The Least Trimmed Quantile Regression

N.M. Neykov[*,a], P. Čížek[b], P. Filzmoser[c], P.N. Neytchev[a]

[a] *National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, 66 Tsarigradsko chaussee, 1784 Sofia, Bulgaria*

[b] *Department of Econometrics & OR, Tilburg School of Economics and Management, Tilburg University, 5000LE Tilburg, The Netherlands*

[c] *Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria*

Abstract

The linear quantile regression estimator is very popular and widely used. It is also well known that this estimator can be very sensitive to outliers in the explanatory variables. In order to overcome this disadvantage, the usage of the least trimmed quantile regression estimator is proposed to estimate the unknown parameters in a robust way. As a prominent measure of robustness, the breakdown point of this estimator is characterized and its consistency is proved. The performance of this approach in comparison with the classical one is illustrated by an example and simulation studies.

Key words: Linear regression, Quantile regression, Least trimmed quantile regression, Breakdown point, Outlier detection

1. Introduction

Consider the multiple linear regression model

$$y_i = x_i^T \beta^0 + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, n, \tag{1}$$

where y_i is an observed response, $x_i^T = (x_{i1}, \ldots, x_{ip})$ is a vector of explanatory variables (covariates, carriers), and β^0 is the underlying value of a $p \times 1$ vector of unknown parameters β. Classically, ε_i, $i = 1, \ldots, n$, are assumed to be independent and identically distributed. Denote by $r_i(\beta) = y_i - x_i^T \beta$ the regression residuals.

Koenker and Bassett (1978) define the quantile regression (QR) estimator as any vector $\hat{\beta}_n(\tau)$ such that

$$\hat{\beta}_n(\tau) := \arg\min_{\beta \in R^p} \sum_{i=1}^{n} \rho_\tau(r_i(\beta)), \tag{2}$$

where

$$\rho_\tau(r(\beta)) = |r(\beta)| \left[\tau 1_{\{r(\beta)\geq 0\}} + (1-\tau) 1_{\{r(\beta)<0\}}\right] = \begin{cases} (\tau-1) r(\beta) & \text{if } r(\beta) < 0, \\ \tau r(\beta) & \text{if } r(\beta) \geq 0, \end{cases}$$

$0 < \tau < 1$, and $1_{\{A\}}$ is the usual indicator function of the set A, which equals 1 if A is true and 0 otherwise.

Different quantile regression estimates $\hat{\beta}_n(\tau)$ can be obtained for different values of τ. This offers the analyst a more complete statistical model than the mean regression, and nowadays, quantile regression has widespread applications. It can be derived as the maximum likelihood (ML) estimator of observations coming from an asymmetric Laplace (double exponential) distribution (e.g., Koenker and Machado, 1999). The

[*] Corresponding author. Tel.: +3592 4624597, Fax: +3592 9744409

Email addresses: Neyko.Neykov@meteo.bg (N.M. Neykov), P.Cizek@uvt.nl (P. Čížek), p.filzmoser@tuwien.ac.at (P. Filzmoser), Plamen.Neytchev@meteo.bg (P.N. Neytchev)

quantile regression estimator is robust to skewed tails and departures from normality. In addition, under very general conditions, the asymptotic distribution of the vector of estimated coefficients is multivariate normal, which permits standard inferences to be carried out (Koenker and Bassett, 1978). The finite-sample distribution of quantile regression was also studied (e.g., Jurečková, 2010). Computational algorithms concerning quantile regression estimation are based on linear programming techniques as discussed in Koenker (2005a), or maximization-minorization techniques considered by Hunter and Lange (2000) and Chen (2004). The package *quantreg* developed in R by Koenker (2005b) (http://www.R-project.org) facilitates the wide use of quantile regression. For more details about quantile regression see Koenker (2005a).

Unfortunately, the quantile regression estimator, like other regression M-estimators, can be highly sensitive to outliers in the explanatory variables, see He et al. (1990). Therefore, many attempts based on the downweighting of distant observations appeared that led to a more robust form of quantile regression (e.g., Hubert and Rousseeuw, 1998, and Giloni et al., 2006). Such procedures were shown to be robust in the regression with a small number of uniformly-distributed or fixed-design regressors (see Giloni et al., 2006, for the case of one and two regressors). The robustness of weighted quantile regression, however, diminishes in general with an increasing number of regressors, and even for a small number of covariates, the robustly weighted quantile regression can be substantially biased by outliers (e.g., Čížek, 2011). This led to the development of alternative estimators of the quantile regression model (1), which are generally based on saddle-point optimization problems (e.g., Rousseeuw and Hubert, 1999, and Adrover et al., 2004). The robustness of these procedures is independent of the complexity of the regression model and is proportional to $\min\{\tau, 1-\tau\}$, where τ refers to the quantile of interest (see Adrover et al., 2004, for an overview). The main disadvantages of these methods – the computational difficulties and non-standard asymptotic distributions – are generally related to their definitions based on nested optimization problems.

In this paper, we consider an alternative approach to robust estimation in the framework of quantile regression, the least trimmed quantile regression (LTQR), which is based on trimming in order to reduce the influence of the outliers in the explanatory variables. The proposed method extends the robust location estimator of the median theoretically studied by Tableman (1994a,b) and the Least Trimmed Absolute deviation (LTA) estimator studied empirically by Hawkins and Olive (1999): we generalize them to the general quantile regression model (1), and additionally, prove the consistency of the proposed method and thus also of LTA. Contrary to existing highly robust methods of quantile regression discussed in the previous paragraphs, the LTA and proposed LTQR estimate the regression quantiles for the data that are not trimmed from the objective function, that is, the quantiles are determined for a subset of data. This allows us to achieve robustness properties independent of the quantile τ of interest. However, the superior robustness properties of the LTQR estimator impose also one constraint: although we can consistently estimate the regression coefficients of all variables in the model (1), the intercept will not be identified (i.e., the constant term will converge to another quantity than the τth quantile of errors ε_i at $x_i = 0$). If the intercept is of importance, it has to be identified by some of the existing procedures.

The paper is organized as follows. Section 2 recalls the generalized trimmed estimator, which renders the proposed LTQR method, and discusses its computation. The LTQR estimator is defined and its breakdown property is discussed in Section 3, while the consistency of the proposed methods is discussed in Section 4. Section 5 demonstrates the different behavior of classical and robust estimation on a simple example. In Section 6, a simulation study is performed to illustrate the effectiveness of the proposed estimator in comparison with the classical quantile regression. Finally, the conclusions are given in Section 8 and proofs are provided in the Appendix.

2. The Generalized Trimmed Estimator

The LTQR estimator will be obtained as a special case of the Generalized Trimmed Estimator (GTE) given by Vandev and Neykov (1998). To introduce it, note that GTE can be defined for any regression model by means of an objective function $f_i : \Theta \to \mathbb{R}^+$, where $\Theta \subseteq \mathbb{R}^q$ is an open set. In particular, the GTE estimator

$\hat{\theta}^{\mathrm{k}}_{\mathrm{n,GTE}}$ of θ is defined as the solution of the optimization problem

$$\hat{\theta}^{\mathrm{k}}_{\mathrm{n,GTE}} := \arg\min_{\theta\in\Theta} \left\{ S_{n,k}(\theta) = \min_{I\in I_k} \sum_{i\in I} w_i f_i(\theta) \right\}, \tag{3}$$

where I_k is the set of all k-subsets of the index set $\{1,\ldots,n\}$ and k is the trimming constant determining the number k of observations and their function values $f_i(\theta)$ kept in the objective function from the total number n of observations. Consequently, the trimming parameter k determines the robustness properties of the GTE as $n-k$ observations with the largest values of $f_i(\theta)$ are excluded from the loss function.

The robustness properties of the GTE can be described, for example, by the finite-sample breakdown point (BDP): it is a global measure of an estimator's robustness characterizing the minimum number of observations that, if arbitrarily modified, can cause the estimates to increase above any bound. The BDP of the GTE is characterized by Theorem 1 of Vandev and Neykov (1998) using the d-fullness technique. Dimova and Neykov (2004) proved that the BDP of the GTE is not less than $\frac{1}{n}\min\{n-k, k-d\}$ if the set $F = \{f_i(\theta) : i = 1,\ldots,n\}$ is d-full ($d = p$ if any p observations are linearly independent, see Section 3 and the appendix for more details). The BDP is maximized for $k = \lfloor (n+d+1)/2 \rfloor$, when it approximately equals $1/2$ for large n. Further, the asymptotic properties of the GTE estimator (3) were studied by Čížek (2008) for the case of twice differentiable functions f.

The optimization problem (3) defining the GTE is of combinatorial nature due to the representation

$$\min_{\theta\in\Theta^p} S_{n,k}(\theta) = \min_{\theta\in\Theta^p} \min_{I\in I_k} \sum_{i\in I} w_i f_i(\theta) = \min_{I\in I_k} \min_{\theta\in\Theta^p} \sum_{i\in I} w_i f_i(\theta). \tag{4}$$

Therefore, it follows that all possible $\binom{n}{k}$ partitions of the set $\{f_1,\ldots,f_n\}$ have to be considered and $\hat{\theta}^{\mathrm{k}}_{\mathrm{n,GTE}}$ is defined by the partition with the minimal value of $S_{n,k}(\theta)$. Hence, an exact computation of the GTE is infeasible for large samples. To get an approximative GTE solution, an algorithm was developed in Neykov et al. (2012). It repeatedly (i) sets $s = 0$, selects a small subset $\{f_{i_1},\ldots,f_{i_{k^*}}\}$ of k^* functions from F and forms $I_s = \{i_1,\ldots,i_{k^*}\}$, (ii) minimizes the objective function $\sum_{i\in I_s} f_i(\theta)$ with respect to θ, and uses the obtained estimate $\hat{\theta}_s$, (iii) sets $s = s+1$, orders the functions of F in ascending order, $f_{\nu(1)}(\hat{\theta}_s) \le f_{\nu(2)}(\hat{\theta}_s) \le \ldots \le f_{\nu(k)}(\hat{\theta}_s) \le \ldots \le f_{\nu(n)}(\hat{\theta}_s)$, where $\nu(.)$ is the permutation of the indices $\{1,2,\ldots,n\}$, and forms $I_s = \{\nu(1),\ldots,\nu(k)\}$; the steps (ii) and (iii) are repeated as long as the newly obtained estimates $\hat{\theta}_s$ produce smaller values of the objective function $\sum_{i\in I_s} f_i(\theta)$.

