < 1$ and $\rho = 0$. Suppose ε_i for $i \in \zeta_n$ are i.i.d. with $\mathsf{E}\varepsilon_i = 0$, $\mathsf{Var}\varepsilon_i = 1$, and infinite support (Assumption 5.1,i,iii). Suppose $\overline{\mathsf{G}}$ has finite expectation $\mu_{\mathsf{G}} = \int_0^\infty \{1 - \overline{\mathsf{G}}(x)\}dx$. Then, $\bar{\mu}$ diverges. Noting that $\varepsilon_{(h_n)} \to \infty$ in probability, we have that*

$$|\varepsilon_{(h_n)}|^{-1}(\bar{\mu} - \mu)/\sigma \xrightarrow{\mathsf{P}} 1 - \lambda > 0.$$

6   Estimation of h

The LTS regression model takes the number of good observations, h, as given. In practice, an investigator has to choose h. We will review a commonly used method and discuss its asymptotic properties. We will also propose some new methods for consistently estimating the rate of outliers. The best of these methods is consistent at a $\log h$ rate.

6.1   The index plot method

A common method for estimating h is to apply a high breakdown point LTS estimator selecting approximately $n/2$ observations, then compute scaled residuals for all observations and keep those observations for which the scaled residuals are less than 2.5 in absolute value. This is conveniently done using an index plot (Rousseeuw and Leroy, 1987; Rousseeuw and Hubert, 1997). More specifically, scaled residuals $\hat{\varepsilon}_i \varsigma$ are plotted against the index i. Here, the consistency factor $\varsigma = 0.615$ is the conditional standard deviation of ε_i given that $1/4 < \Phi(\varepsilon_i) < 3/4$ (Croux and Rousseeuw, 1992). Bands on ± 2.5 are displayed on the plot and observations corresponding to scaled residuals $\hat{\varepsilon}_i \varsigma$ beyond these bands are declared outliers. Hence, the

estimator of h is $\hat{h}_{IP} = \sum_{i=1}^{n} 1_{(|\hat{\varepsilon}_i \varsigma| > 2.5)}$. We note that when scaling the residuals, one has to take a stand on the model for the data. Further, with the index plot method, the choice of consistency factor is based on the assumption that all errors are i.i.d. normal.

We can analyze the asymptotic properties of the index plot method using Theorems 4, 5, 7 in Johansen and Nielsen (2016b). This will show, that in a clean, normal sample, the index plot method, asymptotically, finds $\hat{\gamma}_{IP} = \hat{h}_{IP}/n \to_{\mathsf{P}} \gamma = \mathsf{P}(|\varepsilon_i| > 2.5) = 2\Phi(-2.5) = 1.24\%$ outliers. In the terminology of Hendry and Santos (2010); Castle et al. (2011), γ is the asymptotic *gauge* of the procedure. Thus, the index plot method will on average estimate a non-zero proportion of outliers, even in uncontaminated samples.

6.2   Methods based on the LTS model

We consider methods for consistent estimation of the proportion of 'good' observations in the LTS model. This is a semi-parametric model selection problem where the dimension of the parameter space increases with decreasing h, since the number of G_j functions increases when the number of 'outliers' increases. We consider three estimation methods. In all cases, suppose $h \geq \underline{h}$ where $\underline{h} > \dim x_i$ is a lower bound chosen by the user.

Maximum likelihood. We argue that $\hat{h} = \underline{h}$ is maximum likelihood estimator for h. Let $\mathsf{L}^\epsilon(\beta, \sigma, \zeta, \mathsf{G}_j, h)$ be the ϵ-likelihood given in (4.3) for each h. Let $\hat{\sigma}_h$ denote the maximum likelihood estimator in Theorem 4.1 for each h. By (4.6), the profile likelihood for h is

$$\hat{\mathsf{L}}^\epsilon_{\beta,\sigma,\zeta,\mathsf{G}}(h) = \epsilon^h \{(2\pi e \hat{\sigma}_h^2)^{-h/2} + \mathrm{o}(1)\}.$$

We note that $\hat{\sigma}_h^2 > 0$ for all $h \geq \underline{h}$, so that $\epsilon^{h-h} \hat{\mathsf{L}}_h^\epsilon / \hat{\mathsf{L}}_{\underline{h}}^\epsilon$ is bounded uniformly in h. Thus, $\hat{\mathsf{L}}_h^\epsilon / \hat{\mathsf{L}}_{\underline{h}}^\epsilon \to 0$ as $\epsilon \to 0$ whenever $h > \underline{h}$, so that $\hat{h} = \underline{h}$.

An information criteria. Penalizing $\log \hat{\sigma}_h^2$ gives the information criteria

$$IC_h = \log(\hat{\sigma}_h^2) + f(n)\{(n-h)/n\}, \tag{6.1}$$

for a penalty f. Let $\hat{h}_{IC}$ be the minimizer. Then $\hat{\lambda}_{IC} = \hat{h}_{IC}/n$ estimates the proportion of 'good' observations, $\lambda_\circ$ say. Below, we argue, for the location-scale case, that $\hat{\lambda}_{IC}$ is consistent if the penalty is chosen so that $f(n) \to \infty$ for increasing n, but $f(n) < \lambda_\circ^{-1} \log\log n$, where $0 < \lambda_\circ \leq 1$. If, for instance, it is expected that more than half of the observations are 'good', so that $\lambda_\circ > 1/2$, the penalty can be chosen as $f(n) = 2\log\log n$. The intuition is as follows.

Consider data generating processes as in §5.1 and let h_n° denote the number of 'good' observations, so that $h_n^\circ / n \to \lambda_\circ$. We want to show that $\hat{\lambda}_{IC} = \lambda_\circ$ in the limit. Theorem 5.1 shows that $\hat{\sigma}_{h_n^\circ}^2$ is consistent for σ^2 when $h = h_n^\circ$. Thus, we consider sequences h_n so that $h_n/n \to \lambda \neq \lambda_\circ$ and argue that $IC_{h_n} - IC_{h_n^\circ} \sim \log(\hat{\sigma}_{h_n}^2/\sigma^2) + f(n)(\lambda_\circ - \lambda)$ has a positive limit, so that the minimizer $\hat{\lambda}_{IC} = \lambda_\circ$ in the limit.

Suppose $\lambda < \lambda_\circ$. In that case, one could expect that, asymptotically, the estimated set of 'good' observations is a subset of the good observations, so that $\hat{\zeta}_h \subset \zeta$. Further, $\hat{\sigma}_h^2$ converges to σ_λ^2, say, the variance of a truncated normal distribution as analyzed by Butler (1982). Then, $IC_{h_n} - IC_{h_n^\circ} \sim \log(\sigma_\lambda^2/\sigma^2) + f(n)(\lambda_\circ - \lambda)$, which is positive in the limit when $f(n)$ diverges. Thus, $\hat{\lambda}_{IC} \geq \lambda_\circ$ in the limit.

Suppose $\lambda > \lambda_\circ$. Then $IC_{h_n} - IC_{h_n^\circ} \sim \log(\hat{\sigma}_{h_n}^2/\sigma^2) - f(n)(\lambda - \lambda_\circ)$, where the penalty terms is negative, so that $f(n)$ must be chosen carefully. With the normal LTS model, $\hat{\sigma}_{h_n}^2$ must diverge. The reason is that, for $\lambda > \lambda_\circ$, the estimated set of 'good' observations $\hat{\zeta}_h$ includes both 'good' observations and 'outliers'. The 'outliers' diverge at a $(2\log h_n^\circ)^{1/2}$ rate, see Example 5.1, so that $\hat{\sigma}_{h_n}^2$ must diverge at a $2\log h_n^\circ$ rate. Noting that $h_n^\circ \sim n\lambda_\circ$, we get

$IC_{h_n} - IC_{h_n^\circ} \sim \log \log n - f(n)(\lambda - \lambda_\circ)$. When $1/2 < \lambda_\circ < \lambda \le 1$, we have $\lambda - \lambda_\circ < 1/2$, and the $\log \log n$ term dominates when $f(n) \le 2 \log \log n$. Thus, $\hat\lambda_{IC} \le \lambda_\circ$ in the limit.

Combining these arguments indicates that $\hat\lambda_{IC} = \lambda_\circ$ in the limit. Unfortunately, in simulations that are not reported here, we find that a $2 \log \log n$ penalty grows so slowly in n that for some specifications the consistency is only realized for extremely large samples.

Cumulant based normality test. A more useful estimator for the proportion of 'good' observations λ is the minimizer $\hat h_T$, say, of the normality test statistic

$$T_h = h\hat\kappa_{3,h}^2/6 + h\hat\kappa_{4,h}^2/24, \tag{6.2}$$

based on third and fourth cumulants of the estimated 'good' residuals. We argue that $\hat h_T/n$ is consistent for $\lambda_\circ$. This is supported by a simulation study in §7.2.

The intuition of the consistency argument is similar to that for the information criteria. First, for $h = h_n^\circ$, Theorem 5.1 may be extended to show that T_{h_0} is asymptotically χ_2^2. For $h < h_n^\circ$, the sample moments may converge to the moments of a truncated normal distribution, see Berenguer-Rico and Nielsen (2017). Thus, T_h would diverge as it is normalized using the normal distribution. For $h > h_n^\circ$, the estimated set of 'good' observations $\hat\zeta_h$ contains both 'good' and 'outlying' observations, so that T_h diverges at a logarithmic rate, instead of the iterated logarithmic rate for the information criteria. A formal proof is left for future work.

7 Simulations

7.1 Inference in the location-scale model

In the location-scale model $y_i = \mu + \sigma\varepsilon_i$, we study the finite sample properties of tests for $\mu = 0$ using four different tests statistics and six different data generating processes, which we describe below. We consider sample sizes $n = 25, 100, 400, 1600$ and let $h/n = \lambda = 0.8$.

Tests. We consider four different estimators: LTS, LMS, OLS and SLTS (scale-corrected LTS estimator). For each estimator, s say, we compute t-type statistics, $t_s = (\hat\mu_s - \mu)/\mathsf{se}_s$, with associated asymptotic 95% quantiles q_s.

For the OLS estimator, we have $\hat\mu_{OLS} = n^{-1}\sum_{i=1}^n y_i$ and $\mathsf{se}_{OLS}^2 = \hat\sigma_{OLS}^2/n$ with $\hat\sigma_{OLS}^2 = n^{-1}\sum_{i=1}^n (y_i - \bar y)^2$. The asymptotic quantile q_{OLS} is standard normal.

For the LTS estimator, we apply the estimators $\hat\mu_{LTS}$ and $\hat\sigma_{LTS}$ given in (5.1). By Theorem 5.2 we get that $\mathsf{se}_{LTS}^2 = \hat\sigma_{LTS}^2/h$. The asymptotic quantile q_{LTS} is standard normal.

For the LMS estimator, we apply the estimators $\hat\mu_{LMS}$ and $\hat\sigma_{LMS}$ given in (C.4). Theorem C.2 gives that $\mathsf{se}_{LMS}^2 = \hat\sigma_{LMS}^2/h^2$. The asymptotic quantile q_{LMS} is standard Laplace.

Finally, SLTS uses the usual approach to LTS estimation with consistency and efficiency correction factors arising from truncation in a standard normal distribution as outlined in the end of §2. Let $\varsigma_{h/n}^2 = \int_{-c}^c x^2\varphi(x)dx / \int_{-c}^c \varphi(x)dx$ with c chosen so that $\int_{-c}^c \varphi(x)dx = h/n$, which gives $\varsigma_{0.8}^2 = 0.438$. Then, we have estimators $\hat\mu_{SLTS} = \hat\mu_{LTS}$ and $\hat\sigma_{SLTS}^2 = \hat\sigma_{LTS}^2/\varsigma_{h/n}^2$, while $\mathsf{se}_{SLTS}^2 = \hat\sigma_{SLTS}^2/(n\varsigma_{h/n}^2)$. The asymptotic quantile q_{SLTS} is standard normal.

