ADVANTAGES AND DISADVANTAGES, CHALLENGES AND THREADS OF ROBUST METHODS[1]

Jan Ámos Víšek

Department of Macroeconomics and Econometrics, Institute of Economic Studies,
Faculty of Social Siences, Charles University
Opletalova ulice 26, 110 01 Praha 1, the Czech Republic,
e-mail: visek@fsv.cuni.cz

Abstract. Framed by a patterns of history of robust statistics advantages and disadvantages of the robust processing data are explained and put in the proper context. The discussions are illustrated by series of tables demonstrating the real state of situation, not repeating the statistical folklore e. g. about the loss of efficiency of the robust estimators. Examples of accompanying diagnostic tools and modifications of the "parent" method of the *least wighted squares* for the various situations, e. g. of broken orthogonality condition, are also presented.

Keywords : Robust estimation, regression model, loss and gains of efficiency, instrumental weighted variables, significance of explantory variables, computational problems.

An introduction to the (hi)story of robust statistics

Ronald Aylmer Fisher (1920, 1922) studied at twenties of the past century the behaviour of two of the most frequently employed classical estimators, the estimators of the location and the scale

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{and} \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

He was interested in the change of the variance of these statistics if the data instead of being distributed according to the standard normal distribution $N(0,1)$ they are distributed according to the Student's $t(\nu)$ with ν degrees of freedom. He found that

$$\lim_{n\to\infty} \frac{\mathrm{var}_{N(0,1)}(\overline{x}_n)}{\mathrm{var}_{t(\nu)}(\overline{x}_n)} = 1 - \frac{6}{\nu(\nu+1)} \tag{1}$$

$$\lim_{n\to\infty} \frac{\mathrm{var}_{N(0,1)}(s_n^2)}{\mathrm{var}_{t(\nu)}(s_n^2)} = 1 - \frac{12}{\nu(\nu+1)}, \tag{2}$$

notice $t(\nu)$ as the lower index of "var" in the denominators in (1) and (2). It gives the for the values $\nu = 3, 5$ and 9

TABLE 1

T_n	t_9	t_5	t_3
$\overline{x}_n$	0.93	0.80	0.50
s_n^2	0.83	0.40	0!

[1]Research was supported by grant of the Czech Science Foundation project No. P402/12/G097 "DYME - Dynamic Models in Economics".

Surprisingly for $\nu = 3$ the ratio of variances of s_n^2, with respect to $N(0,1)$ and $t(\nu)$ is asymptotically equal to zero. In other words, the variance of s_n^2 in the case when the data are distributed according to the Student's t with 3 degrees of freedom is (asymptotically) infinitely larger than the variance of the same estimator when data are distributed according to the standard normal distribution. The following figure offers a possibility to make the idea how far are these two distributions each from other.

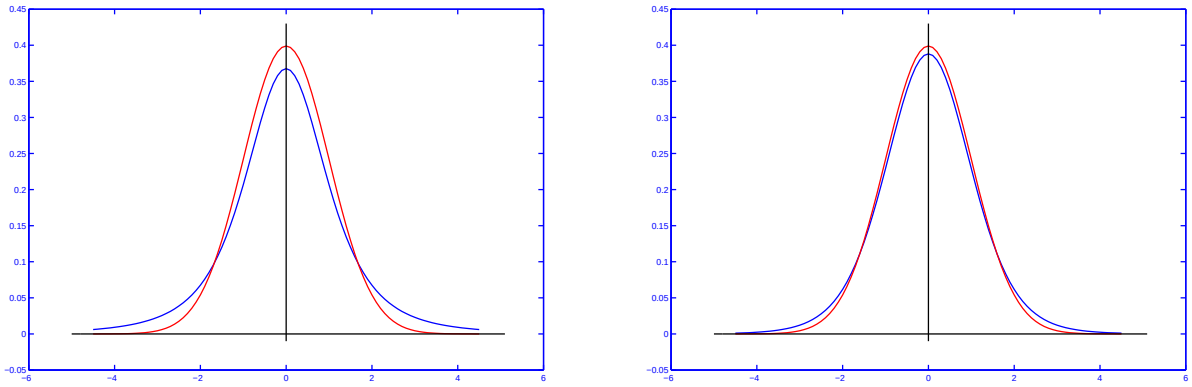


FIGURE 1
The "lower" curves are standard normal while the "upper" (more spiky) ones are the Student's density with 3 and 9 degrees of freedom, respectively.

Fisher's results (see also Hampel et al. (1986)) were hardly acceptable by the classical statistics and hence they were discreetly forgotten for nearly 40 years.

Nevertheless, a new warning arrived by John Tukey (1960):

A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope was drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians.

It was answered firstly by Peter Huber (1964, 1965) and a bit later by Frank Hampel (1968, 1974). The former approach kept in fact the classical statistical approach via solving an extremal problems (like least squares, maximum likelihood) but he proposed to take into account instead of a parameterized family of distribution functions (d.f.), say $\mathcal{F} = \{F_\theta(x)\}_{(\theta\in\Theta)}$, e. g. $\{N(\mu,\sigma^2)\}_{(\mu\in R,\sigma^2\in R^=)}$ (and than to compute e. g. maximum likelihood estimate of θ), to take into account a parameterized family of neighborhoods having the center at the distribution functions from $\mathcal{F}$, i. e. to take into account

$$\mathcal{G} = \left\{G_{(\varepsilon,\theta,H)} = (1-\varepsilon)F_\theta + \varepsilon H\right\}_{(0\le\varepsilon\le\Delta,\,\theta\in\Theta,\,H\ \ a\ distribution\ function)}$$

(and than to compute e. g. maximum likelihood estimate of θ). The latter approach (proposed by Frank Hampel) "redefined" the estimator (as well as the test) as a functional of the empirical distribution function (e. g. $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = I\!E_{F_n(x)}X$ where $F_n(x)$ is the empirical d.f.) and then to study the properties of these functionals by means of the infinitesimal calculus on the vector space of all distribution functions (employing mainly Fréchet's and Gâteaux's derivative of the corresponding functional).

The second half of sixties and especially the seventies than witnessed a bumping development of robust methods - see e. g. Andrews et al. (1972), Hogg (1974), Bickel (1975), Jurečková (1977), Stigler (1977),

Beran (1978), Koenker & Bassett (1978), Maronna et al. (1979) or Yohai & Maronna (1979), to give some among many others. It ends in the famous and still classical monographs Huber (1981), Hampel et al. (1986), Rousseeuw & Leroy (1987), Stahel & Weisberg (1991) or Marazzi (1992) (again giving some, may be the most famous, among many others) and research continues nowadays as a well-established part of statistics (see e. g. http://www.compstat2012.org/sessions.php, http://www.compstat2012.org/COMPSTAT 2012_ BoA.pdf and http://www.compstat2012.org/ Proceedings_ COMPSTAT2012.pdf for the program, abstracts and proceedings of COMPSTAT 2012, respectively)[2]. One feature which is usually typical for a newly established region of statistical research, still overlives in robust statistics - there is still a huge hunt for new and new methods for solving the classical statistical problems as estimating the location, the scale or the regression model. One would assume that nearly half century after the pioneering papers by Peter Huber much larger attention should be paid to development of the accompanying diagnostics of already proposed estimators, to the modifications of them for "nonstandard" situations (e. g. for the situation when the disturbances are correlated with the explanatory variables and hence the least squares as well as the most of robust methods are biased and inconsistent) and to the computational problems.

In 1975 Peter Bickel formulated question of a possibility to establish an estimator of regression coefficients (for the multiple linear model) which would have 50% breakdown point (for definition of the *breakdown point* see the Appendix[3]), i. e. to establish such estimator which would be an analogy, in the regression framework, of median which has 50% breakdown point in the problem of location. The problem appeared to be more complicated than it seemed to be at the first glance. In the meantime the M-estimators became the most popular estimators due to the possibility of the relatively tractable theory (see again Bickel (1975) or Jurečková (1977) and for an overview paper see Davis (1993)) as well as for the reliability of algorithms for their computation, see again Marazzi (1992). The way for a possibility to cope with even intricate problems in proofs was opened by the rediscovery, by Stephan Portnoy (1983), of Skorohod's embedding into Wiener process. Similarly, an amazing trick for the improvement of computation of M-estimators was proposed by Jaromír Antoch (see Antoch & Víšek (1991)). Hence some statisticians hope that appropriately defined M-estimators became the candidate for the solution of Bickel's problem. Let us recall some basic facts about the M-estimators. Prior to do it we need to introduce a basic notations - to avoid a misunderstanding and to keep at least a minimal level of mathematical rigor.

An intermezzo - introducing a bit of the framework

Let $\mathcal{N}$ denote the set of all positive integers and R^p the p-dimensional Euclidean space. Moreover, let $(\Omega, \mathcal{A}, P)$ be a probability space which (for the sake of simplicity and without significant restriction of generality) the all random variables are defined on. The linear regression model

$$Y_i = X_i'\beta^0 + e_i = \sum_{j=1}^{p} X_{ij}\beta_j^0 + e_i, \quad i = 1, 2, ..., n \tag{3}$$

will be considered (all vectors are assumed to be the column vectors) where Y_i's are response variables, X_i's are p-dimensional vector of the explanatory variables and e_i's are disturbances (according to the language of econometrics) or error terms (according to the language of statistics). We will assume hereafter that $\{X_i', e_i\}_{i=1}^{\infty}$ is a sequence of i. i. d. r. v.'s and that e_i's are not correlated with X_i's. If in what follows we will need modify these assumptions, it will be done with an appropriate emphasize.

[2]To be a bit more complete, let us add that recently appeared some new points of view on the robustness, especially in econometrics, see e. g. Hansen & Sargent (2008).

[3]The topics in the appendix are given in the order we come across them in the text of paper.

An introduction to the (hi)story of robust statistics (continued)

The M-estimators were originally defined by

$$\hat{\beta}^{(M,n)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} \rho(Y_i - X_i'\beta) \tag{4}$$

where ρ is usually asked to be symmetric and convex (but not necessarily). As it was already recalled the simplicity of definition (4) allowed to derive some basic results by simple tools. Nevertheless, by definition (4) M-estimators are not *scale-* and *regression-equivariant*[4] and hence the definition was modified to

$$\hat{\beta}^{(GM,n)} = \underset{\beta \in R^p, \sigma^2 \in R^+}{\arg\min} \sum_{i=1}^{n} \rho([Y_i - X_i'\beta)]\,\sigma^{-1})$$

(such estimators are sometimes called *generalized M-estimators*). As the minimization - simultaneously over $\beta \in R^p, \sigma^2 \in R^+$ - need not be simple, they are usually computed in two steps. In the first one we estimate the standard deviation of disturbances, say $\hat{\sigma}$, and then we compute the estimate of regression coefficients. So, we look for

$$\hat{\beta}^{(M,n)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} \rho([Y_i - X_i'\beta)]\,\hat{\sigma}_n^{-1})$$

see Marazzi (1992). Nevertheless, as early as in 1975 Peter Bickel showed that $\hat{\sigma}$ has to be a *scale-equivariant* and *regression-invariant* estimate of scale of error terms (see again the Appendix) to reach the *scale-* and *regression-equivariance* of M-estimator of regression coefficients. To establish such an estimator of the standard deviation of disturbances is not a simple task, see Croux & Rousseeuw (1992), Jurečková & Sen (1993) or Víšek (2010a). Moreover, all of these proposals are in fact based on a preliminary robust *scale-* and *regression-equivariant* estimator of the regression coefficients. **That is the reason why (presumably) it would be better to employ the estimators which are *scale-* and *regression-equivariant* by the definition, i. e. without the necessity of studentization of residuals. It is evident - from what follows, that e. g. $least\ median\ of\ squares$ the $least\ trimmed\ squares$ or the $least\ weighted\ squares$ - see Definitin 1 below - posses this property.**

Finally, it is easy to see that M-estimators have to be solution of *normal equations*

$$\sum_{i=1}^{n} X_i \psi([Y_i - X_i'\beta)]\,\hat{\sigma}_n^{-1})$$

where $\psi(z) = \rho'(z)$. But then it is immediately clear that M-estimators are vulnerable to leverage[5] points. To cope with this is disadvantage we can further generalize the definition to:
Let $w_i \in [0,1],\ i = 1, 2, ..., n$ be weights (generally $w_i = w_i(Y, X)$). The estimator which is solution of normal equations

$$\sum_{i=1}^{n} w_i X_i \psi([Y_i - X_i'\beta)]\,\hat{\sigma}_n^{-1})$$

will be called generalized M-etimators. Unfortunately, even these type of estimators appeared not to fulfill our ideas about an ideal estimator, see a discussion of results by Maronna & Yohai (1981) below.

[4] For an explanation please see Appendix.
[5] See again Appendix.