To fully specify the algorithm, the size and choice of the initial subsets have to be specified (all possible subsets of size $k^* = k$ can be considered to obtain the precise instead of an approximative solution only in very small samples). First, the trial subsample size k^* should be greater than or equal to d, which is necessary for the existence of (3), but the chance to get at least one good subsample of data points is larger if $k^* = d$. Next, the initial subsets of observations are traditionally chosen as random subsamples of size k^*. As this requires a large number of initial subsets to be drawn to obtain a good approximation and because the QR estimator used from Section 3 on possesses some robustness properties if there are no leverage points (cf. Giloni et al., 2006), we combine the random and a deterministic choice of initial subsamples. Specifically, we draw a number of initial subsamples of size k^* randomly, and additionally, n_{init} initial subsamples are taken as the ith observation and its (k^*-1)-nearest neighbors in the space of the explanatory variables, $i = 1,\ldots,n_{init}$. The algorithm could be further accelerated for large data sets by applying the partitioning and nesting techniques as in Rousseeuw and van Driessen (1999, 2006).

3. The Least Trimmed Quantile Regression Estimator

In this section, the Least Trimmed Quantile Regression (LTQR) estimator is introduced and the finite-sample BDP properties of the linear LTQR estimator are discussed. The LTQR estimator is a particular form of the GTE (3) that, contrary to many existing variants, employs a non-differentiable objective function

$f_i(\beta) = \rho_\tau(r_i(\beta))$. The LTQR estimator can thus be defined by

$$\hat{\beta}_n^k(\tau) := \arg\min_\beta \left\{ Q_{n,k}(\beta) = \min_{I \in I_k} \sum_{i \in I} \rho_\tau(r_i(\beta)) \right\}, \tag{5}$$

where I_k is the set of all k–subsets of the set $\{1, \ldots, n\}$, $\rho_\tau(r_i(\beta))$ is defined by (2), and $0 < \tau < 1$.

From this definition, it can be seen that the maximum LTQR estimator is the classical QR estimator calculated for some k-subset of the n cases. Consequently, the LTQR estimator includes the quantile regression estimator (2) as a special case for $k = n$, and the LTA estimator for $\tau = 0.5$. As the linear LTQR estimator is a particular case of the GTE, its finite-sample BDP can be derived from the finite-sample BDP of the GTE.

Theorem 1. *Let $\mathcal{N}(X) = \max_{0 \neq \beta \in \mathbb{R}^p} \text{card}\{i \in \{1, \ldots, n\}; x_i^T \beta = 0\}$. Then the BDP of the linear LTQR estimator equals $\frac{1}{n} \min\{n - k, k - \mathcal{N}(X) - 1\}$. For k such that $\lfloor \{n + \mathcal{N}(X) + 1\}/2 \rfloor \leq k \leq \lfloor \{n + \mathcal{N}(X) + 2\}/2 \rfloor$, the BDP attains its maximum and equals to $\frac{1}{n} \lfloor \{n - \mathcal{N}(X) - 1\}/2 \rfloor$.*

The quantity $\mathcal{N}(X)$, introduced by Müller (1995), provides the maximum number of explanatory variables $x_i \in R^p$ lying in a subspace. If any p observations x_i^T are linearly independent, then $\mathcal{N}(X) = p - 1$, which is the minimal value for $\mathcal{N}(X)$. When the covariates are qualitative variables such as factors with several levels, $\mathcal{N}(X)$ can be much larger.

As $\mathcal{N}(X)$ is bounded and independent of n, the most robust choice of trimming $k = \lfloor \{n + \mathcal{N}(X) + 1\}/2 \rfloor$ guarantees a BDP which will be asymptotically equal to $1/2$ and independent of τ. This is possible because LTQR estimates the quantiles only within the subset of observations that are not trimmed from the objective function, and as shown in the following section, it does not identify the intercept in model (1). On the other hand, the proof of Theorem 1 (in the Appendix) shows that the size of the compact set containing the LTQR estimates in the presence of contamination does depend on τ by means of $\min\{\tau, 1 - \tau\}^{-1}$. Although the BDP can reach $1/2$ for any τ, the maximum bias caused by contamination will be smallest for $\tau = 1/2$, it will increase as τ moves away from $1/2$, and could be arbitrarily large if one requires $\tau \to 0$ or $\tau \to 1$.

4. Consistency of the LTQR estimator

Here it will be shown that the LTQR estimator (5) is a consistent estimator of the slope parameters in model (1). Moreover, the constant identified by the LTQR estimator will be found.

Let us now assume for the sake of simplicity that the distribution function F of the error term ε_i in (1) has an infinite support. Further, as the trimming constant k defining the LTQR estimator generally depends on the sample size n, we will write k_n to indicate this and assume $\lim_{n \to \infty} k_n/n = \lambda \in (0, 1\rangle$ exists. In the location model, that is, in model (1) containing only the constant term, Tableman (1994a) then showed that the LTQR estimator with $\tau = 0.5$ identifies the median on the shortest interval Δ such that $P(y_i \in \Delta) = \lambda$. To formalize this statement in the general case, let us first state assumptions on the data generating process.

Assumption D. The vectors (x_i, ε_i) form a sequence of independent and identically distributed random vectors with the finite $(1 + \delta)$th moments for some $\delta > 0$.

Assumption F. Let the distribution function F be continuous, strictly increasing on its support, and having a differentiable density function f, which is supposed to be unimodal and bounded on its support.

Let us recall that, for an interval $\Delta(a, \lambda) = \langle F^{-1}(a), F^{-1}(a + \lambda)\rangle$, $a \in (0, 1 - \lambda)$, and a fixed $\tau \in (0, 1)$, Tableman (1994a, p. 390) proved in the location model that LTA applied to univariate data following the distribution function F converges to and thus consistently estimates

$$\mu^*(\tau) = F^{-1}(a^*(\tau) + \tau\lambda), \tag{6}$$

where $a^*(\tau) = \arg\min_{a \in (0,1)} \int_{\Delta(a,\lambda)} \rho_\tau(\varepsilon - F^{-1}(a + \tau\lambda)) dF(\varepsilon)$.

Assuming that the intercept is the first element of the parameter vector β, we will now show in the regression case (1) that the LTQR estimator consistently estimates the parameter vector $\beta^*(\tau) =$

$(\mu^*(\tau), 0, \ldots, 0)^T + \beta^0$, where the parameter $\beta^*(\tau)$ obviously equals to β^0 for all its elements, but the first one. The constant term obtained by the LTQR estimator thus corresponds to $\beta_1^0 + \mu^*(\tau)$, where in general $\mu^*(\tau) \neq 0$ ($\mu^*(\tau) = 0$ if F is symmetric and $\tau = 1/2$, for instance).

Theorem 2. *Let Assumptions D and F hold and let $\tau \in (0, 1)$ be fixed. Assuming $\beta^0 \in B$ and $\beta^*(\tau) = (\mu^*(\tau), 0, \ldots, 0)^T + \beta^0$, where $\mu^*(\tau)$ is defined in (6) and B is a compact parametric space, the LTQR estimator defined for $k_n = [\lambda n]$ and $\lambda \in (0, 1\rangle$ consistently estimates $\beta^*(\tau)$, $\hat{\beta}_n^{k_n}(\tau) \to \beta^*(\tau)$ in probability as $n \to \infty$.*

The theorem shows that, under Assumption F, the LTQR estimator correctly identifies the coefficients of the regression variables, but provides a different estimate of the constant term. To obtain the intercept term representing the classical τth quantile, one can use the residuals $r_i(\hat{\beta}_n^{k_n}(\tau))$ from the LTQR estimator fit, compute their empirical τth quantile q_τ, and add q_τ to the LTQR estimator intercept estimate. A possible caveat of such a procedure is its robustness: this newly estimated intercept has (asymptotically) a BDP bounded by $\min\{\tau, 1 - \tau\}$, which is irrelevant for τ close to 0.5, but rather limiting for quantiles τ close to 0 or 1.

5. Examples

Since the trial and refinement steps of the GTE-LTQR algorithm are standard quantile regression procedures, the GTE algorithm can be easily implemented using widely available software. We illustrate this using the package *quantreg* of Koenker (2005), which was developed in R (R Development Core Team, 2011). In particular, we first compare the performance of classical linear quantile regression and the LTQR estimator through a real dataset and a simulation study. Later, some robustness properties of LTQR and existing robust methods are compared.

5.1. Star cluster CYB OB1 dataset

First, the well-known dataset on the star cluster CYB OB1 consisting of 47 observations is considered, which was already analyzed by Adrover et al. (2004) and Rousseeuw and Leroy (1987). In the upper left corners of the plots of Figure 1 one can see four points with high leverage that do not follow the trend of the data majority. The observations are plotted as tiny black bullets on all of the plots. Here we focus on estimating the regression quantiles τ of 0.25, 0.50, and 0.75 by both the classical QR estimator proposed by Koenker and Bassett (1978) and by the LTQR estimator using different trimming percentages. The upper plots show the results of the classical estimator for all data points (upper left) and for a reduced dataset where the four leverage points are deleted (upper right). It is evident that the leverage points have a strong influence on the classical estimator.

The remaining plots in Figure 1 show the results of the LTQR estimator on the original data, with 4%, 9%, 11%, and 17% trimming. This corresponds to trimming 2, 4, 5, and 8 observations, respectively. The trimmed observations are marked by symbols: tiny squares for $\tau = 0.25$, upside-down triangles for $\tau = 0.50$, and normal triangles for $\tau = 0.75$. The corresponding LTQR regression lines are influenced by the leverage points in case the trimming percentage is too low (4%). For 9% trimming the four leverage points are identified as outliers and we obtain practically the same result as for the classical method applied to the reduced data. If the trimming percentage is chosen higher (11%, 17%), additional observations are identified as outliers, but the regression lines are very stable. It is interesting to see that not always the same additional observations are trimmed: this depends on the considered regression quantile τ. This phenomenon is corresponding to the definition of regression outliers, where observations that do not follow the assumed model can be treated as outliers. We can also see that even the LTQR fits for $\tau = 0.75$ with 11% and 17% of trimming are not influenced by the outliers like, for example, the maximum depth estimator in Adrover et al. (2004, Figure 2). The LTQR regression lines are similar to those of the robustified Koenker and Bassett (RobKB) method in Adrover et al. (2004, Figure 1), but look more plausible and stable because the LTQR median regression lines will intersect both with the $\tau = 0.25$ and $\tau = 0.75$ fitted lines for large values of the covariate, whereas the RobKB median regression line intersects with the $\tau = 0.75$ fitted line for small covariate values and with the $\tau = 0.25$ fitted line for large covariate values.

