REGRESSION WITH HIGH BREAKDOWN POINT

JAN ÁMOS VÍŠEK

Abstract. The paper discusses in details various aspects of the point estimation, classic paradigm, Hampel's program and a new paradigm, including reliability of algorithm and its implementation, the role of accompanying procedures and of heuristics. A special attention in paid to the high breakdown point estimation, corresponding prejudices and misleading ideas. It reports theoretical results as well as practical consequences, describes a reliable algorithm for evaluation of *the least trimmed squares* and finally illustrates by the results of analysis of real data how powerful tool the estimators with high breakdown point can be.

Статья подробно обсуждает различные асспекты оценок в класическойпарадигме, в Гампелове и в новойпарадигме, которая включает достоверность алгоритма и тоже имплементации, место сопрождящих процедуре и геуристики. Специальное внимание посвещено оцениванию с высоким бодом отказания, к ним принадлежающые предпассудки и ошибочные представления. Она приносит теоретические результаты точно как и практические следствия. Описывает достоверный алгоритмус для вычисления найменших отрубаных квадратов и потом приносит как илустрацию результаты анализа реальных данных и показывает как мощным инструментом могут быть оценки с высоким бодом отказания.

Introduction and notation

It is sometimes claimed that nearly 95% of statistical applications are from regression analysis. Of course, such a claims have their roots in fact that not only linear regression is taken into account but also nonlinear models, analysis of variance, logit and probit (or generally probability) regression models, regression trees, over seemingly unrelated equations up to, maybe, cointegration analysis. However, it is not important how large the percentage of regression tasks really is, if it is 95% or "only" about 80 %. In any case, a large number of problems solved in the framework of regression analysis indicates that we should pay an attention to the following question:

What features of modern estimator of regression model we should ask for ? What accompanying equipment of the (robust) estimator is to include ?

2000 *Mathematics Subject Classification.* Primary 62F35; Secondary 62J99.

Key words and phrases. Point (robust) estimators, classic paradigm, Hampel's program, a new paradigm, high breakdown point, reliability of evaluation and implementation, diagnostics and sensitivity studies, diversity of estimates, export and FDI..

Research was supported by grant of GA AV ČR no. A 2075803.

The present paper tries to find an answer to the question by discussing related topics, starting with a bit of history over some well-known stories which appeared in developing robust analysis and finishing with an illustrative example.

The regression analysis is for the most of its users still connected with *the least squares* and of course with names of Adrien Marie Legendre (1805) and Carl Friedrich Gauss (1809). Although both Gauss as Legendre used the least squares for fitting models to data, it was Sir Francis Galton (1885) who gave the name to the branch. He used the least squares due to their simplicity in comparison with others methods, however it may be of interest that at least three statistical problems which would by today classified as regression analysis, were solved before Legendre and Gauss and what is even more interesting, they were solved by L_1 technique, see Galilei (1632), Boscovisch (1757) and Laplace (1793).

It is well known that there was a discussion between Ronald Alyner Fisher and Francis Ysidor Edgeworth which method is to be used but further development confirmed Fisher's "solution" who preferred the least squares before L_1-approach. Of course, the main reason was the fact that this method offered a "simple" formula for the estimator and hence it was feasible to compute it. As we shall see later, this is and will be probably one of key requirement for any estimator to be useful, namely an existence of a feasible and reliable algorithm(and better an available implementation of it). A lot of theoreticians in past considered "the process of establishing a new estimator" to be finished when the theoretical properties as the consistency, the asymptotic normality and representation were proved. Nowadays, it is very clear that it is not true. A "discovery" of a reliable algorithm, implementation of it, evaluating the (approximation to the) estimator is a reasonable time, verification that the algorithm gives really good results and developing of the "accompanying" procedures (as alternative estimator for the situations when collinearity or dependence between explanatory variables and disturbances appear and corresponding tests for recognizing that) are also very important. Without all this "equipment" the estimator is handicapped, if not disqualified at all. For further discussion see Víšek (2000 d).

First of all, let us introduce notations. Let N denote the set of all positive integers, R the real line and R^p the p dimensional Euclidean space. We shall consider for any $n \in N$ the linear regression model

$$Y_i = X_i^{\mathrm{T}} \beta^0 + e_i, \quad i = 1, 2, \ldots, n \tag{1}$$

where $Y = (Y_1, Y_2, \ldots, Y_n)^{\mathrm{T}}$ is the response variable, $\{X_i^{\mathrm{T}}\}_{i=1}^n$ ($X_i \in R^p$) is a sequence of vectors of explanatory variables, β^0 is the "true" vector of regression coefficients and $\{e_i\}_{i=1}^n$ ($e_i \in R$) is a sequence of independent and identically distributed random variables, representing random fluctuations (or disturbances, if you wish; since the *disturbances* is shorter, we shall use it). The upper index "$^{\mathrm{T}}$" indicates transposition. (As implicitly follows from this notation, we shall assume all vectors to be column ones.) Finally, let us denote for any $n \in N$ by $X = (X_1^{\mathrm{T}}, X_2^{\mathrm{T}}, ..., X_n^{\mathrm{T}})^{\mathrm{T}}$ the design matrix and by $e = (e_1, e_2, ..., e_n)^{\mathrm{T}}$ the vector of disturbances. Then we can rewrite (1) into (sometimes) more convenient form

$$Y = X\beta^0 + e. \tag{2}$$

We have omitted an indication of the dimension of matrix and of vectors which would presumably unnecessarily burden the notation. Let us notice that in the case that the intercept is included in the model the first coordinates of all vectors X_i's are assumed to be equal to 1. In other words, if the explanatory vectors X_i's are assumed to be random, they have degenerated first coordinate. There are of course, except of special cases, well-known reasons for inclusion of the intercept into the model. Let us realize that in the case when we decide not to include intercept into the model we implicitly assume, in some sense, an absolute character of data and in fact simultaneously give up otherwise natural requirement of scale- and regression-equivariance of the estimator of the regression coefficients. That is why we shall assume in the rest of paper that the intercept is included into the model (there is, of course, at least one other reason for it, for details see Víšek (1997 a, b).

The least squares and classic paradigma of estimation

Let us recall that the (ordinary) least squares estimator of β^0 is given by

$$\hat{\beta}^{(LS,n)} = \operatorname*{arg\,min}_{\beta\in R^p} \sum_{i=1}^{n}(Y_i - X_i^{\mathrm{T}}\beta)^2 = \operatorname*{arg\,min}_{\beta\in R^p} (Y - X\beta)^{\mathrm{T}}(Y - X\beta)$$

which yields

$$\hat{\beta}^{(LS,n)} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$$

where we have assumed that X is of full rank (since in the paper only cross-sectional data will be assumed, this is not substantial restriction of generality).

May be that it is only a statistical folklore but probably already Sir Francis Galton really knew a formula describing sensitivity of the least squares with respect to the deletion of one observation. It reads

$$\hat{\beta}^{(LS,n)} - \hat{\beta}^{(LS,n,\ell)} = \left(X^{\{\ell\}\mathrm{T}}X^{\{\ell\}}\right)^{-1} X_\ell \left(Y_\ell - X_\ell^{\mathrm{T}}\hat{\beta}^{(LS,n)}\right)$$

where $X^{\{\ell\}}$ is design matrix without the ℓ-th row. We shall need it later.

It is clear that one can propose an estimator in thousand ways but then it is necessary to prove that such estimator fulfills some collection of desirable properties. Again one may prefer these and other may ask for another features of estimator but we will presumably agree that there is a set of requirements which we all would ask for. Such collection may be called a *classic paradigm.* It has probably following items

- (unbiasedness),
- ($\sqrt{n}$-)consistency,
- (asymptotic) efficiency,
- asymptotic normality.

The round brackets around the word *unbiasedness* should indicate that for many estimators, especially for modern ones, we are not able to prove unbiasedness. Sometimes we even know that the estimator is not unbiased. In such a case we put up with asymptotic unbiasedness and/or even with consistency (only). A similar facts are indicated by the round brackets in the case of the consistency and the efficiency.

Robust approach and Hampel's program

The very beginning of robust studies are connected with name of John Tukey who in forties began studies of model for contamination of data. Although later there appeared some technical reports by him on problems with the contamination of data, a substantial progress began in middle sixties and is associated with Peter J. Huber and Frank R. Hampel. The first one initiated an approach based on (nearly) strict application of classic statistical principles (as e. g. maximum likelihood or the least favorable pair of distributions) but in his framework instead of (parameterized) families of distribution functions, families of "neighborhoods" of families of distribution functions are taken into account. Converted commas around the word *neighborhoods* hints that we do not have in mind in this case neighborhoods in the topological sense (i. e. open sets) but some sets which contain a central model for disturbances, as an inner point. This central model is usually one of classic stochastic models, e. g. (standard) normal model (an example of such a *neighborhood* will be given below - see definition of min-max bias estimator).

Hampel's approach is based on the interpretation of any estimator as a function(al) of empirical distribution function and the studies of properties of the estimator are performed then by means of the derivative along some trajectory in the space of all distribution functions. To clarify this let us give a simple example, using the most frequently used statistics, the *mean.* To make the explanation correct, we shall use somewhat more complex notation than it is usual.

First of all, let us realize that we "interpret" the *mean* as a sum of n numbers divided by n but, as the statistics, that is sum of n random variables divided by n. So let us consider a sequence of independent identically distributed random variables $\{Z_i(\omega)\}_{i=1}^{\infty}$, defined of course on a basic probability space $(\Omega, \mathcal{C}, P)$. Then for any $\omega \in \Omega$ we have $\bar{Z}(\omega) = \frac{1}{n}\sum_{i=1}^{n} Z_i(\omega)$.

Secondly, let us recall what is the empirical distribution function for the considered sequence of random variables. Let us denote by $I_A(\omega)$ the indicator of the set A, i. e. $I_A(\omega) = 1$ if $\omega \in A$ and $I_A(\omega) = 0$ otherwise. We usually speak about empirical distribution function in the context of having at disposal of $z_1, z_2, ..., z_n$, the *realization* of the first n random variable. Then we consider the empirical distribution function as a step functions, having all steps of magnitude equal to $\frac{1}{n}$ at the points $z_1, z_2, ..., z_n$. We shall denote this empirical distribution function by $F_n(z)$. The *realization* $z_1, z_2, ..., z_n$ is nothing else than the value of the corresponding random variables at some point ω_0, i. e. $z_1 = Z_1(\omega_0), z_2 = Z_2(\omega_0), ..., z_n = Z_n(\omega_0)$. The empirical distribution function, now considered as a statistic, at any $\omega \in \Omega$ is given as

$$F_{(n,\omega)}(z) = \frac{1}{n}\sum_{i=1}^{n} I_{\{Z_i(\omega)\le z\}}(\omega).$$

Finally, let us return to the *mean.* Since we have $z_1 = Z_1(\omega_0), z_2 = Z_2(\omega_0), ..., z_n = Z_n(\omega_0)$, we obtain

$$\bar{Z}(\omega_0) = \frac{1}{n}\sum_{i=1}^{n} Z_i(\omega_0) = \frac{1}{n}\sum_{i=1}^{n} z_i$$

$$= \int_{z\in R} z \mathrm{d}F_n(z) = \int_{z\in R} z \mathrm{d}F_n(z, \omega_0).$$

The verification of the equality needs a few second, if we meet with such arguments (and notation) for the first time but principally is straightforward.