As the median is the middle order statistic, or in other words, the median is the estimate of 50% quantile (see e. g. Anděl (1978)), a part of statisticians hope for a generalization of quantiles into the regression framework, say regression quantiles. The proposal how to establish *regression quantiles* arrived at 1978 by Roger Koenker and David Bassett (1978). The solution of an extremal problem which they defined gives the plains, theoretically parallel to the regression model in question, and the plains were located so that the disturbances fall in one halfspace with (a priori selected) probability α and into the complementary halfspace with probability $1 - \alpha$. Nevertheless, it appeared that contrary to the location framework, the regression quantiles are M-estimators and hence they are not resistant against leverage points.

The situation even worsen when, after the long six years of futile pursuit for a solution of Bickel's problem, Ricardo Maronna & Victor Yohai (1981) showed that in regression framework the generalized M-estimators cannot have the breakdown point larger than $1/p$ (where p is the dimension of model). It indicated that the solution of Bickel's problem need not be very simple, if any.

Finally, in 1982 Andrew Siegel arrived with the *repeated median*[6] which has evidently 50% breakdown point but the definition of the estimator was so complicated that it was never implemented except for the simple regression. However, as it happened in the history of science several times, when a solution of a puzzle appears, it typically opens a door for further solutions. So, immediately the next year Peter Rousseeuw came with the *least median of squares* (LMS[7]) and a few months later with the *least trimmed squares*[8] (LTS). Both these methods, although they were (originally) constructed so to have 50% breakdown point, allowed to select their parameters so to be adjusted for any lower level of robustness in the case when we felt that the contamination is not so high. Peter Rousseeuw proposed also algorithms (*repeated selection of an elemental set*) for computing the estimators and managed their implementation which generously offered to statisticians all over the world. In 1987 in Rousseeuw & Leroy (1987) the whole series of data-sets and their successful processing by LMS and LTS infuse a quite natural idea that the robust methods, especially when we adjust their robustness to be (nearly) 50%, can give us a reliable idea about the underlying model. Of course, the statistical folklore, which was loudly spread by those who fundamentally sticked to the classical statistics, claimed that the price (or sacrifice) for this knowledge is a gross decrease of efficiency. Both ideas - the idea (or may be better word would be *fiction*) of an objective recognition of the underlying model although paying a tax for it in the form of a huge loss of efficiency as well as just this idea (and again a better word would be perhaps a stiff believe) of a huge loss of efficiency - proved to be false. We may say about the former (of course exaggerating a bit), regrettably, and about the latter, fortunately - although, as we shall see later, in the both cases the situation is a bit more complicated than such a simple classification (or appreciation) can induce. Let's briefly describe the downfall of the former idea.

In 1992 Thomas Hettmansperger & Simon Sheather (1992) studied dependence of the number of knocks of spark-ignition engines on their technical parameters (timing of sparks, ratio of air/fuel, temperature of intake, temperature of exhaust). They employed LMS with the algorithm based on the repeated selection of elemental sets. Unfortunately, when they wrote the data (which are usually referred to as Engine Knock Data) into a file in PC, they made an error (they wrote for the variable *ratio of air/fuel* of the second observation the value 15.1 rather than the correct value 14.1). When they had found out the typing error they corrected data and recalculated the estimate of model. *And an unpleasant amazement (or even a shock) took place!* Although they employed robust method with 50% break down point, i. e. the method which

[6] All definitions of estimators can be found in the Appendix. Some of them require notations from the section **The framework (continued)** given below but they are intuitively comprehensible.

[7] The following two estimators are defined below in the section **The framework (continued)**.

[8] The later estimator was proposed so quickly after the LMS due to the fact that it was clear from the very beginning that the convergence of LMS to the true regression coefficients would be (very) slow.

should be able to cope with a high contamination, the difference between the two estimates, for the correct data and the damaged data was significant - see tables in the Appendix; explanation and the discussion of the tables are given at the section **Continuing the story** below). Their "discovery" buried the up-to-those-days prevailing belief that the estimators with the high breakdown point represent a solution of problem how to process the data in a highly safe way to reach objectively an underlying model. Some people were frustrated that after nearly three decade of work the results appeared to be useless and that the sacrifice of efficiency (which we were willing to pay for *"to reach so grand goal as an objective recognition of proper models, i. e. recognition of inherent laws of Nature"*) was in vain. Everything seemed to be demolished and we had to admit that we have had constructed our dreams on the sand as that man, from the famous Gospel story, who built his house on the sand[9]. But similarly as in the Holy Bible after even the darkest picture, a help is always near and the healing can be found. And this is a story we are coming to narrate. Prior to it, let us introduce some further notations.

The framework (continued)

For any $\beta \in R^p$ $r_i(\beta) = Y_i - X_i'\beta$ denotes the i-th residual and $r^2_{(h)}(\beta)$ stays for the h-th order statistic among the squared residuals, i. e. we have

$$r^2_{(1)}(\beta) \leq r^2_{(2)}(\beta) \leq \ ... \ \leq r^2_{(n)}(\beta). \tag{5}$$

Moreover, let $w_i \in [0,1],\ i = 1, 2, ..., n$ be weights. Then

Definition 1 *The solution of the extremal problem*

$$\hat{\beta}^{(LWS,n,w)} = \underset{\beta \in R^p}{arg\,min} \ \sum_{i=1}^{n} w_i r^2_{(i)}(\beta) \tag{6}$$

will be called the least weighted squares (LWS), Víšek (2000a).

Remark 1 *Realize please that Definition 1 - due to the fact that it allows for a whole range of possible weights - establishes in fact a wide, flexible family of estimators. When we select appropriately the weights we obtain the ordinary least squares (OLS), the least median of squares (LMS) or the least trimmed squares (LTS) ($\frac{n}{2} \leq h \leq n$):*

$$\hat{\beta}^{(OLS,n)} = \underset{\beta \in R^p}{arg\,min} \ \sum_{i=1}^{n} r^2_i(\beta), \qquad \hat{\beta}^{(LMS,n,h)} = \underset{\beta \in R^p}{arg\,min} \ r^2_{(h)}(\beta) \tag{7}$$

and

$$\hat{\beta}^{(LTS,n,h)} = \underset{\beta \in R^p}{arg\,min} \ \sum_{\ell=1}^{h} r^2_{(\ell)}(\beta). \tag{8}$$

As $\hat{\beta}^{(LWS,n,w)}$ depends only on the squared residuals, - similarly as the OLS, LMS or LTS - it is scale- and regression-equivariant (see again the Appendix).

The *scale-* and *regression-equivariance* is one of important property which qualifies LWS to be one of candidate for a wide employment in the applications. Another one is of course its flexibility to accommodate to the level and to the character of contamination of the data by proper selecting the weights. Last but not least, we shall see later on the patterns of numerical studies that LWS can cope with even improperly selected weights better than other estimators in (7) or (8).

[9] "The rain came down, the streams rose, and the winds blew and beat against that house, and it fell with a great crash." - Matthew 7_{27}, compare Luke 6_{49}.

Continuing the story

Let us return to Hettmansperger & Sheather's study of *Engine Knock Data* (the data can be found also in Mason et al. (1989)). As it was already recalled, the results achieved by Hettmansperger & Sheather seemed to be shocking. The highly robust method was not able to cope with a relatively small change of one observation. How can such a method be able to cope with large outliers and/or leverage points? *Something seemed to be rotten in the state of Denmark!*

But let us give at first an explanation and a discussion of the tables given in Appendix. We have recalled above that Hettmansperger & Sheather employed LMS with the algorithm based on the repeated selection of elemental sets. They selected $h = 11$ (for h see again the definition of LMS and LTS in (7) and (8)). This selection was appropriate due to the fact that it was proved (see e. g. Rousseeuw & Leroy (1987)) that LMS and LTS reach the optimal properties if we put $h = [\frac{n}{2}] + [\frac{p+1}{2}]$. The Engine Knock Data contains 16 observations and $p = 5$ (the four explanatory variables - timing of sparks, ratio of air/fuel, temperature of intake, temperature of exhaust - plus intercept), hence rightfully $h = 11$. Nevertheless, they results were false. How we found it?

When Hettmansperger & Sheather's results had appeared, we had at hand already an alternative method for computing LMS. The method was invented by Pavel Boček and based on the simplex method, see Boček & Lachout (1993). It proved to be better in the sense of lower values of the minimized h-th order statistics of squared residuals (compare the last row of the second and the third column of table for *air/fuel=14.1* as well as the last row of the first and the second column of table for *air/fuel=15.1*; both tables are in the Appendix) as well as quicker than the method of *repeated selection of elemental sets.* And it was just this struck of good luck that revealed the fact that the surprising results were due to the bad algorithm (results of Hettmansperger & Sheather are denoted below as H-S-LMS while the results computed by Boček's algorithm by Boček-LMS; compare the estimates by Boček's algorithm for the correct and damaged data, again in the Appendix). *So it seemed that the robust statistics is saved. Unfortunately, it was not true.*

We have realized that for Engine Knock Data we can compute LTS exactly. It is due to the fact that LTS coincide with OLS which we apply on some (generally unknown) subset of data containing h observations. So, if we are able to compute OLS for all subsets of original data, all subsets containing h observations, and then we select that subset (and corresponding model) for which the sum of squared residuals is the smallest, we obtain exact LTS. In the case of Engine Knock Data we have $\begin{pmatrix} 16 \\ 11 \end{pmatrix} = 4368$ such subsets and hence (even in the nineties with slow PCs of those days) we were able to compute LTS exactly (in about 3 minutes), see the last columns of the tables. It is evident that the estimated model for the correct and for the damaged data are rather different. So it seemed that the problem opened by Hettmansperger & Sheater (although their results were false) is deeper and may be inherently connected with some type of estimators. The explanation of the puzzle is given by FIGURE 2.

In the left part of figure LTS or LMS selected properly 5 points (including the circle), in the right part LTS or LMS selected, again properly again 5 points. Nevertheless, the difference between two sets of data is the location of the circle, i. e. the difference is a small shift of circle. Notice that the closer the circle would be to axis Y the shift can be smaller, for the theoretical discussion see Víšek (1994)[10] or (2006a).

[10]The paper was submitted only a few months after Hettmansperger & Sheater's paper to the *The American Statistician* but the reviewers claimed that there was nothing new, although any other explanation didn't appear prior to it. Moreover, the paper brought a new method of evaluating LTS for data for which the exact algorithm - described above - cannot be computed due to too large number of subsets containing all h-tuples, see the description of the algorithm below. Nevertheless, the algorithm later appeared to be effective and in a modified forms is still used, see e. g. Hawkins (1994), Hawkins & Olive (1999), Čížek & Víšek (2000),

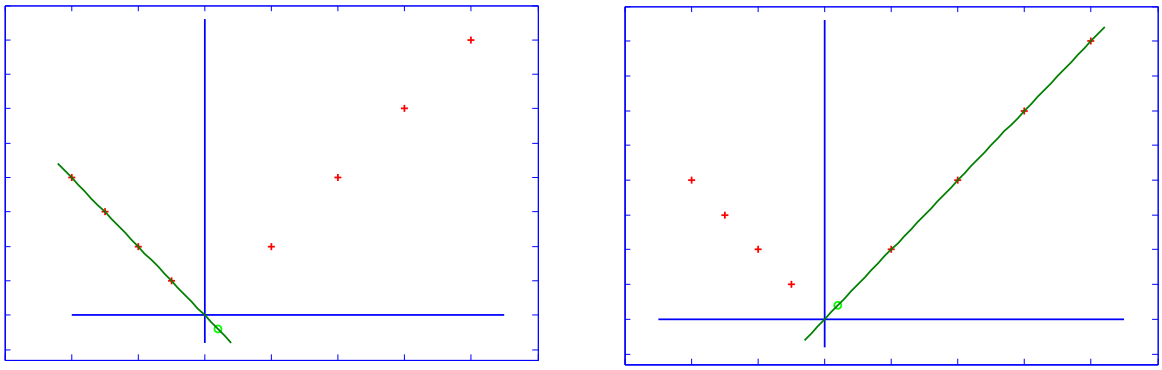


FIGURE 2
Decreasing versus increasing model

We can made conclusions:

- The "switch" property is an inevitable feature of the estimator with a high breakdown point and it can be view as the first disadvantage of the estimators as LMS or LTS.
- The reason for such a large switch of model is the zero-one weights which are prescribed to the residuals, or in the other words, the methods are impertinently sure that they have "discovered" the true underlying model. In other words, the method "without hesitation" decided which observations are "proper" and which are contamination. Such a God-like position is "unhealthy" anytime and here it repaid by e.,g. a high sensitivity to a deletion or a shift of one point. So it can be viewed as the the second disadvantage of the high-breakdown-point estimator.

Unfortunately, the same behaviour (can) exhibit also M-estimators with the discontinuous ψ-function, see Víšek (1996a) and (2002a). Let's briefly discuss both conclusions, starting with the former one.