Data Generating Processes (DGPs). Table 7.1 gives an overview of the DGPs. The first three DGPs are examples of the LTS Model 3.1 and the LMS Model C.1. The proportion of 'good' observations is $\lambda = 80\%$. The 'good' errors ε_i are i.i.d. normal $\mathsf{N}(0,1)$ in DGP1 and DGP2 and i.i.d. uniform $\mathsf{U}[-1,1]$ in DGP3. The 'outlier' errors are defined as in (3.1), so that $\varepsilon_j = (\max_{i\in\zeta}\varepsilon_i + \xi_j)1_{(\xi_j>0)} + (\min_{i\in\zeta}\varepsilon_i + \xi_j)1_{(\xi_j<0)}$, where $\xi_j - \nu^+1_{(\xi_j>0)} + \nu^-1_{(\xi_j<0)}$ are i.i.d. normal $\mathsf{N}(0,1)$ and ν^+ and ν^- separate 'good' and 'outlying' observations. The separators are $\nu^+ = \nu^- = 0$ in DGP1 and $\nu^+ = 3, \nu^- = 1$ in DGP2 and DGP3.

	Model	MLE	'good' error	'outliers'
DGP1	LTS	LTS	$\mathsf{N}(0,1)$	$\nu^+ = \nu^- = 0$
DGP2	LTS	LTS	$\mathsf{N}(0,1)$	$\nu^+ = 3, \nu^- = 1$
DGP3	LMS	LMS	$\mathsf{U}(-1,1)$	$\nu^+ = 3, \nu^- = 1$
DGP4	ϵ-contamination	OLS	$\mathsf{N}(0,1)$	$\mathsf{N}(0,1)$
DGP5	ϵ-contamination	—	$\mathsf{N}(0,1)$	$\mathsf{N}(0,3)$
DGP6	ϵ-contamination	—	$\mathsf{N}(0,1)$	$\mathsf{N}(2,1)$

Table 7.1: Data generating processes. The proportion of good errors is $\lambda = 0.8$. Columns 2 and 3 show the model and indicate which estimator is maximum likelihood. Columns 4 and 5 indicate how the 'good' and the 'outlying' errors are chosen.
For DGP1 – DGP3: λn 'good' observations, 'outliers' have $\xi_j - \nu^+ 1_{(\xi_j > 0)} + \nu^- 1_{(\xi_j < 0)}$ is $\mathsf{N}(0,1)$.
For DGP4 – DGP6: distribution is $\lambda \mathsf{N}(0,1) + (1-\lambda)\mathsf{H}$.

	OLS						LTS					
n	DGP1	2	3	4	5	6	DGP1	2	3	4	5	6
25	0.092	0.084	0.084	0.067	0.066	0.359	0.255	0.081	0.072	0.371	0.337	0.388
100	0.083	0.129	0.199	0.054	0.054	0.887	0.180	0.058	0.055	0.383	0.345	0.518
400	0.100	0.321	0.664	0.051	0.051	1.000	0.110	0.052	0.051	0.389	0.349	0.827
1600	0.159	0.745	0.998	0.050	0.050	1.000	0.071	0.050	0.050	0.390	0.349	0.998
	LMS						SLTS					
n	DGP1	2	3	4	5	6	DGP1	2	3	4	5	6
25	0.720	0.489	0.070	0.641	0.631	0.702	0.011	0.000	0.000	0.034	0.027	0.063
100	0.961	0.785	0.054	0.836	0.831	0.901	0.002	0.000	0.000	0.041	0.028	0.116
400	0.999	0.936	0.051	0.931	0.929	0.982	0.000	0.000	0.000	0.047	0.031	0.373
1600	1.000	0.992	0.050	0.972	0.971	0.999	0.000	0.000	0.000	0.049	0.032	0.941

Table 7.2: Simulated rejection frequencies for nominal 5% tests on intercept.

The last three DGPs are examples of ϵ-contamination (Huber, 1964). We draw n observations from the distribution function $0.8\Phi + 0.2\mathsf{H}$, where Φ is standard normal and H represents contamination. In DGP4, $\mathsf{H} = \Phi$, giving a standard i.i.d. normal model. In DGP5 and DGP6, H is $\mathsf{N}(0,3)$ and $\mathsf{N}(2,1)$, giving symmetric and non-symmetric mixtures, respectively.

We have different maximum likelihood estimators for the different models. These are LTS for DGP1 and DGP2, LMS for DGP3, and OLS for DGP4. None of the considered estimators are maximum likelihood for DGP5 and DGP6.

Table 7.2 reports results from 10^6 repetitions. The Monte Carlo standard error is 0.001.

The OLS statistic is maximum likelihood with DGP4 and performs well. It performs equally well with the symmetric, i.i.d. DGP5. For DGP1 it is slowly diverging, possibly because the absolute sample mean is diverging with n. OLS performs poorly with the non-symmetric DGP2, DGP3 and DGP6.

The LTS statistic is maximum likelihood with DGP1 and DGP2. The asymptotic theory also applies for DGP3. The convergence is slow for DGP1, where there is no separation. The LTS statistic does not perform well with ϵ-contamination in DGP4, DGP5 and DGP6.

The LMS statistic is maximum likelihood with DGP3 and perform well with that DGP, but poorly with all other DGPs. The SLTS statistic is not maximum likelihood for any of the considered models. It is calibrated to be asymptotically unbiased for DGP4 and performs well for that model, but poorly with all other DGPs.

Overall, we see that it is a good idea to apply maximum likelihood but this does require that

| | DGP1 | | | | | | DGP2 | | | | | |
| | $\lambda = 0.7$ | | | $\lambda = 0.8$ | | | $\lambda = 0.7$ | | | $\lambda = 0.8$ | | |
n	$\hat{h}_{bias}$	$\hat{h}_{sd}$	$\hat{h}_r$	$\hat{h}_{bias}$	$\hat{h}_{sd}$	$\hat{h}_r$	$\hat{h}_{bias}$	$\hat{h}_{sd}$	$\hat{h}_r$	$\hat{h}_{bias}$	$\hat{h}_{sd}$	$\hat{h}_r$
25	1.2	2.5		-1.0	2.5		0.3	2.1		-1.5	2.1	
100	9.6	8.6		3.7	6.0		-0.2	4.3		-1.1	3.0	
400	41.3	35.8		4.5	12.4		-0.8	2.1	✓	-0.8	2.0	✓
1600	92.8	143.1		-0.4	3.5	✓	-0.8	2.0	✓	-0.8	2.0	✓
12800	-0.7	4.2	✓	-0.6	4.4	✓	-1.2	3.2	✓	-1.3	3.2	✓

Table 7.3: Estimating h by minimizing the T_h statistic in (6.2). Here, $\hat{h}_{bias}$ and $\hat{h}_{sd}$ report the simulated bias and standard deviation for $\hat{h}$, while $\hat{h}_r$ is ticked if all simulated values of $\hat{h}$ are interior to the range from 60% to 90%.

the model specification is checked. In particular, the LTS estimator is best in DGP1–DGP3, although with some finite sample distortion with DGP1 where 'good' and 'outlying' observations are not well-separated. The LTS estimator does not work well for the ϵ-contaminated models, where a model dependent scale correction is needed. The usual approach of using the normal scale correction as in SLTS does not work well in general. All estimators are poor for asymmetric ϵ-contamination.

7.2 h estimation

Next, we study the finite sample properties of estimating h using the cumulant based normality test statistic T_h in (6.2). Results are reported in Table 7.3, based on 10^3 repetitions.

In each repetition, we compute the T_h statistic for each h in the range from 60% to 90% of n. The estimator of h is the minimizer of T_h over that range.

The data generating processes are DGP1 and DGP2 from above. These are examples of the LTS model, so that DGP1 has symmetric 'outliers' that are not separated from the 'good' observations, while DGP2 has asymmetric 'outliers' that are separated from the 'good' observations. For each DGP, we consider cases with 70% or 80% 'good' observations.

Table 7.3 reports three quantities for each of the DGPs. First, $\hat{h}_{bias}$ and $\hat{h}_{sd}$ are the Monte Carlo average and standard deviation of the estimation error $\hat{h} - h$. Further, $\hat{h}_r$ is a binary variable, which is checked if $\hat{h}$ is in the interior of the range from 60% to 90% of n for all 10^3 simulations. The theory suggests that the proportion h/n of 'good' observations is consistently estimated, whereas h is not consistently estimated. Thus, we would expect $\hat{h}_{bias}$ and $\hat{h}_{sd}$ to grow slower than linearly in n, but not to vanish.

For all the four setups, the simulations confirm that $\hat{h}/n$ is consistent for λ. In DGP1, where there is only little separation between 'good' and 'outlying' observations, the performance differs substantially between the cases $\lambda = 0.7$ and $\lambda = 0.8$. We do not have an explanation for this difference. Nonetheless, the estimation works much better for DGP2, with its separation between 'good' observations and 'outliers', in both cases $\lambda = 0.7$ and $\lambda = 0.8$, matching the theoretical discussion in the above sections.

We also considered the information criteria Φ_h in (6.1), but omit the results. The performance of Φ_h is poor, quite possibly due to the $\log \log n$ rates involved.
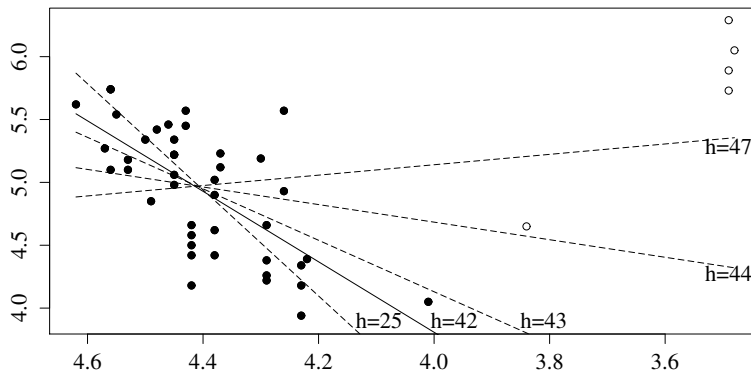
8 Empirical illustrations

Figure 8.1: Star data and fit by LTS with different values of h. Log light intensity against log temperature. Solid dots are estimated 'good' observations for $h = 42$.

We provide two empirical illustrations using the stars data of Rousseeuw and Leroy (1987) and the state infant mortality data of Wooldridge (2015). Both analyses illustrate the estimation of h. The second case also illustrates inference in the LTS model. Therefore, we must determine empirically if the LTS model is appropriate. Indeed, there could be a contamination pattern that is consistent with the ϵ-contamination, with the LTS model, or with neither of those. For both illustrations, we study the source of the data to arrive at reasonable empirical models.

Throughout, we use R version 4.0.2 (R Core Team, 2020) estimating LTS using `ltsReg` from the `Robustbase` package. Before each LTS call we apply `set.seed(0)`.

8.1 The stars data

For this empirical illustration, we consider the data on log light intensity and log temperature for the Hertzsprung-Russell diagram of the star cluster CYG OB1 containing $n = 47$ stars as reported by Rousseeuw and Leroy (1987, Table 2.3). Figure 8.1 shows a cross plot of the variables, where the log temperature axis is reversed. The majority of observations follow a steep band called the main sequence. Rousseeuw and Leroy (1987) refer to the four stars to the top right of Figure 8.1 as outliers.

By consulting the original source of the data in Humphreys (1978), see also (Vansina and De Grève, 1982, Appendix A), we found that the 4 stars to the right of Figure 8.1 are of M-type (observations 11, 20, 30, 34) and they are red supergiants. Further, the fifth star from the right is of F-type (observation 7, called 44 Cyg). The next 31 stars (1 doublet) from the right are of B-type and the remaining 11 stars (1 doublet) furthest to the left are of the O-type. The doublets are not exact doublets in Humphreys' original data, so this in itself should not be seen as evidence against a normality assumption.

