Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **50**, 2 (2011) 45–53

Trimmed Estimators in Regression Framework*

Tomáš JURCZYK

Department of Probability and Mathematical statistics
Faculty of Mathematics and Physics, Charles University Prague
Sokolovská 83, 186 75 Praha 8, Czech Republic
e-mail: jurczyk@karlin.mff.cuni.cz

Dedicated to Lubomír Kubáček on the occasion of his 80th birthday

(Received March 31, 2011)

Abstract

From the practical point of view the regression analysis and its Least Squares method is clearly one of the most used techniques of statistics. Unfortunately, if there is some problem present in the data (for example contamination), classical methods are not longer suitable. A lot of methods have been proposed to overcome these problematic situations. In this contribution we focus on special kind of methods based on trimming. There exist several approaches which use trimming off part of the observations, namely well known high breakdown point method the Least Trimmed Squares, Least Trimmed Absolute Deviation estimator or e.g. regression L-estimate Trimmed Least Squares of Koenker and Bassett. Our goal is to compare these methods and its properties in detail.

Key words: trimmed mean, least trimmed squares, least trimmed absolute deviations, trimmed LSE, regression quantiles

2010 Mathematics Subject Classification: 62J05, 62J20

1 Motivation

We consider the linear regression model $Y_i = \sum_{j=1}^{p} X_{ij}\beta_j^0 + Z_i$, $i = 1, \ldots, n$, where $\beta^0 = (\beta_1^0, \ldots, \beta_p^0)'$ is vector of unknown parameters, $Z_1, \ldots, Z_n$ are independent identically distributed random variables with distribution function F and density f and $EZ_i = 0$. $X = \{X_{ij}\}_{i=1,j=1}^{n,p}$ is the design matrix, $X_i = (X_{1i}, \ldots, X_{ni})'$ is i-th regressor. Suppose that $X_1 = (1, 1, \ldots, 1)'$. Denote $r_i(\hat{\beta}) = Y_i - \sum_{j=1}^{p} X_{ij}\hat{\beta}_j$ the i-th residual ($\hat{\beta}$ stays as estimate of β^0).

*Supported by the grant No. 402/09/0557 of the GA ČR and the project LC06024.

Let us focus first only simple location model, which means $p = 1$ and $Y_i = \beta_1^0 + Z_i$, $i = 1, \ldots, n$. The most-known estimator of β_1^0 in that situation is sample mean as the Least Squares estimator (LS) for the location model. Despite the advantages and optimal properties under normal distributed Z_i, it is well known that sample mean is inefficient when F has heavy tails (see [1] or [10]) and also sensitive to the presence of outliers (see e.g. [4]). Another classical estimate is sample median as a solution $\operatorname{argmin}_{\beta \in \mathcal{R}} \sum_{i=1}^{n} |Y_i - \beta|$. Sample median is less sensitive to heavy tailed distributions but also has its drawbacks, like the sensitivity to the small change of the observation in the center of the data, etc.

To preserve good properties of the sample mean and to ensure good efficiency under wider range of distributions and less sensitivity to outliers, the concept of weighted mean is considered. We are interested especially at the special case called α-trimmed mean (we want to study trimmed estimators) defined as

$$\bar{\beta}_\alpha = \frac{1}{n - 2\,[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} Y_{(i)},$$

where $Y_{(1)} \leq \ldots \leq Y_{(n)}$, $0 < \alpha < \frac{1}{2}$. This means that we remove the α percent smallest and the α percent largest observations from the data before we compute the sample mean. α-trimmed mean was extensively studied for example in [4], let us recall some of its properties: as the linear function of order statistics, it belongs to the L-estimator class, the breakdown point is α (breakdown point is the minimal fraction of changed observations (by arbitrary value) capable of pulling the estimate out of all bounds). Asymptotic representation is $\sqrt{n}(\bar{\beta}_\alpha - \beta^0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_q(Z_i) + o_p(1)$, where $q = F^{-1}(1-\alpha)$ and Ψ_k is Huber's function