The last equality shows that $\bar{Z}(\omega) = T(F_{n,\omega}(z))$, i. e. the empirical mean can be interpreted as a functional T of the empirical distribution function. A theoretical counterpart of it is the fact that the "theoretical" mean is the same functional T of the theoretical distribution function, say F. Hampel's approach then employs derivatives of functionals (as Frechét or Gâteaux) to study the properties of given estimator.

To make Hampel's approach applicable for a wide class of estimators we usually do not ask for existence of Fréchet or Gâteaux derivative but we try to find derivative of corresponding functional along such trajectories which go from the central model to the distribution function degenerated at a point $z \in R$. In other words, the *influence function* is defined as

$$IF(z,F,T) = \lim_{h \to 0} \frac{T((1-h)F + h \cdot z) - T(F)}{h}$$

at the points where the limit exists. After all, as the name of the function hints, the *influence function* indicates an influence of one additional observation, when we put it at point z, on the value of the estimate. That is why some other characteristics of robust estimators are defined utilizing it.

Nowadays the offer of robust procedures is nearly infinite. (Of course, we speak now about theoretical results, not about available implementations.) But it is still possible to claim that M-, L- and R-estimators are the most popular classes. The first inherited the name from the *maximum likelihood estimators* since the most of M-estimators are similar to the *maximum likelihood estimators*, hence *Maximum-likelihood-like estimators.* The second class is based on *Linear combination of order statistics* and the third one employs the *Rank statistics.* Since in what follows we shall need M-estimators, let us recall that they are given by

$$\hat{\beta}^{(\psi,n)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} \rho(Y_i - X_i^{\mathrm{T}}\beta) \tag{3}$$

where the superindex ψ indicates that the derivatives of ρ (which stays in the normal equations which are in turn used for finding the value of M-estimator) is just ψ.

And quite unexpectedly there appeared one problem which was not felt so acute when the classic, we mean maximum likelihood, moment estimators, etc., were used. The problem was that the robust estimators were not generally scale- and regression-equivariant. Since the requirements of the scale- and regression-equivariance represents the fact that it should be irrelevant how the axes and units of measurement were selected, they are not only very natural but from the application point of view nearly unavoidable. The estimators which does not possess these properties are seriously handicapped in the applications.

The statisticians were aware of it but they (tacitly) believed that the studentization of residuals (in the case of robust estimators of regression coefficients) by means of a preliminary scale-equivariant estimator of standard deviation is the remedy. It appeared that the requirements on the estimator of standard deviation have to be enlarged, namely that it must be also regression-invariant - to reach scale- and regression-equivariance of the estimator of regression coefficients, see Bickel (1975) or Jurečková *&* Sen (1993). In other words, (3) has to have the form

$$\hat{\beta}^{(\psi,n)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} \rho\left(\frac{Y_i - X_i^{\mathrm{T}}\beta}{s}\right)$$

where s is an estimate of standard deviation which is scale-equivariant and regression-invariant. However it is not very easy to find (theoretically) such estimator, leaving aside that the evaluation would be complicated. Nowadays there are only a few proposals of such estimators, see again Jurečková *&* Sen (1993) or Víšek (1999 a) (the latter even without a theoretical background, being still only implemented and numerically tested). So it seems that it is preferable to employ the robust estimators which are scale- and regression-equivariant, without any studentization. We shall employ in the rest of this paper two such estimators, having moreover high breakdown point.

The study of the influence function led to establishing some characteristics of robust estimator. Later they were "collected" into a list which became known as *Hampel's program*. It may be viewed as an enlargement of *classic paradigm*. It reads

- acceptably low *gross-error sensitivity*,
- maximal attainable *efficiency*,
- not very large *local shift sensitivity*,
- preferably *finite rejection point*,
- as high as possible *breakdown point*.

Before giving definitions and an explanation of items of *Hampel's program* let us remark that the word "enlargement" is of course meant in a somewhat vague way. E. g. for most of robust estimators we are not able to prove unbiasedness. As we have already argued, they are defined (typically) by an extremal problem and hence there is not usually any (simple) formula for them. It makes a proof of unbiasedness nearly impossible. We usually put up with consistency.

DEFINITION 1. *Gross-error sensitivity is given as*

$$\gamma(T,F) = \sup_{z\in R} |IF(z,F,T)|\,. \tag{4}$$

An idea which is behind the definition of the *gross-error sensitivity* is given by relation

$$\sqrt{n}\left(\hat\beta^{(\psi,n)} - \beta^0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF(z_i,F,T) + \ \textit{remainder term} \tag{5}$$

which can be derived for M-estimators, see Hampel et al. (1986), p. 85. So $IF(z_i,F,T))$ is a contribution of the observation z_i to the value of the estimator. On the other hand, we may consider the *contribution of the i-th observation* to be the *sensitivity of the estimator* to the i-th observation. On the other hand, as the *sensitivity of the estimator* to the i-th observation one usually considers the value of the change of estimator when the i-th observation is deleted from data.

Does it coincide with the idea which is "behind" the gross-error sensitivity?

The answer is: **Sometimes yes sometimes no.**

The asymptotic representation of the change of M-estimator of regression coefficients is generally given by

$$n\left(\hat\beta^{(\psi,n)} - \hat\beta^{(\psi,n,\ell)}\right) = -\kappa^{-1}Q^{-1}X_\ell\psi\left(\frac{Y_\ell - X_\ell^{\mathrm T}\hat\beta^{(\psi,n)}}{\hat\sigma_n}\right) + \mathcal{R}_n \ \textit{as}\ n\to\infty \tag{6}$$

where

$$\kappa = \hat\sigma_n I\!\!E\psi'\left(\frac{e}{\hat\sigma_n}\right) + \sum_{k=1}^{s} f(\sigma r_k)\left[\psi(r_k+) - \psi(r_k-)\right]$$

with $r_1, r_2, ..., r_s$ being points of discontinuity (if any) of corresponding ψ-function and

$$\mathcal{R}_n = \left(W(\sum_{i=1}^{n} \mu_{i1}^{(j)}(n,t,u)), W(\sum_{i=1}^{n} \mu_{i2}^{(j)}(n,t,u)), ..., W(\sum_{i=1}^{n} \mu_{ip}^{(j)}(n,t,u)) \right).$$

where W denotes Wiener process and $\mu_{ik}^{(j)}(n,t,u)$ some appropriate stopping times, see Víšek (1996 a). The last term of (6) is present in the formula (6) in the case when ψ-function is discontinuous.

Before proceeding further let us add that the influence function is in fact linearly proportional to ψ-function which generated respective M-estimator. It means that (5) can be rewritten as

$$\sqrt{n}\left(\hat{\beta}^{(\psi,n)} - \beta^0\right) = \frac{C_n^{(2)}}{\sqrt{n}} \sum_{i=1}^{n} X_i \psi(z_i) + \textit{ remainder term} \tag{7}$$

(notice that parameters F and T of the influence function (seemingly) disappeared; it is due to fact that the influence function of M-estimator does not depend of the underlying stochastic model and T is in fact indicated by ψ). It means that in the case when ψ-function is continuous (5) (or if you wish, (7)) and (6) give the same indication of the magnitude of contribution of observation z_i to the value of estimator. However, in the case when ψ-function is discontinuous (5) and (6) do not "agree", i. e. the change of the estimate when we delete one point from data may be much larger than the gross-error sensitivity indicates. What is however much worse, is the fact that we cannot control, by an upper bound of the absolute values of the discontinuous ψ-function, the maximal value of the norm $\left\| n\left(\hat{\beta}^{(\psi,n)} - \hat{\beta}^{(\psi,n,\ell)}\right)\right\|$, for more details see Víšek (1998 c). **It hints that in the case when[1] we decide to employ M-estimator, it is preferable to use a continuous redescending ψ-function**. The last requirement, namely that ψ-function should be redescending, is given by the fact that the M-estimators generated by redescending ψ-functions are able to cope, at least partially, with leverage points.

Of course, the second point of Hampel's program has no definition but it is nearly evident that it means. First of all, let us recall once again that usually we have at hand no formula for robust estimator, only an extremal problem which defines the estimator. So to prove, that the estimator in question, is the best one among all estimators for all distribution from some Huber model of contaminacy (of some central classic stochastic model), is nearly impossible. So, if we define efficiency, as it is usually done, as the ratio of (asymptotic) variances, e. i. ratio of that minimal attainable variance and the variance of the estimator in question, it is not usually possible to evaluate it. That is why we usually only estimate efficiency in more or less vague way, see Huber (1981).

DEFINITION 2. *The local shift sensitivity is given as*

$$\sup_{z,v \in R} \left| \frac{IF(z,F,T) - IF(v,F,T)}{z - v} \right|. \tag{8}$$

The idea which inspired the definition of *local shift sensitivity* is transparent (and moreover it is hinted by its name). It of course indicates how the estimator reacts on (a large number of) small changes of observations, so, except of others, it also

[1] Despite of all what was said about the difficulties with a preliminary estimate of scale for studentization of residuals.

describes how the estimator takes into account that we do not measure the variables in question "precisely". Let us realize that except of discrete variables all others are always measured as rounded to that level of preciseness which is given by smallest amount of given entity which our instrument can still measure.

DEFINITION 3. *The rejection point is defined as*

$$\inf_{z \in R} \; \{|z| : \forall (v, \; |v| \geq |z|) \; : \; IF(v, F, T) = 0\} \, . \tag{9}$$

We have already mentioned that it is preferable, if employing M-estimators, to use redescending ψ-function. The requirement that they are to be directly zero, instead of (e. g.) converging (steeply) to zero, is somewhat, more or less of technical character since it is easier to cope with such estimators from both point of views, theoretical and evaluationary.

Breakdown point

The most prominent and probably the most misunderstood item of Hampel's program is the *breakdown point.*

Let us denote Prohorov distance by π.