Changing the disadvantages into an advantages

Employing the "switch" property for a profit

It is known from the history that some disadvantage may turn into advantage if we interpret it in a non-usual way. Although not fully analogous but surely appreciable is the story of king Leonidas of Sparta who led the alliance of Greek city-states (approximately 7000 men) in 480 BC to the battle of Thermopylae against the second Persian invasion of king Xerxes (one million men). The Greeks lost the battle but if not treachery of local resident Ephialtes it is not clear what the result of the fight in the mountain defile would have been. Nevertheless, the most appealing is the answer of king Leonidas on Xerxes verbal assault that the number of Persian archers is such that the Greeks would not see the sun due to the Persian arrows. He replied: *"We will have an advantage, we will fight in shade."*. Similarly, we can use as an advantage the high breakdown point estimator as a diagnostic tool. Firstly, we can process data by a classical method, as OLS and then by highly robust estimator. If the results are similar, let's thank the God - the data have a sense and they are (in the worst case) slightly contaminated. If the results are (significantly?) different, we should try to find why. We are going to give an example of such a utilization.

Rousseeuw & van Driessen (2005), to give some among others.

In 1995 we have looked for a model for the export of the Czech economy into the EU. We have data containing 91 industries (of the Czech economy), the response variable was export and the explanatory variables were the following ones:

X	export	BAL	Balasa index
S	sales	DP	price development after opening-up
US	number of university students	IRS	increasing return from scale
HS	number of high school students	FDI	foreigner direct investment
VA	value added	W	wages
K	capital	R&D	research and development
CR3	market power (concentration in given industry)	TFPW	total factor productivity related to wages

Due to the fact that the production of tobacco was monopolistic, we have excluded it from research. The ordinary least squares gave models (for various combination of explanatory variables) with very low coefficients of determination. We employed the least trimmed squares, starting with $h = 48$ and increased it in every step about one up to $h = 59$, the results see in TABLE A2 below in Appendix. The following figure offers the development of the estimates of some coefficients in a graphic way.

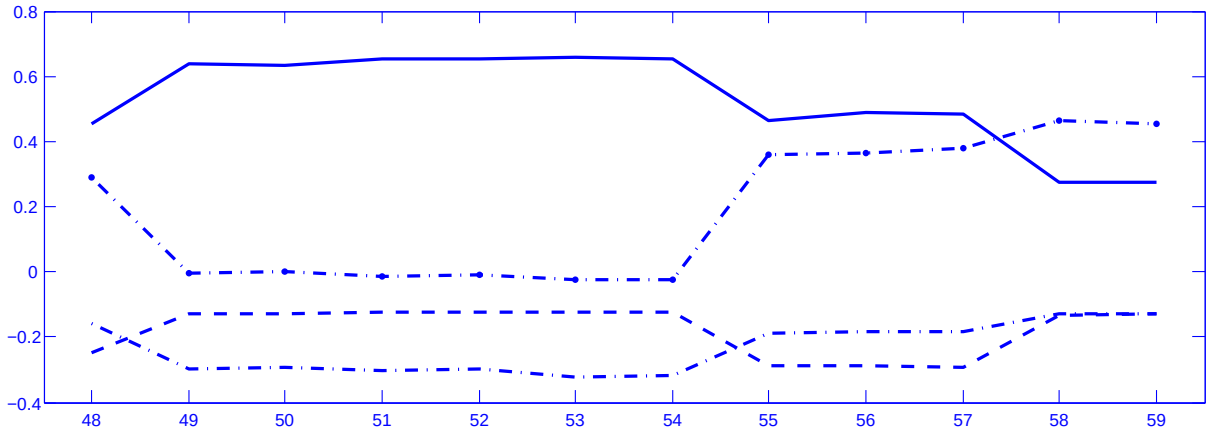


FIGURE 3
The development of the estimates of regression coefficients. The solid line represents $\hat{\beta}_1^{(LTS,n,h)}$ (down-scaledby $\frac{1}{10}$), the upper dashed-and-dot is $\hat{\beta}_3^{(LTS,n,h)}$, the dashed is $\hat{\beta}_4^{(LTS,n,h)}$ and lower dashed-and-dot is $\hat{\beta}_6^{(LTS,n,h)}$ (down-scaled again by $\frac{1}{10}$).

It is clear that a break in a smooth development appeared in $h = 55$. Moreover, up to $h = 54$ the selected subpopulations were nested, i. e. when we increased h about one, LTS selected subpopulation which consists of the subpopulation from the previous step plus some additional industry. However, starting from $h = 55$ it does not hold. LTS selected for $h = 55$ a subpopulation which contains many industries from the subpopulation for $h = 54$ but not all. Finally - for $h = 54$ - we arrived at a model

$$\frac{X_k}{S_k} = 4.64 - 0.032 * \frac{US_k}{VA_k} - 0.022 * \frac{HS_k}{VA_k} - 0.124 * \frac{K_k}{VA_k} + 1.035 * CR3_k$$
$$-3.199 * TFPW_k + 1.048 * BAL_k + 0.452 * DP_k + e_k \qquad k = 1, 2, ..., 54 \tag{9}$$

with the coefficient of determination 0.974. Moreover, the residuals were approximately normally distributed. Let's call this subpopulation the *Main*. Then we applied LTS on the rest of 36 industries, let's call this subpopulation of the original data the *Complementary* subpopulation. As one can see from TABLE A3 we encounter again a break in a development (in this case not so smooth as in the previous step of establishing the *Main subpopulation*) of the estimates of regression coefficients between 33 and 34. The other facts were practically the same as above: up to 33 the subpopulations in the successive steps were nested and the residuals were approximately normally distributed (the normality even improved with increasing h but much worsen for $h = 34$). So, finally we arrived to the model for the *Main subpopulation*

$$\frac{X_\ell}{S_\ell} = -0.634 + 0.089 * \frac{US_\ell}{VA_\ell} + 0.235 * \frac{HS_\ell}{VA_\ell} + 0.249 * \frac{K_\ell}{VA_\ell} + 1.174 * CR3_\ell$$
$$+0.690 * TFPW_\ell + 2.691 * BAL_\ell - 0.051 * DP_\ell + e_\ell, \qquad \ell = 1, 2, ..., 33$$

with coefficient of determination 0.932.

A good sense of this decomposition of the whole Czech economy can be further supported by the mutual relation of the Capital (normalized by Wages, K/W) and Labour (normalized by Sales, L/S). Although for the whole population of all industries it gives no reasonable picure (see the next figure), for the *Main subpopulation* it confirms that the Cobb-Douglas production function (see Cobb & Douglas (1928)) holds for K/W and L/S, in other words, the product

$$\frac{K}{W} \cdot \frac{L}{S} \approx const.$$

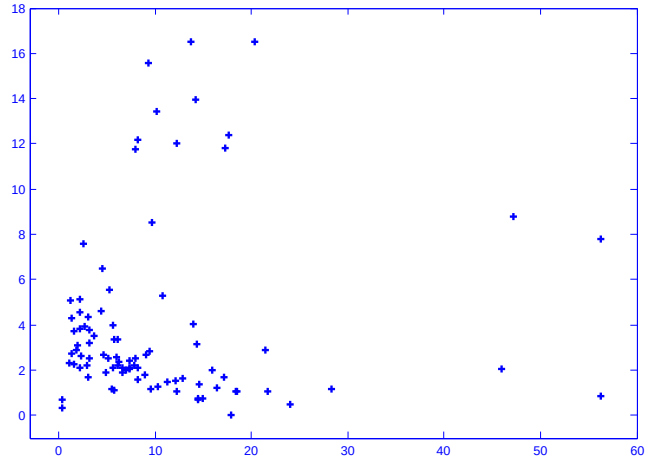
It indicates that the *Main subpopulation* collects the industries which were already in 1993 reoriented on the market economy. Contrary to it, the same figure for the mutual relation of K/W and L/S for the *Complementary subpopulation* shows that there is approximately a linear dependence between K/W and L/S, i. e. that for some a and b

$$\frac{K}{W} \approx a \cdot \frac{L}{S}^{-1} + b.$$

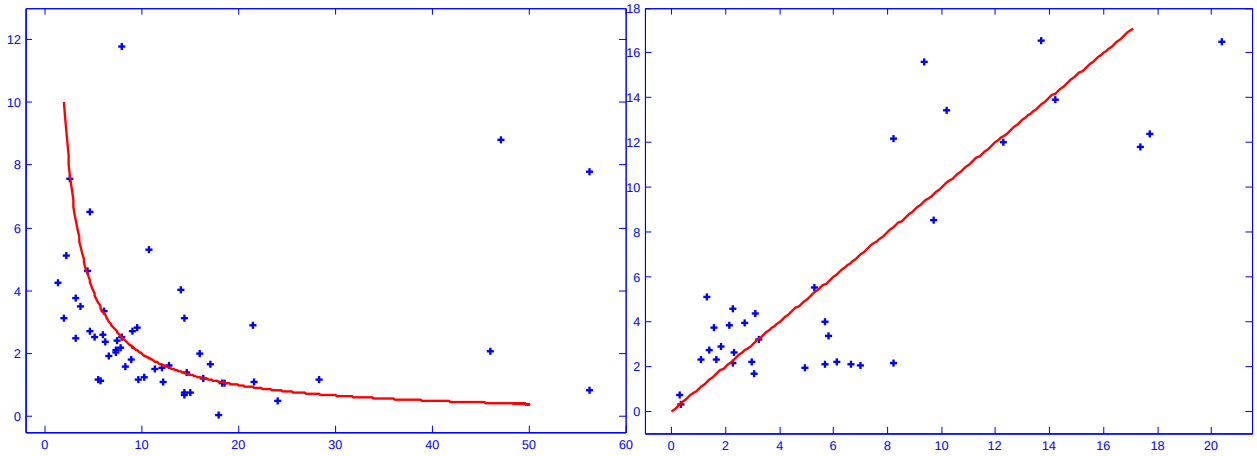
It means that

$$\frac{\frac{L}{S}}{\frac{K}{W}} \approx const_1$$

which hints that in such industries the larger investment of capital requires the larger need of labour to produce still the same amount of products. Exaggerating a bit, one can say that it resembles the socialistic, planned economy. In other words, the *Complementary subpopulation* collects the industries which up to the end of 1993 didn't reorient on the market economy and still run in the habits of previous socialistic (dis)organization of the czech economy.



All industries of the Czech economy



Main subpopulation *Complementary subpopulation*

FIGURE 4
The first graph exhibits the mutual relation of K/W and L/S for the all industries of the Czech economy for the year 1993 while the other two for the *Main* and *Complementary subpopulation*.

So, we evidently employed the disadvantage as an advantage.

Returning from Heaven to the Earth - giving up the omniscience

Let us turn to the second disadvantage, the possibility to assign to the observation only zero-one weights. Zero in the case when we (or the method implicitly) assume that the observation represent contamination or - even if belonging properly into the "clean" population - the observation is so atypical that it influences the result in a not lucky way. We assign the weight 1 when we think the observation is O.K.. The remedy is straightforward - let us allow for possibility to assign the weights from the whole interval $[0, 1]$. The idea which then immediately appears is that we should assign these weights to the order statistics of the squared residuals rather than to the squared residuals directly. And the LWS comes to reality, see Definition 1. Let's briefly summarize the pros nd , see Definition 1.

Properties inherited from LTS, pros

- $\sqrt{n}$-consistency (even under heteroscedasticity) (will be recalled below).
- Scale- and affine-equivariance.
- Adaptable breakdown point.
- Quick and reliable algorithm (will be recalled below).
- Efficiency of the estimation when data are not contaminated and we, in an adaptive way, properly adjust the robustness of the estimator (will be mentioned in the conclusion).

Properties inherited from LTS, cons

- Still lacking a "coefficient of determination"

New features, improving properties and fulfilling the gaps

- Diagnostics for finite sample size - significance of explanatory variables (will be briefly touched below).
- Modifications for nonstandard situations, e.g.:
  - the instrumental weighted variables, (will be briefly touched below),
  - the total least weighted squares,
  - the fixed or random weighted effects[11].
- Low sensitivity to the inliers.
- More diagnostic tools (Durbin-Watson test[12], White test[13], Hausman test[14], etc.).
- Applicability for panel data.
- "Coping automatically" with heteroscedasticity of data - an experimental experience.

The least weighted squares - consistency

We are going to give a few ideas about the theory and applicability of $\hat{\beta}^{(LWS,n,w)}$. Although Definition 1 is transparent and it directly hints how LWS would work, it is evident that the definition itself is not convenient for searching e. g. for the consistency of the estimator. Hence, prior to searching for the properties and/or an algorithm we make some modification of Definition 1.

First of all, we usually generate the weights by means of a weight function for which we ask:

Conditions $\mathcal{C}1$ *The weight function* $w(u)$ *is continuous nonincreasing,* $w : [0,1] \to [0,1]$ *with* $w(0) = 1$. *Moreover,* w *is Lipschitz in absolute value, i.e. there is* L *such that for any pair* $u_1, u_2 \in [0,1]$ *we have* $|w(u_1) - w(u_2)| \leq L \cdot |u_1 - u_2|$.