$$\Psi_k(x) = \begin{cases} -k/(1-2\alpha) \text{ if } x < -k \\ x/(1-2\alpha) \text{ if } -k \leq x \leq k, \\ k/(1-2\alpha) \text{ if } x > k \end{cases}$$

which is the same as the influence function for Huber's M-estimate (recall that influence function in point x describes the effect of infinitesimal contamination with value x on the estimate—for more details see [4]). Asymptotic distribution of $\sqrt{n}(\bar{\beta}_\alpha - \beta^0)$ is normal, centered, with the variance

$$\sigma^2(\alpha, F) = (1-2\alpha)^{-2} \left(2\alpha q^2 + \int_{-q}^{q} x^2 dF(x) \right).$$

As the sample mean has the Least Squares method (LS) as its direct generalization for regression model and the sample median the Least Absolute Deviation estimate (LAD), there was a great afford to find some regression analog also for α-trimmed mean. We will see several proposals in the next section.

2 Estimators based on trimming

Before we define trimmed estimators we need to recall the concept of regression quantiles. α-regression quantile $\hat{\beta}(\alpha) = \operatorname{argmin}_{\beta \in \mathcal{R}^p} \sum_{i=1}^{n} r_i(\beta)\,(\alpha - I\{r_i(\beta) < 0\})$,

where I denotes indicator. The most desirable property of α-regression quantile is that exactly α percent of observations are under the regression hyperplane based on $\hat{\beta}(\alpha)$. The regression quantiles are therefore the generalization of quantiles for regression (for more details about regression quantiles see [7] or [8]).

Now, we are ready to present methods based on trimming which we are going to investigate in the rest of the paper. Simple representatives of trimmed estimators are two-step procedures of Koenker and Bassett [7] (denoted as $\hat{\beta}_{\alpha,RQ}$) and Ruppert and Carroll's proposal [10] ($\hat{\beta}_{\alpha,PE}(\hat{\beta}^0)$). Different and more complicated approach is represented by well-known high breakdown point method the Least Trimmed Squares ($\hat{\beta}_{\alpha,LTS}$) proposed in [9]. The Least Trimmed Absolute Deviation estimator ($\hat{\beta}_{\alpha,LTA}$ is also based on the same principle of the implicit residual weighting, see e.g. [12].

Definition 1 The Trimmed Least Squares estimator of Koenker and Bassett $\hat{\beta}_{\alpha,RQ}$ (in the following we will call this method shortly RQ as the method based on regression quantiles) is the LS estimate calculated after the removal of all observations that satisfy $Y_i - X_i'\hat{\beta}(\alpha) \leq 0$ or $Y_i - X_i'\hat{\beta}(1-\alpha) \geq 0$.

Definition 2 $\hat{\beta}_{\alpha,PE}(\hat{\beta}^0)$ is the LS estimate from the data where the observations with the $[n\alpha]$ smallest and $[n\alpha]$ largest residuals based on the preliminary estimate $\hat{\beta}^0$ are removed (we will use the shortcut PE for this method).

Definition 3 The Least Trimmed Squares estimator (LTS) is defined as

$$\hat{\beta}_{\alpha,LTS} = \underset{\beta\in\mathcal{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n-2[n\alpha]} |r(\beta)|^2_{(i)} ,$$

where $|r(\beta)|_{(1)} \leq |r(\beta)|_{(2)} \leq \ldots \leq |r(\beta)|_{(n)}$.

Definition 4 The Least Trimmed Absolute Deviations estimator (LTA) is defined as

$$\hat{\beta}_{\alpha,LTA} = \underset{\beta\in\mathcal{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n-2[n\alpha]} |r(\beta)|_{(i)} .$$

Remark 1 Notice please that RQ and PE method choose the observations which are used by external rule, while LTS and LTA choose them in an implicit way (the situation is much more complicated because the ordering may be different for different β).