We fitted the linear model *log light* $= \beta_1 + \beta_2$ *log temperature* $+ \sigma \varepsilon_i$. Table 8.1, left panel, shows LTS estimates for different h values. For $h = n = 47$ the LTS estimator is the full sample OLS estimator. The β estimates are the 'raw coefficients' found by `ltsReg` while $\hat{\sigma}$ is computed directly from (2.3) without any consistency correction. Figure 8.1 shows LTS fits for selected values of h. It is seen that the fits rotate when h increases. Table 8.1 also reports the T_h criterion as a function of h. It is minimized for $h = 42$ pointing at five 'outliers': The four M-stars and the F-star. This is supported by a traditional index plot, see §6.1. Figure 8.1 indicates estimated 'good' observations and 'outliers' with solid and open dots.

The estimation of h by the statistic T_h appears a little shaky in Table 8.1, left panel. The lowest value is obtained for $h = 42$, while the value for $h = 44$ is nearly as low. Table 8.1, right

	Full sample				Sub sample			
h	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}$	T_h	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}$	T_h
25	-13.62	4.22	0.18	1.72	-13.62	4.22	0.18	1.72
36	-11.49	3.71	0.27	1.98	-11.49	3.71	0.27	1.98
37	-9.00	3.16	0.28	2.49	-9.00	3.16	0.28	2.49
40	-8.58	3.07	0.31	2.13	-8.58	3.07	0.31	2.13
41	-8.50	3.05	0.33	1.26	-8.50	3.05	0.33	1.26
42	-7.40	2.80	0.37	0.39	-7.40	2.80	0.37	0.39
43	-4.06	2.05	0.40	0.69	7.88	-0.65	0.49	2.57
44	1.89	0.70	0.49	0.49	7.74	-0.62	0.51	2.76
45	7.34	-0.53	0.51	2.94	7.58	-0.59	0.53	2.73
46	6.92	-0.44	0.53	2.74	7.12	-0.49	0.55	2.83
47	6.79	-0.41	0.55	2.75				

Table 8.1: Estimates by LTS and T_h criterion for selecting h. Left panel has full sample. Right panel excludes the F-type star.
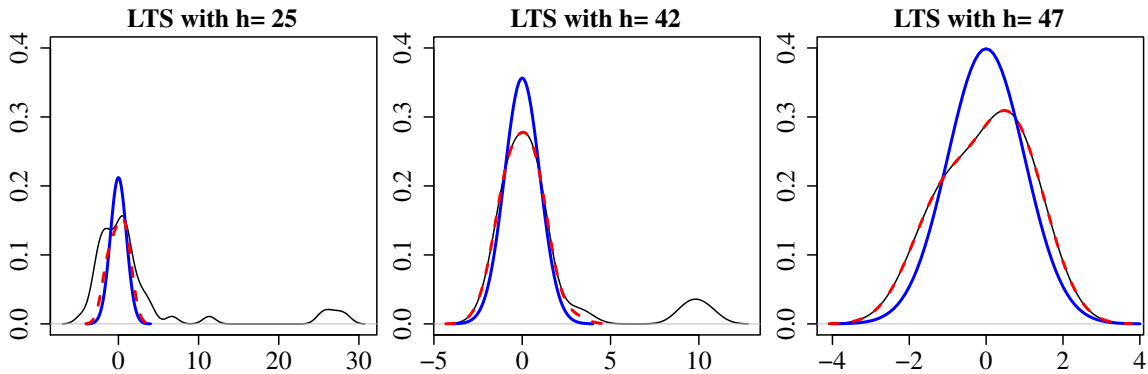

Figure 8.2: Scaled LTS residuals for $h = 25, 42, 47$. Kernel densities for all residuals (black, thin) and for 'good' residuals (red dashed) with fitted standard normal density (blue, thick).

panel, presents results corresponding to those in the left panel when dropping the F star from the sample. The results are the same for $h \leq 42$ as for the full sample. We see that $h = 42$ is now a clear minimum identifying the four M-type stars as 'outliers'. The M stars appear to have a masking effect, where after their deletion, the F star emerges as very *influential* in the sense of Rousseeuw and Leroy (1987, p. 81).

Figure 8.2 shows kernel density plots for the scaled residuals for $h = 25, 42, 47$. The black, thin lines gives kernel densities for the full sample. The red, dashed lines gives kernel densities for the estimated 'good' observations. The standard normal distribution is shown with a blue, thick line. For $h = 42$, the red, blue and part of the black lines coincide, which indicates the normality of the 'good' observations. The full sample kernel density has a probability mass in the right tail corresponding to the four giants. There is a slight discrepancy between the full sample and the 'good' kernel densities in the region from 2 to 4 corresponding to the F star. By construction this 43rd residual will be outside the range of the 42 'good' residuals, but not by far. Kernel densities are very sensitive to the choice of kernel and bandwidth. For illustration we chose a Gaussian kernel and bandwidth $1.5h^{-1/5}$ to get the best match of the red and blue curves for $h = 42$. With that choice, we can more clearly see discrepancies between the kernel density for the 'good' observations and the normal curve for $h = 25, 47$, that is, for the LTS
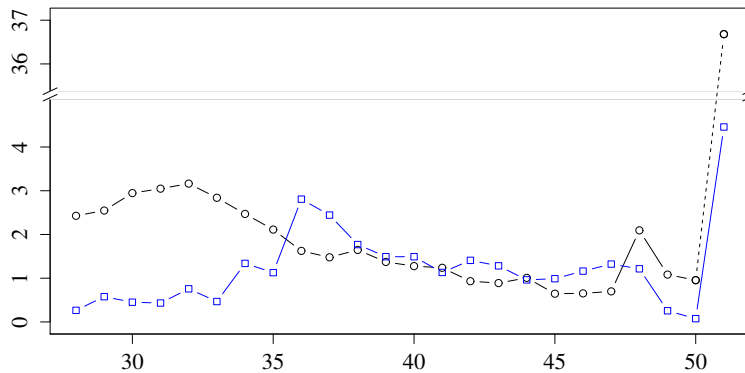
Figure 8.3: State infant mortality rates. T_h criterion as function of h.
Model for *infmrt* shown with (black) circles. Model for log*infmrt* shown with (blue) squares.

with a high breakdown point and the OLS estimator, respectively.

8.2 State infant mortality rates

We consider 1990 data for the the United States on infant mortality rates by state, including Washington DC, which has a particularly high infant mortality rate (Wooldridge, 2015, p. 299). The data are available as `infmrt` in the R package `wooldridge` and are taken from U.S. Bureau of the Census (1994). We analyze two models.

The first model follows Wooldridge. It is a linear regression of the number of deaths within the first year per 1,000 live births, *infmort*, on the log of per capita income, *lpcinc*, the log of physicians per 100,000 members of the civilian population, *lphysic*, and the log of population in thousands, *lpopul*. In Figure 8.3, the graph using circle symbols shows the cumulant based criteria T_h as function of h. The OLS regression is obtained for $h = n = 51$ and has a rather large value of T_h - notice the gap in the T_h-axis - indicating that the full sample model is mis-specified. Choosing $h = 50$ would lead to one 'outlier', which is Washington DC. However, the T_h function is minimized at $h = 45$ indicating six 'outliers': Delaware, Washington DC, Georgia, Texas, California and Alaska (U.S. Bureau of the Census, 1994, Table 123). The interpretation is not obvious and could be due to a missing regressor.

The second model differs in two respects. It applies logarithms to the regressand to stabilise rates close to zero as well as for Washington DC. It also includes a regressor for the log proportion of black people in the population, *lblack* (U.S. Bureau of the Census, 1992, Table 255), since infant mortality is quite different for white and black infants in most states (U.S. Bureau of the Census, 1994, Table 123). In Figure 8.3, the graph using square symbols shows T_h versus h for this model. The minimizer is $h = 50$ with Washington DC as the 'outlier'. The minimum of 0.08 is small compared to a χ^2_2 distribution, so no evidence against the LTS model. We note that the T_h function is also quite low for h in the left side of the plot, albeit not as low as for $h = 50$. The estimated LTS model for $h = 50$ is

$$\log \widehat{infmrt} \underset{(s.e.)}{=} \underset{(0.98)}{4.91} - \underset{(0.128)}{0.104} \, lpcinc - \underset{(0.093)}{0.251} \, lphysic - \underset{(0.019)}{0.012} \, lpopul + \underset{(0.014)}{0.093} \, lblack,$$

$$\hat{\sigma} = 0.0973. \tag{8.1}$$

The standard errors are the usual OLS standard errors as the LTS model appears to apply. We conclude that the variables *lpcinc* and *lpopul* are not significant.

Changing the regressand to be measured on the original scale introduces a second 'outlier', South Dakota, which has one of the lowest infant mortality rates for the black population.

17

9   Discussion

Other models of the LTS type. New models and estimators can be generated by replacing the normal assumption in the LTS models with some other distribution. For instance, the uniform leads to the Least Median of Squares (LMS) estimator analyzed in Appendix C, while the Laplace distribution leads to the Least Trimmed sum of Absolute deviations (LTA) estimator (Hawkins and Olive, 1999; Hössjer, 1994; Dodge and Jurečková, 2000, §2.7). Ron Butler has suggested to us that the approach in this paper could also be applied to the Minimum Covariance Determinant (MCD) estimator for multivariate location and scale (Rousseeuw, 1985; Butler et al., 1993).

Inference requires a model for both 'good' and 'outlying' observations. In the presented theory, the 'good' and the 'outlying' observations are separated. The traditional approach, as advocated by Huber (1964), is to consider mixture distributions formed by mixing a reference distribution with a contamination distribution. Any subsequent inference on the regression parameter β would require a specific formulation of the ϵ-contaminated distribution. This is a non-trivial practical problem. Instead, there has been a focus on showing that the bias of estimators is bounded under contamination (Huber and Ronchetti, 2009, §4) while inference is conducted using the asymptotic distribution that assumes all observations are i.i.d. normal, implying that the 'good' observations are truncated normal. This gives a different distribution theory for inference compared to the one presented here and simulations in §7.1 indicate that this will not control size in general. It is therefore of interest to formulate models allowing 'outliers' under which consistency can be proved. The LTS model in this paper does so. The presented simulations show that the two inferential theories are really very different. In practice, LTS estimation should therefore be evaluated in the context of a particular model and inference should be conducted accordingly.

Estimation of h. The suggested procedure for estimation of h seems to work well in the two empirical illustrations. It helps in finding a satisfactory LTS model for the data. It is designed to be consistent within the framework of LTS models where, once 'outliers' have been removed, the remaining 'good' observations look normal. We suspect that it will not be so helpful in situations where the data do not adhere to the LTS model, like ϵ-contamination models where the error follow some non-normal, but regular distribution.

Alternative estimators of h. We have proposed consistent estimators for h/n, but it would be useful to investigate their performance further. In a regression context, it may be worth considering the Forward Search algorithm (Atkinson et al., 2010). Omitted regressors and 'outliers' may confound each other, so a simultaneous search over these may be useful as in the Autometrics algorithm (Hendry and Doornik, 2014; Castle et al., 2020). Some asymptotic theory for these algorithms are provided in Johansen and Nielsen (2016a,b).

Misspecification tests can be developed for the present model. The asymptotic theory developed here shows that standard normality tests can be applied to the set of estimated 'good' observations. Other tests could also be investigated, in particular those that are concerned with functional form or omission of regressors.

More 'outliers' than 'good' observations. This is allowed in the LTS model and supported by the asymptotic analysis under regularity conditions for the 'outliers'. This lends some support to the practice of starting the Forward Search algorithm by an LTS estimator for fewer than half of the observations (Atkinson et al., 2010). Whether it makes more sense to model the 'outliers' or the 'good' observations as normally distributed in this situation must rest on a careful consideration of the data and the substantive context.

References

Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Stat.*, 7:226–248.

Atkinson, A. C., Riani, M., and Cerioli, A. (2010). The forward search: Theory and data analysis (with discussion). *J. Korean Statist. Soc.*, 39:117–163.