DEFINITION 4. *The breakdown point of an estimator* T_n *of the parameter* $\theta \in \Theta$

$$\varepsilon^* = \sup_{0 \leq \varepsilon \leq 1} \; \{\varepsilon \; : \; \exists K(\varepsilon) \subset \Theta, K(\varepsilon) \; compact \; : \\ \pi(F, G) < \varepsilon \Rightarrow G(T_n \in K(\varepsilon)) \to 1 \; for \; n \to \infty\} \, .$$

The idea which led to the definition of *breakdown point* is easy to trace out. It says that we try to learn how high level of contamination data could be, without breaking a "reasonable" behaviour of the estimator. The mathematical formalization of course seems at the first glance somewhat strange since it "reflects" reasonable behaviour as *not escaping from a (fixed) compact set.* As however the definition is given in an asymptotic form, it is correct. For finite samples it gives the ratio of "good" and "bad" observation which does not imply explossion or implossion of the estimator. Let us add that it is usual to give the value of breakdown point in "%", not as it would follow from definition as a number from the interval $[0, 0.5]$.

As an example of estimators, even from the classic statistics, which possess the lowest and highest possible breakdown point are *mean* and *median.* An arbitrarily large change of the former may be caused by one observation placed sufficiently far away from other, however the latter can be substantially shifted only by moving at least half of observations. On the other hand, just opposite situation is what concerns the *local shift sensitivity*, which is the lowest for the median and the highest for the mean.

Enlarging Hampel's program

As we have already mentioned the *breakdown point* is one of the most discussed points and, as from these discussion follows, probably the most misunderstood characteristics of robust estimators. After all, an example of misunderstanding the behaviour of the estimators with high breakdown point, will be given below. Hence before discussing properties of one of the estimator of regression coefficients with possibly high (in fact even controllable) breakdown point we are going to make an excursion into the history of such estimation and evaluation of corresponding estimators.

What is really worthwhile of our attention is the fact that building the theory of robust statistics was always accompanied by "an empirical" studies made mostly on some simulations, see e. g. Andrews et al. (1972), Huber (1973), Lax (1975), Schweingruber (1980), Ruppert *&* Carroll (1980) and some papers in *Directions in Robust ...*. What is however sorrowful is the fact that most of the programs which were used for these empirical studies were not included into any commercially available statistical package, so that there is evident lack of (reliable) implementations of efficient algorithms.

We have already discussed the *mean*, with zero breakdown point, and the *median*, with 50% breakdown point. In the regression framework a lot of estimators (of model) were established nevertheless no with very high breakdown point. Moreover the result of Maronna (1976) (see also Maronna et al. (1979)) brought a disappointment since it appeared that the breakdown point of M-estimators cannot exceed $\frac{1}{p}$. On the other hand, the existence of the estimator of location having breakdown point as high as 50% was a challenge for statisticians to find also in the regression framework an estimator with such a high breakdown point (or to prove that it is not possible). The challenge can be viewed at least from two standpoints. The first one is to interpret it just as a purely mathematical challenge to reach a boundary of possibility (realize please that 50% breakdown point is the maximal possible, in some sense). The second standpoint is to see the problem of reaching 50% breakdown point as the problem inspired by a hope that such an estimator (even probably losing a lot of efficiency) can hint what is the *true model behind the data*. And since the idea of a *true model behind the data* implicitly assumed that such a *true model* is independent from contamination of data (any under 50%), the high breakdown point was believed to be something which guarantees stability of estimator with respect to nearly any change and/or damage of data.

Of course, this hope was tacit from the very beginning because the idea of some *true, "objectively existing" model behind the data* is hardly justifiable and hence misleading at all. Nevertheless, a disappointment which arrived when it appeared that the estimators with high breakdown point work in other way than it was wrongly and senselessly assumed is still at the roots of prevailing part of criticism of the estimators with high breakdown point. We shall return to this discussion later when we will be able understand better the arguments. Now let us perform the promised excursion into the history.

As a winner of the contest for the construction of an estimator with 50% breakdown point is usually assumed the *repeated median* by Siegel (1982), although already in Hampel (1975) can be found an idea which led to *the least median of squares*, (mainly) for the framework of location parameter in details studied in Rousseeuw (1984). We shall not recall the definition of the *repeated median* since it (the definition as well as the estimator per se) is complicated and probably it was never implemented, maybe not even for the experimental purposes. Fortunately, a bit later appeared *the least median of squares* by Peter J. Rousseeeuw. The definition in this case is quite transparent.

DEFINITION 5. *The least median of squares is given as*

$$\hat{\beta}^{(LMS,n)} = \arg\min_{\beta \in R^p} \operatorname{med}\left\{r_i^2(\beta)\right\}. \tag{10}$$

This is an original definition which gave, for the evident reasons, the name to the estimator. Immediately however it was generalized on *the least h-th order statistic of squared residuals* but this name is not used and the old one overlived. To be able to introduce it, let us denote for any $\beta \in R^p$ the i-th residual by $r_i(\beta) = Y_i - \sum_{j=1}^p X_{ij}\beta_j$ and the order statistics of squared residuals by $r^2_{(i)}(\beta)$, it means that

$$r^2_{(1)}(\beta) \ \le \ r^2_{(2)}(\beta) \ \le \ \dots \ \le \ r^2_{(h)}(\beta) \ \le \ r^2_{(h+1)}(\beta) \ \le \ \dots \ \le \ r^2_{(n)}(\beta).$$

Now, a modified definition of $\hat\beta^{(LMS,n)}$ reads as follows.

DEFINITION 6. *For an h, $\frac{n}{2} \le h \le n$ the least median of squares is given as*

$$\hat\beta^{(LMS,n,h)} = \ \underset{\beta\in R^p}{\arg\min} \ r^2_{(h)}(\beta)^2. \tag{11}$$

As it was already noted the estimator was studied in details for location parameter in Rousseeuw (1984) and for regression framework in Rousseeuw *&* Leroy (1987). This reference contains also a short passage devoted to *the least trimmed squares.* They are defined as

DEFINITION 7. *For an h, $\frac{n}{2} \le h \le n$ the least trimmed squares are given as*

$$\hat\beta^{(LTS,n,h)} = \ \underset{\beta\in R^p}{\arg\min} \ \sum_{i=1}^{h} r^2_{(i)}(\beta)^2. \tag{12}$$

Both $\hat\beta^{(LMS,n,h)}$ as well as $\hat\beta^{(LTS,n,h)}$ have for $h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right]$ the highest attainable breakdown point (for the scale- and regression-equivariant) estimators, namely $\varepsilon^* = \left(\left[\frac{n-p}{2}\right] + 1\right)/n$. Later there appeared a lot of estimators of regression coefficients with high breakdown point as $S-estimators$, *minimum distance estimators, minimum volume estimators, minimum determinant estimators, min-max bias estimators,* etc., to name at least a few among many others. The last ones, namely *min-max bias estimators* were studied in Martin et al. (1989). Let us stop for a while with them.

Instead of giving a definition of the *min-max bias estimators*, let us try to describe them in words, since it will clarify immediately the idea which led to their proposal. We shall keep the framework for explanation as simple as possible.

So, let us consider a probability measure on the real line, say in a form of distribution function F, e. g. standard normal distribution $\mathcal{N}(0,1)$. Denote by $\mathcal{H}$ the set of all distribution functions on the real line. Further, for a fix $\delta \in [0,1]$ denote

$$\mathcal{F}_{(F,\delta)} = \left\{G, \ G(x) = (1-\varepsilon)\cdot F(x) + \varepsilon \cdot H(x), \ H \in \mathcal{H}, \ 0 \le \varepsilon \le \delta\right\}.$$

Moreover, denote by $\mathcal{B} = \left\{\hat\beta \ : \ I\!\!E_F\hat\beta = \beta^0\right\}$. Now, consider a fix estimator, say $\hat\beta^{(1)} \in \mathcal{B}$, and find the distribution function $G^{(\hat\beta^{(1)})} \in \mathcal{F}_{(F,\delta)}$ for which the bias of the estimator $\hat\beta^{(1)}$ from β^0 is maximal. It means that

$$G^{(\hat\beta^{(1)})} = \ \underset{G\in\mathcal{F}_{(F,\delta)}}{\arg\max} \ \left\| I\!\!E_G\hat\beta^{(1)} - \beta^0 \right\|$$

and put (maximal bias) $MB(\hat\beta^{(1)}) = \left\| I\!\!E_{G^{(\hat\beta^{(1)})}}\hat\beta^{(1)} - \beta^0\right\|$. Finally, put

$$\hat\beta^{(MinMaxBias)} = \ \underset{\beta\in\mathcal{B}}{\arg\min} \ MB(\hat\beta). \tag{13}$$

DEFINITION 8. *The estimator given in (13) is called min-max bias estimator.*

Of course, it is not simple to evaluate such estimator but in special cases it is possible. Since under some technical conditions they are equivalent with S-estimators, we may find them as estimators which minimize scale (of residuals). It means, except of others that we need not to specify $\mathcal{F}_{(F,\delta)}$, $\mathcal{B}$. etc., we just only minimize scale of residuals. The following picture shows an example of their work (notice please that except of data depicted by crosses there is one datum at the origin, given by a small circle).

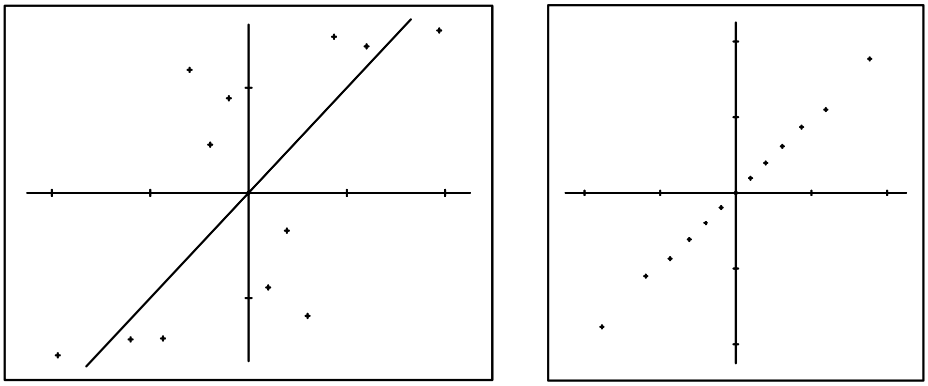


Figure 1 *"Pure" data - S-estimate and normal plot of residuals.*

The same or only indiscriminately different estimate of model we obtain by the least trimmed squares, by the least median of squares, by M-estimator with Hampel's ψ-function etc.