Remark 2 PE estimate depends on the preliminary estimate. We use in our work 3 choices of $\hat{\beta}^0$, the same as in [10], i.e. LS estimate, LAD estimate and $\hat{\beta}_{RQ} = \frac{1}{2}(\hat{\beta}(\alpha) + \hat{\beta}(1-\alpha))$.

Remark 3 An alternative of PE estimate with asymmetric trimming can be defined in such way that we remove $[2n\alpha]$ observations with the largest absolute value of residuals. Such proposal could be found in [10], for symmetric distributions it has the same asymptotic properties as PE defined in Definition 2.

3 Comparison of the methods

The comparison between the estimators $\hat{\beta}_{\alpha,RQ}$ and $\hat{\beta}_{\alpha,PE}(\hat{\beta}^0)$ has been made to some extend in paper [10]. Also some marginal comparisons of presented methods with the LS method can be found in [10], [7], [3] and [12]. Nevertheless, the comparison of $\hat{\beta}_{\alpha,RQ}$ or $\hat{\beta}_{\alpha,PE}(\hat{\beta}^0)$ with $\hat{\beta}_{\alpha,LTS}$ and $\hat{\beta}_{\alpha,LTA}$ is still missing in the literature. This is what we are going to investigate in the following part of the paper.

Before we focus on asymptotic properties let us recall that all methods are regression, scale and affine equivariant and the estimates are unbiased when the distribution of Z_i is symmetric.

3.1 Asymptotic properties

We start the comparison of introduced methods with their asymptotic properties. Under some regularity conditions (for more details see [10], [7], [3] and [12]) all methods are $\sqrt{n}$-consistent. Only to sum up the very basic conditions, we suppose continuous and symmetric distribution of Z_i. We have $\sqrt{n}(\hat{\beta} - \beta^0) = \frac{1}{\sqrt{n}} Q^{-1} \sum_{i=1}^{n} x_i \psi(Z_i) + o_p(1)$, where $\lim_{n\to\infty} n^{-1} X'X = Q$ is supposed to be positive definite. The shape of ψ functions for each estimate[1] can be seen in Table 2. These ψ functions are at the same time the influence functions of corresponding estimators. We can see that LS and PE with preliminary estimate LS have unbounded influence, other estimates have bounded influence, LTS and LTA[2] have even zero influence for large values. Compare the shape of influence functions with corresponding M-estimation theory (see [4]). We also notice that RQ method has the same asymptotic representation as α-trimmed mean which makes from it the regression analog of α-trimmed mean ($\hat{\beta}_{\alpha,PE}$ in location model has also the same asymptotic distribution, but for regression it is only the case for $\hat{\beta}_{\alpha,PE}(\hat{\beta}_{RQ})$ and only under symmetric f).

We can see from the formulas for asymptotic variances in Table 2 that some direct comparison of variances or simple rules for describing the situations where one estimate outperforms others are not possible. Therefore, we try to compare asymptotic variances graphically. We tried classical symmetric continuous distributions and also two setups of symmetric contaminating distributions: the first one $F(x) = (1-\epsilon)\Phi(x) + \epsilon\Phi(x/b)$ (the same as in [10], Φ denotes the distribution function of $N(0,1)$) and the second mixture (used e.g. in [3]) where $F = (1-2\epsilon)N(0,1) + \epsilon N(c,1) + \epsilon N(-c,1)$. We have tried different choices of constants b, c and ϵ. The comparison of these asymptotic variances gives us several foundings about estimates (see examples for chosen distributions in Figure 1):

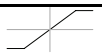
- The performance of $\hat{\beta}_{PE}$ strongly depends on preliminary estimate $\hat{\beta}^0$ (as was already mentioned in [10]).

[1] $\hat{\beta}_{\alpha,PE}$ has different representations for location and slope parameters, see Table 2 or [10].

[2] Rigorous proofs for asymptotic representation of LTA method have only been given in the location model, see [11] and [12].