Secondly, it is only a technicality to show that $\hat{\beta}^{(LWS,n,w)}$ has to be one of the solutions of the **normal equations**

$$I\!N E_{Y,X,n}(\beta) = \sum_{i=1}^{n} w\left(F_{\beta}^{(n)}(|r_i(\beta)|)\right) X_i \left(Y_i - X_i'\beta\right) = 0. \tag{10}$$

[11] For details see Víšek (2012c).
[12] For details see Víšek (2003) and Kalina (2004).
[13] For details see Víšek (2002b).
[14] For details see Víšek (1998).

where $F_\beta^{(n)}(r)$ is the empirical d. f. of the absolute values of residuals $|r_j(\beta)|$ (the definition of $r_j(\beta)$ is given a few lines above Definition 1), i. e.

$$F_\beta^{(n)}(r) = \frac{1}{n}\sum_{j=1}^{n} I\left\{|r_j(\beta)| < r\right\} = \frac{1}{n}\sum_{j=1}^{n} I\left\{|Y_j - X_j'\beta| < r\right\}. \tag{11}$$

For the consistency we will need two additional conditions.

Conditions $\mathcal{C}2$ *The sequence* $\left\{(X_i', e_i)'\right\}_{i=1}^{\infty}$ *is sequence of independent* $p+1$*-dimensional random variables (r.v.'s) distributed according to distribution functions (d.f.)* $F_{X,e_i}(x,r) = F_X(x)\cdot F_{e_i}(r)$ *where* $F_{e_i}(r) = F_e(r\sigma_i^{-1})$ *with* $I\!E e_i = 0$*, var* $(e_i) = \sigma_i^2$ *and* $0 < \liminf_{i\to\infty} \sigma_i \le \limsup_{i\to\infty} \sigma_i < \infty$*. Moreover,* $F_e(r)$ *is absolutely continuous with density* $f_e(r)$ *bounded by* U_e*. Finally, there is* $q > 1$ *so that* $I\!E\,\|X_1\|^{2q} < \infty$ *(as* $F_X(x)$ *doesn't depend on* i*, the sequence* $\{X_i\}_{i=1}^{\infty}$ *is sequence of independent and identically distributed (i.i.d.) r.v.'s).*

Conditions $\mathcal{C}3$ *There is the only solution of*

$$I\!E\left[w\left(F_\beta(|r(\beta)|)\right) X_1\left(e - X_1'\beta\right)\right] = 0 \tag{12}$$

namely $\beta^0 = 0$ *(the equation (21) is assumed as a vector equation in* $\beta \in R^p$*). Moreover* $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\sigma_i = 1$.

Then we can prove:

Theorem 1 *Let* ***Conditions*** $\mathcal{C}1$*,* $\mathcal{C}2$ *and* $\mathcal{C}3$ *be fulfilled. Then any sequence* $\left\{\hat\beta^{(LWS,n,w)}\right\}_{n=1}^{\infty}$ *of the solutions of sequence of normal equations* $I\!N E_{Y,X,n}(\hat\beta^{(LWS,n,w)}) = 0,\ \ n = 1, 2, ...,$ *is weakly consistent.*

For the proof see Víšek (2011b). The main trick of the proof is an approximation of the e.d.f. $F_\beta^{(n)}(r)$ by its theoretical counterpart

$$\overline{F}_{n,\beta}(v) = \frac{1}{n}\sum_{i=1}^{n} F_{\beta,i}(v), \qquad \text{where} \quad F_{\beta,i}(v) = P\left(|Y_i - X_i'\beta| < v\right),$$

realize that **Conditions** $\mathcal{C}2$ allow for heteroscedasticity, i. e. $F_{\beta,i}(v)$'s are of the same type but may have different variances, the corresponding lemma is in the Appendix.

If we add moreover

Conditions $\mathcal{NC}1$ *The derivative* $f_e'(r)$ *exists and is bounded in absolute value by* B_e*. The derivative* $w'(\alpha)$ *exists and is Lipschitz of the first order (with the corresponding constant* J_w*). Moreover, for any* $i \in \mathcal{N}$

$$I\!E\left[w'(\overline{F}_{n,\beta^0}(|e_i|))\left(f_e(|e_i|) - f_e(-|e_i|)\right)\cdot e_i\right] = 0.$$

Finally, for any $j, k, \ell = 1, 2, ..., p$ $I\!E\,|X_{1j}\cdot X_{1k}\cdot X_{1\ell}| < \infty$ *(as* $F_X(x)$ *does not depend on* i*, the sequence* $\{X_i\}_{i=1}^{\infty}$ *is sequence of independent and identically distributed* p*-dimensional r.v.'s).*

we can prove

Theorem 2 *Let Conditions $\mathcal{C}1$, $\mathcal{C}2$, $\mathcal{C}3$ and $\mathcal{NC}1$ hold. Then any sequence $\left\{\hat{\beta}^{(LWS,n,w)}\right\}_{n=1}^{\infty}$ of solutions of the normal equations (**??**) $I\!N E_{Y,X}^{(n)}(\hat{\beta}^{(LWS,n,w)}) = 0$ is $\sqrt{n}$-consistent, i. e.* $\forall(\varepsilon > 0)\ \ \exists(K_\varepsilon < \infty)\ \ \forall(n \in \mathcal{N})$

$$P\left(\left\{\omega \in \Omega\ :\ \sqrt{n}\left\|\hat{\beta}^{(LWS,n,w)} - \beta^0)\right\| < K_\varepsilon\right\}\right) > 1 - \varepsilon.$$

For the proof see Víšek (2010b).

The least weighted squares - the algorithm

We are going to start with the explanation of the algorithm for the *least trimmed squares* as it is a bit simpler. The idea which underlies the algorithm is that the data in question have such a structure that it gives a hope for a reasonable processing them so that the results would have a sense. Formally the algorithm looks as follows:

- **The first step :** We select randomly from our n observations as small number of observations as we need just to determine a plain in the p-dimensional Euclidean space, say P_1 (lower index indicate the number of cycle of iterations we are just in, so put $i = 1$; please remember it).

- **The second step :** Calculate squared residuals of all observations - residuals with respect to just established plain P_1, select the h observations having the smallest squared residuals (denote it as $h\text{-}tuple_i$), sum up the squared residuals of these observations and denote this sum as $S_i^2 = S_1^2$.

- **The third step :** Apply the *ordinary least squares* (OLS) on the observations in $h\text{-}tuple_i$, selected in the previous step. We obtain some new plain in p-dimensional Euclidean space, say P_{i+1}. The sum of squared residuals of observations in $h\text{-}tuple_i$, residuals with respect to P_{i+1} is smaller or equal to S_i because the plain P_{i+1} is optimal in the sense of minimal sum of squares for the observations in $h\text{-}tuple_i$ - remember we have applied OLS. But there can be some observations from the whole data which were not in the $h\text{-}tuple_i$ but which are now closer to the plain P_{i+1} than some observations from $h\text{-}tuple_i$ (remember that the plain P_i changed into the plain P_{i+1}). So select the h observations which have the smallest squared residuals (among all n observations) - residuals with respect to P_{i+1} and so create $h\text{-}tuple_{i+1}$. Now, calculate sum of squared residuals of observations which are in $h\text{-}tuple_{i+1}$ and denote it S_{i+1}^2. As we have explained $S_i^2 \geq S_{i+1}^2$.

- **The fourth step - question of cycle :** $\boldsymbol{Is}$

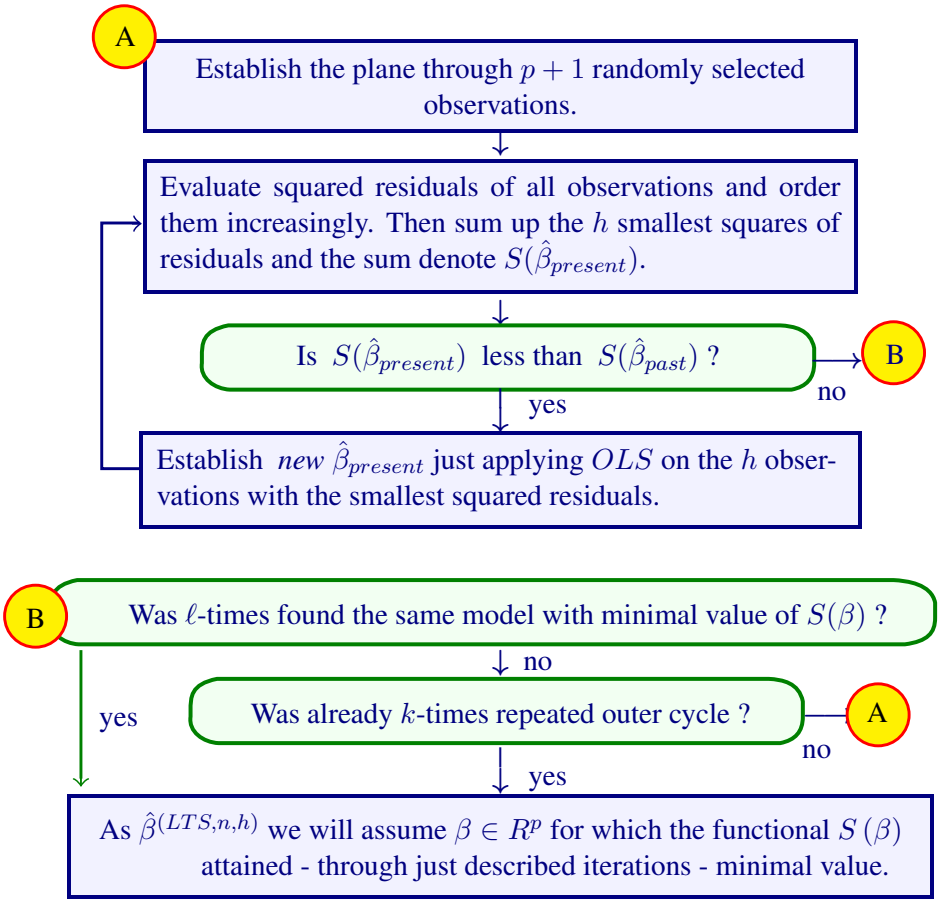
$$\boldsymbol{S_i^2 > S_{i+1}^2\ ?}$$

  - If YES, increase $\boldsymbol{i}$ to $\boldsymbol{i+1}$ and goto to the third step (this cycle will be called *inner*),
  - if not, i. e. $\boldsymbol{S_i^2 = S_{i+1}^2}$ stop the *inner cycle* and go to the fifth step.

- **The fifth step :** Save the value of sum of squares attained in just stopped *inner cycle* and go to the first step. This create an *outer cycle* for which we need some stopping rule. We will save (say) 20 smallest values of sum of squared residuls attained in the previous *inner cycles* in a *list of minimal squares*, say $\boldsymbol{\mathcal{S}}$. So whenever we stop the *inner cycle*, we look wheather just attained sum of squared residuals is smaller that any value in $\boldsymbol{\mathcal{S}}$ and if yes, we put it on the corresponding place (it means that we at first delete from $\boldsymbol{\mathcal{S}}$ the largest value, then we shift the other values "up" in $\boldsymbol{\mathcal{S}}$ and on the

empty position we record the present value of the sum of squared residuals). If $\boldsymbol{\mathcal{S}}$ contains already 20 same values we stop the *outer cycle* and say that the model which corresponds to this value of sum of squared residuals is the solution. The experience with hundreds of sets of data indicate that there were no two different models with exactly same values of sum of squared residuals. Of course sometimes it may happen that we don't find after a large number of iterations of outer cycle 20 same models. That is why we need also some additional stopping rule which stops the outer cycle if a priori selected number of these cycles was exhausted, say 10 000.

As somebody may prefer an explanation of the algorithm in the form of block-graph, here it is.



As it was already recalled the algorithm was firstly presented in the manuscript of the paper[15] which was submitted only a few months after Hettmansperger & Sheater's paper into the *The American Statistician*. The paper was not accepted but the algorithm proved to be efficient and in a various versions it is still employed and discussed, see e. g. Hawkins (1994), Hawkins & Olive (1999), Čížek & Víšek (2000), Rousseeuw & van Driessen (2005), to give some among others[16].

[15]The paper was later published in the form of Víšek (1994), however due to the space restriction in the proceedings, the paper focused on the explanation of the "switch effect" of the high breakdown point estimator and hence the algorithm and the results computed by it appeared later in Víšek (1996b) and others, e. g. Víšek (2000b).