Berenguer-Rico, V. and Nielsen, B. (2017). Marked and weighted empirical processes of residuals with applications to robust regressions. Discussion Paper 841, Dept. Econ., Oxford.

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley & Sons, New York.

Butler, R. (1982). Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule. *Ann. Statist.*, 10:197–204.

Butler, R., Davies, P. L., and Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *Ann. Statist.*, 21:1385–1400.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics*, 3. Article 8.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2020). Robust discovery of regression models. Discussion paper, Nuffield College.

Čížek, P. (2005). Least trimmed squares in nonlinear regression under dependence. *J. Statist. Plann. Inference*, 136:3967–3988.

Croux, C. and Rousseeuw, P. J. (1992). A class of high-breakdown scale estimators based on subranges. *Comm. Statist. Theory Methods*, 21:1935–1951.

Csörgő, M. (1983). *Quantile Processes with Statistical Applications*, volume 42 of *CBMS-NFS Regional Conference Series in Applied Mathematics*. SIAM.

Dodge, Y. and Jurečková, J. (2000). *Adaptive Regression*. Springer, New York.

Doornik, J. A. (2016). An example of instability: Discussion of the paper by Søren Johansen and Bent Nielsen. *Scand. J. Stat.*, 43:357–359.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. A*, 222:309–368.

Gissibl, N., Klüppelberg, C., and Lauritzen, S. (2020). Identifiability and estimation of recursive max-linear models. *Scand. J. Stat.* To appear.

Gumbel, E. J. and Keeney, R. D. (1950). The extremal quotient. *Ann. Math. Statist.*, 21:523–538.

Hald, A. (2007). *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*. Springer, New York.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.*, 42:1887–1896.

Hawkins, D. M. and Olive, D. J. (1999). Improved feasible solution algorithms for high break down estimation. *Comput. Statist. Data Anal.*, 30:1–11.

Hendry, D. F. and Doornik, J. A. (2014). *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*. MIT Press, London.

Hendry, D. F. and Santos, C. (2010). An automatic test of super exogeneity. In Watson, M. V., Bollerslev, T., and Russell, J., editors, *Volatility and Time Series Econometrics*. Oxford University Press, Oxford.

Hössjer, O. (1994). Rank-based estimates in the linear model with high breakdown point. *J. Amer. Statist. Assoc.*, 89:149–158.

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

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Humphreys, R. M. (1978). Studies of luminous stars in nearby galaxies. I. Supergiants and O stars in the Milky Way. *Astrophysical Journal Supplement Series*, 38:309–350.

Johansen, S. (1978). The product limit estimator as maximum likelihood estimator. *Scand. J. Stat.*, 5:195–199.

Johansen, S. and Nielsen, B. (2016a). Analysis of the forward search using some new results for martingales and empirical processes. *Bernoulli*, 22:1131–1183. Corrigendum (2019) 25, 3201.

Johansen, S. and Nielsen, B. (2016b). Asymptotic theory of outlier detection algorithms for linear time series regression models (with discussion). *Scand. J. Stat.*, 43:321–81.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27:887–906.

Knight, K. (2017). On the asymptotic distribution of the l_∞ estimator in linear regression. Mimeo.

Leadbetter, M. R., Lindgreen, G., and Rootzén, H. (1982). *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Riani, M., Atkinson, A. C., and Perrotta, D. (2014). A parametric framework for the comparison of methods of very robust regression. *Statist. Sci.*, 29:128–143.

Rousseeuw, P. (1985). Least median of squares regressions. In Grossmann, W., Pflug, G., Vincze, I., and Wertz, W., editors, *Mathematical Statistics and Applications*, pages 283–297. Reidel, Dordrecht.

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

Rousseeuw, P. J. and Hubert, M. (1997). Recent developments in PROGRESS. In Dodge, Y., editor, L_1-*Statistical Procedures and Related Topics*, volume 31 of *Lecture Notes–Monograph Series*, pages 201–214. Institute of Mathematical Statistics, Beachwood, OH.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, Hoboken, NJ.

Rousseeuw, P. J., Perrotta, D., Riani, M., and Hubert, M. (2019). Robust monitoring of time series with application to fraud detection. *Econometrics and Statistics*, 9:108–121.

Rousseeuw, P. J. and van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps. In Gaul, W., Opitz, O., and Schader, M., editors, *Data Analysis: Scientific Modeling and Practical Application*, pages 335–346. Springer Verlag.

Scholz, F. W. (1980). Towards a unified definition of maximum likelihood. *Canad. J. Statist.*, 8:193–203.

U.S. Bureau of the Census (1992). *1990 Census of Population: General Population Statistics, CP-1-1*. Washington DC.

U.S. Bureau of the Census (1994). *Statistical Abstract of the United States: 1994*. Washington DC.

Vansina, F. and De Grève, J. P. (1982). Close binary systems before and after mass transfer. III. Spectroscopic binaries. *Astrophysics and Space Science*, 87:377–401.

Víšek, J. A. (2006). The least trimmed squares; part III: Asymptotic normality. *Kybernetika*, 42:203–224.

Wooldridge, J. M. (2015). *Introductory Econometrics*. Cengage Learning, Boston, MA, 6 edition.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Math. Statist.*, 15:642–656.

A Derivation of the LTS likelihood

We prove that $\mathsf{P}(z_j - \epsilon < y_j \leq z_j | y_i$ for $i \in \zeta) = \Delta^\epsilon \mathsf{G}_j(\tilde{z}_j^{\beta\sigma}) = \mathsf{G}_j(\tilde{z}_j^{\beta\sigma\epsilon}) - \mathsf{G}_j(\tilde{z}_j^{\beta\sigma})$.

First, we prove that $\mathsf{P}_0 = \mathsf{P}(y_j \leq z_j | \cdot) = \mathsf{G}(\tilde{z}_j^{\beta\sigma})$, where the dot indicates the conditioning set. Since $\varepsilon_j = (y_j - \beta' x_j)/\sigma$ and $z_j^{\beta\sigma} = (z_j - \beta' x_j)/\sigma$, we have that $\mathsf{P}_0 = \mathsf{P}(\varepsilon_j \leq z_j^{\beta\sigma} | \cdot)$.

Recall that the 'outlier' errors are $\varepsilon_j = (\max_{i \in \zeta} y_i^{\beta\sigma} + \xi_j) 1_{(\xi_j > 0)} + (\min_{i \in \zeta} y_i^{\beta\sigma} + \xi_j) 1_{(\xi_j < 0)}$, see (3.1) and since $y_i^{\beta\sigma} = \varepsilon_i$. Further, ξ_j has distribution function G_j, which is continuous at zero. As a consequence, $\mathsf{P}(\min_{i \in \zeta} y_i^{\beta\sigma} \leq \varepsilon_j \leq \max_{i \in \zeta} y_i^{\beta\sigma} | \cdot) = 0$ and $\mathsf{P}(\varepsilon_j \leq \min_{i \in \zeta} y_i^{\beta\sigma}) = \mathsf{P}(\xi_j < 0) = \mathsf{G}_j(0)$. Thus, $\mathsf{P}_0 = \mathsf{P}(\varepsilon_j \leq z_j^{\beta\sigma} | \cdot) = \mathsf{G}_j(0)$ for $\min_{i \in \zeta} y_i^{\beta\sigma} \leq z_j^{\beta\sigma} \leq \max_{i \in \zeta} y_i^{\beta\sigma}$.

Let $\mathsf{P}_0 = \mathsf{P}_1 + \mathsf{P}_2$, where $\mathsf{P}_1 = \mathsf{P}\{(\varepsilon_j \leq z_j^{\beta\sigma}) \cap (\xi_j < 0) | \cdot\}$ and $\mathsf{P}_2 = \mathsf{P}\{(\varepsilon_j \leq z_j^{\beta\sigma}) \cap (\xi_j > 0) | \cdot\}$.

By (3.1), we have $(\varepsilon_j \leq z_j^{\beta\sigma}) = (\xi_j \leq z_j^{\beta\sigma} - \min_{i \in \zeta} y_i^{\beta\sigma})$, so that P_1 can be written as $\mathsf{P}\{\xi_j \leq \min(z_j^{\beta\sigma} - \min_{i \in \zeta} y_i^{\beta\sigma}, 0) | \cdot\}$. Hence,

$$\mathsf{P}_1 = \begin{cases} \mathsf{G}_j(z_j^{\beta\sigma} - \min_{i \in \zeta} y_i^{\beta\sigma}) & \text{if } z_j^{\beta\sigma} < \min_{i \in \zeta} y_i^{\beta\sigma}, \\ \mathsf{G}_j(0) & \text{if } z_j^{\beta\sigma} > \min_{i \in \zeta} y_i^{\beta\sigma}. \end{cases}$$

Similarly, $\mathsf{P}_2 = \mathsf{P}\{(\xi_j \leq z_j^{\beta\sigma} - \max_{i \in \zeta} y_i^{\beta\sigma}) \cap (\xi_j > 0) | \cdot\}$ by (3.1). If $z_j^{\beta\sigma} < \max_{i \in \zeta} y_i^{\beta\sigma}$, then the intersection is empty. If instead $z_j^{\beta\sigma} > \max_{i \in \zeta} y_i^{\beta\sigma}$, then, the intersection is the set $(0 < \xi_i \leq z_j^{\beta\sigma} - \max_{i \in \zeta} y_i^{\beta\sigma})$. Hence,

$$\mathsf{P}_2 = \begin{cases} 0 & \text{if } z_j^{\beta\sigma} < \max_{i \in \zeta} y_i^{\beta\sigma}, \\ \mathsf{G}_j(z_j^{\beta\sigma} - \max_{i \in \zeta} y_i^{\beta\sigma}) - \mathsf{G}_j(0) & \text{if } z_j^{\beta\sigma} > \max_{i \in \zeta} y_i^{\beta\sigma}. \end{cases}$$

Note also that if $z_j^{\beta\sigma} < \min_{i \in \zeta} y_i^{\beta\sigma}$, then $z_j^{\beta\sigma} < \max_{i \in \zeta} y_i^{\beta\sigma}$. And, if $z_j^{\beta\sigma} > \max_{i \in \zeta} y_i^{\beta\sigma}$, then $z_j^{\beta\sigma} > \min_{i \in \zeta} y_i^{\beta\sigma}$. In combination, we have

$$\mathsf{P}_0 = \mathsf{P}_1 + \mathsf{P}_2 = \begin{cases} \mathsf{G}_j(z_j^{\beta\sigma} - \min_{i \in \zeta} y_i^{\beta\sigma}) & \text{if } z_j^{\beta\sigma} < \min_{i \in \zeta} y_i^{\beta\sigma}, \\ \mathsf{G}_j(0) & \text{if } \min_{i \in \zeta} y_i^{\beta\sigma} \leq z_j^{\beta\sigma} \leq \max_{i \in \zeta} y_i^{\beta\sigma}, \\ \mathsf{G}_j(z_j^{\beta\sigma} - \max_{i \in \zeta} y_i^{\beta\sigma}) & \text{if } z_j^{\beta\sigma} > \max_{i \in \zeta} y_i^{\beta\sigma}. \end{cases}$$

Recall the notation $\tilde{z}_j^{\beta\sigma} = (z_j^{\beta\sigma} - \min_{i \in \zeta} \varepsilon_i) 1_{(z_j^{\beta\sigma} < \min_{i \in \zeta} \varepsilon_i)} + (z_j^{\beta\sigma} - \max_{i \in \zeta} \varepsilon_i) 1_{(z_j^{\beta\sigma} > \max_{i \in \zeta} \varepsilon_i)}$. We then get $\mathsf{P}_0 = \mathsf{P}(y_j < z_j | \cdot) = \mathsf{G}_j(\tilde{z}_j^{\beta\sigma})$.

Second, similarly, $\mathsf{P}(y_j < z_j - \epsilon | \cdot) = \mathsf{G}_j(\tilde{z}_j^{\beta\sigma\epsilon})$. Combining, the desired result follows.