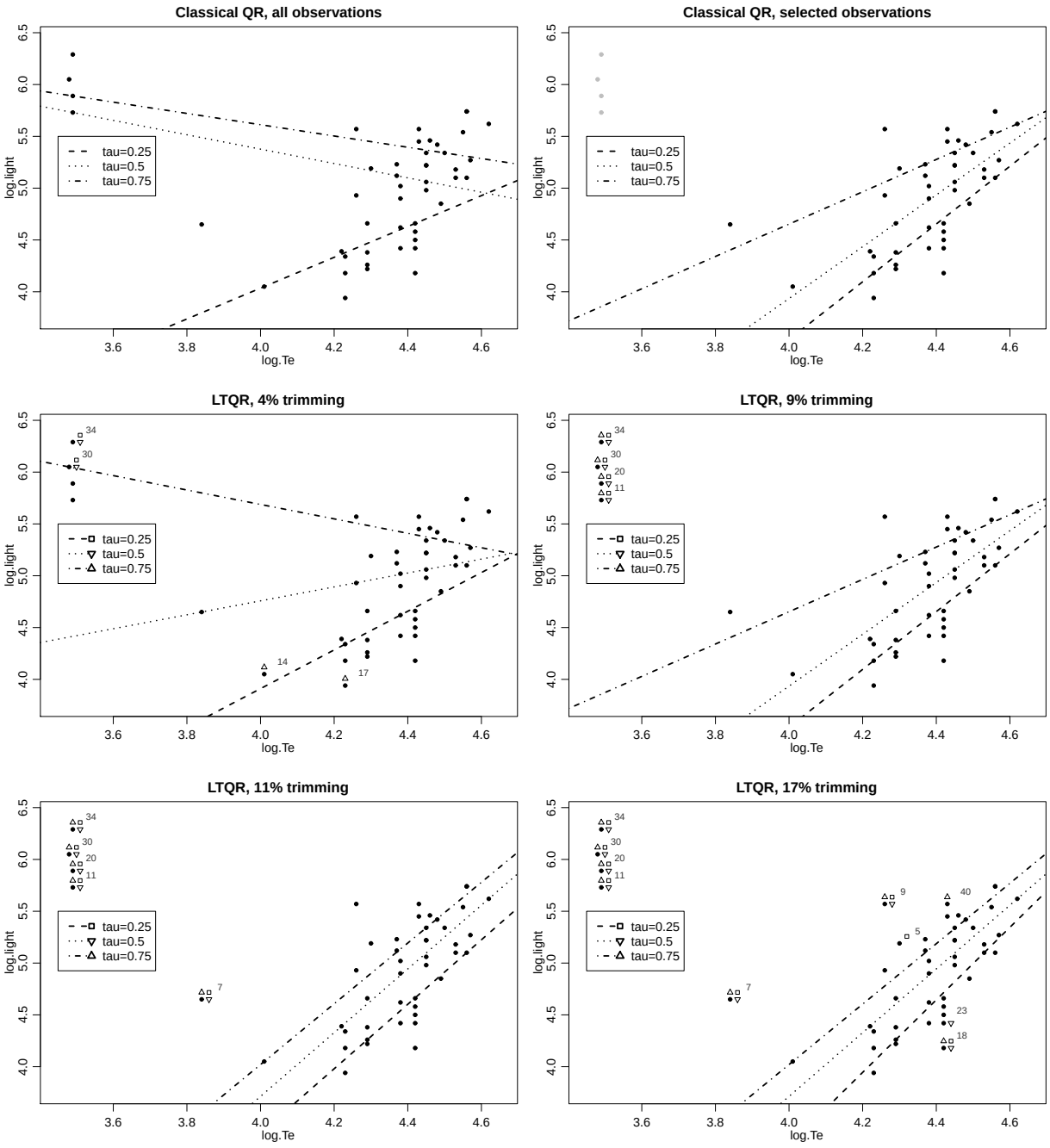

Figure 1: Star data: 0.25, 0.50, and 0.75 regression quantiles from Koenker and Bassett estimate, based on whole data (upper left) and on data without the four extreme points (upper right); LTQR fits with 4%, 9%, 11% and 17% of trimming (remaining plots).

5.2. Simulation experiments

We compare the performance of the QR and LTQR estimators through a simulation study within the classical heteroscedastic multiple linear regression model. The data were generated according to the model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \sigma_i \varepsilon_i \quad \text{for} \quad i = 1, \ldots, 100, \\ \sigma_i &= \sqrt{\exp(0.11(x_{i1} + x_{i2}))}, \end{aligned}$$

where $\beta_0 = \beta_1 = \beta_2 = 0$ can be chosen without loss of generality because of the regression equivariance of LTQR, $\varepsilon_i \sim \mathrm{N}(z_\alpha, 1)$, and z_α is the α–quantile of N(0,1). (Note that this traditional form of heteroscedasticity implies a slight nonlinearity of the QR regression lines for $\tau \neq 0.5$.)

Two distribution types for the covariates are considered: in the *1st experiment*, the covariates x_{i1} and x_{i2} are uniformly distributed on the interval $[-10, 10]$, that is, $x_{ij} \sim \mathrm{U}[-10, 10]$ for $j = 1, 2$; in the *2nd experiment*, the covariates x_{i1} and x_{i2} are normally distributed, that is, $x_{ij} \sim \mathrm{N}(0, 1)$ for $j = 1, 2$. Data contamination is introduced by modifying the first $m = \lfloor 100\epsilon \rfloor$ observations for $\epsilon = 0.1, 0.2, 0.3$ as follows ($r = 2, 3, 4$): in the *1st experiment*, $x_{ij} \sim \mathrm{U}[-30, -20]$ for $j = 1, 2$ and $y_i \sim \mathrm{U}[-10r, -10r + 10]$; in the *2nd experiment*, $(x_{i1}, x_{i2}, y_i)^T \sim \mathrm{N}_3(\mu, \Sigma)$ where $\mu = (-10, -10, -10r)^T$ and $\Sigma = 3I_3$ for $i = 1, \ldots, m$. In this way all those generated outliers are bad leverage points of different magnitude. As the results are similar across different choices of r, we present the *1st experiment* with $r = 2$ and the *2nd experiment* with $r = 3$.

All simulation experiments were replicated 1000 times to explore the small sample behavior of the classical QR and LTQR estimators for the different quantile values $\tau = (0.5, 0.75, 0.90)$ and different trimming percentages over the clean and contaminated data. Subsequently, the simulated estimates were obtained and summarized in boxplots, see Figures 2–9. The plot panels for the upper rows of the figures show the results for the intercept term β_0, while the middle and bottom rows present the results of the slope parameters β_1 and β_2, respectively. The "correct" trimming percentages are indicated by dotted vertical lines, and the true simulated parameters are indicated by horizontal dashed lines.

Figures 2 and 3 demonstrate the performance in the uncontaminated case for both uniformly and normally distributed regressors (QR corresponds to 0% trimming). One can see that, when increasing the trimming percentage, the intercept estimates are unbiased for $\tau = 0.5$ (the error distribution is symmetric), but the bias for the regression quantiles $\tau = 0.75$ and $\tau = 0.90$ increases with the amount of trimming. The reason for this effect was given in Section 4, where we noted that LTQR identifies the sum of the intercept and $\mu^*(\tau) = F^{-1}(a^*(\tau) + \tau\lambda)$ (see equation (6)), which depend both on the quantile τ and the amount of trimming $\lambda = \lim_{n\to\infty} k_n/n$. The estimates of the slopes are presented on the lower plot panels. Both QR and LTQR estimators are unbiased in agreement with the theory and perform well, although the variability of the estimates increases for larger amounts of trimming. This is caused by LTQR using less and less observations due to a higher amount of trimmed data points.

Figures 4–9 present the results for the *1st* and *2nd experiment* corresponding to an increasing proportion of outliers $\epsilon = 0.1, 0.2, 0.3$. When choosing the same trimming percentage as the contamination level, the resulting estimates are very precise – comparable to the uncontaminated case. Similarly, if the trimming is chosen higher than the contamination level (i.e., $1 - \lambda \geq \epsilon$), we observe essentially the same picture as for the uncontaminated case. On the other hand, the use of smaller trimming percentages (i.e., $1 - \lambda < \epsilon$) has an immediate effect on the quality of the estimates and this becomes more severe for high contamination levels. In such cases, both bias and variance of the estimates increase dramatically because the resulting procedure has not sufficient robustness.

Further, these boxplots on Figures 4–9 also reveal that the variation in any panel depends on the chosen trimming percentage. In general, the smallest variation is obtained by choosing the exact trimming percentage corresponding to the contamination level in the data. In practice, it is preferable to be conservative, and in case of doubts, choose a higher trimming proportion than necessary (and thus a bit higher variance of estimates) to prevent a substantial bias caused by the lack of robustness.