[16]There appeared also modifications which do not work or contain even evident errors, see e. g. Er-Wei Bai (2003) with the correction made by Klouda (2007). Also some discussions, that the algorithm is not consistent, were published, see e. g. Hawkins & Olive (2003) or (2011), which seems misunderstanding what is the algorithm for - see also Hubert et al.(2002). The study if

The modification of the algorithm (which was described above for the *least trimmed squares*) is straightforward. Instead of selecting in every run of *inner cycle* the h-$tuple$ of the observations with the smallest squared residuals, we reorder all observations according to the order of values of their squared residuals. Then we take into account instead of sum of squared residuals (of the corresponding h-$tuple$) the sum of *weighted squared residuals* of all observations.

To illustrate how the algorithm works we include a table.

TABLE 2

True coeffs β^0	**6.3**	**-0.9**	**-5.2**	**6.8**	**3.1**
Heteroscedastic disturbances, independent from explanatory variables					
Data without contamination					
$\hat{\beta}^{OLS}_{(\mathrm{var}(\hat{\beta}^{OLS}))}$	$6.30_{(0.02)}$	$-0.91_{(0.02)}$	$-5.20_{(0.02)}$	$6.80_{(0.02)}$	$3.10_{(0.02)}$
$\hat{\beta}^{LWS}_{(\mathrm{var}(\hat{\beta}^{LWS}))}$	$6.29_{(0.02)}$	$-0.90_{(0.02)}$	$-5.20_{(0.02)}$	$6.80_{(0.02)}$	$3.11_{(0.02)}$
Intently contaminated data					
Outliers: for 10% of observations with the largest absolute value of response variable we put $Y_i = -2 * Y_i$					
$\hat{\beta}^{OLS}_{(\mathrm{var}(\hat{\beta}^{OLS}))}$	$-1.67_{(2.26)}$	$0.21_{(3.37)}$	$1.54_{(2.32)}$	$-1.49_{(2.29)}$	$-0.94_{(3.10)}$
$\hat{\beta}^{LWS}_{(\mathrm{var}(\hat{\beta}^{LWS}))}$	$6.27_{(0.03)}$	$-0.90_{(0.02)}$	$-5.17_{(0.03)}$	$6.77_{(0.03)}$	$3.08_{(0.02)}$
Leverage points: for 2% of observations with the largest norm of vector of explanatory variables we put $\tilde{X}_i = 10 \cdot X_i$ and $Y_i = -\tilde{X}_i' \cdot \beta^0$					
$\hat{\beta}^{OLS}_{(\mathrm{var}(\hat{\beta}^{OLS}))}$	$0.27_{(48.2)}$	$-0.01_{(52.6)}$	$-0.32_{(48.8)}$	$0.31_{(47.2)}$	$0.21_{(49.8)}$
$\hat{\beta}^{LWS}_{(\mathrm{var}(\hat{\beta}^{LWS}))}$	$6.30_{(0.02)}$	$-0.89_{(0.02)}$	$-5.20_{(0.02)}$	$6.80_{(0.02)}$	$3.09_{(0.02)}$
Outliers (10%) & leverage points (2%) were generated in the same way as above					
$\hat{\beta}^{OLS}_{(\mathrm{var}(\hat{\beta}^{OLS}))}$	$-6.79_{(40.7)}$	$1.13_{(54.8)}$	$5.75_{(44.0)}$	$-7.33_{(40.0)}$	$-3.71_{(49.0)}$
$\hat{\beta}^{LWS}_{(\mathrm{var}(\hat{\beta}^{LWS}))}$	$6.25_{(0.03)}$	$-0.89_{(0.02)}$	$-5.16_{(0.03)}$	$6.76_{(0.03)}$	$3.08_{(0.02)}$

the algorithm - which can be of course viewed as a stochastic process - is consistent, has the same roots and practical relevance as the complains that the Central Limit Theorem doesn't work on computers or that the d. f.'s of the sums of i.i.d. r. v.'s distributed according to the d. f. with heavy tails have to have also heavy tails while due to CLT their d. f.'s converge to the normal d. f. with light tails.

Significance of explanatory variables

One of very precarious debt of robust statistics was (and still mostly is) a lack of characteristics of quality of the estimated model, as significance of the individual explanatory variables - let us say as a "partial" characteristic - and coefficient of determination - as a global characteristic of quality. Due to the special form of LWS, it is possible to solve the former problem at least in a numerical way.

Definition 2 *Let for a* $k \in N$ *and* $\sigma_i \neq 0, i = 1, 2, ..., k$ *the sequence of random variables* $\{\xi_i\}_{i=1}^k$ *be normally distributed with zero means and variances* σ_i^2*. Then the distribution of the random variable* $\tau = \sum_{i=1}^k \xi_i^2$ *will be called generalized* χ^2 *with k degrees of freedom. Moreover, let the random variable* η *be indpendent from* τ *and have the standard normal distribution. Then the distribution of random variable* $\eta/\sqrt{\frac{\tau}{k}}$ *will be called the generalized t-distribution with k degrees of freedom.*

We will need some conditions which a bit differ from the previous ones.

Conditions $\mathcal{S}1$ *The sequence* $\{(X_i', e_i)'\}_{i=1}^{\infty}$ *is sequence of independent and identically distributed* $(p+1)$*-dimensional random variables (i.i.d. r.v.'s) with distribution function* $F_{X,e}(v,u) = F^{(1)}(v^{(1)}) \cdot F_{X,e}^{(2)}(v,u)$ *where* $F^{(1)}(v^{(1)}) : R^1 \to [0,1]$ *is d. f. degenerated at 1 and* $F_{X,e}^{(2)}(v^{(2)},u) = F_X(v^{(2)}) \cdot F_e(u)$ *is absolutely continuous and* $F_X(v^{(2)}) : R^{p-1} \to [0,1]$ *and* $F_e(u) : R^1 \to [0,1]$ *denote the corresponding marginals.*

Theorem 3 *Under* **Conditions** $\mathcal{C}1$ *and* $\mathcal{S}1$ *we have*

$$\mathcal{L}\left(\frac{\hat{\beta}_\ell^{(LWS,n,w)} - \beta_\ell^0}{\Sigma_{\ell\ell}^{\frac{1}{2}}} \cdot \left[\frac{\sum_{i=1}^n w_i\,(1-d_{ii})}{\tilde{r}'\left(\hat{\beta}^{(WLS,n)}\right)\cdot\tilde{r}\left(\hat{\beta}^{(WLS,n,w)}\right)}\right]^{\frac{1}{2}}\right) = t_{generalized}\,(n-p)\,.$$

It seems (nearly?) impossible to treat analytically $t_{generalized}\,(n-p)$ but we can simulate its quantiles. The next table offers them for a collection of sample sizes.

TABLE 3

The simulated quantiles for 5%.

n	20	30	40	50	60	70
$\hat{t}_{0.975}^{LWS}(n)$	2.148 (0.047)	2.087 (0.040)	2.056 (0.046)	2.027 (0.045)	2.017 (0.046)	2.012 (0.045)
$t_{0.975}(n)$	2.085	2.043	2.022	2.009	2.000	1.995

n	80	90	100	110	120	130
$\hat{t}_{0.975}^{LWS}(n)$	2.008 (0.040)	1.999 (0.041)	1.992 (0.040)	1.991 (0.041)	1.990 (0.040)	1.988 (0.040)
$t_{0.975}(n)$	1.990	1.987	1.984	1.982	1.980	1.978

n	140	150	160	170	180	190
$\hat{t}_{0.975}^{LWS}(n)$	1.986 (0.043)	1.989 (0.041)	1.975 (0.035)	1.974 (0.035)	1.973 (0.035)	1.973 (0.035)
$t_{0.975}(n)$	1.977	1.976	1.975	1.974	1.974	1.973

For further details see Víšek (2012b).

The instrumental weighted variables - consistency and algorithm

The classical method of *instrumental variables* is employed in the situation when the explanatory variables are correlated with disturbances which implies that the OLS are biased and inconsistent. The method, due to the fact that the economic (and generally social sciences) data suffer by this correlation, became the most requently used estimating method in econometrics. Their definition is as follows.

Definition 3 *For any sequence of p-dimensional random vectors* $\{Z_i\}_{i=1}^{\infty}$ *the solution(s) of the (vector) equation*

$$\sum_{i=1}^{n} Z_i \left(Y_i - X_i'\beta\right) = Z' \left(Y - X\beta\right) = 0 \tag{13}$$

will be called the estimator obtained by means of the method of instrumental variables (or instrumental variables, for short) and denoted by $\hat{\beta}^{(IV,n)}$.

Their robustified version can be defined in an analogy of moving from OLS to LWS (compare the **normal equations** for OLS

$$\sum_{i=1}^{n} X_i \left(Y_i - X_i'\beta\right) = 0$$

with the **normal equations** for LWS, see (10).

Definition 4 *For any sequence of p-dimensional random vectors* $\{Z_i\}_{i=1}^{\infty}$ *the solution(s) of the (vector) equation*

$$\sum_{i=1}^{n} w\left(F_{\beta}^{(n)}(|r_i(\beta)|)\right) Z_i \left(Y_i - X_i'\beta\right) = 0 \tag{14}$$

will be called the estimator obtained by means of the method of instrumental weighted variables (or instrumental weighted variables (IWV), for short) and denoted by $\hat{\beta}^{(IWV,n,w)}$*, see Víšek (2004), (2006c).*

To be able to prove the consistency we need to enlarge up-to-here introduced notations. In what follows we shall denote the joint d. f. of explanatory variables, of instrumental variables and of disturbances by $F_{X,Z,e}(x,z,r)$ and of course the marginal d. f.'s by $F_{X,Z}(x,z)$, $F_{X,e}(x,r)$, $F_X(x), F_Z(z)$ etc. We will need also the following notation. For any $\beta \in R^p$ the distribution of the product $\beta' Z X' \beta$ will be denoted $F_{\beta' Z X' \beta}(u)$, i. e.

$$F_{\beta' Z X' \beta}(u) = P(\beta' Z_1 X_1' \beta < u) \tag{15}$$

and similarly as in (**??**), the corresponding empirical distribution will be denoted $F^{(n)}_{\beta' Z X' \beta}(u)$, so that

$$F^{(n)}_{\beta' Z X' \beta}(u) = \frac{1}{n} \sum_{j=1}^{n} I\left\{\beta' Z_j X_j' \beta < u\right\}. \tag{16}$$

For any $\lambda \in R^+$ and any $a \in R$ put

$$\gamma_{\lambda,a} = \sup_{\|\beta\|=\lambda} F_{\beta Z X' \beta}(a). \tag{17}$$

Notice please that due to the fact that the surface of ball $\{\beta \in R^p, \|\beta\| = \lambda\}$ is compact, there is $\beta_\lambda \in \{\beta \in R^p, \|\beta\| = \lambda\}$ so that

$$\gamma_{\lambda,a} = F_{\beta_\lambda' Z X' \beta_\lambda}(a). \tag{18}$$

For any $\lambda \in R^+$ let us denote

$$\tau_\lambda = - \inf_{\|\beta\| \le \lambda} \beta' I\!E \left[Z_1 X_1' \cdot I\{\beta' Z_1 X_1' \beta < 0\}\right] \beta. \tag{19}$$

Notice please that due to the indicator $I\{\beta' Z_1 X_1' \beta < 0\}$ in (19) $\tau_\lambda \geq 0$ and that again due to the fact that the ball $\{\beta \in R^p, \|\beta\| \leq \lambda\}$ is compact, the infimum is finite, since there is a $\tilde{\beta} \in \{\beta \in R^p, \|\beta\| \leq \lambda\}$ so that

$$\tau_\lambda = -\tilde{\beta}' I\!\!E \left[Z_1 X_1' \cdot I\{\tilde{\beta}' Z_1 X_1' \tilde{\beta} < 0\} \right] \tilde{\beta}. \tag{20}$$

The classical regression analysis accepted the assumption that $I\!\!E Z_1 X_1'$ is regular and $I\!\!E \{e_1 | Z_1\} = 0$ (see e. g. Bowden, Turkington (1984) or Judge et al. (1985)) to be able to prove consistency of the estimator obtained by the method of *instrumental variables*. We need to assume similar ones.