B Proofs of asymptotic theory for the LTS model

We start with some preliminary extreme value results, which then allows analysis of the case with more 'good' observations than 'outliers'. Then some results on empirical processes follows, which are needed for the general case.

B.1 Extreme values

For a distribution function F define the quantile function $\mathsf{Q}(\psi) = \inf\{c : \mathsf{F}(c) \geq \psi\}$.

Lemma B.1. *Suppose* $\mathsf{F}(c) = 0$ *for* $c < 0$. *Let* $\psi_n = o_\mathsf{P}(1)$. *Then* $\mathsf{Q}_n(\psi_n) = O_\mathsf{P}(1)$.

Proof. Let a small $\epsilon > 0$ be given. Then a finite $x \geq 0$ exists to that $f = \mathsf{F}(x) \geq 1 - \epsilon$. We show $\mathcal{P}_n = \mathsf{P}(A_n) \leq 2\epsilon$ where $A_n = \{\mathsf{Q}_n(\psi_n) \geq x\}$. Applying F_n, we get $A_n = \{\psi_n \geq \mathsf{F}_n(x)\}$. By the Law of Large Numbers, $\mathsf{F}_n(x) = f + \mathsf{o}_\mathsf{P}(1)$. Hence, if $B_n = \{\mathsf{F}_n(x) > f - \epsilon\}$ then $\mathsf{P}_n(B_n) > 1 - \epsilon$ for large n. Since $A_n = (A_n \cap B_n) \cup (A_n \cap B_n^c)$, we have $A_n \subset (A_n \cap B_n) \cup B_n^c$. Here, $\mathsf{P}(B_n^c) \leq \epsilon$ by construction. Moreover, $A_n \cap B_n \subset (\psi_n > f - \epsilon)$ where $\mathsf{P}(\psi_n > f - \epsilon) \leq \epsilon$ for large n since $\psi_n = \mathsf{o}_\mathsf{P}(1)$ by assumption. Thus, $\mathcal{P}_n \leq 2\epsilon$. □

Lemma B.2. *Suppose $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. with $\mathsf{E}|\varepsilon_i|^q < \infty$ for some $q > 0$. Then $\varepsilon_{(n)} = \mathsf{o}_\mathsf{P}(n^{1/p})$ for any $p < q$.*

Proof. We show $\mathcal{P}_n = \mathsf{P}\{\varepsilon_{(n)} \geq \epsilon n^{1/p}\} \to 0$ for any $\epsilon > 0$. Write $\mathcal{P}_n = \mathsf{P} \cup_{i=1}^n (\varepsilon_i \geq \epsilon n^{1/p})$. Boole's inequality gives $\mathcal{P}_n \leq \sum_{i=1}^n \mathsf{P}(\varepsilon_i \geq \epsilon n^{1/p}) = n\mathsf{P}(\varepsilon_1 \geq \epsilon n^{1/p})$. Markov's inequality gives $\mathcal{P}_n \leq n^{1-q/p}\mathsf{E}|\varepsilon_i|^q$, which vanishes for $p < q$. □

B.2 Fewer 'outliers' than 'good' observations

We consider the LTS estimator under the sequence of data generating processes defined in §5.1. The main challenge is to show that $\hat{\delta}$ is close to the $\mathsf{Binomial}(n - h_n, \rho)$-distributed variable $\delta_n = \sum_{j \in \zeta_n} 1_{(\varepsilon_j < \min_{i \in \zeta_n} \varepsilon_i)}$. In the following lemmas, we condition on the sequence δ_n, so that the randomness stems from 'good' errors $\varepsilon_{(\delta_n+1)}, \dots, \varepsilon_{(\delta_n+h_n)}$ and the magnitudes of the 'outliers', $\underline{\varepsilon}_{(\delta_n+1-j)}$ for $j \leq \delta_n$ and $\overline{\varepsilon}_{(j-\delta_n-h_n)}$ for $j > \delta_n + h_n$. The unconditional statements in the Theorems about $\hat{\delta}$ are then derived as follows. If $\mathsf{P}(\hat{\delta} - \delta_n \in I_n | \delta_n) \to 1$ for an interval I_n and a sequence δ_n then by the law of iterated expectations

$$\mathsf{P}(\hat{\delta} - \delta_n \in I_n) = \mathsf{E}\{\mathsf{P}(\hat{\delta} - \delta_n \in I_n | \delta_n)\} \to 1, \tag{B.1}$$

due to the dominated converges theorem, because $\mathsf{P}(\hat{\delta} - \delta_n \in I_n | \delta_n)$ is bounded.

We give detailed proofs for the case $\hat{\delta} > \delta_n$, so that some of the small 'good' observations are considered left 'outliers' and some of the small right 'outliers' are considered 'good'. The case $\hat{\delta} < \delta_n$ is analogous, since we can multiply all observations by -1 and relabel left and right. When considering $\hat{\delta} > \delta_n$ we note that $\hat{\delta} - \delta_n \leq \overline{n}$. Due to the binomial construction of δ_n then $\overline{n} = 0$ *a.s.* when $\rho = 1$, so that the event $\hat{\delta} > \delta_n$ is a null set. Thus, when analysing $(\hat{\delta} - \delta_n)/h_n$ it suffices to consider $\hat{\delta} > \delta_n$ and $\rho < 1$.

The next lemma concerns cases where $\hat{s} = \hat{\delta} - \delta_n$ is small. We show that the LTS estimators are close to the infeasible OLS estimators for the 'good' observations. We note that, by Lemma B.2 and Assumption 5.1(ii) that $\mathsf{E}|\varepsilon_i|^{2+\omega} < \infty$, we can find a small $\eta > 0$ so that $\varepsilon_{(\delta_n+1)}, \varepsilon_{(\delta_n+h_n)} = \mathsf{o}_\mathsf{P}(h_n^{1/2-\eta})$. We write $\varepsilon_{(\hat{\delta}+i)}^p$ for $\{\varepsilon_{(\hat{\delta}+i)}\}^p$.

Lemma B.3. *Consider the LTS estimator under Assumption 5.1(ii). Let $\hat{\delta} = \delta_n + \mathsf{O}_\mathsf{P}(h_n^\eta)$ for some small $\eta > 0$ defined above. Then, $h_n^{1/2}(\hat{\mu} - \hat{\mu}_{\delta_n})$ and $\hat{\sigma}^2 - \hat{\sigma}_{\delta_n}^2$ are $\mathsf{o}_\mathsf{P}(1)$.*

Proof. The estimators $\hat{\mu}$ and $\hat{\sigma}^2$ are formed from the sample moments of $\varepsilon_{(\hat{\delta}+1)}, \dots, \varepsilon_{(\hat{\delta}+h_n)}$. Let $\mathcal{E}_{np} = \sum_{i=1}^{h_n} \varepsilon_{(\hat{\delta}+i)}^p - \sum_{i=1}^{h_n} \varepsilon_{(\delta_n+i)}^p$. We show that $\mathcal{E}_{n1} = \mathsf{o}_\mathsf{P}(h_n^{1/2})$ and $\mathcal{E}_{n2} = \mathsf{o}_\mathsf{P}(h_n)$. As remarked above, we condition on δ_n and consider $\hat{\delta} > \delta_n$ while assuming $\rho < 1$. Then

$$\mathcal{E}_{np} = \sum_{i=1}^{h_n+\delta_n-\hat{\delta}} \varepsilon_{(\hat{\delta}+i)}^p + \sum_{i=h_n+\delta_n-\hat{\delta}+1}^{h_n} \varepsilon_{(\hat{\delta}+i)}^p - \sum_{i=1}^{\hat{\delta}-\delta_n} \varepsilon_{(\delta_n+i)}^p - \sum_{i=\hat{\delta}-\delta_n+1}^{h_n} \varepsilon_{(\delta_n+i)}^p.$$

The first and the fourth sum cancel. In the second sum, change summation index $j = i - h_n - \delta_n + \hat{\delta}$, so that $\hat{\delta}_n + i = \delta_n + h_n + j$, and replace $\varepsilon_{(\delta_n+h_n+j)} = \varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(j)}$. This gives

$$\mathcal{E}_{np} = \sum_{i=1}^{\hat{\delta}-\delta_n} \{\varepsilon_{(\delta_n+h_n+i)}^p - \varepsilon_{(\delta_n+i)}^p\} = \sum_{i=1}^{\hat{\delta}-\delta_n} [\{\varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(i)}\}^p - \varepsilon_{(\delta_n+i)}^p].$$

Here $\varepsilon_{(\delta_n+h_n)}$ is the maximum and $\varepsilon_{(\delta_n+i)}$ is the ith order statistic of the 'good' errors.

For $p = 1$ we find $\varepsilon_{(\delta_n+i)} \geq \varepsilon_{(\delta_n+1)}$ and $\overline{\varepsilon}_{(i)} \leq \overline{\varepsilon}_{(\hat{\delta}-\delta_n)}$, so that

$$\mathcal{E}_{n1} \leq (\hat{\delta} - \delta_n)\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)} + \overline{\varepsilon}_{(\hat{\delta}-\delta_n)}\}.$$

Here, $\overline{\varepsilon}_{(\hat{\delta}-\delta_n)}$ is the $(\hat{\delta}-\delta_n)/\overline{n}$ empirical quantile in the $\overline{\mathsf{G}}$ distribution. By assumption $\hat{\delta} - \delta_n = \mathsf{O}_\mathsf{P}(h_n^\eta)$ and $\overline{n}/h_n \to \overline{\omega} = (1-\lambda)(1-\rho)/\lambda > 0$ so that $(\hat{\delta} - \delta_n)/\overline{n} = \mathsf{O}_\mathsf{P}(h_n^{\eta-1}) = \mathsf{o}_\mathsf{P}(1)$. Lemma B.1 then shows $\overline{\varepsilon}_{(\hat{\delta}-\delta_n)} = \mathsf{O}_\mathsf{P}(1)$. Further, Lemma B.2 using Assumption 5.1(ii) that $\mathsf{E}|\varepsilon_i|^{2+\omega} < \infty$ shows that $\varepsilon_{(\delta_n+1)}, \varepsilon_{(\delta_n+h_n)}$ are $\mathsf{o}_\mathsf{P}(h_n^{1/2-\eta})$ for some $\eta < 0$. In combination, $\mathcal{E}_{n1} = \mathsf{O}_\mathsf{P}(h_n^\eta)\{\mathsf{o}_\mathsf{P}(h_n^{1/2-\eta}) + \mathsf{O}_\mathsf{P}(1)\} = \mathsf{o}_\mathsf{P}(n^{1/2})$.

For $p = 2$ we find similarly, using the inequality $(x+y)^2 \leq 2(x^2 + y^2)$,

$$\mathcal{E}_{n2} \leq \sum_{i=1}^{\hat{\delta}-\delta_n} \{2\varepsilon_{(\delta_n+h_n)}^2 + 2\overline{\varepsilon}_{(i)}^2 - \varepsilon_{(\delta_n+i)}^2\} \leq 2(\hat{\delta} - \delta_n)\{\varepsilon_{(\delta_n+h_n)}^2 + \overline{\varepsilon}_{(\hat{\delta}-\delta_n)}^2\}.$$

Apply the above bounds $\hat{\delta} - \delta_n = \mathsf{O}_\mathsf{P}(h_n^\eta)$, $\varepsilon_{(\delta_n+h_n)} = \mathsf{o}_\mathsf{P}(h_n^{1/2-\eta})$ and $\overline{\varepsilon}_{(\hat{\delta}-\delta_n)} = \mathsf{O}_\mathsf{P}(1)$ to get that $\mathcal{E}_{n2} = \mathsf{O}_\mathsf{P}(h_n^\eta)[\mathsf{o}_\mathsf{P}\{h_n^{(1/2-\eta)2}\} + \mathsf{O}_\mathsf{P}(1)] = \mathsf{o}_\mathsf{P}(h_n)$. $\qquad\square$

The next Lemma is the main ingredient to showing consistency of $\hat{\delta}$, when less than half of the observations are 'outliers'. By Assumption 5.1(iii) we have $\varepsilon_{(\delta_n+h_n)}^{-1} = \mathsf{o}_\mathsf{P}(1)$. It is convenient to define sequences

$$\underline{s}_n = h_n^\eta, \qquad \overline{s}_n = |\varepsilon_{(\delta_n+h_n)}|^{-1/2} h_n. \tag{B.2}$$

We note that for large h_n then $0 < \underline{s}_n < \overline{s}_n$ where $\overline{s}_n/h_n = \mathsf{o}_\mathsf{P}(1)$.