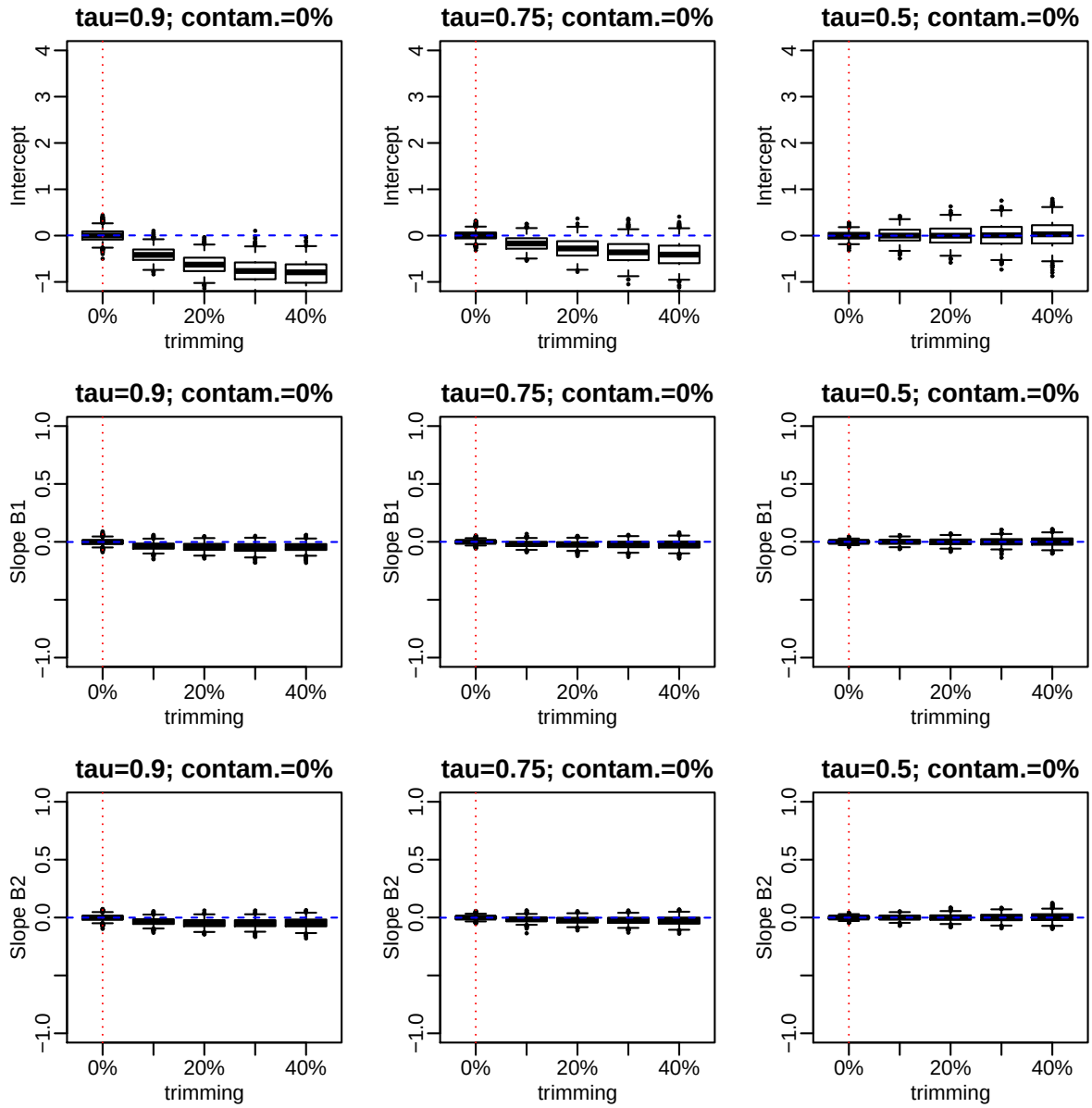

Figure 2: Boxplots of the estimates based on the originally generated data (0% contamination) and uniformly distributed covariates.

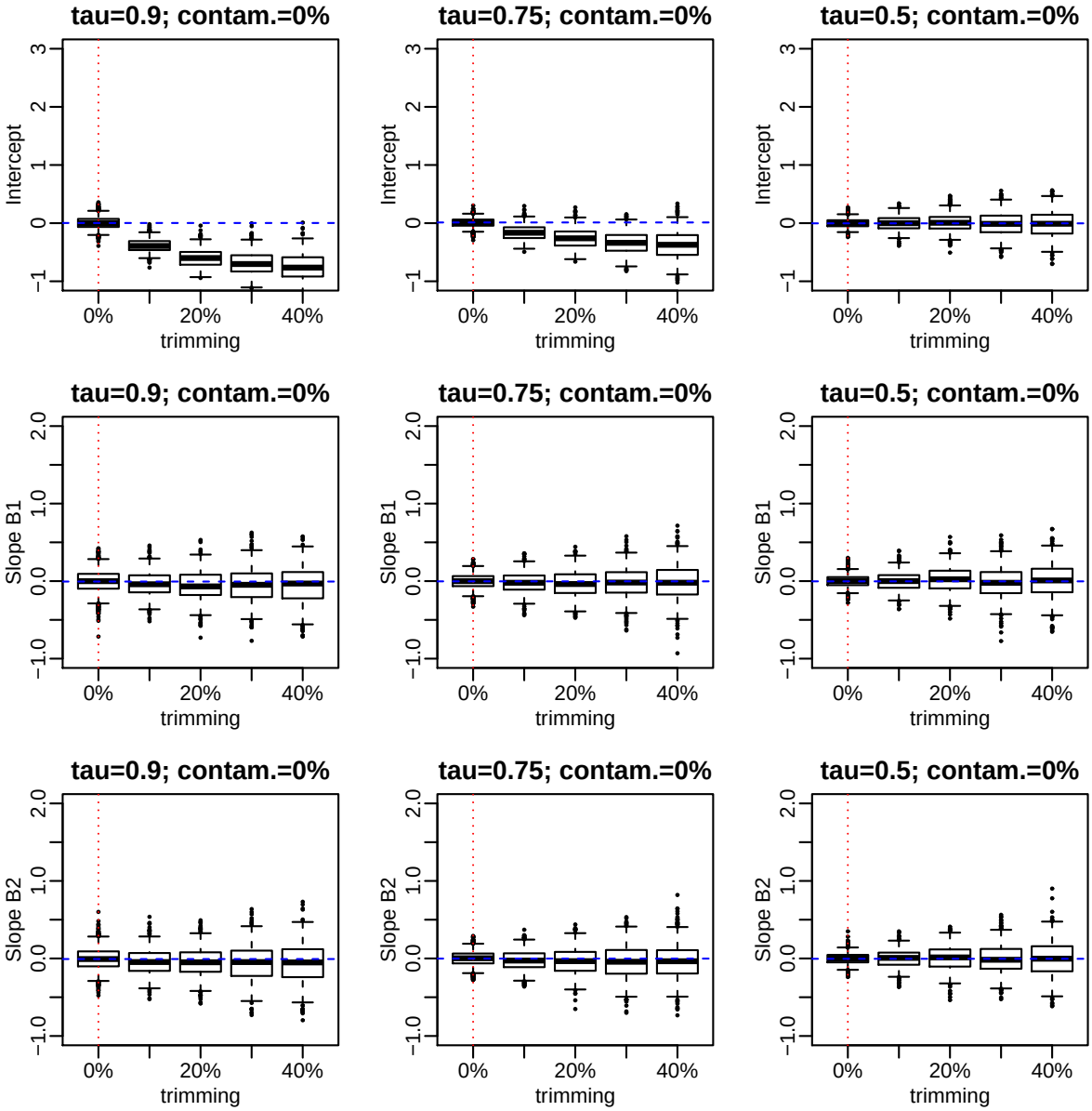

Figure 3: Boxplots of the estimates based on the originally generated data (0% contamination) and normally distributed covariates.

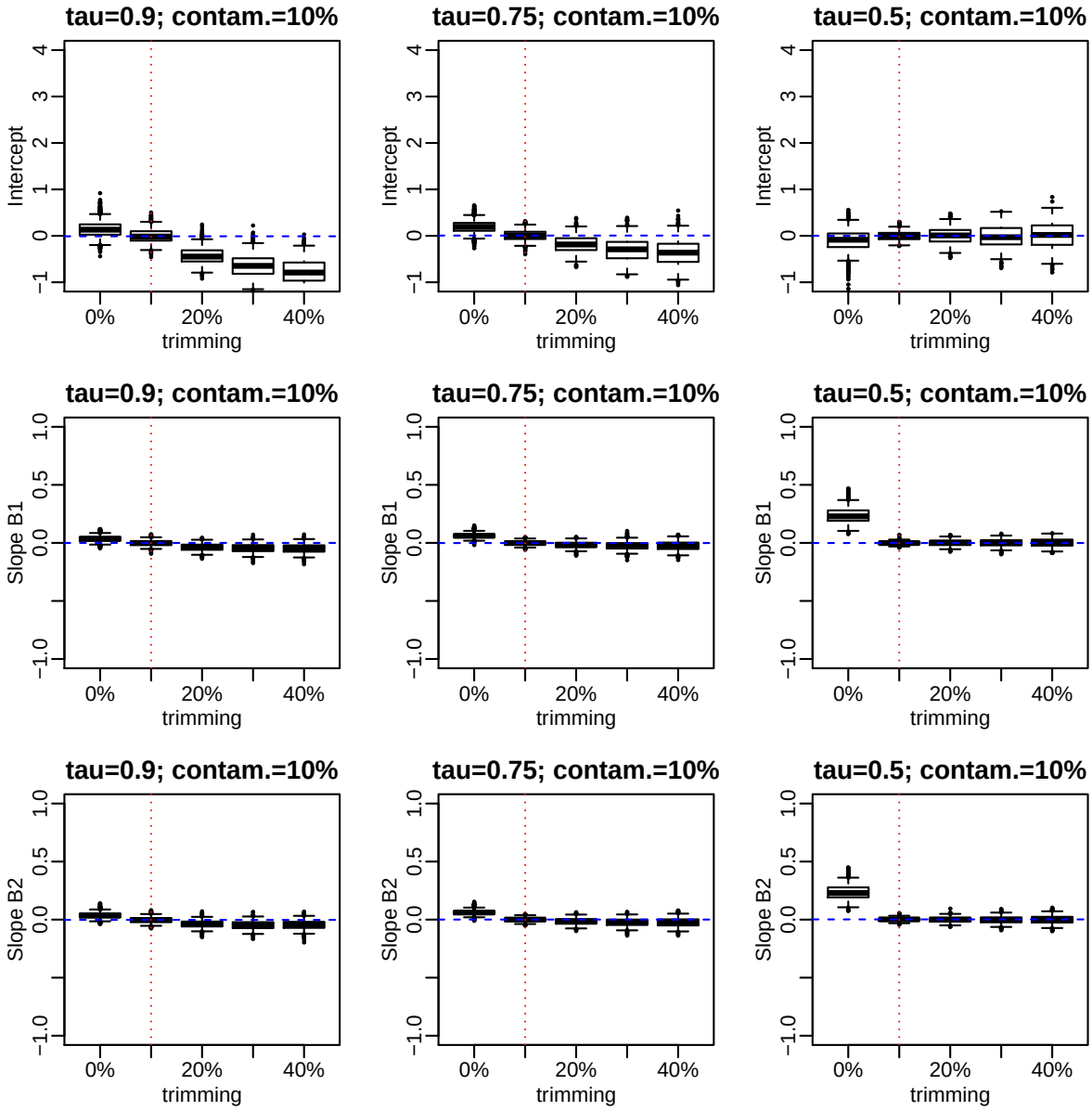

Figure 4: Boxplots of the estimates for the *1st experiment* with 10% contamination and uniformly distributed covariates.