Now, let us shift the datum given by circle (at the origin) somewhat right and low

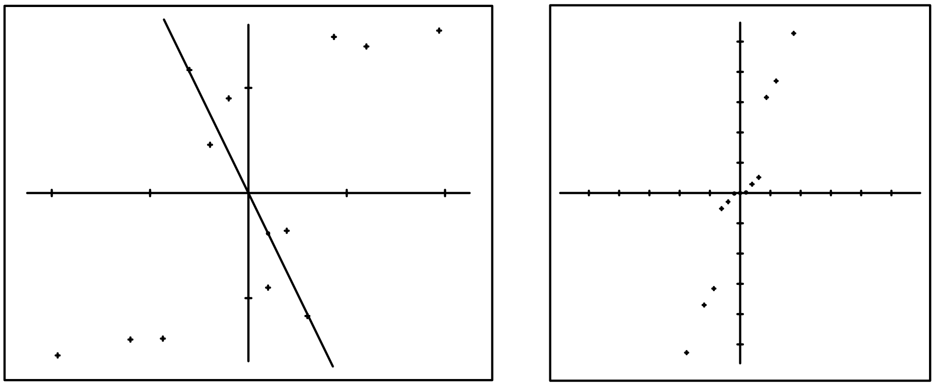


Figure 2 *"Contaminated" data - S-estimate and normal plot of residuals.*
(Notice the different scale of the right picture of Figure 1 and 2)

A large change of S-estimate quite contrasts with the other mentioned estimates, namely the least trimmed squares, the least median of squares and M-estimate with Hampel's ψ-function, which remain nearly the same as they were for data given at Figure 1. The normal plots of these estimates also remain nearly as the plot given in the right hand side of Figure 1 while the normal plot of S-estimate (given on the right hand side of Figure 2) is quite unsatisfactory. All these facts hint that for the "contaminated" data a "true" model (or if you wish, reasonable model) is still that one given in at the left hand side of Figure 1. But it means that min-max bias estimate is considerably biased, much more than the other estimates.

What is a reason for it ?

The reason is simple. The theory built up for min-max bias estimator in Martin et al. (1989) assumed that the regression model is unknown but fix, let us say just "true", and that the distribution function belongs in $\mathcal{F}_{(F,\delta)}$ for some F and $\delta \in [0,1]$. In fact δ was just the only restriction which was prescribed *a priori.* But when the estimator was applied on data it "took" into account several, let us stress, very different regression models and according to the corresponding sets of residuals made a conclusion about the underlying regression model and the distribution function of disturbances. So Martin et al. (1989) were betrayed by the assumption (practically always (unconsciously) accepted by most of us when applying any statistical method) that "behind the data" is an *objectively existing, unique true model* and if we could increase number of observations above all limits, we inevitably arrive at it. And that *objectively existing true models* is hence included - or imprint, if you wish - (in a strange way ?) into the world around us. The problems and traps which such an approach implicitly contains were already known to Emmanuel Kant and in modern philosophy of mathematical modeling they were discussed by Ilya Prigogine and Isabella Stengers, see e. g. (1977) or (1984).

Our example demonstrates that S-estimators assumes quite different "true" model from the "true" models assumed by other estimators (for "contaminated data", i. e. for data with one datum shifted). And for this model, S-estimator gave a model which is minimally biased from "its true". It may seem strange (at the first glance) but as we shall see later it may be even formalized and then we conclude that for one, fix sequence of data two (strongly) consistent estimators may give quite different model for any size of data. Nevertheless, the conclusion from just closed discussion is that (of course) any new proposal of an estimator is to have a heuristic background, to be easier acceptable in applications, **but** the heuristics are to be well thought-out, i. e. all pros and cons are to be taken into account to arrive at heuristics which really work (especially for finite samples).

We have already mentioned how important is for any new (robust) estimator a reliable algorithm for its evaluation. Of course, it is much better, if there is an available implementation of this algorithm. This "feature" of estimator was "underrated" in the past due to the fact that there was usually available a formula for the estimator. Of course, even then we may have some problems with an implementation, see e. g. how much effort was spent to solve in a reliable way the problem of evaluating the inverse matrix in the formula for the least squares, see Antoch *&* Vorlíčková (1992).

We have also recalled that there is still a lot of misunderstandings or misbeliefs connected with the robust methods. One very deeply rooted but completely misleading was already discussed, namely erroneously assumed a large loss of efficiency. Another one is a tacit belief that the estimators with high breakdown point are stable under any circumstances. In other words, statistical folklore still assumes that since the estimators with high breakdown point are able to cope with a high level of contamination of data, it would not change too much in any small change of data.

Let us give an example which will illustrate the problems discussed in last two paragraphs and which will indicate how to cope with them. Previous to the example let us recall once again that the robust estimators are defined typically as a solution of an extremal problem. Since in the most cases we are not able to evaluate precise solution, we put up with a (tight) approximation. We shall return to this problem in details later.

T. P. Hettmansperger *&* S. J. Sheather published in 1992 results of processing data (originally studied by Mason et al. (1989)) recording dependence of number of engine knocks on the spark timing, on the air/fuel ratio, on the intake temperature and on the exhaust temperature. They reported that when they included the data into memory of PC, they wrote wrongly 15.1 for the value of air/fuel ratio of the second observation rather than the correct value 14.1. (In what follows we shall call data with the wrong value 15.1 as *damaged data* and the data with the correct value 14.1 as *correct data.*) When they noticed the error they recalculate the results with an expectation that the new results would differ from the previous ones slightly. Let us add that they used $\hat{\beta}^{(LMS,n,h)}$, i. e. the estimator with (asymptotically) 50% breakdown point. However, to their great surprise, the change of values of the estimate of regression coefficients was large, see next table. Unfortunately, they did not write by which program (i. e. by which algorithm and by which implementation) they evaluated results. Nevertheless the results, they gave in paper, were up to all given decimal digits the same as the results, the program PROGRESS (see Rousseeuw *&* Leroy (1987)) returns. (We are grateful to Peter Rousseeuw and Annick Leroy who sent us a diskette with fortran source of PROGRESS, for the possibility to use it.) The same values gives also S-PLUS (version 3.2 which was available at those days; I am afraid that nothing changed). Both programs are based on algorithm which can be called *resampling algorithm.* The algorithm randomly selects an elemental set of p points, then it fits a regression plane to them and then performs (repeatedly) its shift and rotation to decrease the value of the minimized order statistic. This step of program has its justification in the geometric characterization of $\hat{\beta}^{(LMS,n,h)}$ which was found by Joss *&* Marazzi (1990). It says that at least $p+1$ points are at the same distance from the regression plane, given by any solution of (11). So reaching that requirement this step of program is stopped and the whole procedure, of selecting another elemental set of p points, is repeated. Of course, the LMS criterion, see (11), is monitored and at the end, given by a stopping rule (e. g. by performance of *a priori* given number of repetitions), program returns as estimate that vector of regression coefficient for which the criterion was minimal.

Table 1
Estimates of regression coefficients for Engine Knock Data given in Hettmansperger & Sheather (1992)

	Estimates for correct data	Estimates for damaged data
Intercept	30.08	-86.50
Spark	0.211	4.586
Air/Fuel	2.905	1.209
Intake	0.555	1.468
Exhaust	-0.009	0.069

By a stroke of good luck, at the same time appeared an algorithm based on a dual version of linear programming problem and corresponding form of simplex method, later described in Boček & Lachout (1995) (let us call it *BL-algorithm*). Firstly, *BL-algorithm* is (many times) quicker then the *resampling algorithm*. Secondly, we did not yet found any set of data for which it gives larger value of the minimized order statistics than the other methods, see Víšek (1996 b) and (2000 a). We would like to express our gratitude to Pavel Boek who is the author of the implementation for an offer to use his software. In the next table results of both algorithms are gathered. (The abbreviations are nearly self-explaning, nevertheless let us say that $\hat{\beta}_R^{(LMS,n,h)}$ is $\hat{\beta}^{(LMS,n,h)}$ evaluated by PROGRESS while $\hat{\beta}_{BL}^{(LMS,n,h)}$ is $\hat{\beta}^{(LMS,n,h)}$ yielded by Boček's implementation; finally $r^2_{(h:n)}$ is h-th order statistics among the squared residuals, $h = 11$).

Table 2
Estimates of regression coefficients given by resampling algorithm and BL-algorithm

	Estimates for correct data		Estimates for damaged data	
Estimator	$\hat{\beta}_R^{(LMS,n,h)}$	$\hat{\beta}_{BL}^{(LMS,n,h)}$	$\hat{\beta}_R^{(LMS,n,h)}$	$\hat{\beta}_{BL}^{(LMS,n,h)}$
Intercept	30.08	30.04	-86.50	48.38
Spark	0.211	0.144	4.586	-0.732
Air/Fuel	2.905	3.078	1.209	3.393
Intake	0.555	0.460	1.468	0.195
Exhaust	-0.009	-0.007	0.069	-0.011
$r^2_{(h:n)}$	0.103	0.053	0.328	0.203

So, the conclusion is that "Hettmansperger-Sheather effect" was caused by poor algorithm they used (and which was not yet abandon). It again underlines the importance of availability of the *reliable algorithm* which is of course acceptably quick (or not very slow, if you wish) to evaluate the estimates in a reasonable time. The last requirement is also very important because when applying robust regression we need to "experiment". As we shall see later by the promised example we need to evaluate the estimates for various sets of explanatory variables, as it is also usual in the classic least squares analysis, but also for various h's, i. e. for various numbers of observations which the least median of squares or the least trimmed squares take into account. Let us add that for solution of the corresponding extremal problems (which defines the robust estimators) the routine methods for finding the extrema are not suitable due to the large number of local minima in the extremal problem of type (11) or (12). Moreover the extrema are sometimes deep sometimes rather flat. It means that in fact for every single robust estimator we have to find a tailored approach which consists of two steps. Firstly, it is necessary to invent a new algorithm (i. e. a trick) for finding an approximation to the theoretical value of the estimator in question. Secondly, another trick has to be found for checking that the algorithm gives really tight approximation.

But it is not yet the end of story. Even when we have verified that a new algorithm gives good approximation to the theoretical solution of given extremal problem, we should equip the estimator by some accompanying procedures. As we have already briefly mentioned, it is clear that similarly as in the classic analysis performed by the least squares we can meet with collinearity (hence we need a "ridge" version of the estimator in question), we may get into a situation when the disturbances are dependent with the explanatory variables (and hence we need a version of the method of instrumental variables for given robust estimator and of Hausman test), we can obtain $AR(p)$ structure of residuals (and hence we need a version of Durbin-Watson statistics for given estimator and a "remedy", i. e. some version of Praiss-Winston or Cochran-Orcutt transformations) etc. It is sufficient to look into a monograph on classic regression to learn how much accompanying procedures the (ordinary) least squares have to be able to cope with (any ?) situations in which some assumptions are distorted (of course, the handicap of such monographs is that they assume distortion of all assumptions except of the assumption of the (strict) normality). Nowadays, there is not any robust estimator fully equipped by such accompanying procedures. Moreover, it seems that there is not (even) a systematic research in that field (unfortunately).