- Estimator $\hat{\beta}_{RQ}$ outperforms other estimators when the trimming proportion is high for almost all choices of F.
- With low contamination, LTS[3] and LTA methods are inefficient for large trimming proportion.
- LTS and LTA have feasible efficiency level with the trimming proportion a bit higher than the proportion of contamination.
- There should be the same kind of relationship between LTA and LTS variance as between LAD and LS (comparison of asymptotic variances shows that LAD has greater statistical efficiency than LS if $f(0) > \frac{1}{2\sigma}$), i.e. LTA is more efficient than LTS for more peaked distributions, see also [6].
- Only LTA has reasonable variance if we choose lower trimming proportion than the contamination percentage. LTA is based on absolute values of residuals, therefore is not so much sensitive to normal contamination.
- Variance of LTA is hardly better than the classical $\hat{\beta}_{LAD}$. Rewriting the inequality for asymptotic variances $\sigma^2_{LTA}(\alpha, F) < \sigma^2_{LAD}(F)$ gives quadratic inequality $f^2(q) - 2f(0)f(q) + 2\alpha f^2(0) > 0$, for unimodal f it is equivalent to $\frac{f(q)}{1-\sqrt{1-2\alpha}} < f(0)$, which means that LTA is better for distributions with lighter tails, even lighter than normal.

Table 2: Asymptotic representation of estimates: $\sqrt{n}(\hat{\beta} - \beta^0) = \frac{1}{\sqrt{n}} Q^{-1} \sum_{i=1}^{n} x_i \psi(Z_i) + o_p(1)$ with its asymptotic variance $Q^{-1}\sigma^2(\alpha, F)$, where $Q = \lim \frac{1}{n} X'X$, $q = F^{-1}(1-\alpha)$, $a = 2qf(q)$, $I_q = \int_{-q}^{q} x dF(x)$ and $I_q^2 = \int_{-q}^{q} x^2 dF(x)$.

estimator	$\sigma^2(\alpha, F)$	ψ function
$\hat{\beta}_{LS}$	$\sigma^2 = varZ_1$	
$\hat{\beta}_{LAD} = \hat{\beta}(0.5)$	$(2f(0))^{-2}$	
$\hat{\beta}_{\alpha,RQ}$	$(1-2\alpha)^{-2}[2\alpha q^2 + I_q^2]$	
$\hat{\beta}_{\alpha,PE}$ - location	$(1-2\alpha)^{-2}[2\alpha q^2 + I_q^2]$	
$\hat{\beta}_{\alpha,PE}(\hat{\beta}_{LS})$ - slope	$(1-2\alpha)^{-2}[a^2\sigma^2 + (1+2a)I_q^2]$	
$\hat{\beta}_{\alpha,PE}(\hat{\beta}_{LAD})$ - slope	$(1-2\alpha)^{-2}[(a/f(0))^2 + aI_q/f(0) + I_q^2]$	
$\hat{\beta}_{\alpha,PE}(\hat{\beta}_{RQ})$ - slope	$(1-2\alpha)^{-2}[2\alpha q^2 + I_q^2]$	
$\hat{\beta}_{\alpha,LTS}$	$(1-2\alpha-a)^{-2} I_q^2$	
$\hat{\beta}_{\alpha,LTA}$	$(1-2\alpha)[2(f(0)-f(q))]^{-2}$	

[3] In [3] was pointed out that for unimodal symmetric distributions $1-2\alpha-a>0$ should hold. We can prove this statement simply: $1-2\alpha-a = \int_{-q}^{q}[f(x)-f(q)]dx$ and $f(x) > f(q),\ x \in (-q, q)$.