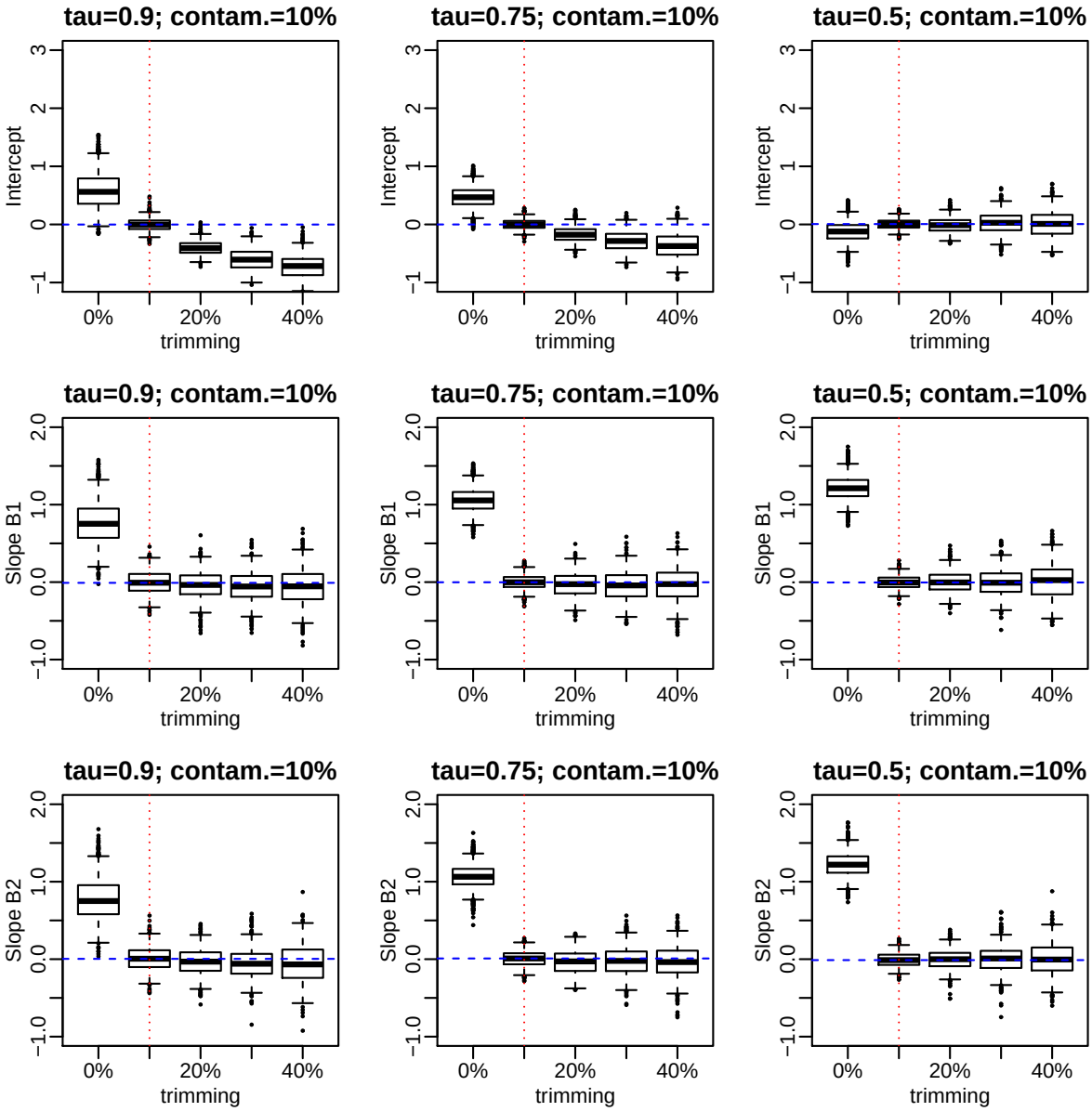

Figure 5: Boxplots of the estimates for the *2nd experiment* with 10% contamination and normally distributed covariates.

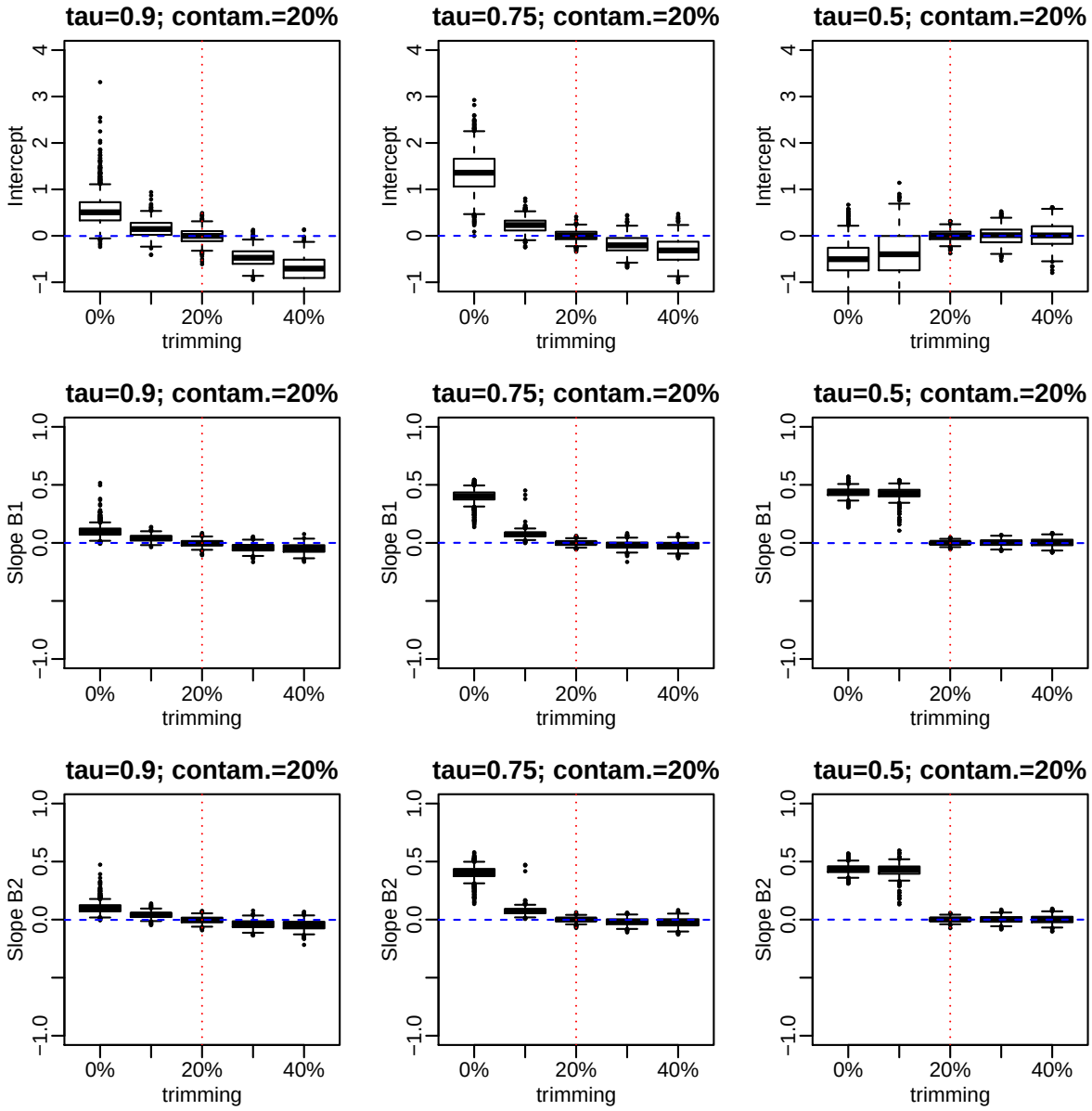

Figure 6: Boxplots of the estimates for the *1st experiment* with 20% contamination and uniformly distributed covariates.