But it is not yet the end of story. Accomplishing all these steps, we should "sell" new estimator to users (who are conservative and not too much eager to use anything new, at least up to a moment when old methods evidently fail). And that is task for the heuristics which initiated the proposal of the estimator. It is to be so easy acceptable that it should seem to user that the estimator is as "natural" as the least squares (converted commas of course indicates that there is nothing natural on the least squares except of the fact that they are intrinsically connected with Euclidian geometry which we have accepted as natural). Clearly, a few last sentences somewhat played down the role of the heuristics. In fact, its role is, of course, important also for the following reason. As it was already said, most of our results are of asymptotic type. So the heuristic is inevitably also a support that the estimator will work for the finite samples in an appropriate way, too.

It implies that we should establish a new paradigm (according to present level of knowledge) to

- *consistency*,
- *asymptotic normality*,
- reasonably high *efficiency*,
- *scale-* and *regression-equivariance*,
- quite low *gross-error sensitivity*,

- low *local shift sensitivity*,
- preferably *finite rejection point*,
- controllable *breakdown point*,
- available *diagnostics*, *sensitivity studies* and *accompanying procedures*,
- existence of an *algorithm* (better, of *an implementation with accep-*) *table complexity and reliability of evaluation*,
- an efficient and acceptable *heuristics*.

The least trimmed squares - the theory

Now we are going to present somewhat more details about the least trimmed squares $\hat{\beta}^{(LTS,n,h)}$, then to describe an efficient algorithm for the evaluation and employing the algorithm to illustrate and then to discuss some, at the first glance, strange effect of (high breakdown point) estimation. Finally, we give an example demonstrating the employment of $\hat{\beta}^{(LTS,n,h)}$ in the applications. To be able to do it we will need some assumptions.

First of all, denote $G(z)$ the distribution function of e_1^2. For some $\alpha \in [0, \frac{1}{2})$, u_α^2 will be the upper α-quantile of $G(z)$, i. e. $P(e_1^2 > u_\alpha^2) = 1 - G(u_\alpha^2) = \alpha$. Further, denote by $[a]$ the integer part of a and for any $n \in N$ put $h_n = [(1-\alpha)n]$. Finally, for an arbitrary sequence $k = \{k_i\}_{i=1}^\infty$ ($k_i \in \{0,1\}$) such that $\sum_{i=1}^n k_i = h_n$ put $Q_n(k) = \frac{1}{n}\sum_{i=1}^n k_i X_i X_i^{\mathrm{T}}$. The promised assumptions are as follows.

Assumptions $\mathcal{A}$. The sequences $\{X_i\}_{i=1}^\infty$ ($X_i \in R^p$) is a fix sequence of nonrandom vectors from R^p. Further, the sequence $\{e_i\}_{i=1}^\infty$ ($e_i \in R$) is a sequence of independent identically distributed random variables. The distribution function $F(z)$ of random fluctuation e_1 is symmetric and absolutely continuous with a bounded density $f(z)$. The density is positive on $(-\infty, \infty)$ and has bounded derivative. Moreover,

$$\sum_{i=1}^n \|X_i\|^3 = \mathcal{O}(n), \tag{14}$$

$$I\!\!E e_1^2 = \sigma_{e_1}^2 \in (0, \infty). \tag{15}$$

Uniformly with respect to k (i. e. uniformly with respect to any sequence $k = \{k_i\}_{i=1}^\infty$)

$$\lim_{n\to\infty} Q_n(k) = Q \tag{16}$$

where Q is a regular matrix (and convergence is of course assumed coordinatewise).

REMARK 1. *The assumption (16) is somewhat stronger than the usually accepted*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i X_i^{\mathrm{T}} = Q.$$

Since however we cannot guarantee which indeces will be selected by $\hat{\beta}^{(LTS,n,h)}$ into the subsample which is finally taken into account, it has to be given in this form. It is easy to give an example of data demonstrating that without (16), $\hat{\beta}^{(LTS,n,h)}$ need not be consistent.

REMARK 2. *It follows from Assumptions $\mathcal{A}$ that we shall consider the setup with nonrandom carriers (or explanatory variables, if you wish). The theory for the setup with random carriers requires some modifications what concerns the assumptions (orthogonality and sphericality conditions), see Víšek (1999 b). However, what concerns the results comparing those in Víšek (1999 b) and those given in this paper, one concludes that they are nearly identical.*

The absolute continuity of F seems at a first glance rather strong assumption. However, let us realize that firstly the (ordinary) least squares are optimal (among all estimators) only under *strict* normality of random fluctuations e_i's. The argument that without normality the least squares are still optimal among all linear estimators is true but misleading, since the restriction on the class of linear estimators is drastic. Secondly, any study of order statistics assumes the absolute continuity of the underlying distribution, since without this assumption we have got into some technical troubles as the probability that two order statistics attain the same value need not be zero.

Also assumption that the density is bounded and has bounded derivative everywhere may be considered somewhat strong. As we shall see in the next, we shall need (except of other) to estimate the probability

$$P\left(u_\lambda \le e_i \le u_\lambda + n^{-\frac{1}{2}} x_i^{\mathrm{T}} t\right) \tag{17}$$

(of course for the case when $x_i^{\mathrm{T}} t > 0$). Then it is clear that we need some assumptions on $\|X_i\|$ and on $F(z)$. If we assume that for some $K < \infty$

$$\sup_{i \in N} \|X_i\| < K, \tag{18}$$

it is evidently sufficient to assume existence of bounded derivative of density in the neighborhood of u_λ and of $-u_\lambda$. However, the assumption (18) is considered by some statisticians as inadmissibly restricting while they are willing to accept the assumptions of type (14). Then of course, the norms $\|X_i\|$, $i = 1, 2, ..., n$ are not uniformly bounded and hence to be able to achieve the equality $P\left(u_\lambda \le e_i \le u_\lambda - n^{-\frac{1}{2}} x_i^{\mathrm{T}} t\right) = \|X_i\| \mathcal{O}(n^{-\frac{1}{2}})$, we need some assumption(s) about $F(z)$ to be fulfilled on the whole support of $F(z)$. Of course, under (18) as well as under (14), it is possible to estimate probability (17), nevertheless in the former case it is straightforward while in the latter it requires rather involving considerations. Moreover, Lemma A.1 shows that from the practical point of view, there is not considerable difference between (14) and (18). Finally, results in Chatterjee, Hadi (1988), Zvára (1989) or Víšek (1996 b), (1997 a, b) (2000 b) indicate that in the case when the norm of some explanatory vectors is out of control, we cannot guarantee anything about subsample sensitivity (see also Theorem 3 below). That is why we shall also assume an alternative version of assumption .

Assumptions $\mathcal{B}$. The sequences $\{X_i\}_{i=1}^{\infty}$ ($X_i \in R^p$) is a fix sequence of nonrandom vectors from R^p. Moreover, (16) holds for some regular matrix Q. Further for any $n \in N$

$$\max_{1 \le i \le n,\ 1 \le j \le p} |X_{ij}| = \mathcal{O}(1).$$

The sequence $\{e_i\}_{i=1}^{\infty}$ ($e_i \in R$) is a sequence of independent identically distributed random variables with absolutely continuous distribution function $F(z)$. There are *neighbourhoods of u_α and of $-u_\alpha$ in which the distribution $F(z)$ has a bounded density $f(z)$ which is positive and has bounded derivative. Finally, (15) holds.*

THEOREM 1. *Let Assumptions $\mathcal{A}$ or $\mathcal{B}$ hold. Then $\hat{\beta}^{(LTS,n,h)}$ is $\sqrt{n}$-consistent, i. e.*

$$\sqrt{n}\left(\hat{\beta}^{(LTS,n,h)} - \beta^0\right) = \mathcal{O}_p(1) \quad \text{as } n \to \infty.$$

THEOREM 2. *Let Assumptions $\mathcal{A}$ or $\mathcal{B}$ be fulfilled and $(1-\alpha) - u_\alpha\left(f(u_\alpha) + f(-u_\alpha)\right) \neq 0$. Then*

$$\sqrt{n}\left(\hat{\beta}^{(LTS,n,h)} - \beta^0\right) = n^{-\frac{1}{2}} Q_n^{-1}\left[(1-\alpha) - u_\alpha(f(u_\alpha) + f(-u_\alpha))\right]^{-1} \times$$

$$\times \sum_{i=1}^{n}\left(Y_i - X_i^{\mathrm{T}}\beta^0\right) X_i \cdot I\left\{e_i^2 \le u_\alpha^2\right\} + o_p(1)$$

and $\hat{\beta}^{(LTS,n,h)}$ is asymptotically normal with mean value equal to β^0 and covariance matrix

$$V(\hat{\beta}^{(LTS,n,h)}, F) = Q_n^{-1}\left[(1-\alpha) - u_\alpha(f(u_\alpha) + f(-u_\alpha))\right]^{-2} \int_{-u_\alpha}^{u_\alpha} z^2 \mathrm{d}F(z),$$

i. e.

$$\mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(LTS,n,h)} - \beta^0\right)\right) \to \mathcal{N}(0, V(\hat{\beta}^{(LTS,n,h)}, F)) \qquad \text{as } n \to \infty.$$