Conditions $\mathcal{C}3'$ *For any* $n \in \mathcal{N}$ *the vector equation*

$$\beta' \sum_{i=1}^{n} I\!\!E \left[w\left(\overline{F}_{n,\beta}(|r_i(\beta)|) \right) Z_i \left(e_i - X_i' \left(\beta - \beta^0 \right) \right) \right] = 0 \tag{21}$$

in the variable $\beta \in R^p$ *has unique solution* $\beta^0 = 0$. Moreover, we need also

Conditions $\mathcal{C}4$ *The instrumental variables* $\{Z_i\}_{i=1}^{\infty}$ *are independent and identically distributed with distribution function* $F_Z(z)$. *Moreover, they are independent from the sequence* $\{e_i\}_{i=1}^{\infty}$. *Further, the joint distribution function* $F_{X,Z}(x,z)$ *is absolutely continuous with a density* $f_{X,Z}(x,z)$ *bounded by* $U_{ZX} < \infty$. *Further* $I\!\!E \left\{ w(F_{\beta^0}(|e_1|)) Z_1 X_1' \right\}$ *as well as* $I\!\!E Z_1 Z_1'$ *are positive definite (one can compare C3 with Víšek (1998a) where we considered instrumental* M*-estimators and the discussion of assumptions for* M*-instrumental variables was given) and there is* $q > 1$ *so that* $I\!\!E \{\|Z_1\| \cdot \|X_1\|\}^q < \infty$. *Finally, there is* $a > 0$, $b \in (0,1)$ *and* $\lambda > 0$ *so that*

$$a \cdot (b - \gamma_{\lambda,a}) \cdot w(b) > \tau_\lambda \tag{22}$$

for $\gamma_{\lambda,a}$ *and* τ_λ *given by (17) and (19).*

Theorem 4 *Let* **Conditions** $\mathcal{C}1$, $\mathcal{C}2$, $\mathcal{C}3'$ *and* $\mathcal{C}4$ *be fulfilled. Then any sequence* $\left\{ \hat{\beta}^{(IWV,n,w)} \right\}_{n=1}^{\infty}$ *of the solutions of normal equations* $I\!N E_{Z,n}(\hat{\beta}^{(IWV,n,w)}) = 0$ *is weakly consistent.*

For the proof see Víšek (2012a) where one can find also a lot of references on the classical *instrumental variables*.

The modification of the algorithm (which was described above for the *least weighted squares* (LWS)) is straightforward. Instead of calculating the ıweighted least squares we calculate the *weighted instrumental variables*. The algorithm was presented in Víšek (2006b), see also Víšek (2004) or (2006b) for further details.

Again to illustrate how the algorithm works we include a table.

TABLE 4

True coeffs $\boldsymbol{\beta^0}$	**3.5**	**-1.1**	**8.4**	**5.2**	**9.8**

Heteroscedastic disturbances, correlated with explanatory variables

TABLE 4 *(continued)*

Outliers: for 20% of observations with the largest absolute value of response variable we put $Y_i = -2 * Y_i$					
$\hat{\beta}^{OLS}_{(\mathrm{var}(\hat{\beta}^{OLS}))}$	$-3.87_{(4.87)}$	$0.56_{(5.30)}$	$-8.07_{(4.74)}$	$-5.44_{(5.13)}$	$-8.98_{(4.14)}$
$\hat{\beta}^{IV}{}_{(\mathrm{var}(\hat{\beta}^{IV}))}$	$-3.02_{(27.1)}$	$1.25_{(28.1)}$	$-7.29_{(24.4)}$	$-4.87_{(26.4)}$	$-8.44_{(25.6)}$
$\hat{\beta}^{LWS}_{(\mathrm{var}(\hat{\beta}^{LWS}))}$	$3.94_{(0.04)}$	$-0.57_{(0.04)}$	$8.77_{(0.06)}$	$5.61_{(0.05)}$	$10.1_{(0.07)}$
$\hat{\beta}^{IWV}_{(\mathrm{var}(\hat{\beta}^{IWV}))}$	$3.03_{(2.91)}$	$-1.03_{(2.12)}$	$7.37_{(10.7)}$	$4.63_{(3.17)}$	$8.65_{(12.1)}$
Leverage points: for 4% of observations with the largest norm of vector of explanatory variables we put $\tilde{X}_i = 10 \cdot X_i$ and $Y_i = -\tilde{X}_i' \cdot \beta^0$					
$\hat{\beta}^{OLS}_{(\mathrm{var}(\hat{\beta}^{OLS}))}$	$-2.69_{(151.7)}$	$0.96_{(148.6)}$	$-6.20_{(139.7)}$	$-3.32_{(147.6)}$	$-7.32_{(135.4)}$
$\hat{\beta}^{IV}{}_{(\mathrm{var}(\hat{\beta}^{IV}))}$	$-3.07_{(435.9)}$	$0.16_{(440.4)}$	$-5.68_{(401.9)}$	$-3.74_{(425.9)}$	$-6.45_{(397.7)}$
$\hat{\beta}^{LWS}_{(\mathrm{var}(\hat{\beta}^{LWS}))}$	$4.15_{(1.19)}$	$-0.35_{(1.18)}$	$8.93_{(1.08)}$	$5.83_{(1.11)}$	$10.2_{(0.82)}$
$\hat{\beta}^{IWV}_{(\mathrm{var}(\hat{\beta}^{IWV}))}$	$3.49_{(0.68)}$	$-1.21_{(0.72)}$	$8.34_{(0.66)}$	$5.16_{(0.60)}$	$9.72_{(0.68)}$
Outliers (20%) & leverage points (4%) were generated in the same way as above					
$\hat{\beta}^{OLS}_{(\mathrm{var}(\hat{\beta}^{OLS}))}$	$-6.92_{(139.0)}$	$2.38_{(146.7)}$	$-16.1_{(117.4)}$	$-9.97_{(135.4)}$	$-18.3_{(119.8)}$
$\hat{\beta}^{IV}{}_{(\mathrm{var}(\hat{\beta}^{IV}))}$	$-5.76_{(601.6)}$	$0.14_{(599.8)}$	$-13.9_{(566.7)}$	$-7.39_{(572.7)}$	$-15.1_{(516.3)}$
$\hat{\beta}^{LWS}_{(\mathrm{var}(\hat{\beta}^{LWS}))}$	$3.97_{(0.07)}$	$-0.59_{(0.06)}$	$8.84_{(0.08)}$	$5.65_{(0.07)}$	$10.22_{(0.09)}$
$\hat{\beta}^{IWV}_{(\mathrm{var}(\hat{\beta}^{IWV}))}$	$3.34_{(1.18)}$	$-1.13_{(1.12)}$	$8.08_{(0.91)}$	$5.00_{(1.05)}$	$9.41_{(1.39)}$

The paper (Víšek (2012a) from which we have borrowed the previous table) contains also a large numerical study which confirmed a good estimating quality of IWV but also addressed the problem of optimal shape of the weight function in (6) and (14). The surprising result indicating that even in the situation when data contain only a few leverage points, say up to 2%, the weight function should decrease from one to zero rather slowly. The result is presented by the next figure.

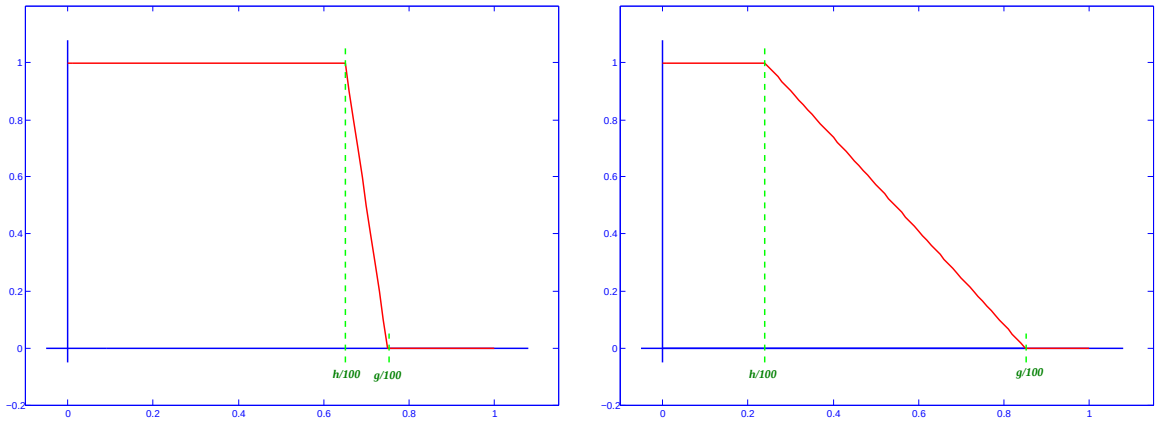


FIGURE 5
An intuitively optimal and by simulations approved the optimal weight function (left and right frame, respectively) for the contamination represented by 10% of outliers and 4% of leverage points.

Conclusions

On the example of the search for the model of export from the Czech republic to EU in 1993 we have seen that the disadvantage of the "switch" effect can be utilized as a diagnostic tool and in the situation when data are in fact a mixture of two (or several) subpopulations we can use the high breakdown point estimator as means for decomposition of data sets on reasonable subpopulations. On the other hand, when data are "only" contaminated and/or contain some (highly) influential observations it is better to depress the influence of these observations in a smooth way, just by weighting down the influence of some observations by assigning the weights to the order statistics of squared residuals. This way proved to be more efficient than to try by means of some external (usually geometric) rule to discover the suspicious observations and then to depress the influence of their residuals.

The important property of the estimator of regression model is its independence on the scale of units of device (by which we measure the variables) and on the coordinate system. Mathematically it means that the estimator is scale- and regression-equivariant. Some robust estimators, e. g. the M-estimators in its simplest definition, posses these properties only after the studentization of the residuals by scale-invariant and regression-equivarint estimator of standard deviation of disturbances. It is not simple to calculate such estimators (of standard deviation of disturbances) and moreover they are usually based on some preliminary estimator of regression model which is (directly) scale- and regression-equivariant. The remedy is to use either generalized M-estimators (where we estimate simultaneously standard deviation of disturbances and the regression coefficients) or to employ estimators which are "automatically" scale- and regression-equivariant, as LMS, LTS or LWS. The former are again computationally intensive and the algorithms are not simple to implement (although the complexity of algorithm and a tedious implementation need not necessarily hamper the employment of the method - if once implemented, reviewed in a good computational

journal and acknowledged as effective, the user can employ it if the the implementation is user friendly). The latter can be therefore preferable.

Let us return once again to the "switch" property of the high breakdown point estimators. As FIGURE 2 indicates the property is an inevitable consequence of the requirement of high robustness in the sense of the breakdown point. But it immediately inspires a question whether the breakdown point is the good, appropriate characteristic of the robustness of the estimator. The answer is yes and no simultaneously. If we are aware of consequences which this requirement (to have high breakdown point) brings and we employ the estimator with a corresponding care and caution, everything is O.K.. However, a "defect on beauty" (at least implicitly felt) it is. So, the field for further ideas and research is opened.

At the total end of paper let as address one of the most important and unfortunately still misunderstood question. It is still one topic of the statistical folklore, spread especially among the fundamentalists, that the robust procedures pay for their robustness by the considerable decrease of efficiency. The following tables show that it is not so.

We have considered the model

$$Y_i = 1 \cdot X_{i1} - 2 \cdot X_{i1} + 3 \cdot X_{i3} - 4 \cdot X_{i4} + 5 \cdot X_{i5} + \varepsilon_i, \quad i = 1, 2, ..., 100,$$

under various situations. We generated 1000 data-sets containing 100 observations and computed $\hat{\beta}^{(OLS,n)}$, $\hat{\beta}^{(LWS,n,w)}$ and $\hat{\beta}^{(LTS,n,w)}$, so that we obtained, say

$$\left\{\hat{\beta}^{(index,k)} = (\hat{\beta}_1^{(index,k)}, \hat{\beta}_2^{(index,k)}, ..., \hat{\beta}_5^{(index,k)})'\right\}_{k=1}^{1000}$$

where the abbreviations *OLS, LWS* and *LTS* at the position of *"index"* indicate (in the next table) the method employed for the computation. The empirical means and empirical mean square errors of estimates of coefficients (over these 1000 repetitions) were computed, i. e. we report values (for $j = 1, 2, 3, 4$ and 5)

$$\hat{\beta}_j^{(index)} = \frac{1}{1000} \sum_{k=1}^{1000} \hat{\beta}_j^{(index,k)} \quad \text{and} \quad \widehat{\text{MSE}}\left(\hat{\beta}_j^{(index)}\right) = \frac{1}{1000} \sum_{k=1}^{1000} \left(\hat{\beta}_j^{(index,k)} - \beta_j^0\right)^2.$$

The following table reports the situation when the data were not contaminated. We have started with rather hight robustness of LWS and LTS - because we should assume that we don't know whether the data are contaminated or not, see the first subtable of TABLE 5. One can notice that the efficiency of LWS and LTS is lower than the efficiency of OLS but the decrease - although, let us repeat, the level of robustness of LWS and LTS was nearly 50% - was not dramatic. By the way, it shows that even in the case when we adjust the level of robustness of the estimators to be higher than it is necessary, they (LWS and LTS) do not lose so much efficiency as the fundamentalist are used to claim. Notice that LWS works a bit better than LTS - see MSE in the parentheses.

Then - after we learnt that data are less contaminated than we assumed - we decreased successively the level of robustness of LWS and of LTS and the resulting efficiensy - see last subtable of TABLE 5 - is the same as the efficiency of OLS (although LWS and LTS are still able to cope with several percent of contamination).