Lemma B.4. *Suppose Assumption 5.1(i, iii, iv, v) and let $\rho < 1$. Then, for some constant $C > 0$, we have $\min_{\underline{s}_n \leq s \leq h_n - \overline{s}_n} h_n^\eta(\hat{\sigma}_{\delta_n+s}^2 - \hat{\sigma}_{\delta_n}^2) \to \infty$ in probability.*

Proof. We condition on δ_n, which is the number of left 'outliers'. Recall that the ordered 'good' observations are $\varepsilon_{(\delta_n+1)}, \ldots, \varepsilon_{(\delta_n+h_n)}$. As $\rho, \lambda < 1$ then $\overline{n}/n \to \overline{\omega} = (1-\rho)(1-\lambda) > 0$. Let $\varepsilon_{(\delta_n+h_n+j)} = \varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(j)}$ for $1 \leq j \leq \overline{n}$. Expand, see §D in Supplementary material,

$$S_s = (\hat{\sigma}_{\delta_n+s}^2 - \hat{\sigma}_{\delta_n}^2)/\sigma = (s/h_n)\{1 - (s/h_n)\}\varepsilon_{(\delta_n+h_n)}^2 + A_n, \tag{B.3}$$

with coefficients $A_n = A_{n1} - A_{n2} + 2A_{n3} - 2A_{n4}$ where

$$A_{n1} = \frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}^2 - \{\frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\}^2, \quad A_{n3} = \frac{1}{h_n}\sum_{j=1}^{s}\overline{\varepsilon}_{(j)}\frac{1}{h_n}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\},$$

$$A_{n2} = \frac{1}{h_n}\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}^2 - \{\frac{1}{h_n}\sum_{i=1}^{s}\varepsilon_{(\delta_n+i)}\}^2, \quad A_{n4} = \frac{1}{h_n}\sum_{i=s+1}^{h_n}\varepsilon_{(\delta_n+i)}\frac{1}{h_n}\sum_{i=1}^{s}\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\}.$$

23

We find a lower bound for A_n. Notice $A_{n1}, A_{n3} \geq 0$. Further, $A_{n2} \leq h_n^{-1} \sum_{i=1}^{s} \varepsilon_{(\delta_n+i)}^2 = B_{n2}$ say. For A_{n4}, use Jensen's inequality, add further summand and use the Law of Large Numbers using Assumption 5.1(i) for the unordered normal 'good' ε_{δ_n+i} to get

$$|\frac{1}{h_n} \sum_{i=s+1}^{h_n} \varepsilon_{(\delta_n+i)}|^2 \leq \frac{1}{h_n} \sum_{i=s+1}^{h_n} \varepsilon_{(\delta_n+i)}^2 \leq \frac{1}{h_n} \sum_{i=1}^{h_n} \varepsilon_{(\delta_n+i)}^2 = \frac{1}{h_n} \sum_{i=1}^{h_n} \varepsilon_{\delta_n+i}^2 \overset{\mathsf{P}}{\to} 1. \qquad (B.4)$$

Further, we have $h_n^{-1} \sum_{i=1}^{s} \{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\} \leq (s/h_n)\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)}\} = B_{n4}$ say, so that $|A_{n4}| \leq B_{n4}\{1 + o_{\mathsf{P}}(1)\}$, where the remainder term from (B.4) is uniform in s. In combination,

$$S_s \geq (s/h_n)(1 - s/h_n)\varepsilon_{(\delta_n+h_n)}^2 - B_{n2} - 2B_{n4}\{1 + o_{\mathsf{P}}(1)\}, \qquad (B.5)$$

where the $o_{\mathsf{P}}(1)$ term is uniform in s. We analyze separately for $s \leq \bar{s}_n$ and $\bar{s}_n \leq s$.

1. *Consider* $\bar{s}_n \leq s \leq h_n - \bar{s}_n$ *where* $\bar{s}_n/h_n = |\varepsilon_{(\delta_n+h_n)}|^{-1/2}$. Since the function $x(1 - x)$ is concave with roots at $x = 0$ and $x = 1$, so that, for $x_0 < x < 1 - x_0$ with $0 < x_0 < 1/2$, we get $x(1 - x) \geq x_0/2$. Thus, $2(s/h_n)(1 - s/h_n) \geq \bar{s}_n/h_n = |\varepsilon_{(\delta_n+h_n)}|^{-1/2}$ on the considered range. We bound $B_{n2} \leq 1 + o_{\mathsf{P}}(1)$ uniformly in s as in (B.4). For B_{n4}, we use $s/h_n \leq 1$. Thus, (B.5) reduces to $2S_s \geq |\varepsilon_{(\delta_n+h_n)}|^{-1/2}\varepsilon_{(\delta_n+h_n)}^2 - 2\{1 + 2\varepsilon_{(\delta_n+h_n)} - 2\varepsilon_{(\delta_n+1)}\}\{1 + o_{\mathsf{P}}(1)\}$. Since $\varepsilon_{(\delta_n+h_n)} \to \infty$ in probability and $\varepsilon_{(\delta_n+1)}/\varepsilon_{(\delta_n+h_n)} = -1 + o_{\mathsf{P}}(1)$ by Assumption 5.1(iv) we get that $\min_{\bar{s}_n \leq s \leq h_n - \bar{s}_n} 2S_s \geq |\varepsilon_{(\delta_n+h_n)}|^{3/2}\{1 + o_{\mathsf{P}}(1)\}$. In particular, $\min_{\bar{s}_n \leq s \leq h_n - \bar{s}_n} h_n^{\eta} S_s \to \infty$ in probability.

2. Consider $\underline{s}_n \leq s \leq \bar{s}_n$ *where* $\underline{s}_n = h_n^{1-\eta}$ *for any* $\eta > 0$ *and* $\bar{s}_n = |\varepsilon_{(\delta_n+h_n)}|^{-1/2}h_n$, see (B.2). We find lower bounds for the $B_{n\ell}$ terms in (B.5).

First, $B_{n2} = h_n^{-1} \sum_{i=1}^{s} \varepsilon_{(\delta_n+i)}^2$. Write $B_{n2} = h_n^{-1}\{\sum_{i=1}^{r_n} \varepsilon_{(\delta_n+i)}^2 + \sum_{i=r_n+1}^{s} \varepsilon_{(\delta_n+i)}^2\}$, where $r_n = h_n^{1-2\eta}$. Since the squared order statistics decrease with increasing index we can bound $B_{n2} \leq h_n^{-1}\{r_n\varepsilon_{(\delta_n+1)}^2 + (s - r_n)\varepsilon_{(\delta_n+r_n)}^2\}$. Bounding $(s - r_n) \leq s$ and $r_n/s \leq r_n/\underline{s}_n = h_n^{-\eta}$ we get $B_{n2} \leq (s/h_n)\{h_n^{-\eta}\varepsilon_{(\delta_n+1)}^2 + \varepsilon_{(\delta_n+r_n)}^2\}$. By Assumption 5.1($iv,v$) then $\varepsilon_{(\delta_n+1)}/\varepsilon_{(\delta_n+h_n)} = -1 + o_{\mathsf{P}}(1)$ and $\varepsilon_{(\delta_n+r_n)}/\varepsilon_{(\delta_n+1)} \leq C_\eta + o_{\mathsf{P}}(1)$ for some $C_\eta < 1$. Then, $B_{n2} \leq (s/h_n)\varepsilon_{(\delta_n+h_n)}^2[h_n^{-\eta}\{1 + o_{\mathsf{P}}(1)\} + \{C_\eta + o_{\mathsf{P}}(1)\}\{1 + o_{\mathsf{P}}(1)\}]$, which reduces to $(s/h_n)\varepsilon_{(\delta_n+h_n)}^2\{C_\eta + o_{\mathsf{P}}(1)\}$.

Second, $B_{n4} = (s/h_n)\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)}\}$ by definition.

Insert bounds for B_{n2}, B_{n4} in (B.5) along with $s/h_n \leq \bar{s}_n/h_n = |\varepsilon_{(\delta_n+h_n)}|^{-1/2}$ to get

$$S_s \geq (s/h_n)[\varepsilon_{(\delta_n+h_n)}^2 - |\varepsilon_{(\delta_n+h_n)}|^{3/2} - \varepsilon_{(\delta_n+h_n)}^2\{C_\eta + o_{\mathsf{P}}(1)\} - 2\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+1)}\}\{1 + o_{\mathsf{P}}(1)\}].$$

Using that $\varepsilon_{(\delta_n+1)}/\varepsilon_{(\delta_n+h_n)} = -1 + o_{\mathsf{P}}(1)$ and $\varepsilon_{(\delta_n+h_n)}^{-1} = o_{\mathsf{P}}(1)$ by Assumption 5.1(iii) gives that $S_s \geq (s/h_n)\varepsilon_{(\delta_n+h_n)}^2\{1 - C_\eta + o_{\mathsf{P}}(1)\}$. Further $s \geq \underline{s}_n$ where $\underline{s}_n/h_n = h_n^{-\eta}$. Thus, $\min_{\underline{s}_n \leq s \leq \bar{s}_n} h_n^{\eta} S_s \to \infty$ in probability. $\qquad \square$

Proof of Theorem 5.2. First, we show that $\hat{s} = \hat{\delta} - \delta_n = o_{\mathsf{P}}(h_n^{\eta})$ for any $\eta > 0$. It suffices to show $\hat{s} = O_{\mathsf{P}}(h_n^{\eta})$ for each η. We consider $\rho < 1$ and $1/2 < \lambda < 1$ so that $\overline{\omega} = (1 - \rho)(1 - \lambda)/\lambda$.

Choose $\underline{s}_n = h_n^{\eta}$ and $\bar{s}_n = |\varepsilon_{(\delta_n+h_n)}|^{1/2}$ as in (B.2). We have $\hat{s} \leq \overline{n} \leq n - h_n$, since there are $\overline{n}$ right 'outliers' and $n - h_n$ 'outliers' in total. Here, h_n satisfies $h_n/n \to \lambda$ by (5.2), while $\overline{n}/h_n \to \overline{\omega}$. Thus, for large n, we have that $\hat{s}/h_n \leq 1 - \bar{s}_n/h_n$ and it suffices to show $\mathsf{P}(\underline{s}_n \leq \hat{s} \leq h_n - \bar{s}_n)$ vanishes. Now, $\hat{s}$ is the minimizer of $S_s = \hat{\sigma}_{\delta_n+s}^2 - \hat{\sigma}_{\delta_n}^2$. Since $S_0 = 0$, then $\mathsf{P}(\underline{s}_n \leq \hat{s} \leq h_n - \bar{s}_n) \to 0$ if $\min_{\underline{s}_n \leq s \leq h_n - \bar{s}_n} h_n^{\eta}(\hat{\sigma}_{\delta_n+s}^2 - \hat{\sigma}_{\delta_n}^2) \to \infty$ in probability. This follows from Lemma B.4 using Assumption 5.1(i, iii, iv, v).

Second, since $\hat{\delta} - \delta_n = o_{\mathsf{P}}(h_n^{\eta})$ then Lemma B.3 using Assumption 5.1(ii) shows that $h_n^{1/2}(\hat{\mu} - \hat{\mu}_{\delta_n})$, $\hat{\sigma}^2 - \hat{\sigma}_{\delta_n}^2$ are $o_{\mathsf{P}}(1)$.