THEOREM 3. *Let*

$$\mathcal{R}_n =_{\mathcal{D}} W(\sum_{i=1}^{n}(\tilde{\tau}_i^+ + \tilde{\tau}_i^-))$$

where

$$\tilde{\tau}_i^+ = \text{time for Wiener process } W(s) \text{ to exit the interval } (-\tilde{a}_i^+, \tilde{b}_i^+)$$

with

$$(-\tilde{a}_i^+, \tilde{b}_i^+) = (u_\alpha x_i^{\mathrm{T}} u[1-\tilde{\pi}_i^+], -u_\alpha x_i^{\mathrm{T}}\delta\tilde{\pi}_i^+) \text{ if } x_i^{\mathrm{T}}\delta \le 0,$$
$$(-\tilde{a}_i^+, \tilde{b}_i^+) = (-u_\alpha x_i^{\mathrm{T}}\delta\tilde{\pi}_i^+, u_\alpha x_i^{\mathrm{T}}\delta[1-\tilde{\pi}_i^+]) \text{ if } x_i^{\mathrm{T}}\delta > 0$$

and

$$\tilde{\tau}_i^- = \text{time for Wiener process } W(s) \text{ to exit the interval } (-\tilde{a}_i^-, \tilde{b}_i^-)$$

with

$$(-\tilde{a}_i^-, \tilde{b}_i^-) = (u_\alpha x_i^{\mathrm{T}}\delta[1-\tilde{\pi}_i^-], -u_\alpha x_i^{\mathrm{T}}\delta\tilde{\pi}_i^-) \text{ if } x_i^{\mathrm{T}}\delta \le 0,$$
$$(-\tilde{a}_i^-, \tilde{b}_i^-) = (-u_\alpha x_i^{\mathrm{T}} u\tilde{\pi}_i^-, u_\alpha x_i^{\mathrm{T}}\delta[1-\tilde{\pi}_i^-]) \text{ if } x_i^{\mathrm{T}}\delta > 0$$

and where

$$\delta = n\left(\hat{\beta}^{(LTS,n,h)} - \hat{\beta}^{(LTS,n-1,h,\ell)}\right),$$

$$\tilde{\pi}_i^+ = P(I\left\{r_i^2(\hat{\beta}^{(LTS,n-1,\ell)}) \le r_{(h)}^2(\hat{\beta}^{(LTS,n-1,\ell)})\right\} >$$
$$I\left\{r_i^2(\hat{\beta}^{(LTS,n,h)}) \le r_{(h)}^2(\hat{\beta}^{(LTS,n,h)})\right\})$$

and

$$\tilde{\pi}_i^- = P(I\left\{r_i^2(\hat{\beta}^{(LTS,n-1,\ell)}) \le r_{(h)}^2(\hat{\beta}^{(LTS,n-1,\ell)})\right\}$$
$$< I\left\{r_i^2(\hat{\beta}^{(LTS,n,h)}) \le r_{(h)}^2(\hat{\beta}^{(LTS,n,h)})\right\}).$$

Moreover, let $\left[(1-\alpha) - u_\alpha \left(f(u_\alpha) + f(-u_\alpha)\right) - \mathcal{R}_n\right]^{-1} = \mathcal{O}_p(1)$. *Then under Assumptions* $\mathcal{B}$ *we have*

$$n\left(\hat{\beta}^{(LTS,n,h)} - \hat{\beta}^{(LTS,n-1,h,\ell)}\right) = Q_n^{-1}\left[(1-\alpha) - u_\alpha \left(f(u_\alpha) + f(-u_\alpha)\right) - \mathcal{R}_n\right]^{-1} \times \tag{19}$$

$$\times \left(Y_\ell - X_\ell^{\mathrm{T}} \hat{\beta}^{(LTS,n,h)})\right) X_\ell + o_p(1) \qquad \textit{as } n \to \infty.$$

More details can be found in Víšek (2000 c) and in (1999 b).

As follows from Theorem 3, $n\left(\hat{\beta}^{(LTS,n,h)} - \hat{\beta}^{(LTS,n-1,\ell)}\right)$ is, except of others, proportional to $\mathcal{R}_n$ which, as we have seen, is a random variable obtained from Wiener process by plugging in appropriate stopping times. Such variable is bounded in probability but we cannot control upper bound of its absolute value by an *a priori* selected parameter. It implies that change of the estimates of regression coefficients evaluated at first for the whole data and then for the data from which the ℓ-th observation was deleted may be considerably large. The asymptotic representation (19) of $n\left(\hat{\beta}^{(LTS,n,h)} - \hat{\beta}^{(LTS,n-1,h,\ell)}\right)$ is nearly the same as the asymptotic representation of the difference $n\left(\hat{\beta}^{(M,n)} - \hat{\beta}^{(M,n-1,\ell)}\right)$, i. e. the difference of M-estimators generated by discontinuous ψ-functions, see (6) and more details in Víšek (1996 a).

May be that it is not possible to grasp it immediately and only from results which concern $\hat{\beta}^{(LTS,n,h)}$ (which were given a few lines ago) but from the proofs it is straightforward that the root of that behavior is in the "sharp" (or complete, if you wish) rejection of some observations. Hence a first conjecture may be:

CONJECTURE 1. *An estimator which comply with the new paradigm (except of the equipment by the accompanying procedures) is the least weighted squares given as*

$$\hat{\beta}^{(LWS,n,h)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{h} w_i \; r_{(i)}^2(\beta)$$

($\frac{n}{2} \le h \le n$) *for appropriately selected weights* w_i.

REMARK 3. *Notice please that again, similarly as for* $\hat{\beta}^{(LTS,n,h)}$, *the order of words is substantial, i. e. this estimator differs from the classic weighted least squares. For the former the weights are assigned to observations implicitly by the estimator itself while for the latter the weights are generated by an external rule.*

It is well-known that applying the ordinary least squares we may get in troubles when the collinearity of explanatory variables takes place. It is nowadays also well-known that in 1970 Hoerl and Kennard proposed *ridge regression* as a possible solution of the problem. The corresponding estimator is biased but it has, under some technical condition, smaller mean squared error than the classic least squares, see e. g. Zvára (1989). The ridge regression estimator is a special case of estimators with linear constraints, see e. g. Víšek (1997 a). The estimators with linear constraints are in some monographs offered as one of classic solution of collinearity, see e. g. Judge et al. (1985) (another one is the regression on the main components). It is not difficult to see that the solution of (12) can be found (theoretically) by a successive application of the least squares on all h-tuples which are subsample of data. Practically this approach is feasible only for data having not more 20 observations. Nevertheless, the algorithm which will be described below for evaluating a (tight) approximation to the solution of (12) applies also in an iterative way the least trimmed squares. Hence we may get due to collinearity of explanatory variables into the same troubles as the ordinary least squares got in. That is why we have studied also the least trimmed squares under linear constraints. The following theorem brings the asymptotic representation of such an estimator. First of all we have to give a definition of the estimator.

DEFINITION 9. *Let C be a matrix of the type $\ell \times p$ and full rank. Moreover let $\gamma \in R^\ell$. For an h, $\frac{n}{2} \le h \le n$ the least trimmed squares with the linear constraints given by matrix C are given as*

$$\hat\beta^{(LTS,C,n,h)} = \arg\min_{\beta \in R^p} \left\{ \sum_{i=1}^{h} r^2_{(i)}(\beta)^2 \text{ with } C\beta = \gamma \right\}. \tag{20}$$

THEOREM 4. *Let Assumptions $\mathcal{A}$ or $\mathcal{B}$ be fulfilled and $(1-\alpha) - u_\alpha\left(f(u_\alpha) + f(-u_\alpha)\right) \neq 0$. Moreover, let C be a matrix of the type $\ell \times p$ and full rank. Moreover let $\gamma \in R^\ell$. Denote*

$$\tilde{Q}^{-1} = Q^{-1} - Q^{-1}C^{\mathrm{T}}\left\{CQ^{-1}C^{\mathrm{T}}\right\}^{-1}C^{\mathrm{T}}Q^{-1}.$$

Then

$$\sqrt{n}\left(\hat\beta^{(LTS,C,n,h)} - \beta^0\right) = n^{-\frac{1}{2}}\tilde{Q}_n^{-1}\left[(1-\alpha) - u_\alpha(f(u_\alpha) + f(-u_\alpha))\right]^{-1} \times$$

$$\times \sum_{i=1}^{n}\left(Y_i - X_i^{\mathrm{T}}\beta^0\right)X_i \cdot I\left\{e_i^2 \le u_\alpha^2\right\} + o_p(1)$$

and $\hat\beta^{(LTS,n,h)}$ is asymptotically normal with mean value equal to β^0 and covariance matrix

$$\tilde{V}(\hat\beta^{(LTS,C,n,h)}, F) = \tilde{Q}_n^{-1}\left[(1-\alpha) - u_\alpha(f(u_\alpha) + f(-u_\alpha))\right]^{-2}\int_{-u_\alpha}^{u_\alpha} z^2 \mathrm{d}F(z),$$

i. e.

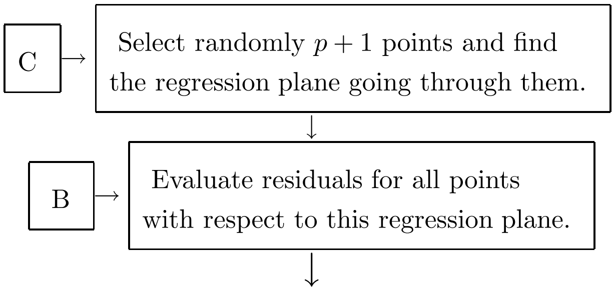
$$\mathcal{L}\left(\sqrt{n}\left(\hat\beta^{(LTS,C,n,h)} - \beta^0\right)\right) \to \mathcal{N}(0, \tilde{V}(\hat\beta^{(LTS,C,n,h)}, F)) \qquad \text{as } n \to \infty.$$

The least trimmed squares - the algorithm

Now we are going to describe algorithm which is suitable for evaluation of an approximation to the precise solution of (12) and the way how we have confirmed that the approximation is tight.

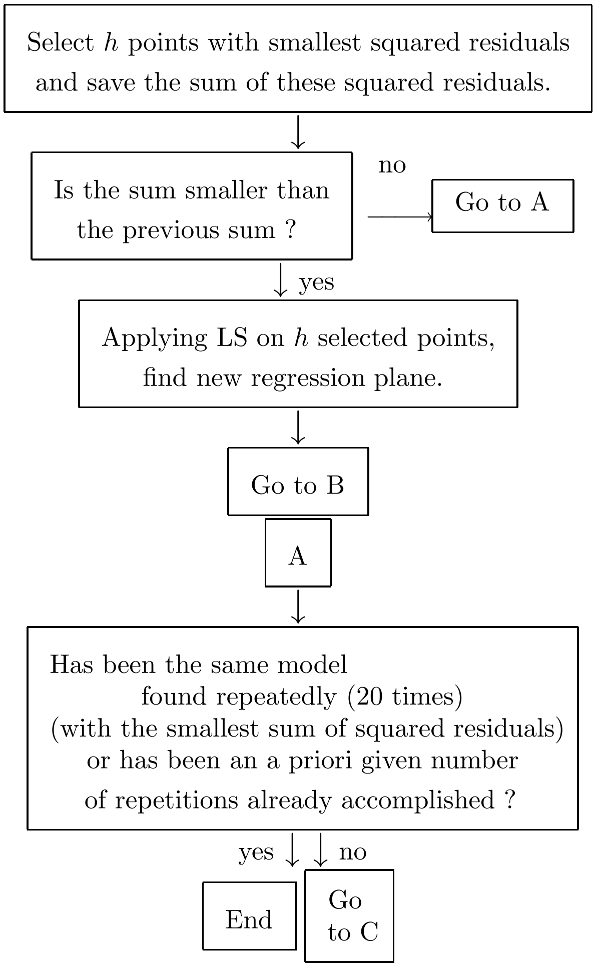
The algorithm is given by the following scheme:

Figure 3



Select randomly $p+1$ points and find the regression plane going through them.