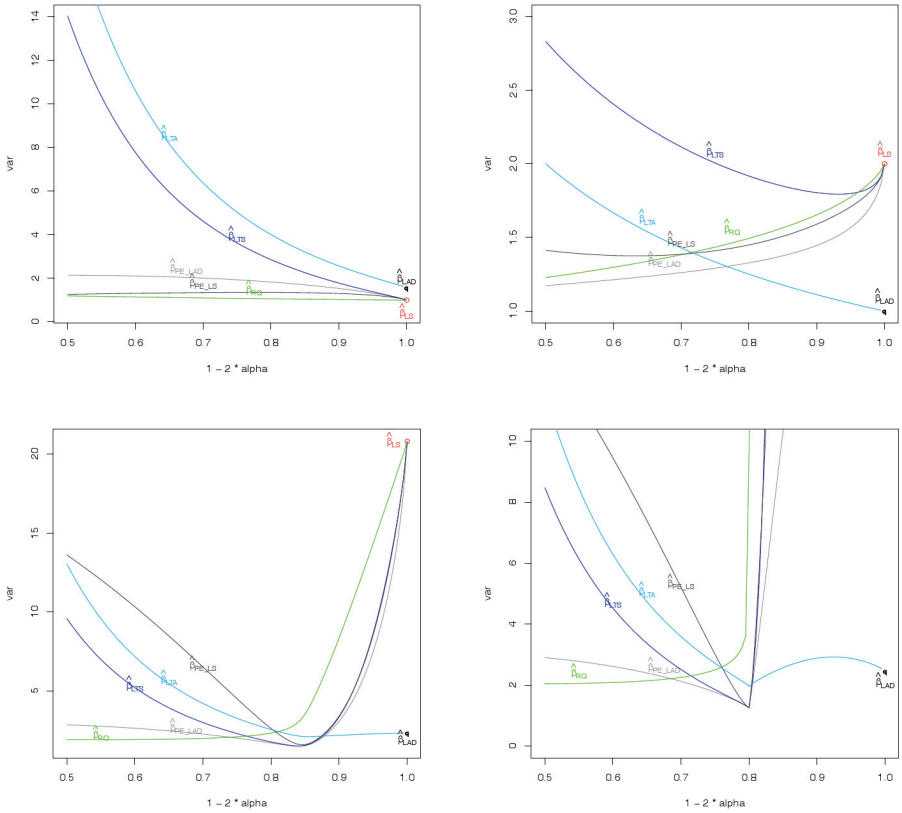


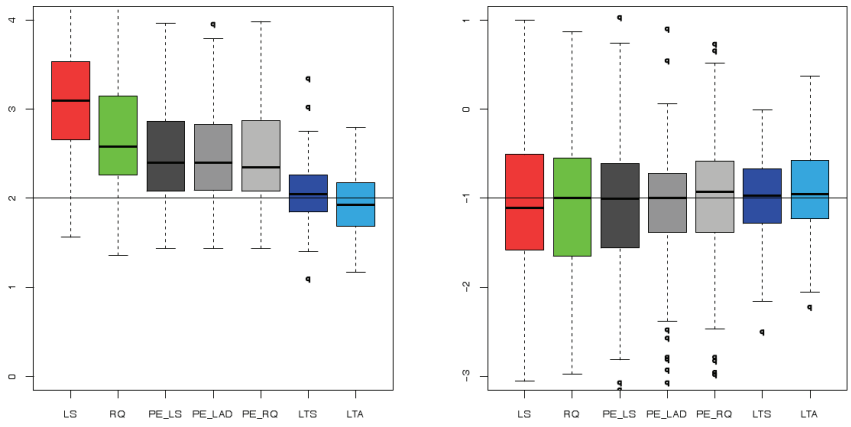
Figure 1: Several examples of theoretical asymptotic variances $\sigma^2(\alpha, F)$ in dependence on $1 - 2\alpha$. $\sigma^2(\alpha, F)$ for $\hat{\beta}_{\alpha,PE}(\hat{\beta}_{RQ})$ is not depicted because it is the same as for $\hat{\beta}_{\alpha,RQ}$.