Third, the i.i.d. Law of Large Numbers and Central Limit Theorem using Assumption 5.1(i) show that $h_n^{1/2}(\hat{\mu}_{\delta_n} - \mu)/\sigma$ is asymptotically normal while $\hat{\sigma}_{\delta_n}^2$ is consistent for σ^2. $\qquad \square$

B.3 Marked empirical processes evaluated at quantiles

We start with some preliminary results on marked empirical processes evaluated at quantiles.

Consider random variables ε_i for $i = 1, \ldots, n$ and define the marked empirical distribution and its expectation, for $c \geq 0$, by

$$\mathsf{F}_n^p(c) = n^{-1} \sum_{i=1}^n \varepsilon_i^p 1_{(\varepsilon_i \leq c)}, \qquad \overline{\mathsf{F}}^p(c) = \mathsf{E}\mathsf{F}_n^p(c) = \mathsf{E}\varepsilon_1^p 1_{(\varepsilon_1 \leq c)}. \tag{B.6}$$

For $p = 0$, let $\mathsf{F}_n^0 = \mathsf{F}_n$. We also define the quantile function $\mathsf{Q}(\psi) = \inf\{c : \mathsf{F}(c) \geq \psi\}$ and the empirical quantiles $\mathsf{Q}_n(\psi) = \inf\{c : \mathsf{F}_n(c) \geq \psi\}$. The first result follows from the theory of empirical quantile processes.

Lemma B.5. *Suppose F is regular (Definition 5.1). Then, for all $\zeta > 0$,*
(a) $n^{1/2}[\mathsf{F}_n\{\mathsf{Q}(\psi)\} - \psi]$ converges in distribution on $D[0,1]$ to a Brownian bridge;
(b) $\sup_{0 \leq \psi \leq 1} |n^{1/2}[\mathsf{F}\{\mathsf{Q}_n(\psi)\} - \psi] + n^{1/2}[\mathsf{F}_n\{\mathsf{Q}(\psi)\} - \psi]| \overset{a.s.}{=} \mathrm{o}(n^{\zeta-1/4})$.

Proof. (a) This is Billingsley (1968, Theorem 16.4).

(b) Let $D(\psi) = \mathsf{f}\{\mathsf{Q}(\psi)\} n^{1/2}\{\mathsf{Q}_n(\psi) - \mathsf{Q}(\psi)\}$ and write the object of interest as the sum of $n^{1/2}[\mathsf{F}\{\mathsf{Q}_n(\psi)\} - \psi] - D(\psi)$ and $n^{1/2}[\mathsf{F}_n\{\mathsf{Q}(\psi)\} - \psi] + D(\psi)$. These two terms are $\mathrm{o}(n^{\zeta-1/4})$ *a.s.* by Csörgő (1983, Corollaries 6.2.1, 6.2.2), uniformly in ψ. $\qquad\square$

We need the following Glivenko-Cantelli result.

Lemma B.6. *Let ε_i be i.i.d. continuous, positive random variables with $\mathsf{E}|\varepsilon_i|^p < \infty$. Then $\sup_{c>0} |\mathsf{F}_n^p(c) - \overline{\mathsf{F}}^p(c)| = \mathrm{o}_\mathsf{P}(1)$.*

Proof. We note that $\overline{\mathsf{F}}^p$ is non-decreasing with $\overline{\mathsf{F}}^p(\infty) < \infty$. Since $\overline{\mathsf{F}}^p$ is continuous, then for any $\delta > 0$ exists a finite integer $K \in \mathbb{N}$ and chaining points $-\infty = c_0 < c_1 < \cdots < c_K = \infty$ so that $\max_{1 \leq k \leq K}\{\overline{\mathsf{F}}^p(c_k) - \overline{\mathsf{F}}^p(c_{k-1})\} \leq \delta$.

Since F_n^p and $\overline{\mathsf{F}}^p$ are both non-decreasing we get for $c_{k-1} < c \leq c_k$ the bounds

$$\mathsf{F}_n^p(c) - \overline{\mathsf{F}}^p(c) \leq \{\mathsf{F}_n^p(c_k) - \overline{\mathsf{F}}^p(c_k)\} + \{\overline{\mathsf{F}}^p(c_k) - \overline{\mathsf{F}}^p(c_{k-1})\},$$
$$\mathsf{F}_n^p(c) - \overline{\mathsf{F}}^p(c) \geq \{\mathsf{F}_n^p(c_{k-1}) - \overline{\mathsf{F}}^p(c_{k-1})\} - \{\overline{\mathsf{F}}^p(c_k) - \overline{\mathsf{F}}^p(c_{k-1})\}.$$

In combination, $|\mathsf{F}_n^p(c) - \overline{\mathsf{F}}^p(c)| \leq \max_{k-1,k} |\mathsf{F}_n^p(c_k) - \overline{\mathsf{F}}^p(c_k)| + \{\overline{\mathsf{F}}^p(c_k) - \overline{\mathsf{F}}^p(c_{k-1})\}$. The last term is bounded by δ, so that $\sup_{c>0} |\mathsf{F}_n^p(c) - \overline{\mathsf{F}}^p(c)| \leq \max_{1 \leq k \leq K} |\mathsf{F}_n^p(c_k) - \overline{\mathsf{F}}^p(c_k)| + \delta$. For each k then $\mathsf{F}_n^p(c_k) - \overline{\mathsf{F}}^p(c_k) = \mathrm{o}_\mathsf{P}(1)$ by the Law of Large Numbers, requiring ε_i to be i.i.d. with $\mathsf{E}|\varepsilon_i|^p < \infty$. Since K is finite then the maximum over k also vanishes almost surely. By choosing δ sufficiently small the overall bound is seen to be $\mathrm{o}_\mathsf{P}(1)$. $\qquad\square$

The next result is inspired by Johansen and Nielsen (2016a, Lemma D.11).

Lemma B.7. *Let $p \in \mathbb{N}_0$. Suppose ε_i is positive, regular and $\mathsf{E}\varepsilon_i^q < \infty$ for some $q > 2p$. Then, $\sup_{1/(n+1) \leq \psi \leq n/(n+1)} |\mathsf{F}_n^p\{\mathsf{Q}_n(\psi)\} - \overline{\mathsf{F}}^p\{\mathsf{Q}(\psi)\}| = \mathrm{o}_\mathsf{P}(1)$.*

Proof. For $p = 0$, then $\phi_n = n^{1/2}[\mathsf{F}_n^p\{\mathsf{Q}_n(\psi)\} - \overline{\mathsf{F}}^p\{\mathsf{Q}(\psi)\}]$ satisfies $\phi_n = n^{1/2}[\mathsf{F}\{\mathsf{Q}_n(\psi)\} - \psi]$. Lemma B.5 shows that ϕ_n converges in distribution to a Brownian bridge as a process in ψ. By Billingsley (1968, p. 142-143) then $\sup_{0 \leq \psi \leq 1} \phi_n$ converges in distribution so that $\sup_{0 \leq \psi \leq 1} \phi_n = \mathrm{O}_\mathsf{P}(1)$ and the result follows. For $p \in \mathbb{N}$, add and subtract $\overline{\mathsf{F}}_n^p\{\mathsf{Q}_n(\psi)\}$ to get

$$\mathsf{F}_n^p\{\mathsf{Q}_n(\psi)\} - \overline{\mathsf{F}}^p\{\mathsf{Q}(\psi)\} = [\mathsf{F}_n^p\{\mathsf{Q}_n(\psi)\} - \overline{\mathsf{F}}^p\{\mathsf{Q}_n(\psi)\}] + [\overline{\mathsf{F}}^p\{\mathsf{Q}_n(\psi)\} - \overline{\mathsf{F}}^p\{\mathsf{Q}(\psi)\}].$$

Term 1. This is $o_P(1)$ since $\sup_{c>0} |\mathsf{F}_n^p(c) - \overline{\mathsf{F}}^p(c)| = o_P(1)$ by Lemma B.6.

Term 2. Write $S_n(\psi) = \overline{\mathsf{F}}^p\{\mathsf{Q}_n(\psi)\} - \overline{\mathsf{F}}^p\{\mathsf{Q}(\psi)\}$ as integral. First, change variable $u = \mathsf{F}(c)$, $du = \mathsf{f}(c)dc$, so that $c = \mathsf{Q}(u)$. Then apply the mean value theorem, so that

$$S_n(\psi) = \int_{\mathsf{Q}(\psi)}^{\mathsf{Q}_n(\psi)} c^p \mathsf{f}(c)dc = \int_{\psi}^{\mathsf{F}\{\mathsf{Q}_n(\psi)\}} \{\mathsf{Q}(u)\}^p du = \{\mathsf{Q}(\psi^*)\}^p n^{-1/2}\phi_n.$$

for an intermediate point ψ^* so that $\mathsf{Q}(\psi^*)$ belongs to the interval from $\mathsf{Q}(\psi)$ to $\mathsf{Q}_n(\psi)$. Since $\sup_{0\le\psi\le 1} \phi_n = O_P(1)$, we must show that $\{\mathsf{Q}(\psi^*)\}^p = o_P(n^{1/2})$. It suffices to show that $\mathsf{Q}(\psi)$ and $\mathsf{Q}_n(\psi)$ are $o_P\{n^{1/(2p)}\}$ for $\psi = n/(n+1)$.

Consider $\mathsf{Q}_n(\psi)$. Write $\mathsf{Q}_n(\psi) = \max_{1\le i\le n} \varepsilon_i$. Lemma B.2 shows $\mathsf{Q}_n(\psi) = o_P\{n^{1/(2p)}\}$ for $2p < q$ since $\mathsf{E}\varepsilon_i^q < \infty$ by assumption.

Consider $\mathsf{Q}(\psi)$. Bound $|\mathsf{Q}(\psi)| \le c_n$, where c_n satisfies $(n+1)^{-1} = \mathsf{P}(\varepsilon_1 > c_n)$. We must show $c_n = o\{n^{1/(2p)}\}$. It suffices that $c_n^q = O(n)$ for $q > 2p$. By the Markov inequality $\mathsf{P}(|\varepsilon_1| > c_n) \le c_n^{-q}\mathsf{E}\varepsilon_i^q$, so that $c_n^q \le (n+1)/\mathsf{E}\varepsilon_i^q = O(n)$. $\qquad\square$

B.4 More 'outliers' than 'good' observations

The next lemma is needed when there are more than half of the observations are 'outliers'. As $\hat{\sigma}_\delta^2$ is not diverging, additional regularity conditions are needed to ensure that $\hat{\sigma}_\delta^2 > \hat{\sigma}_{\delta_n}^2$.

Lemma B.8. *Suppose Assumption 5.2(i) holds. Let $1 \le \overline{\omega} = (1-\rho)(1-\lambda)/\lambda < \infty$. Recall $\overline{s}_n = (2\log h_n)^{-1/4}h_n$ from (B.2). Then, conditional on δ_n, an $\epsilon > 0$ exists so that $\min_{h_n-\overline{s}_n\le s\le\overline{n}}(\hat{\sigma}_{\delta_n+s}^2 - \hat{\sigma}_{\delta_n}^2) \ge \epsilon + o_P(1)$ for large n.*

Proof. The errors $\varepsilon_{(\delta_n+i)}$ are standard normal order statistics for $1 \le i \le h_n$ and $\varepsilon_{(\delta_n+h_n+j)} = \varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(j)}$ for $1 \le j \le \overline{n}$, where $\overline{\varepsilon}_j$ is $\overline{\mathsf{G}}$-distributed. We let $\overline{\varepsilon}_0 = 0$.

It suffices to show that $\hat{\sigma}_{\delta_n+s}^2/\sigma^2 \ge 1 + \epsilon + o_P(1)$ uniformly in s, since $\hat{\sigma}_{\delta_n}^2/\sigma^2 = 1 + o_P(1)$ by the Law of Large Numbers using Assumption 5.1(i) applied to the sample variance of the 'good' errors. We consider separately the cases $h_n \le s \le \overline{n}$ and $h_n - \overline{s}_n \le s < h_n$.