Evaluate residuals for all points with respect to this regression plane.


Select h points with smallest squared residuals
and save the sum of these squared residuals.
Is the sum smaller than
the previous sum ?
no
Go to A
yes
Applying LS on h selected points,
find new regression plane.
Go to B
A
Has been the same model
found repeatedly (20 times)
(with the smallest sum of squared residuals)
or has been an a priori given number
of repetitions already accomplished ?
yes
no
End
Go
to C

Let us add that an implementation by Pavel Čížek is possible to access on INTERNET in XPLORE which is a statistical package organized by Wolfganag Hardle from (and on) Humboldt University. By a request a version applicable under DOS is available from present author and soon a WINDOWS-application will be available (also from present author).

Due to the fact that a lot of attention was paid to the study of methods of evaluation of the least squares, the steps in the previous algorithm which employ the least squares do not represent any problem. Nevertheless it does not guarantee that the algorithm is reliable. The next lines bring an information which approves that the algorithm does give a tight approximation to the precise solution of the extremal problem (12).

We have already mentioned that Hettmansperger *&* Sheather analyzed in 1992 data (let us call them the *Engine Knock Data*) which contained only 16 observations (we have already described what the data recorded). For such data we may find precise solution of the extremal problem (12) by means of a complete inspection over all subsamples of size $11 = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right]$ Since there is "only" 4368 subsamples of size 11, we obtain a solution of the problem in a few minutes. We have also mentioned that not too long after the time when Hettmansperger's *&* Sheather's results appeared we had at hand Boček-Lachout's algorithm and we evaluated estimates (after all we have presented them already in previous). Let us give them once again, now in one table with precise results for $\hat{\beta}^{(LTS,n,h)}$.

Example 1. Engine Knock Data, $h = 11$ ($\hat{\beta}^{(LTS,n,h)}$ - precise values.)

Table 3
Correct data

Method	Interc.	SPARK	AIR	INTK	EXHS.	$\sum_{i=1}^{h} r_{(i)}^2(\beta)$	$r_{(h)}^2(\beta)$
$\hat{\beta}^{(LMS,n,h)}$	30.04	0.144	3.078	0.460	-0.007	0.423	0.053
$\hat{\beta}^{(LTS,n,h)}$	35.11	-0.028	2.949	0.477	-0.009	0.271	0.096

Table 4
Damaged data

Method	Interc.	SPARK	AIR	INTK	EXHS.	$\sum_{i=1}^{h} r_{(i)}^2(\beta)$	$r_{(h)}^2(\beta)$
$\hat{\beta}^{(LMS,n,h)}$	48.4	-.732	3.39	.195	-0.011	1.432	0.203
$\hat{\beta}^{(LTS,n,h)}$	-88.7	4.72	1.06	1.57	0.068	0.728	0.291

Let us turn our attention to the last two columns (successively in both tables, i. e. for correct as well as for damaged data). The last but one gives the sums of eleven smallest squared residuals in respective models. We see that smaller are the sums which stay in tables for the least trimmed squares than the sums of the eleven smallest squared residuals which belong to the least median of squares. But it is O. K., since the method of *the least trimmed squares* is an "expert" for finding such models in which corresponding sums of the h smallest squared residuals are minimal. (Moreover in this case we have at hand precise solution of (12), so the sum is really the smallest possible. After all, we see that the sum which corresponds to *the least median of squares* is nearly two times larger.) So, since we found by the throughout inspection "really good" model, we can expect that all residuals (up to the eleventh one) are also small. Nevertheless, comparing items in the last column we notice that Boček-Lachout's algorithm gave even smaller eleventh order statistics among the squared residuals. It hints that the Boček-Lachout's algorithm is really efficient.

In the case when the number of observations is however larger than, say 20, we are not able to perform inspection of all subsamples of size h and we have to employ the algorithm, we have just described. It is the case of the next example.

Example 2. Salinity Data (Ruppert, Carroll 1980), 21 cases, $h = 16$.
Concentration of salt in water in North Carolina's Pamlico Sound probably depends on

- on salinity lagged by two weeks,
- on number of biweekly periods elapsed since ......,
- on the volume of river discharge into the sound.

Table 5

Method	Interc.	SAL.LAG	TREND	DISCHR.	$\sum_{i=1}^{h} r_{(i)}^2(\beta)$	$r_{(h)}^2(\beta)$
$\hat{\beta}^{(LMS,n,h)}$	37.4	.362	-.086	-1.33	.874	.315
$\hat{\beta}^{(LTS,n,h)}$	36.7	.389	-.114	-1.31	.698	.379

The items given in table show that the situation is analogous to that one described in previous. Again the sum of the sixteen smallest squared residuals is smaller for

$\hat{\beta}^{(LTS,n,h)}$ than for $\hat{\beta}^{(LMS,n,h)}$ while the sixteenth order statistics of squared residuals is smaller for $\hat{\beta}^{(LMS,n,h)}$ than that one for $\hat{\beta}^{(LTS,n,h)}$.

The next example only confirm that the algorithms behave in the same way as in previous also for somewhat larger number of observations.

Example 3. Educational Data (Rousseeuw, Leroy 1987), 50 cases, $h = 27$. Expenditure on public education (per capita) in 50 U. S. states depends on

- on number of residents per thousand residing in urban areas in 1970,
- on personal income (per capita) in 1973,
- on number of residents per thousand under 18 years of age in 1974.

Table 6

Method	Interc.	RESID.	INCM.	YOUNG	$\sum_{i=1}^{h} r^2_{(i)}(\beta)$	$r^2_{(h)}(\beta)$
$\hat{\beta}^{(LMS,n,h)}$	-272.4	.090	.034	.962	3734.8	16.78
$\hat{\beta}^{(LTS,n,h)}$	-143.5	.043	.035	.639	3414.5	19.04

Some other results of processing "famous" data sets may be found in Víšek (1996 b) and (2000 a).

Let us return to Hettmansperger and Sheather's study once again. We have said that they expected, after correcting the error in data, that recalculated results would be only somewhat different from the initial ones but they where surprised by the magnitude of their change. We have also showed that the change of their results was only miseffect caused by the bad algorithm they employed for evaluation of the approximation to the solution of (11). Nevertheless, why they were surprised ? Why they expected that the change of results would be small ? It was due to the (well-spread) idea that the estimators with high breakdown point, due to the fact that they are able to cope with a high contamination, are to be stable under any change of data.

We are going to demonstrate that this idea is wrong. We are going to show that it depends on the (character of the) change of data whether the change of estimates will be small or large, even in a rather small change of data. We shall use once again *Engine Knock Data.*

Example 4. Engine Knock Data, $h = 11$.

(Please remember that in this case estimates evaluated by $\hat{\beta}^{(LTS,n,h)}$ are precise values of estimates, no approximation.)

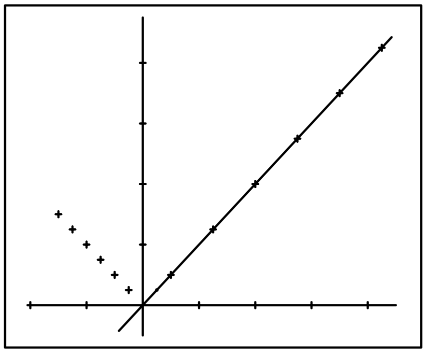
Table 7

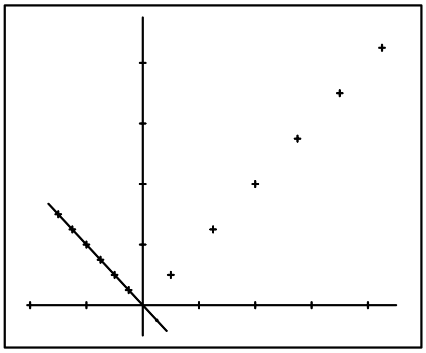
Data	Interc.	SPARK	AIR	INTK	EXHS.
Correct data (with 14.1)	35.11	-0.028	2.949	.477	-.009
Damaged data (with 15.1)	-88.7	4.72	1.06	1.57	.068

So, we see that a small change, even of one datum, may cause a large change of estimate.

To enlighten what is behind this behavior let us give an *academic example*. As any other *academic example* also this is unrealistic, quite senseless, etc. but has one advantage, it is immediately clear what was the reason for the behavior of $\hat{\beta}^{(LTS,n,h)}$ described by previous table.

"Decreasing true" model **"Increasing true" model**

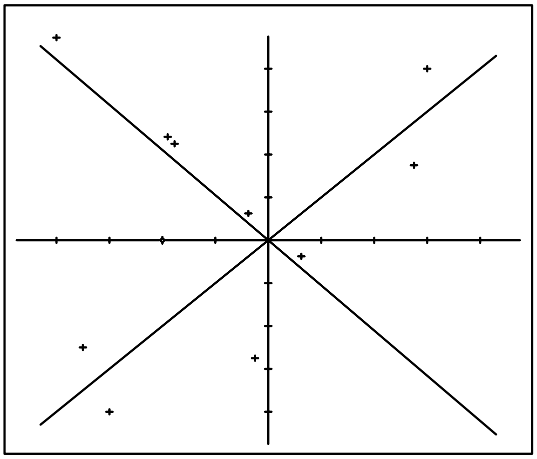




The pictures hint that a small change, even of one datum, may cause a large change of (our ideas about) the "underlying" model.

Returning to the table for damaged data of Example 1 we notice that *the least trimmed squares* and *the least median of squares* offered quite different estimates of underlying model. Let us recall that we gave arguments why we may expect that for the extremal problem (11) we have at hand a tight approximation of its precise solution. For *the least trimmed squares* we have at hand even the precise value of the estimator. Keeping that in mind we may be surprised that the estimators gave so different estimates of coefficients.

An academic example (again in the form of a picture) may enlighten the situation so that it is immediately clear what is the reason for a "strange" behavior of estimators.



$\hat{\beta}^{(LMS,n,h)}$ *versus* $\hat{\beta}^{(LTS,n,h)}$

The fact that two consistent estimators may give rather different (even orthogonal) estimates of underlying model was already studied, formalized and in known as the *diversity of estimates*. Since the formalization is in easy available sources, see Víšek (1996 b), (1997 c) or (2000 a), we shall not repeat it. Let us only add that the formalization demonstrates that two consistent estimators may give completely different estimates of regression coefficients and an increase of sample size need not help. In fact, the phenomenon may appear both for strong as well as for weak consistency and for any size of data.

Now let us turn to the promised example about the Czech economy.