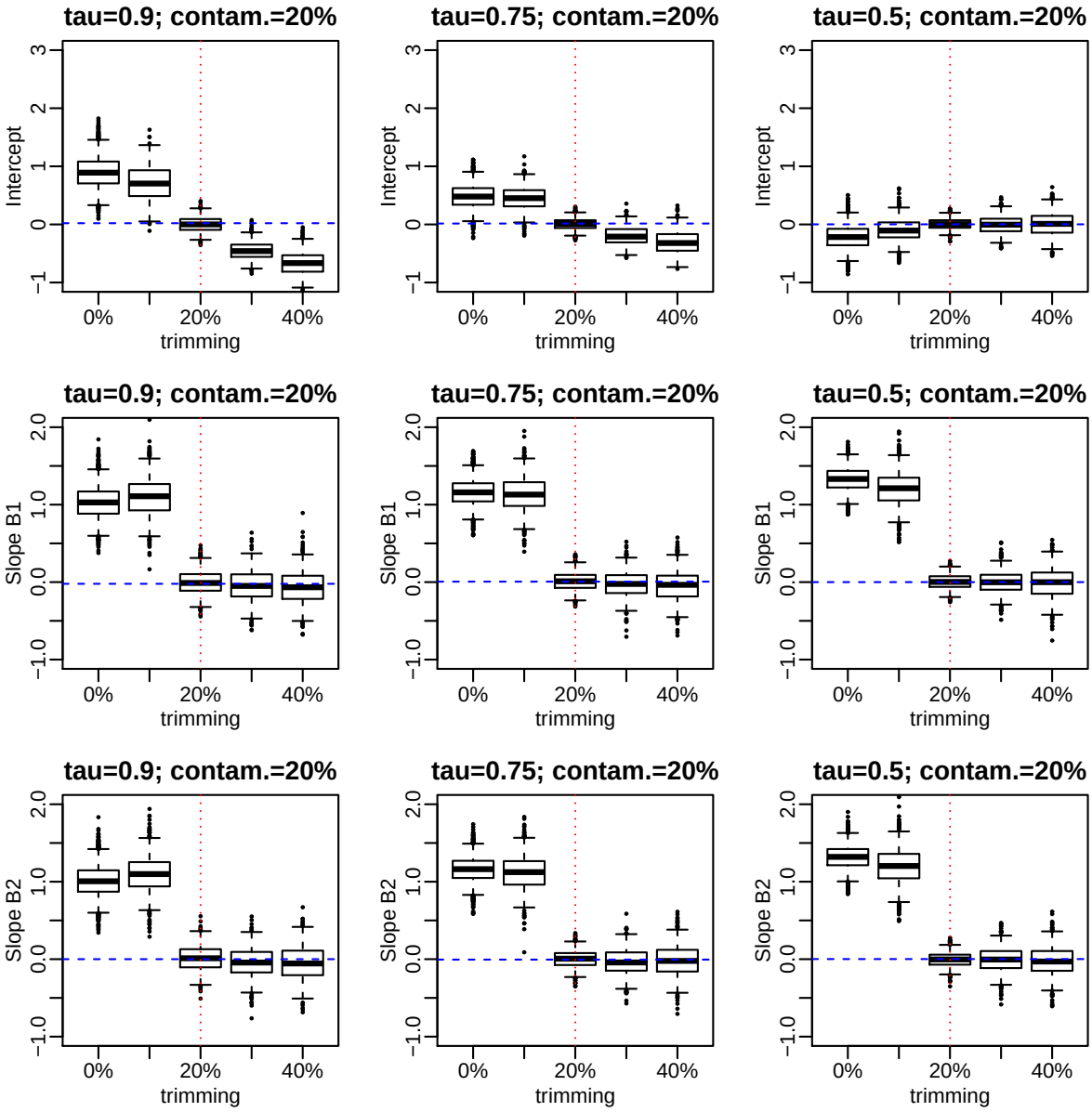

Figure 7: Boxplots of the estimates for the *2nd experiment* with 20% contamination and normally distributed covariates.

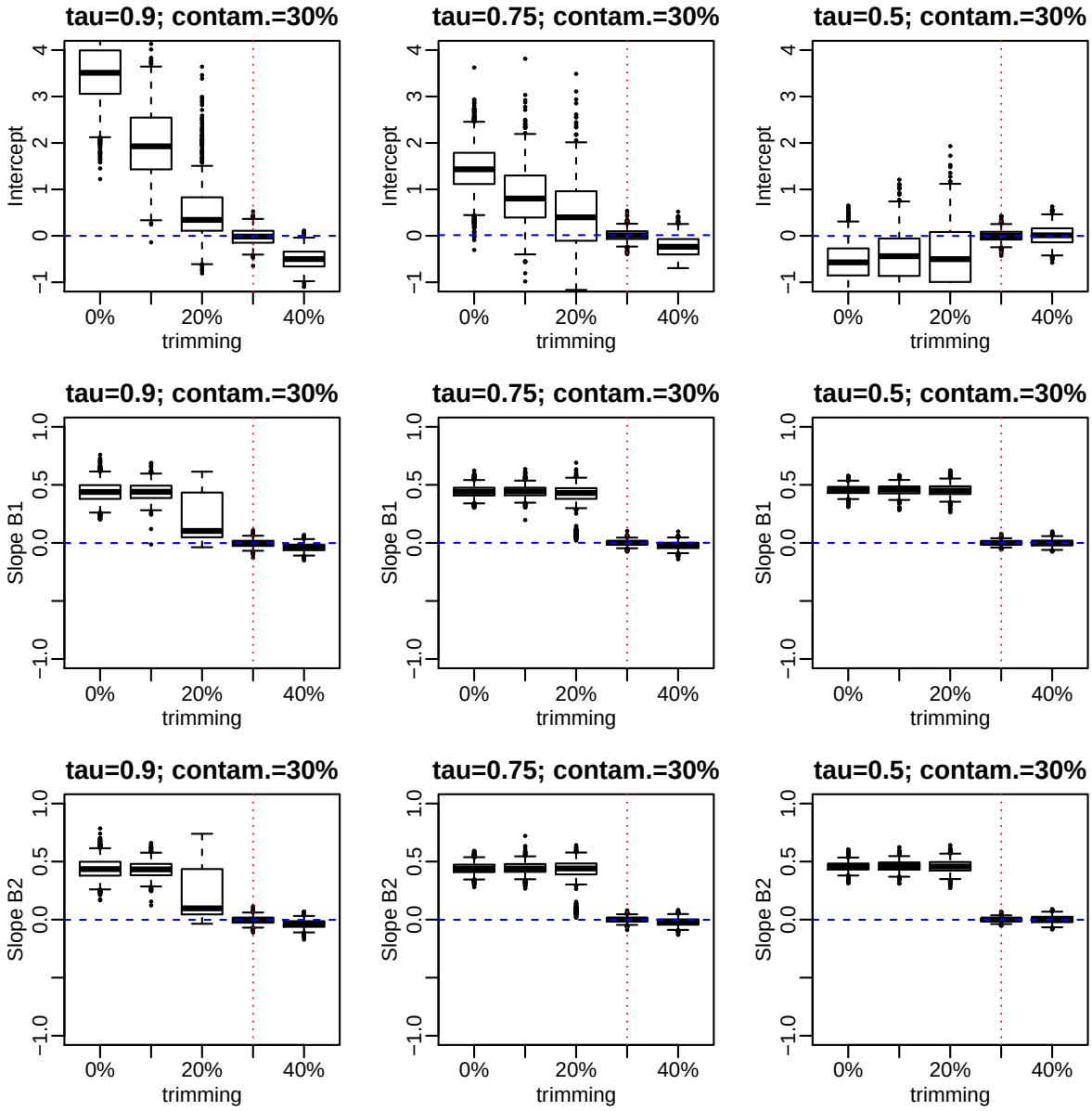

Figure 8: Boxplots of the estimates for the *1st experiment* with 30% contamination and uniformly distributed covariates.

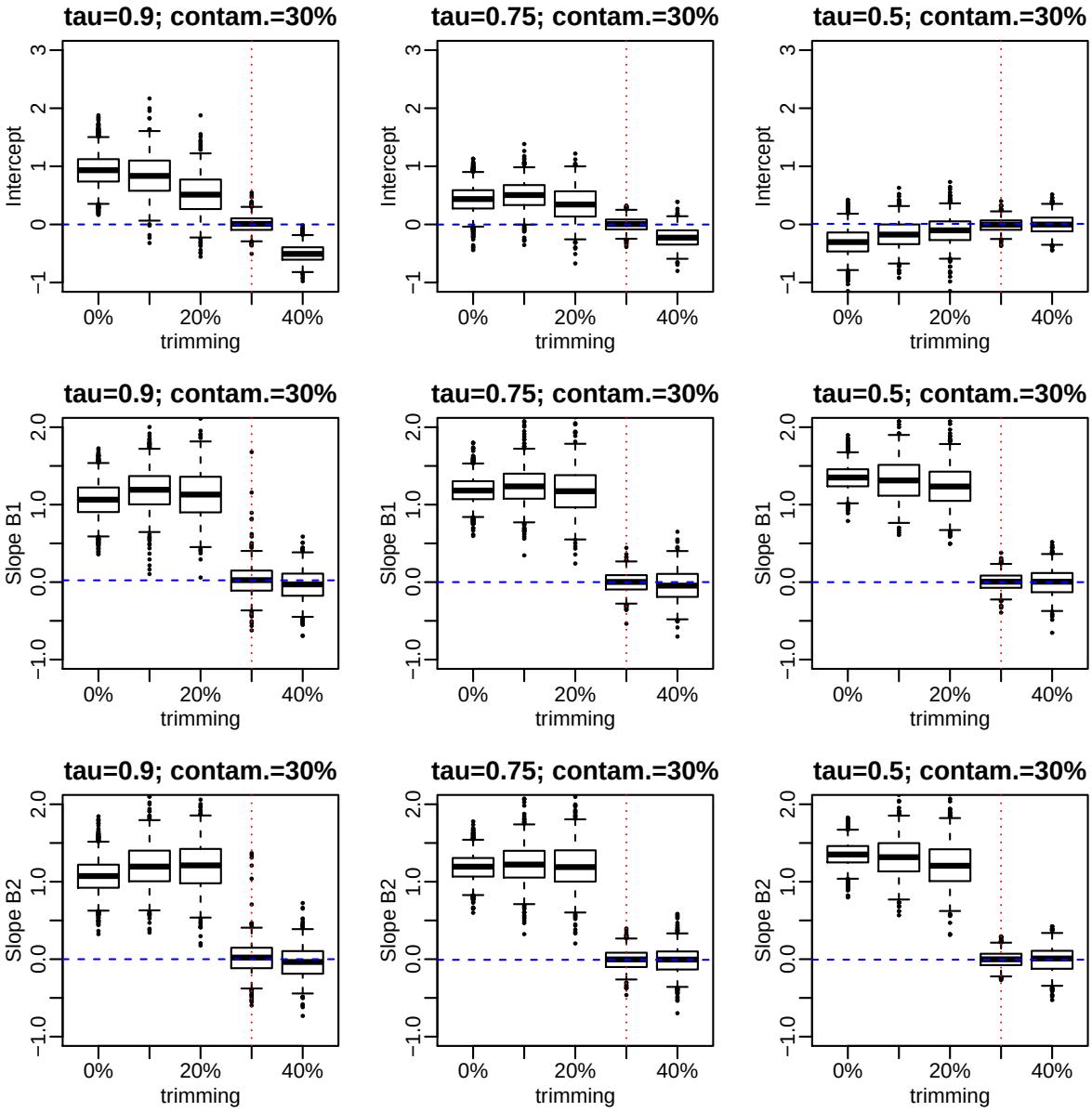

Figure 9: Boxplots of the estimates for the *2nd experiment* with 30% contamination and normally distributed covariates.

The percentage of outliers in real data is of course unknown or can be only roughly estimated. A technique for automatically selecting the trimming parameter $k_n = [\lambda n]$ or the trimming percentage $[(1-\lambda)n]100\%$ can be developed for LTQR in a straightforward way by mimicking the procedure of Čížek (2010) for the (reweighted) LTS regression estimator, which is in turn an adaptation of the method of Gervini and Yohai (2002). One only needs to take the asymmetric Laplace distribution instead of the normal one.