1. *Consider $h_n \le s \le \overline{n}$.* In this case, $\hat{\sigma}_{\delta_n+s}^2$ is the sample variance of $\varepsilon_{(\delta_n+s+j)} = \varepsilon_{(\delta_n+h_n)} + \overline{\varepsilon}_{(s-h_n+j)}$ for $1 \le j \le h_n$. Sample variances are invariant to the location, so that $\hat{\sigma}_{\delta_n+s}^2 = h^{-1}\sum_{j=1}^{h} \overline{\varepsilon}_{(s-h_n+j)}^2 - \{h^{-1}\sum_{j=1}^{h} \overline{\varepsilon}_{(s-h_n+j)}\}^2$. Let $\mathcal{A}_{s/\overline{n}} = \{s/\overline{n} - \overline{\omega}^{-1} < \overline{\mathsf{G}}(\overline{\varepsilon}_1) \le s/\overline{n}\}$.

We argue that $\min_{h_n\le s\le\overline{n}} \hat{\sigma}_{\delta_n+s}^2/\sigma^2 \ge \overline{v}$, where $\overline{v} = \min_{\overline{\omega}^{-1}\le\varsigma\le 1} \mathsf{Var}(\overline{\varepsilon}_1|\mathcal{A}_\varsigma)$. This suffices, as $\overline{v} > 1$ by Assumption 5.2(i). Write $\sum_{j=1}^{h_n} \overline{\varepsilon}_{(s-h_n+j)}^p = \sum_{k=1}^{\overline{n}} \overline{\varepsilon}_k^p 1_{\{\overline{\varepsilon}_{(s-h_n)}<\overline{\varepsilon}_k\le\overline{\varepsilon}_{(s)}\}}$ and let $\overline{\mathsf{G}}_n^p(c) = \overline{n}^{-1}\sum_{i=1}^{\overline{n}} \overline{\varepsilon}_i^p 1_{(\varepsilon_i\le c)}$ and $\overline{\mathsf{G}}_n^{-1}(\psi) = \inf\{c : \overline{\mathsf{G}}(c) \ge \psi\}$, so that $\overline{\varepsilon}_{(k)} = \overline{\mathsf{G}}_n^{-1}(k/\overline{n})$. Then,

$$\overline{n}^{-1}\sum_{j=1}^{h_n} \overline{\varepsilon}_{(s-h_n+j)}^p = \overline{\mathsf{G}}_n^p\{\overline{\mathsf{G}}_n^{-1}(s/\overline{n})\} - \overline{\mathsf{G}}_n^p[\overline{\mathsf{G}}_n^{-1}\{(s-h_n)/\overline{n}\}].$$

Apply Lemma B.7 with $\mathsf{F} = \overline{\mathsf{G}}$, $n = \overline{n}$, requiring the 4+ moments and regularity of Assumption 5.2(i), so that, uniformly in $h_n \le s \le \overline{n}$,

$$\overline{n}^{-1}\sum_{j=1}^{h_n} \overline{\varepsilon}_{(s-h_n+j)}^p = \mathsf{E}\overline{\varepsilon}_1^p 1_{\{s/\overline{n}-h_n/\overline{n}<\overline{\mathsf{G}}(\overline{\varepsilon}_1)\le s/\overline{n}\}} + o_P(1).$$

Now, $h_n/\overline{n} \to \overline{\omega}^{-1}$, where $\overline{\omega} \ge 1$ by construction. Then, $\mathsf{E}\overline{\varepsilon}_1^p 1_{\{s/\overline{n}-h_n/\overline{n}<\overline{\mathsf{G}}(\overline{\varepsilon}_1)\le s/\overline{n}\}} = \mathsf{E}\overline{\varepsilon}_1^p 1_{\mathcal{A}_{s/\overline{n}}} + o_P(1)$ uniformly in $h_n \le s \le \overline{n}$. Noting that $h_n/\overline{n} = \overline{n}^{-1}\sum_{j=1}^{h_n} \overline{\varepsilon}_{(s+j)}^0 = \mathsf{E}1_{\mathcal{A}_{s/\overline{n}}} + o_P(1)$, we get

$$h_n^{-1}\sum_{j=1}^{h_n} \overline{\varepsilon}_{(s+j)}^p = \frac{\mathsf{E}\overline{\varepsilon}_1^p 1_{\mathcal{A}_{s/\overline{n}}}}{\mathsf{E}1_{\mathcal{A}_{s/\overline{n}}}} + o_P(1) = \mathsf{E}(\overline{\varepsilon}_1^p|\mathcal{A}_{s/\overline{n}}) + o_P(1),$$

so that $\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = \mathsf{E}(\bar{\varepsilon}_1^2|\mathcal{A}_{s/\overline{n}}) + o_\mathsf{P}(1) - \{\mathsf{E}(\bar{\varepsilon}_1|\mathcal{A}_{s/\overline{n}}) + o_\mathsf{P}(1)\}^2$. Since $\mathsf{E}\bar{\varepsilon}_1 1_{\mathcal{A}_{s/\overline{n}}} \leq \mathsf{E}\bar{\varepsilon}_1 < \infty$ and $\mathsf{E}1_{\mathcal{A}_{s/\overline{n}}} = \overline{\omega}^{-1} > 0$ uniformly in s, we get $\mathsf{E}(\bar{\varepsilon}_1|\mathcal{A}_s) \leq \overline{\omega}\mathsf{E}\bar{\varepsilon}_1 < \infty$. Thus,

$$\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = \mathsf{E}(\bar{\varepsilon}_1^2|\mathcal{A}_{s/\overline{n}}) - \{\mathsf{E}(\bar{\varepsilon}_1|\mathcal{A}_{s/\overline{n}})\}^2 + o_\mathsf{P}(1) = \mathsf{Var}(\bar{\varepsilon}_1|\mathcal{A}_{s/\overline{n}}) + o_\mathsf{P}(1).$$

Since the errors are uniform in s, we get $\min_{h_n \leq s \leq \overline{n}} \hat{\sigma}^2_{\delta_n+s}/\sigma^2 \geq \overline{v} + o_\mathsf{P}(1)$ as desired.

2. *Consider* $h_n - \overline{s}_n \leq s < h_n$ *where* $\overline{s}_n = (2\log h_n)^{-1/4} h_n$, see (B.2). In this case, we have $h_n - \overline{s}_n$ 'outliers' and $\overline{s}_n$ 'good' observations. Expand,

$$\hat{\sigma}^2_{\delta_n+s}/\sigma^2 = A_n = A_{n1} + A_{n2} + A_{n3} + 2A_{n4}, \tag{B.7}$$

see §D in Supplementary material, where

$$A_{n1} = h_n^{-1}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+i)} - \varepsilon_{(\delta_n+h_n)}\}^2 - [h_n^{-1}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+i)} - \varepsilon_{(\delta_n+h_n)}\}]^2,$$

$$A_{n2} = (\frac{s}{h_n})^2[\frac{1}{s}\sum_{j=1}^{s}\bar{\varepsilon}_{(j)}^2 - \{\frac{1}{s}\sum_{j=1}^{s}\bar{\varepsilon}_{(j)}\}^2], \qquad A_{n3} = \frac{s}{h_n}(1 - \frac{s}{h_n})\frac{1}{s}\sum_{j=1}^{s}\bar{\varepsilon}_{(j)}^2$$

$$A_{n4} = [h_n^{-1}\sum_{i=s+1}^{h_n}\{\varepsilon_{(\delta_n+h_n)} - \varepsilon_{(\delta_n+i)}\}]\{h_n^{-1}\sum_{j=1}^{s}\bar{\varepsilon}_{(j)}\}.$$

We note that $A_{n1}, A_{n2}, A_{n3}, A_{n4} \geq 0$. Therefore, $\hat{\sigma}^2_{\delta_n+s}/\sigma^2 \geq A_{n2}$.

We argue, as in part 1, that $A_{n2} \geq \overline{v} + o_\mathsf{P}(1)$. Indeed, since $1 > s/h_n \geq 1 - \overline{s}_n/h_n \to 1$, then $\overline{n}^{-1}\sum_{j=1}^{h_n-\overline{s}_n}\bar{\varepsilon}_{(j)}^p \leq \overline{n}^{-1}\sum_{j=1}^{s}\bar{\varepsilon}_{(j)}^p \leq \overline{n}^{-1}\sum_{j=1}^{h_n}\bar{\varepsilon}_{(j)}^p$. Both bounds equal $\mathsf{E}\bar{\varepsilon}_1^p 1_{\mathcal{A}_{\overline{\omega}^{-1}}} + o_\mathsf{P}(1)$, so that we can proceed as in part 1. $\square$

Proof of Theorem 5.3. We will show that $\hat{s} = \hat{\delta} - \delta_n = o_\mathsf{P}(h_n^\alpha)$ for any $\alpha > 0$. It suffices to consider the case where $\mathsf{P}(\hat{\delta} - \delta_n > h_n^\alpha) \to 0$ and where $\rho < 1$ as remarked in §5.1. We consider $\lambda, \rho < 1$ so that $\overline{\omega} = (1-\rho)(1-\lambda)/\lambda$ satisfies $0 < \overline{\omega}$. The case $\overline{\omega} < 1$ was covered in the proof of Theorem 5.2. Thus, suppose $\overline{\omega} \geq 1$.

Recall $\underline{s}_n, \overline{s}_n$ from (B.2). In particular, $h_n^\alpha > \underline{s}_n$ for large n, so that $\mathsf{P}(\hat{s} > h_n^\alpha) \leq \mathsf{P}(\hat{s} \geq \underline{s}_n)$. We show, that the latter probability vanishes. Note that $\hat{s} \leq \overline{n}$.

We have that $\hat{s}$ is the minimizer to $\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}$, which is zero for $s = \delta - \delta_n = 0$. The Lemmas B.4, B.8 using Assumption 5.2(i) show that $\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}$, asymptotically, has a uniform, positive lower bound on each of the intervals $\underline{s}_n \leq \hat{s} \leq h_n - \overline{s}_n$ and $h_n - \overline{s}_n \leq \hat{s} \leq \overline{n}$. Thus, $\hat{\sigma}^2_{\delta_n+s} - \hat{\sigma}^2_{\delta_n}$ is bounded away from zero on $s \geq \underline{s}_n$ so that $\mathsf{P}(\hat{s} \geq \underline{s}_n) \to 0$.

A similar argument applies for $\hat{\delta} - \delta_n < -h_n^\alpha$ using Assumption 5.2(ii). The limiting results for $\hat{\mu}, \hat{\sigma}^2$ then follow as in the proof of Theorem 5.2. $\square$

B.5 The OLS estimator in the LTS model

Proof of Theorem 5.4. The sample average satisfies $(\bar{\mu} - \mu)/\sigma = n^{-1}\sum_{i=1}^{n}\varepsilon_i$. Since $\rho = 0$ there are only right 'outliers'. Separate the 'good' observations ε_i for $i = 1, \ldots, h_n$ with maximum $\varepsilon_{(h_n)}$ and 'outliers' $\varepsilon_{h_n+j} = \varepsilon_{(h_n)} + \bar{\varepsilon}_j$ for $j = 1, \ldots, n - h_n$, to get

$$\frac{\bar{\mu} - \mu}{\sigma} = \frac{1}{n}\sum_{i=1}^{h_n}\varepsilon_i + (\frac{n-h_n}{n})\varepsilon_{(h_n)} + \frac{1}{n}\sum_{j=1}^{n-h_n}\bar{\varepsilon}_j. \tag{B.8}$$

Under Assumption 5.1(i, iii) the first sum vanishes by the Law of Large Numbers, while $\varepsilon_{(h_n)} \to \infty$ in probability. Further, $(n - h_n)/n \to 1 - \lambda$. The second sum converges to $(1-\lambda)\mu_\mathsf{G}$ by the Law of Large Numbers. Combine to see that $|\varepsilon_{(h_n)}|^{-1}(\hat{\mu}_{OLS} - \mu)/\sigma = 1 - \lambda + o_\mathsf{P}(1)$. $\square$