3.2 Simulation study and robustness properties

To show not only asymptotic variance but also bias and finite sample properties, as well as the behaviour under asymmetric contamination in Y as well as in X values, we run the simulation study. The setup is following: for simplicity two regressors (intercept is one of them), 20 observations, $\beta^0 = (2, -1)$. We have computed presented estimates for 100 runs for each setup and each choice of distribution (constants: ϵ was chosen consequently 0.05, 0.1, 0.2, 0.25, c was 3 or 10, α was 0.05, 0.1, 0.2 or 0.25). Simulations are performed in R-software, source code is available upon request.

We have performed the experiment for all symmetric distributions mentioned above, it only verified the same relationship among variances as in the previous section. We have tried the performance under asymmetric contamination in the data as well. These results concern robustness properties (see Figure 2):

10% outliers in Y, $F = 0.9N(0,1) + 0.1N(10,1)$



10% outliers in X_2, distribution of Z is $N(0,1)$, with the probability 0.1 the value X_{i2} is changed by addition of the random variable $N(10,1)$

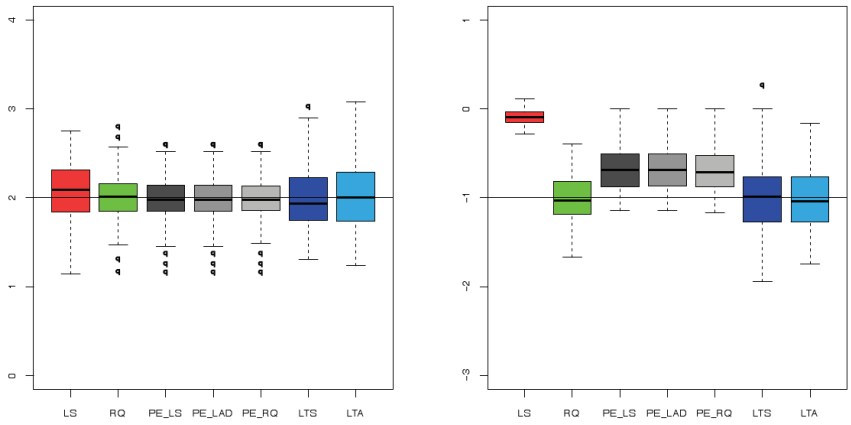


Figure 2: Examples of contaminating distributions. Boxplots of estimates based on 100 runs of simulated data sets (intercept coefficient—left pictures, slope parameter—right pictures).

Table 3: Breakdown point properties

estimator	location model	regression (X_{ij} random variable)
$\hat{\beta}_{LS}$	0	0
$\hat{\beta}_{LAD} = \hat{\beta}(0.5)$	0.5	0
$\hat{\beta}_{\alpha,RQ}$	$\min\{\alpha, 0.25\}$	0
$\hat{\beta}_{\alpha,PE}(\hat{\beta}_{LS})$	0	0
$\hat{\beta}_{\alpha,PE}(\hat{\beta}_{LAD})$	$\min\{2\alpha, 0.5\}$	0
$\hat{\beta}_{\alpha,PE}(\hat{\beta}_{RQ})$	$\min\{\alpha, 0.25\}$	0
$\hat{\beta}_{\alpha,LTS}$	$\min\{2\alpha, 0.5\}$	$\min\{2\alpha, 0.5\}$
$\hat{\beta}_{\alpha,LTA}$	$\min\{2\alpha, 0.5\}$	$\min\{2\alpha, 0.5\}$

- Under asymmetric contamination (outliers in Y direction) represented by model $F(x) = (1 - 2\epsilon)N(0, 1) + 2\epsilon N(c, 1)$, all methods except LTS and LTA have biased estimates of intercept. Which is not so surprising because other methods are based either on LS or on quantile regression, which are biased.
- $\hat{\beta}_{PE}$ estimates are biased under asymmetric contamination in X_2 (outliers in X_2 direction are generated in the following way: original value X_{i2} is changed by addition of random variable $N(c, 1)$ with the probability 2ϵ).