TABLE 5

The disturbances are homoscedastic and independent from explanatory variables. Data are not contaminated (but we do not know it - hence 4 successive tables with decreasing level of robustness of the estimators).

TABLE 5$_{(\ continued)}$

The first one contains results when we took measures against an unknown level of contamination. The number of observations h taken into account by LTS was 55% of n, the weight function w had $h = 55\%$ and $g = 85\%$ of n.					
$\hat{\beta}^{OLS}_{(\mathrm{MSE}(\hat{\beta}^{OLS}))}$	$1.00_{(0.001)}$	$-2.00_{(0.001)}$	$3.00_{(0.001)}$	$-4.00_{(0.001)}$	$5.00_{(0.001)}$
$\hat{\beta}^{LWS}_{(\mathrm{MSE}(\hat{\beta}^{LWS}))}$	$1.00_{(0.004)}$	$-2.00_{(0.004)}$	$3.00_{(0.004)}$	$-4.00_{(0.004)}$	$5.00_{(0.004)}$
$\hat{\beta}^{LTS}_{(\mathrm{MSE}(\hat{\beta}^{LTS}))}$	$1.00_{(0.008)}$	$-2.00_{(0.007)}$	$3.00_{(0.008)}$	$-4.00_{(0.008)}$	$5.00_{(0.008)}$

The second one contains results when we decreased level of robustness of LTS and LWS. The number of observations h taken into account by LTS was 75% of n, the weight function w had $h = 75\%$ and $g = 95\%$ of n (OLS would give the same results as in the previous table).					
$\hat{\beta}^{LWS}_{(\mathrm{MSE}(\hat{\beta}^{LWS}))}$	$1.00_{(0.004)}$	$-2.00_{(0.004)}$	$3.00_{(0.004)}$	$-4.00_{(0.004)}$	$5.00_{(0.004)}$
$\hat{\beta}^{LTS}_{(\mathrm{MSE}(\hat{\beta}^{LTS}))}$	$1.00_{(0.004)}$	$-2.00_{(0.004)}$	$3.00_{(0.004)}$	$-4.00_{(0.004)}$	$5.00_{(0.004)}$

The third one contains results when we again decreased level of robustness of LTS and LWS. The number of observations h taken into account by LTS was 95% of n, the weight function w had $h = 95\%$ and $g = 99\%$ of n (OLS again omitted).					
$\hat{\beta}^{LWS}_{(\mathrm{MSE}(\hat{\beta}^{LWS}))}$	$1.00_{(0.002)}$	$-2.00_{(0.002)}$	$3.00_{(0.002)}$	$-4.00_{(0.002)}$	$5.00_{(0.002)}$
$\hat{\beta}^{LTS}_{(\mathrm{MSE}(\hat{\beta}^{LTS}))}$	$1.00_{(0.002)}$	$-2.00_{(0.002)}$	$3.00_{(0.002)}$	$-4.00_{(0.002)}$	$5.00_{(0.002)}$

The fourth one contains results when we decreased level of robustness once again. The number of observations h taken into account by LTS was 99% of n, the weight function w had $h = 99\%$ and $g = 100\%$ of n (OLS again omitted).					
$\hat{\beta}^{LWS}_{(\mathrm{MSE}(\hat{\beta}^{LWS}))}$	$1.00_{(0.001)}$	$-2.00_{(0.001)}$	$3.00_{(0.001)}$	$-4.00_{(0.001)}$	$5.00_{(0.001)}$
$\hat{\beta}^{LTS}_{(\mathrm{MSE}(\hat{\beta}^{LTS}))}$	$1.00_{(0.001)}$	$-2.00_{(0.001)}$	$3.00_{(0.001)}$	$-4.00_{(0.001)}$	$5.00_{(0.001)}$

So, let us conclude by a question:

What is a tax we pay for robustness of the procedures?

In fact we have to pay, nothing in the world is gratis. **We pay by the complexity of the proofs of the corresponding assertions, by the complexity of the algorithms which have to be taylored for the methods individually, by the difficulties when we implement the algorithms and when we verify their quality and, the last but surely not least, that the employment requires much better instructed and experienced user than the classical methods, much more carefull interpretation.** But it seems quite natural - driving the space shuttle need more knowledge than driving the car. Unfortunately, larger and larger abilities and user friedliness of commercially supplied packages (which is by itself a good trend) encourage users to apply methods without learning their advantages and disadvanteges, challenges and threads. All after James Heckman wrote: *"More and more thesis employ with less and less understanding and negligible practice econometrical methods, requiring deep comprehension."*, see Heckman (2007).

Appendix

Breakdown point

Definition 5 *Let F, G be d.f. and for any $A = (-\infty, a)$, put $A^\varepsilon = (-\infty, a + \varepsilon)$. Then*

$$\pi(F, G) = \inf_{\varepsilon \in [0,1]} \{\varepsilon : F(A) \leq G(A^\varepsilon) + \varepsilon \quad \textit{and} \quad G(A) \leq F(A^\varepsilon) + \varepsilon\}$$

is called Prokhorov distance of F and G.

Definition 6

$$\varepsilon^* = \sup_{\varepsilon \in [0,1]} \left\{\varepsilon : \exists K_\varepsilon \subset R^p, K_\varepsilon \; is \; compact \; and \; \pi(F, G) < \varepsilon \Rightarrow G(\hat{\beta}^{(n)} \in K_\varepsilon) \to 1 \; for \; n \to \infty\right\}.$$

is called breakdown point of the estimator T at the d.f. F.

Since the previous definition is not very transparent, frequently another one is given.

Definition 7 *Let F_n be e.d.f. corresponding to the data (Y, X). Now let m rows of these data are corrupted in an arbitrary way, i.e. replaced by any values of response variable and of explanatory variables. Denote these corrupted data by $(\tilde{Y}, \tilde{X})$ and the bias by*

$$\textit{bias}(m, T, (Y, X), (\tilde{Y}, \tilde{X})) = \left\|\hat{\beta}^{(n)}(Y, X) - \hat{\beta}^{(n)}(\tilde{Y}, \tilde{X})\right\|.$$

The empirical breakdown point is given as

$$\varepsilon_n^* = \textit{min}\left\{\frac{m}{n} : \textit{bias}(m, T, (Y, X), (\tilde{Y}, \tilde{X})) \; infinite\right\}.$$

If the limit

$$\hat{\varepsilon}^* = \lim_{n \to \infty} \varepsilon_n^*$$

exists, we call it asymptotic empirical breakdown point.

Outliers and leverage points

We speak about the *outlier* if the value of the response variable Y is atypical in the data, i. e. the value is outlying with respect to other observations in "given region" of explanatory space.
We speak about the *leverage point* if the observations is outlying in the space of explanatory variables.

Siegel's *repeated median*

Definition 8 *Let for the points $(Y_{i_1}, X'_{i_1})', (Y_{i_2}, X'_{i_2})', ..., (Y_{i_{p+1}}, X'_{i_{p+1}})'$ of $(p+1)$-dimensional Euclidean space which are in general position $\beta(i_1, i_2, ..., i_{p+1})$ be the $(p+1)$-dimensional vector of regression coefficients of the uniquely given p-dimensional plane and $\beta_j(i_1, i_2, ..., i_{p+1})$ its j-th coordinate. Then the repeated median estimator is given as*

$$\hat{\beta}_j^{(R,n)} = \underset{i_1}{med} \left(... \left(\underset{i_{p-1}}{med} \left(\underset{i_p}{med} \; \beta^{(j)}(i_1, i_2, ..., i_{p+1}) \right) \right) ... \right) \tag{23}$$

for $j = 1, 2, ..., p$.

Engine Knock Data - Hettmansperger & Sheather

TABLE A1

	Correct data air/fuel of the 2nd observation = 14.1			Damaged data air/fuel of the 2nd observation = 15.1		
Predictor	H-S-LMS	Boček-LMS	exact LTS	H-S-LMS	Boček-LMS	exact LTS
intercept	30.0826	30.0404	35.1134	-86.4967	48.38094	-88.7289
Spark	0.2107	0.1441	-0.0275	4.5864	-0.73189	4.7194
Air	2.9049	3.0784	2.9490	1.2087	3.39255	1.0576
Intake	0.5555	0.4600	0.4774	1.4678	0.19479	1.5693
Exhaust	-0.0092	-0.0069	-0.0091	0.0687	-0.01090	0.0676
11^{th} res.	0.3209	0.2278	0.3071	0.4930	0.45058	0.5392

Scale- and regression-equivariance of the estimator of regression coefficients

The estimator $\hat{\beta}$ of β^0 (employing data (Y, X) where Y is n-dimensional random vector, i. e. $Y : \Omega \rightarrow R^n$ and X is a random matrix of type $(n \times p)$, rows of which are p-dimensional random vectors which are independent and identically distributed) is scale-equivariant, if for any $c \in R^+$ and for any matrix X of type $n \times p$ we have

$$\hat{\beta}(cY, X) = c\hat{\beta}(Y, X)$$

and regression-equivariant if for any $b \in R^p, Y \in R^n$ and for any matrix X of type $n \times p$

$$\hat{\beta}(Y + Xb, X) = \hat{\beta}(Y, X) + b.$$

Scale-equivariance and regression-invariance of the estimator of standard deviation of disturbances

The estimator $\hat{\sigma}$ of standard deviation σ of error terms is said to be scale-equivariant if for any $c \in R^+, Y \in R^n$ and $X \in M(n, p)$

$$\hat{\sigma}(cY, X) = c \cdot \hat{\sigma}(Y, X)$$

and regression-invariant if for any $b \in R^p, Y \in R^n$ and $X \in M(n, p)$

$$\hat{\sigma}(Y + Xb, X) = \hat{\sigma}(Y, X).$$

Results of the analysis of export from the Czech republic to EU

TABLE A2

The development of regression coefficients, of the residual sums of squares and the estimates of regression coefficients when we increase number of observation taking into account $h = 48, 49, ..., 59$ (S_R^2 denotes the sum of squared residuals and R^2 the coefficient of determination).

h	48.000	49.000	50.000	51.000	52.000	53.000
$\hat{\beta}_1^{(LTS,n,h)}$	4.554	6.372	6.328	6.550	6.508	6.586
$\hat{\beta}_2^{(LTS,n,h)}$	-0.040	-0.029	-0.026	-0.028	-0.028	-0.032
$\hat{\beta}_3^{(LTS,n,h)}$	0.292	-0.004	-0.001	-0.014	-0.009	-0.026
$\hat{\beta}_4^{(LTS,n,h)}$	-0.249	-0.127	-0.127	-0.124	-0.125	-0.124
$\hat{\beta}_5^{(LTS,n,h)}$	1.500	0.922	0.890	0.917	0.949	1.003
$\hat{\beta}_6^{(LTS,n,h)}$	-1.581	-2.955	-2.933	-3.026	-2.978	-3.244
$\hat{\beta}_7^{(LTS,n,h)}$	1.264	1.156	1.139	1.127	1.136	1.041
$\hat{\beta}_8^{(LTS,n,h)}$	0.019	0.395	0.391	0.373	0.339	0.484
S_R^2	7.262	7.865	8.574	9.367	10.279	11.192
R^2	0.980	0.986	0.985	0.984	0.982	0.981
h	54.000	55.000	56.000	57.000	58.000	59.000
$\hat{\beta}_1^{(LTS,n,h)}$	6.547	4.641	4.883	4.838	2.771	2.748
$\hat{\beta}_2^{(LTS,n,h)}$	-0.032	-0.039	-0.041	-0.043	-0.021	-0.027
$\hat{\beta}_3^{(LTS,n,h)}$	-0.022	0.357	0.363	0.378	0.462	0.454
$\hat{\beta}_4^{(LTS,n,h)}$	-0.124	-0.288	-0.289	-0.295	-0.132	-0.127
$\hat{\beta}_5^{(LTS,n,h)}$	1.035	1.233	1.163	1.211	1.220	1.157
$\hat{\beta}_6^{(LTS,n,h)}$	-3.199	-1.861	-1.835	-1.830	-1.295	-1.291
$\hat{\beta}_7^{(LTS,n,h)}$	1.048	1.080	1.031	1.074	1.029	1.014
$\hat{\beta}_8^{(LTS,n,h)}$	0.452	0.083	-0.041	-0.064	0.025	0.054
S_R^2	13.485	13.485	14.566	15.863	17.209	18.486
R^2	0.974	0.974	0.979	0.978	0.976	0.974

TABLE A3

The development of regression coefficients, of the residual sums of squares and the estimates of regression coefficients when we increase number of observation taking into account $h = 29, 30, ..., 36$ for the complementary subpopulation containing 36 industries (S_R^2 denotes again the sum of squared residuals and R^2 the coefficient of determination).