5.3. Comparison with other robust regression quantile estimators

In order to study the small sample behavior of the proposed estimator and to compare it with some robust regression quantile estimators considered by Rousseeuw and Hubert (1999) and Adrover et al. (2004), we estimated the maximum effect of the point contamination on the LTQR estimator in terms of the mean squared errors. The simulation experiment closely follows that of Adrover et al. (2004). Each experiment consists of $n = 50$ data points. The data generation is based on a multiple linear regression model $y_i = \beta_0^T \mathbf{x}_i + u_i$ for $i = 1, \ldots, n$, where $\mathbf{x}_i^T := (1, x_{i1}, \ldots, x_{i,p-1})$, the $p-1$ covariates follow independent unit normals, $u_i \sim N(z_\alpha, 1)$, z_α is the α–quantile of $N(0,1)$, and $\beta_0 = 0$, which can be chosen without loss of generality because of the regression equivariance of LTQR. The point contamination is introduced via replacement of the last $m = \lfloor \epsilon n \rfloor$ observations by the following outliers: $y_i := y_0 = 5b$ and $\mathbf{x}_i^T := \mathbf{x}_0^T = (1, 5\mathbf{e}_1^T) \in R^p$ for $i = n-m+1, \ldots, n$ for various values of ϵ (see Table 1), where $\mathbf{e}_1$ is the first element of the canonical basis of R^{p-1}. As in Adrover et al. (2004), the contamination slope b varies over a large grid from 0 to 10 with step 0.1 in order to search for the least favorable situations, and the experiment was performed for the number of regressors $p = 2$ and $p = 5$. Each simulation experiment was replicated 500 times. The trimming parameter k_n of the LTQR estimator is set equal to $\lfloor (1-\epsilon)n \rfloor$ and also to a slightly lower value $\lfloor (1 - \epsilon - 0.1 * (1-\tau))n \rfloor$. The maximum mean of the squared errors $\|\hat{\beta}_n^{k_n} - \beta_0\|^2$ (MaxMSE) is used as an error criterion and its values for different quantiles are presented in Table 1.

In the case of LTQR, MaxMSE is computed for the slope coefficients as well as for all coefficients including intercept: as LTQR does not consistently estimate the intercept, the MaxMSE computed for the whole coefficient vector necessarily reflects also its intercept bias, which is not directly related to the bias due to contamination. Additionally, we include results for the LTQR using larger percentages of trimming $1-\lambda = \epsilon + 0.1*(1-\tau)$ than the contamination levels. This is done because in practice the exact contamination level is unknown (even though it can be estimated as discussed in the previous section) and thus a more realistic procedure is to use a larger trimming than expected contamination. These values for LTQR can be compared to the results for the classical Koenker–Bassett (K-B), the robustified Koenker–Bassett (RobKB) estimator of Adrover et al. (2004), and other alternatives such as the maximum depth estimator (MaxDep) of Rousseeuw and Hubert (1999). We implemented those estimators, but – because of the limited information concerning the estimation algorithms – we also report the original results of Adrover et al. (2004), who report the median of the squared errors. The FORTRAN software developed by Van Aelst et al. (2002) was used to handle the MaxDep computations.

As the LTQR estimator for $1 - \lambda = 0$, that is for $k_n = n$, reduces to the classical Koenker–Bassett estimator, its values for no contamination case $\epsilon = 0$ can be used as a reference for the mean squared errors of QR, see Table 1. For the positive levels of contamination, the LTQR performance is proportional to the level of contamination ϵ, but does not depend substantially on the estimated quantile. This is most visible if the contamination level $\epsilon \in (0.10, 0.12)$ is considered: the MaxMSE of LTQR increases only by 50% if $\tau = 0.50$ grows to 0.90. On the other hand, the most robust alternative RobKB (cf. Adrover et al., 2004) increases its bias $\epsilon \in (0.10, 0.12)$ by 50–100% if $\tau = 0.50$ changes to 0.75 and grows above any bound if $\tau = 0.90$ as it reaches its breakdown point 10% at this point. Thus, LTQR generally outperforms the other methods for $\tau > 0.5$ and higher contamination levels as a consequence of its breakdown point independent of the actual quantile (although BDP itself does not quantify the bias of an estimator). Additionally, LTQR performs better than competing methods at higher quantiles such as $\tau = 0.90$ irrespective of the contamination level. On the other hand, RobKB and MaxDep perform better for quantiles closer to $\tau = 0.50$ and lower contamination levels. This can be expected especially in the case of RobKB as methods based on the pairwise differences or comparisons of observations typically outperform the corresponding methods minimizing plain sums of functions of individual observations (e.g., compare least trimmed squares and least trimmed differences estimator of Stromberg et al., 2000).

Table 1: The Monte Carlo maximum mean squared errors based on the LTQR estimator of the slope (LTQR 'no intercept'), LTQR estimator of the intercept and slope (LTQR with 'intercept'), maximum depth estimator (MaxDep), robustified Koenker-Bassett (RobKB) estimator, and Koenker-Bassett (K-B) estimator in samples of $n = 50$ observations. For LTQR, the trimming parameter equals $k_n = \lfloor \lambda n \rfloor$, where the fraction of trimmed observations $1 - \lambda$ equals exactly to the contamination level ϵ or is slightly larger $\epsilon + \delta$, where $\delta = 0.1 * (1 - \tau)$.

MaxMSE			LTQR [trimming $1-\lambda$]				MaxDep	RobKB	K-B	MaxDep	RobKB
			no intercept		intercept					results of	
p	τ	ϵ	$[\epsilon]$	$[\epsilon+\delta]$	$[\epsilon]$	$[\epsilon+\delta]$				Adrover et al. (2004)	
2	0.50	0.00	0.04	0.05	0.08	0.09	0.10	0.09	0.07	0.10	0.10
		0.10	0.45	0.39	0.53	0.48	0.19	0.18		0.21	0.20
		0.20	1.03	0.94	1.20	1.11	0.84	0.81		0.79	0.97
	0.75	0.00	0.05	0.06	0.09	0.12		0.11	0.07	0.10	0.11
		0.06	0.28	0.25	0.34	0.31		0.16		0.24	0.21
		0.12	0.66	0.55	0.76	0.64		0.38		1.82	0.57
		0.20	1.07	0.94	1.24	1.10		1.43			
	0.90	0.00	0.07	0.08	0.15	0.15		0.16	0.11	0.14	0.14
		0.02	0.15	0.12	0.22	0.19		0.17		0.23	0.16
		0.04	0.25	0.20	0.31	0.28		0.26		0.77	0.26
		0.08	0.53	0.42	0.61	0.52		1.16			
		0.10	0.64	0.54	0.74	0.63		117.			
5	0.50	0.00	0.17	0.20	0.21	0.24	0.26	0.29	0.19	0.37	0.38
		0.10	0.78	0.77	0.87	0.87	0.65	0.61		0.74	0.66
		0.20	2.64	2.65	2.93	2.95	4.14	2.89		4.60	2.40
	0.75	0.00	0.20	0.23	0.25	0.29		0.38	0.21	0.36	0.36
		0.06	0.54	0.48	0.61	0.55		0.46		0.77	0.54
		0.12	1.22	1.10	1.33	1.21		0.89		2.87	1.70
		0.20	2.91	2.78	3.24	3.07		4.13			
	0.90	0.00	0.30	0.31	0.37	0.42		0.46	0.34	0.48	0.38
		0.02	0.40	0.34	0.47	0.45		0.52		0.71	0.44
		0.04	0.54	0.46	0.61	0.56		0.66		3.00	0.87
		0.08	0.98	0.81	1.08	0.93		2.86			
		0.10	1.24	1.04	1.35	1.17		119.			

6. Summary and conclusions

A robust version of the linear quantile regression estimator is introduced which is based on the idea of trimming. The breakdown point and the consistency of the proposed estimator are characterized. The computation of the estimator is taking advantage of the same technology as used for its classical counterpart, but here the estimation is based on subsamples only. The used algorithm consisting of a trial and a refinement step (Neykov et al., 2012) follows the ideas of the FAST-LTS and FAST-MCD algorithms of Rousseeuw and van Driessen (1999, 2006) and Neykov and Müller (2003). The new estimator generally performs very well, which is confirmed by an example, by simulation studies, and by a comparison to other proposals.

An important choice for estimators based on trimming is the trimming percentage. In the numerical experiments, it has been shown that a trimming percentage lower than the contamination level can lead to very poor estimates, but any higher trimming percentage gives very reasonable results. Therefore, a general rule is to work with a conservative choice of the trimming percentage or to estimate the amount of trimming similarly to Čížek (2010) and Gervini and Yohai (2003).

R code of our method will soon be made available as a CRAN package.

Acknowledgment

The authors are very grateful to Prof. P. Rousseeuw for his valuable comments on an earlier draft of the paper and Prof. R. Maronna for providing us by the GAUSS code of the simulation design in Adrover et al. (2004). We would like to thank the associate editor and two anonymous referees for their constructive comments on our work that greatly improved the presentation of these results. N. Neykov and P. Neytchev are thankful to the Vienna University of Technology and the ESF (COST Action IC0702) for supporting their stay in Vienna during June 2011.

Proofs

The BDP will be derived using the d-fullness technique proposed by Vandev (1993). According to Vandev and Neykov (1998), the set $F = \{f_i(\theta); i = 1, \ldots, n\}$ is called d-full if the function $g(\theta) = \max_{j \in J} f_j(\theta)$, $\theta \in \Theta$, is subcompact for every subset $J \subset \{1, \ldots, n\}$ of cardinality d. A function $g : \Theta \to \mathbb{R}$, $\Theta \subseteq \mathbb{R}^q$, is called subcompact if its Lebesgue set $L_g(C) = \{\theta \in \Theta : g(\theta) \leq C\}$ is contained in a compact set for every real constant C. The d-fullness index ensures the existence of a solution and provides positive BDP of the optimization problem (3) at any subset of functions with size $k \geq d$.