We did not think about presented estimates as about robust methods at the first time but if we have the data where there are some amount of contaminating points do not following the model, we clearly expect from the trimmed methods not to use exactly these contaminating points. Therefore, we expect implicitly the robustness of the methods. We have also talked about robustness properties when we compared the influence functions. Another robustness characteristics, which we have not discuss yet, is the breakdown point. Table 3 summarizes breakdown points for our methods. For the location model or for the model when we do not allow replacement of the regressor's values, only $\hat{\beta}_{LS}$ and $\hat{\beta}_{\alpha,PE}(\hat{\beta}_{LS})$ have zero breakdown point. But when regressor's values are also expected to be changed, then all methods except LTS and LTA fail already when small amount of points is changed. This is caused by the fact that regression quantiles have problems with bad leverage points (see [8]).

4 Conclusion

We have compared the asymptotic behaviour of presented trimmed estimates, as well as finite sample behaviour through the simulation study. Robustness properties have been shown as well. From these results we can make some conclusions and recommendations about usage of the estimates.

At first we must recall that more complicated LTS and LTA methods are more robust than other methods but with not correctly chosen trimming pro-

portion can be quite inefficient (proposal of procedure for adaptive choice of h can be found e.g. in [2]). The choice of large trimming proportion should be avoided unless there is strong reason for it (trimming proportion near to the $2\alpha = 0.5$ can also cause unstable subsample behaviour and also high sensitivity to the small change of data (see [5]), which is negative consequence of high breakdown point). Koenker and Bassett's $\hat{\beta}_{RQ}$ outperforms Ruppert and Carroll's $\hat{\beta}_{PE}$ and also is regression analog to α-trimmed mean. If we do not expect gross errors but only the different distribution of the error therms, then $\hat{\beta}_{RQ}$ is a good choice.

We can continue with our comparison further but due to the limits on the scope of this paper let us add only last remark about computational aspect. LTS and LTA are computationally much more intensive than other methods (for more details see e.g. [6]), which induces problems for large data sets, nevertheless such one-step highly robust procedures can be used profitably as some preliminary estimate for different robust approaches.

References

[1] Andrews, D. F.: Robust Estimates of Location: Survey and Advances. *Princeton University Press*, Princeton, N.Y., 1972.

[2] Atkinson, A. C., Cheng, T. C.: *Computing least trimmed squares regression with forward search.* Statistics and Computing **9** (1998), 251–263.

[3] Čížek, P.: *Asymptotics of the trimmed least squares.* Journal of Statistical Planning and Inference, CentER DP series **2004/72** (2004), 1–53.

[4] Hampel, F. R. et al.: Robust Statistics: The Approach Based on Influence Functions. Wiley Series in Probability and Statistics, *Wiley*, 1986.

[5] Hettmansperger, T. P., Sheather, S. J.: *A Cautionary Note on the Method of Least Median Squares.* The American Statistician **46** (1991), 79–83.

[6] Hawkins, D. M., Olive, D.: *Applications and algorithms for least trimmed sum of absolute deviations regression.* Computational Statistics & Data Analysis **32**, 2 (1999), 119–134.

[7] Koenker, R., Bassett, G.: *Regression quantiles.* Econometrica **46** (1978), 466–476.

[8] Koenker, R.: Quantile Regression. *Cambridge University Press*, Cambridge, 2005.

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

[10] Ruppert, D., Carroll, J.: *Trimmed Least Squares Estimation in the Linear Model.* Journal of the American Statistical Association**75** (1980), 828–838.

[11] Tableman, M.: *The influence functions for the least trimmed squares and the least trimmed absolute deviations estimators.* Statistics & Probability Letters **19** (1994), 329–337.

[12] Tableman, M.: *The asymptotics of the least trimmed absolute deviations (LTAD) estimator.* Statistics & Probability Letters **19** (1994), 387–398.