h	29.000	30.000	31.000	32.000	33.000	34.000	35.000	36.000
$\hat{\beta}_1^{(LTS,n,h)}$	2.495	-2.144	-1.836	-1.959	-0.634	0.077	-2.482	-1.840
$\hat{\beta}_2^{(LTS,n,h)}$	0.063	-0.037	-0.081	0.040	0.089	0.055	0.018	-0.030
$\hat{\beta}_3^{(LTS,n,h)}$	0.064	0.260	0.163	0.337	0.235	0.114	0.570	0.456
$\hat{\beta}_4^{(LTS,n,h)}$	-0.049	0.293	0.313	0.257	0.249	0.250	0.340	0.354
$\hat{\beta}_5^{(LTS,n,h)}$	1.714	1.541	1.767	1.136	1.174	1.423	0.268	0.484
$\hat{\beta}_6^{(LTS,n,h)}$	-0.940	1.251	1.173	1.351	0.690	0.517	1.235	1.087
$\hat{\beta}_7^{(LTS,n,h)}$	2.051	2.689	2.599	2.809	2.691	2.618	1.704	1.502
$\hat{\beta}_8^{(LTS,n,h)}$	0.333	0.350	0.507	-0.013	-0.051	0.055	-0.251	-0.136
S_R^2	14.805	18.960	23.734	28.730	34.205	43.493	73.457	73.457
R^2	0.958	0.961	0.955	0.942	0.932	0.919	0.864	0.864

Generalization of Kolmogorov-Smirnov result

Lemma 1 *Let the **Conditions** $\mathcal{C}2$ hold. For any $\varepsilon > 0$ there is a constant K_ε and $n_\varepsilon \in \mathcal{N}$ so that for all $n > n_\varepsilon$*

$$P\left(\left\{\omega \in \Omega : \sup_{v \in R^+} \sup_{\beta \in R^p} \sqrt{n}\left|F_\beta^{(n)}(v) - \overline{F}_{n,\beta}(v)\right| < K_\varepsilon\right\}\right) > 1 - \varepsilon.$$

For the **proof** of lemma see Víšek (2011a), the main technical tool is the Skorochod embedding into the Wiener process, see Portnoy (1983)

References

[1] Anděl, J. (1978): *Matematická statistika*, SNTL & ALFA, Praha, Bratislava.

[2] Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, J. W. Tukey (1972): *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton, N. J.

[3] Antoch, J., J. Á. Víšek (1991): Robust estimation in linear models and its computational aspects. *Contributions to Statistics: Computational Aspects of Model Choice, Springer Verlag, (1992), ed. J. Antoch, 39 - 104.*

[4] Beran, R. (1978): An efficient and robust adaptive estimator of location. *Ann. Statist. 6, 292–313.*

[5] Bickel, P. J. (1975): One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc. 70, 428–433.*

[6] Boček, P., P. Lachout (1993): Linear programming approach to LMS-estimation. *Memorial volume of Comput. Statist. & Data Analysis 19(1995), 129 - 134.*

[7] Bowden, R. J., D. A. Turkington (1984): *Instrumental variables.* Cambridge: Cambridge Univ. Press.

[8] Čížek, P., J. Á. Víšek (2000): Least trimmed squares. *XPLORE, Application Guide, 49 - 64.* Springer Verlag, (2000), Berlin, eds. W. Hărdle, Z. Hlavka, S. Klinke.

[9] Cobb, C., P. H. Douglas (1928): A Theory of Production. *American Economic Review, 18:1, Supplement, March, 139-165.*

[10] Croux, C., Rousseeuw, P. J. (1992): *A class of high-breakdown scale estimators based on subranges.* Communications in Statistics - Theory and Methods 21, 1935 - 1951.

[11] Davis, P. L.(1993): Aspects of robust linear regression. *Ann. Statist. 21, 1843 - 1899.*

[12] Er-Wei Bai (2003): A random least-trimmed-squares identification algorithm. *Automatica, 39, 1651-1659.*

[13] Fisher, R. A. (1920): A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean squares error. *Mon. Not. Roy. Astr. Soc. vol. 80, 758–770.*

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

[15] Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel (1986): *Robust Statistics – The Approach Based on Influence Functions.* New York: J.Wiley *&* Son.

[16] Hampel, F. R. (1968): *Contributions to the theory of robust estimation.* Ph. D. thesis. University of California, Berkeley.

[17] Hampel, F. R. (1974): The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69, No. 364, 383–393.*

[18] Hansen, L. P., T. Sargent (2008): *Robustness*. Princeton University Press, Princeton.

[19] Hawkins, D. M. (1994): The feasible solution algorithm for least trimmed squares regression. Computational Statistics and Data Analysis 17, 185 - 196.

[20] Hawkins, D. M., D. J. Olive (1999): Improved feasible solution algorithms for breakdown estimation. *Computational Statistics & Data Analysis, Volume 30, Number 1, 1 - 12.*

[21] Hawkins, D. M., D. J. Olive (2003): Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm. *Journal of the American Statistical Association, 97, 136-159.*

[22] Hawkins, D. M., D. J. Olive (2011): Practical High Breakdown Regression. *Preprint.* bibitem Heckman, J. J. (2007): The American high school graduation rate : Trend and levels. MIT Press 29, 244 - 262.

[23] Hettmansperger, T. P., S. J. Sheather (1992): A Cautionary Note on the Method of Least Median Squares. *The American Statistician 46, 79–83.*

[24] Hogg, R. V. (1974): Adaptive robust procedures: A partial review and some suggestions for future applications and theory (with discussion). *J. Am. Statist. Assoc. 69, 909–927.*

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

[26] Huber, P. J. (1965): A robust version of the probability ratio test. *Ann. Math. Statist. 36, 1753–1758.*

[27] Huber, P.J.(1981): *Robust Statistics.* New York: J.Wiley *&* Sons.

[28] Hubert, M., P. J. Rousseeuw, S. Van Aelst (2002): [Inconsistency of Resampling Algorithms for High-Breakdown Regression Estimators and a New Algorithm]: Comment *Journal of the American Statistical Association, 97, 151-153.*

[29] Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, T. C. Lee (1985): *The Theory and Practice of Econometrics.* New York: J.Wiley *&* Sons(second edition).

[30] Jurečková, J. (1977): Asymptotic relations of M-estimates and R-estimates in linear regression model. *Ann. Statist. 5, 464–472.*

[31] Jurečková, J., P. K. Sen (1993): Regression rank scores scale statistics and studentization in linear models. *Proceedings of the Fifth Prague Symposium on Asymptotic Statistics, Physica Verlag, 111-121.*

[32] Kalina, J. (2004): Durbin-Watson Test for Least Weighted Squares. *Proceedings of COMPSTAT 2004, Physica-Verlag/Springer, 1287 - 1294.*

[33] Klouda, K. (2007): *Algorithms for computing robust regression estimates*. Diploma thesis, Czech Technical University, Faculty of Nuclear Scinces and Physical Engeneering.

[34] Koenker,R., G. Bassett (1978): Regression quantiles. *Econometrica, 46, 33-50.*

[35] Marazzi, A. (1992): *Algorithms, Routines and S Functions for Robust Statistics.* Wadsworth & Brooks/Cole Publishing Company, Belmont, California, 1992.

[36] Maronna, R. A., O. H. Bustos, V. J. Yohai (1979): Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation. Eds. T. Gasser and M. Rosenblatt, New York: Springer-Verlag, 91 - 116.*

[37] Maronna, R. A., V. J. Yohai (1981): The breakdown point of simultaneous general M-estimates of regression and scale. *J. of Amer. Statist. Association, vol. 86, no 415, 699 - 704.*

[38] Mason, R. L., R. F. Gunst, J. L. Hess (1989): *Statistical Design and Analysis of Experiments*, New York: J.Wiley & Sons.

[39] Portnoy, S. (1983): Tightness of the sequence of empiric c.d.f. processes defined from regression fractiles. In *Robust and Nonlinear Time-Series Analysis (J. Franke, W. Hărdle, D. Martin, eds.), 231 - 246. Springer-Verlag, New York, 1983.*

[40] Rousseeuw, P.J., K. van Driessen (2005): Computing LTS Regression for Large Data Sets. *Data Mining and Knowledge Discovery, 12, 29 - 45, 2006.*

[41] Rousseeuw, P. J., A. M. Leroy (1987): *Robust Regression and Outlier Detection.* New York: J.Wiley & Sons.

[42] Siegel, A. F. (1982): Robust regression using repeated medians. *Biometrica, 69, 242 - 244.*

[43] Stahel, W., S. Weisberg, eds. (1991): *Directions in Robust Statistics and Diagnostics.* New York, Springer-Verlag.

[44] Stigler, S. M. (1977): Do robust estimator work with real data? *Annals of Statistcs 5, 1055 - 1098*.

[45] Tukey, J. W. (1960): A survey of sampling from contaminated distribution. In: *Contributions to probability and statistics.Ed. J. Olkin. Stanford University Press, Stanford, California, 448 - 485.*

[46] Víšek, J. Á. (1994): A cautionary note on the method of Least Median of Squares reconsidered, *Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Lachout, P., Víšek, J.Á. (eds), Academy of Sciences of the Czech Republic, Prague, 1994, 254 - 259.*

[47] Víšek, J. Á. (1996a): Sensitivity analysis of M-estimates. *Annals of the Institute of Statistical Mathematics, 48(1996), 469-495.*

[48] Víšek, J. Á. (1996b): On high breakdown point estimation. *Computational Statistics (1996) 11:137-146, Berlin*.

[49] Víšek, J. Á. (1998): Robust specification test. *Proceedings of Prague Stochastics'98 (eds. Marie Hušková, Petr Lachout & Jan Ámos Víšek, published by Union of Czechoslovak Mathematicians and Physicists), 1998, 581 - 586.*

[50] Víšek, J. Á. (2000a): Regression with high breakdown point. Proceedings of *ROBUST 2000, 324 - 356.*

[51] Víšek, J. Á. (2000b): On the diversity of estimates. *Computational Statistics and Data Analysis 34, (2000) 67 - 89.*

[52] Víšek, J. Á. (2002a): Sensitivity analysis of M-estimates of nonlinear regression model: Influence of data subsets. *Annals of the Institute of Statistical Mathematics, Vol. 54, No.2, 261 - 290, 2002*.

[53] Víšek, J. Á. (2002b): White test for the least weigthed squares. *COMPSTAT 2002, Berlin, Proceedings of the Conference CompStat 2002 - Short Communications and Poster (CD), eds. S. Klinke, P. Ahrend, L. Richter*.

[54] Víšek, J. Á. (2003): Durbin-Watson statistic for the least trimmed squares. *Bulletin of the Czech Econometric Society, vol. 8, 14/2001, 1 - 40.*

[55] Víšek, J. Á. (2004): Robustifying instrumental variables. *Proceedings of COMPSTAT'2004. Physica-Verlag/Springer. ISBN 3-7908-1554-3. 1947 - 1954.* Víšek, J. Á. (2006a): The least trimmed squares. Sensitivity study. *Proceedings of the Prague Stochastics 2006, Prague, 21.8.-25.8.2006, eds. Marie Hušková & Martin Janžura, matfyzpress, ISBN 80-86732-75-4, 728-738.* Víšek, J. Á. (2006b): Instrumental Weighted Variables - algorithm. *Proceedings of the COMPSTAT 2006, eds. A. Rizzi & M. Vichi, Physica-Verlag (Springer Company) Heidelberg 2006, 777-786.* Víšek, J. Á. (2006c): Instrumental weighted variables. *Austrian Journal of Statistics, vol 35(2006), No. 2& 3, 379 - 387.*

[56] Víšek, J. Á. (2010a): Robust error-term-scale estimate. *IMS Collections. Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: Festschrift for Jana Jurečková, Vol. 7(2010), 254 - 267.*

[57] Víšek, J. Á. (2010b): Weak $\sqrt{n}$-consistency of the least weighted squares under heteroscedasticity. *Acta Universitatis Carolinae, Mathematica et Physica, 2/51, 2010, 71 - 82.*

[58] Víšek, J. Á. (2011a): Empirical distribution function under heteroscedasticity. *Statistics, 45, 5, 497-508.*

[59] Víšek, J. Á. (2011b): Consistency of the least weighted squares under heteroscedasticity. *Kybernetika 47 , 179-206, 2011.*

[60] Víšek, J. Á. (2012a): Instrumental weighted variables under heteroscedasticity. Part I. Consistency. Part II. Numerical study. *Preprint.*

[61] Víšek, J. Á. (2012b): Diagnostics of robust identification of model, submitted to *Quality Engineering*.

[62] Víšek, J. Á. (2012c): Robust estimation of model with the fixed and random effects. *Proceedings of the COMPSTAT 2012, ISBN: 978-90-73592-32-2, The International Statistical Institute/International Association for Statistical Computing, eds. Colubi A., Fokianos K., Gonzalez-Rodriguez G., Kontoghiorghes E. J., 855 - 865.*

[63] Yohai, V. J., Maronna, R. A. (1979): Asymptotic behaviour of M-estimators for the linear model. *Ann. Statist. 7, 248–268.*