Proof of Theorem 1: As the linear LTQR estimator is a particular case of the GTE, its finite-sample BDP equals to $\frac{1}{n} \min\{n-k, k-d\}$, provided the set of functions $F = \{\rho_\tau(r_i(\beta)); i = 1, \ldots, n\}$ is d-full, Dimova and Neykov (2004). Now we are ready to show that the set F is $d = \mathcal{N}(X) + 1$-full for any fixed $\tau \in (0, 1)$, where $\mathcal{N}(X)$ is defined as $\mathcal{N}(X) = \max_{0 \neq \beta \in \mathbb{R}^p} \text{card}\{i \in \{1, \ldots, n\}; x_i^T \beta = 0\}$. Let D be an arbitrary constant and

τ be fixed. Then for any subset $J \subset \{1,\ldots,n\}$ of cardinality $\mathcal{N}(X)+1$, the set

$$
\begin{aligned}
& \{\beta \in R^p : \max_{j\in J} \rho_\tau(x_j^T\beta - y_j) \le D\} \\
= & \left\{\beta \in R^p : \max_{j\in J} \left[\left|x_j^T\beta - y_j\right| \left(\tau 1_{\{x_j^T\beta - y_j \ge 0\}} + (1-\tau)\, 1_{\{x_j^T\beta - y_j < 0\}}\right)\right] \le D\right\} \\
\subseteq & \left\{\beta \in R^p : \min(\tau, 1-\tau) \max_{j\in J} \left|x_j^T\beta - y_j\right| \le D\right\} \\
= & \left\{\beta \in R^p : \max_{j\in J} \left|x_j^T\beta - y_j\right| \le \frac{D}{\min(\tau, 1-\tau)}\right\} \\
\subseteq & \left\{\beta \in R^p : \max_{j\in J} \left(\left|x_j^T\beta\right| - |y_j|\right) \le \frac{D}{\min(\tau, 1-\tau)}\right\} \\
\subseteq & \left\{\beta \in R^p : \max_{j\in J} \left|x_j^T\beta\right| \le \frac{D}{\min(\tau, 1-\tau)} + \max_{j\in J} |y_j|\right\} \\
\subseteq & \left\{\beta \in R^p : \frac{1}{N(X)+1}\beta^T \sum_{j\in J} x_j x_j^T \beta \le \left[\frac{D}{\min(\tau, (1-\tau))} + \max_{j\in J} |y_j|\right]^2\right\}
\end{aligned}
$$

is contained in a compact set. Indeed, the last set is bounded because J is of cardinality $\mathcal{N}(X)+1$ and the definition of $\mathcal{N}(X)$ implies that the matrix $\sum_{j\in J} x_j x_j^T$ has full rank. This last set is closed as the quadratic form $\beta^T \sum_{j\in J} x_j x_j^T \beta$ is a continuous function in β. Hence it is compact because it is closed and bounded.

Therefore, the BDP of the linear LTQR estimator is $\frac{1}{n}\min\{n-k, k-\mathcal{N}(X)-1\}$. This BDP is maximized for $\lfloor\{n+\mathcal{N}(X)+1\}/2\rfloor \le k \le \lfloor\{n+\mathcal{N}(X)+2\}/2\rfloor$ and equals to $\frac{1}{n}\lfloor\{n-\mathcal{N}(X)-1\}/2\rfloor$. □

Proof of Theorem 2: Let $Q_\lambda(\beta) = E[\rho_\tau(r_i(\beta)) \cdot 1_{\{\rho_\tau(r_i(\beta)) \le G_\beta^{-1}(\lambda)\}}]$ be the asymptotic form of $Q_{n,k_n}(\beta)$ defined in (5), where G_β and G_β^{-1} are the distribution and quantile functions of $\rho_\tau(r_i(\beta))$ (the uniform convergence of $Q_{n,k_n}(\beta)$ to $Q_\lambda(\beta)$ under Assumptions D and F is derived in Lemmas 2.1 and A.1 of Čížek, 2008).

Now, considering an interval $\Delta(a,\lambda) = \langle F^{-1}(a), F^{-1}(a+\lambda)\rangle$ for $a \in (0, 1-\lambda)$ and a fixed $\tau \in (0,1)$, Tableman (1994a, p. 390) proved in the location model that $Q_\lambda(\mu)$ applied to univariate data following the distribution function F has a unique minimum at $\mu^*(\tau) = F^{-1}(a^*(\tau)+\tau\lambda)$, where $a^*(\tau) = \arg\min_{a\in(0,1)} \int_{\Delta(a,\lambda)} \rho_\tau(\varepsilon - F^{-1}(a+\tau\lambda))dF(\varepsilon)$ (the proof is given for $\tau = 0.5$ by an argument, which directly applies also to general $\tau \in (0,1)$).

This result can be employed in the regression model (1). Conditionally on a given value of covariates x, the distribution function of the response y equals $F_{y|x}(t) = F(t - x^T\beta^0), t \in R$, and the corresponding quantile function equals $F_{y|x}^{-1}(u) = F^{-1}(u) + x^T\beta^0$. Therefore, $Q_{\lambda|x}(\beta) = E[\rho_\tau(r_i(\beta)) \cdot 1_{\{\rho_\tau(r_i(\beta)) \le G_\beta^{-1}(\lambda)\}}|x]$ is minimized at $\mu^*(\tau) + x^T\beta^0$ (conditionally on x). Consequently, $Q_\lambda(\beta)$ is minimized at $\beta^*(\tau) = (\mu^*(\tau), 0, \ldots, 0)^T + \beta^0$ unconditionally if the intercept is supposed to be the first element of the parameter vector β.

The limit $Q_\lambda(\beta)$ of the LTQR estimator objective function identifies the parameter vector $\beta^*(\tau)$. To prove the consistency of the LTQR estimator, we can apply now Theorem 3.1 of Čížek (2008): since we verified the identification condition for β^*, assume Assumptions D and F, and $\{\rho_\tau(r_i(\beta)) = \max\{\tau r_i(\beta), -(1-\tau)r_i(\beta)\} | \beta \in R^p\}$ forms a Vapnik-Chervonenkis class of functions (van der Vaart and Wellner, 1996, Lemmas 2.6.15 and 2.6.18), we only need to check Assumption D3 of Čížek (2008). This is however verified under Assumptions D and F by Lemma 2 of Čížek (2006). Thus, Theorem 3.1 of Čížek (2008) implies the claim of the theorem. □

References

Adrover, J., Maronna, R.A. and Yohai, V.J., 2004. Robust regression quantiles. Journal of Statistical Planning and Inference 122, 187–202.

Chen, C. (2004). An adaptive algorithm for quantile regression. In: Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Theory and Applications of Recent Robust Methods. Birkhäuser, Basel, 39–48.
Čížek, P., 2006. Least trimmed squares in nonlinear regression under dependence. Journal of Statistical Planning and Inference 136, 3967–3988.
Čížek, P., 2008. General trimmed estimation: Robust approach to nonlinear and limited dependent variable models. Econometric Theory 24, 1500–1529.
Čížek, P., 2010. Reweighted least trimmed squares: an alternative to one-step estimators. CentER Discussion Paper 2010/91, Tilburg University, The Netherlands.
Čížek, P., 2011. Semiparametrically weighted robust estimation of regression models. Computational Statistics & Data Analysis 55, 774–786.
Dimova, R. and Neykov, N.M., 2004. Generalized d-fullness technique for breakdown point study of the trimmed likelihood estimator with applications. In: Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Theory and Applications of Recent Robust Methods. Birkhäuser, Basel, 83–91.
Gervini, D. and Yohai, V.J., 2002. A class of robust and fully efficient regression estimators. The Annals of Statistics 30, 583–616.
Giloni, A., Simonoff, J.S. and Sengupta, B., 2006. Robust weighted LAD regression. Computational Statistics & Data Analysis 50, 3124–3140.
Hawkins, D.M. and Olive, D., 1999. Applications and algorithms for least trimmed sum of absolute deviations regression. Computational Statistics & Data Analysis 32, 119–134.
He, X., Jurečkova, J., Koenker, R. and Portnoy, S., 1990. Tail behavior of regression estimators and their breakdown points. Econometrics 58, 1195–1214.
Hubert, M. and Rousseeuw, P.J., 1998. The catline for deep regression. Journal of Multivariate Analysis 66, 270–296.
Hunter,D. and Lange, K. (2000) Quantile regression via an MM. J. of Computational and Graphical Statistics 9, 60–77.
Jurečková, J., 2010. Finite-sample distribution of regression quantiles. Statistics and Probability Letters 80, 1940–1946.
Koenker, R.W., 2005a. Quantile Regression. Cambridge University Press, Cambridge.
Koenker, R.W., 2005b. Quantile Regression in R. http://cran.R-project.org/doc/packages/quantreg/quantreg.pdf
Koenker, R.W. and Bassett, G. Jr., 1978. Regression quantiles. Econometrica 84, 33–50.
Koenker, R. and Machado, J., 1999. Goodness of fit and related inference processes for quantile regression. J. Am. Statist. Ass. 94, 1296–1309.
Müller, Ch.H., 1995. Breakdown points for designed experiments. J. Statistical Planning and Inference. 45, 413–427.
Neykov, N.M. and Müller, Ch.H., 2003. Breakdown point and computation of trimmed likelihood estimators in generalized linear models. In: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P.J. (Eds.), Developments in robust statistics. Physica-Verlag, Heidelberg, 277–286.
Neykov, N.M., Filzmoser, P. and Neytchev, P.N., 2012. Robust joint modeling of mean and dispersion through trimming. Computational Statistics & Data Analysis 56, 34-48.
R Development Core Team, 2011. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
Rousseeuw, P.J. and Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.
Rousseeuw, P.J. and Hubert, M., 1999. Regression depth. Journal of the American Statistical Association 94, 388–402.
Rousseeuw, P.J. and Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley & Sons, New York.
Rousseeuw, P.J. and Van Driessen, K., 2006. Computing LTS regression for large data sets. Data Mining and Knowledge Discovery 12(1), 29–45.
Stromberg, A.J., Hössjer, O. and Hawkins, D.M., 2000. The Least Trimmed Differences Regression Estimator and Alternatives. Journal of the American Statistical Association 95, 853–864.
Tableman, M., 1994a. The asymptotics of the least trimmed absolute deviations (LTAD) estimator. Statistics and Probability Letters 19, 387–398.
Tableman, M., 1994b. The influence functions for the least trimmed squares and the least trimmed absolute deviations estimator. Statistics and Probability Letters 19, 329–337.
Vandev, D.L., 1993. A note on breakdown point of the least median squares and least trimmed squares. Statistics and Probability Letters 16, 117–119.
Vandev, D.L. and Neykov, N.M., 1998. About regression estimators with high breakdown point. Statistics 32, 111–129.
Van Aelst,S., Rousseeuw, P.J., Hubert, M. and Struyf, A., 2002. The deepest regression method. J. Multivariate Analysis 81, 138–166.
Van der Vaart, A.W. and Wellner, J.A., 1996. Weak Convergence and Empirical Processes With Applications to Statistics. Springer, New York.