Example of analysis by high breakdown point estimator

In the study of data describing the Czech economy in 1994 we have looked for factors having a significant influence on the export and on the foreigner direct investment, only. We shall present below results for foreigner direct investment. The reason is that the paper is already rather long and moreover in both cases, as we shall after all mention it in the next text, the conclusion were very similar. At the beginning of study we have took into account the whole set of raw data recording the export (X)[2], the import, the total sail (S), the labour (L), the quality of labour given by number of the university educated workers and other labour, the total production, total profit, the value added (VA), the wages (W), total expenditures on labour (cost of labour), the capital (K), percentage of production sailed by 3 largest producers (and so representing concentration in given industry), the development of prices (DP) (given by weighted mean of inflation in given industry), Ballasa index (difference of export and import divided by sum of export and import), the debts, an increasing return to scale (IRS), negative externalities (waste, chemicals, etc.), the total factor productivity (TFPW) (related to the level of wages) in the Czech republic and Germany, the foreigner direct investments (FDI), the research and development (R&D), sensitive products (there are industries which, due to a better organized lobby, such as in textile, clothing, agricultural products, steel, automobiles and some chemicals, are subject to a higher protection tariffs), energy intensity (including, coal, gas, oil and electricity), depreciations, kilogram prices, etc. were collected for 92 industries. Finally, we have tried to use also some explanatory variables which were derived from these (which were just given) as unit labour cost, capital per labour, profit per value added, etc.

It is not necessary to be expert on the Czech economy to be aware that such industries as *Tobacco* (due to Philip-Morris) and *Energy production* (due to ČEZ) are completely atypical. Since both have a special status one can justifiably conclude that they may damage the study and try to exclude them at first. It appeared that it is sufficient to exclude *Tobacco* since the foreigner direct investment was completely atypical, namely zero at respective year.

Moreover, it is clear that the industries have different sizes and that is why we need some standardization of them. It need not be clear at the first glance that there was no suitable unique variable for standardization of all variables. E. g. *value added* may seem to be appropriate for such role but empirical results demonstrated that it is not the case. So at the end we have standardized different factors (explanatory variables) as well as the response variable by different variable (see model given

[2] Abbreviations are given only for those variables which are used in below given reported results of regression analysis.

below). Maybe that in stabilized market economy with prices representing at least vaguely real information about situation on the market, value added may serve well for this purpose.

Of course, we have started with the least squares and due to not very large number of observations (91) and a limited number of explanatory variables (about 30) we could experiment with a lot of combination of them. We were not satisfied by the results. Even the coefficients of determination were not very large leaving aside that other characteristics (as e. g. normality of residuals) were also rather poor.

That was the reason we have applied the least trimmed squares (why just the least trimmed squares is clear from the previous). In this case we were somewhat limited in experimenting by time since the algorithm described in previous is of course a bit slower than that one for evaluating the least squares. Nevertheless we have tried more than twenty combinations of explanatory variables for several transformation of response variable. Moreover, we had to try to fit the model for various value of *h*. For each combination we started with *h* equal to 45 and increased it. At we end we have arrived to the conclusion that the model given below is the best one. The reason was not only the good statistical characteristics but also the fact that after fluctuations of estimates of coefficients for several starting values of *h* we arrived to an interval of values, approximately from 48 to 56, for which the corresponding subsamples of data, i. e. corresponding collections of industries were nested and fluctuations of the estimates of coefficients of regression models were small (in the next text we call these subpopulations as *main* while the complementary ones are denoted as *complementary*). Also the increase of the estimate of variance of disturbances was acceptable. Outside this interval of values of *h* we met with (much) more “wild” behaviour of all items in question.

$$\log\left(\frac{FDI_i}{W_i}\right)=\beta_0+\beta_1\cdot\frac{X_i}{W_i}+\beta_2\cdot\frac{VA_i}{W_i}$$

$$+\beta_3\cdot\frac{R\&D_i}{VA_i}+\beta_4\cdot IRS_i+\beta_5\cdot DP_i+e_i \text{ for i} = 1,2,...,\text{h}, \tag{21}$$

The next table is a pattern of results we have collected. (In all tables which follow *Estimates* and *Standard* mean *Estimates of regression coefficients* and *Estimates of standard errors of estimates of regression coefficients*, respectively.)

Table 8

Estimates of coefficients for main subpopulation for model (21)

91 cases h = 48

Item	Estimates	Standard	t-value	P-value
Intercept	-11.4878	0.4606	-24.9422	0
X/W	0.1785	0.0445	4.016	0.00024
VA/W	0.6082	0.0567	10.7189	0
R&D/PH	0.001	0.0003	3.4167	0.001419
IRS	4.5787	0.2182	20.9841	0
DP	0.304	0.0831	3.6601	0.000699

Having gathered tables of results for $h = 45, 46, ..., 62$ we collected estimates of coefficients and of other characteristics of models in the next two tables (of course, for $h = 48, 49, ..., 55$ - see arguments given in previous).

Table 9

The estimates of coefficients for all models for $h = 48, 49, ..., 55$.

Number of cases	48	49	50	51	52	53	54	55
Intercpt	-11.49	-11.02	-11.40	-11.41	-11.16	-11.15	-11.13	-10.88
signif.	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
X/W	0.1785	0.1592	0.19	0.204	0.2057	0.2142	0.2256	0.2194
signif.	(0.0002)	(0.0014)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0003)
PH/W	0.6082	0.5661	0.5571	0.5517	0.5412	0.5504	0.5577	0.5661
signif.	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
RD/PH	0.001	0.0009	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003	-0.0003
signif.	(0.0014)	(0.0027)	(0.0131)	(0.0192)	(0.0136)	(0.0233)	(0.0312)	(0.0297)
IRS	4.5787	4.4467	4.5885	4.6169	4.5003	4.4717	4.4377	4.3186
signif.	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
DP	0.304	0.2953	0.3553	0.3394	0.3454	0.3259	0.3114	0.3066
signif.	(0.0007)	(0.0015)	(0.0003)	(0.0006)	(0.0007)	(0.0018)	(0.0035)	(0.0056)

Table 10

Values of sum of squares (SS), estimates of variance, coefficients of determination, Durbin-Watson (DW) and χ^2-statistics

Number of cases	48	49	50	51	52	53	54	55
SS	14.01	15.76	17.29	18.94	20.60	22.62	24.59	27.44
$\hat{\sigma}^2$	0.334	0.367	0.393	0.421	0.448	0.481	0.512	0.560
R^2	0.941	0.936	0.942	0.937	0.933	0.926	0.920	0.911
DW	1.947	1.809	1.843	1.820	1.763	1.611	1.548	1.641
χ^2 [10]	8.84(8)	6.42(7)	9.01(9)	7.12(9)	5.48(9)	8.33(8)	6.97 (8)	6.31 (8)

Now let us turn to *complementary* subpopulations. In the same way as described in previous we have found also for complementary subpopulations regression models. We concluded that the model

$$\log\left(\frac{FDI_i}{W_i}\right) = \beta_0 + \beta_1 \cdot \frac{VA_i}{W_i} + \beta_2 \cdot \frac{R\&D_i}{VA_i} + \beta_3 \cdot IRS_i + \beta_4 \cdot DP_i + e_i$$

[10] The degrees of freedom (7, 8 and 9) were given as the results of the application of the χ^2 test procedure which automatically divided residuals into cells so that to have at every cell at least 5 in.

$$\text{for } \mathrm{i} = 1, 2, ..., \mathrm{n} - \mathrm{h} - \ell, \tag{22}$$

is the best fitting to the complementary subpopulations.

Table 11

Estimates of coefficients for complementary subpopulation for model (22)

41 cases h = 37

Item	Estimates	Standard	t-value	P-value
Inter	3.099	1.6823	1.8422	0.07473
VA/W	1.5768	0.5000	3.1539	0.003493
RD/VA	0.0034	0.0007	5.0475	0.000017
IRS	-2.9364	0.5448	-5.3896	0.000006
TFPW	-3.3955	1.5883	-2.1378	0.040269

Table 12

Other characteristics of models, estimates of coefficients of which are given in Table 11

Number of cases / size of subsample, i. e. h	38/36	41/37
Sum of squares	74.2522	63.2336
Estimate of scale	2.3952	1.976
Coefficient of determination	0.5988	0.6263
Durbin - Watson	1.9453	1.8932
χ^2	6.69(4)	9.28(5)

There is of course (legitimate and interesting) question:

Is such division on the main and the complementary subpopulations justifiable ?

The answer is **Yes**.

At the late forties P.H. Douglas studied the question whether there is a model for the production based on the labor and capital, and on the base of some empirical material proposed the function which is nowadays commonly known as *Cobb-Douglas production function*, see Douglas (1948) or Kmenta(1986). For further details see e. g. Arrow et al. (1961), Greene (1993) or Judge et al. (1985)). It may be written as

$$Q_i = \mu L_i^{\lambda} K_i^{1-\lambda}$$

where Q_i is the output in the i-th industry (and L and K are labor and capital, respectively, as already denoted in previous). Moreover, the results given in tables in previous text hint that there is a positive influence of the characteristic which is called *increasing return to scale* in the *main* subpopulation and negative in the *complementary*, we may try at the first analysis to fit this function to the subpopulation given by our division. Assuming e. g. that $\lambda = 1$ and taking into account (again) that industries are of different magnitudes (and hence we have to standardize corresponding items), we may try to estimate the coefficients of the regression model

$$\frac{K_i}{W_i} = \alpha_1 + \alpha_2 \cdot \frac{S_i}{L_i} + \nu_i \tag{23}$$

where of course ν_i are some disturbances. Prior to reporting the results of such experiment, let us say that we were successful for the *main* subpopulations but the model was unsuitable for the *complementary* ones. Again, after some experimentation we arrived to the conclusion that for the *complementary* subpopulations the best fit is evidently achieved by the model

$$\frac{K_i}{W_i} = \gamma_1 + \gamma_2 \cdot \frac{L_i}{S_i} + \kappa_i. \tag{24}$$

The corresponding coefficients of determination are given in Tables 13 and 14 below.

Table 13
Coefficients of determination for main subpopulations

Cases	48	49	50	51	52	53	54	55
Model (23)	0.7436	0.7442	0.7399	0.7408	0.7432	0.7446	0.7456	0.7467
Model (24)	0.1676	0.1694	0.1807	0.1816	0.1681	0.1375	0.1307	0.1319

Table 14
Coefficients of determination for complementary subpopulations

Cases	43	42	41	40	39	38	37	36
Model (24)	0.0084	0.0343	0.1047	0.1092	0.1105	0.1162	0.1221	0.1288
Model (23)	0.5444	0.5588	0.5572	0.5492	0.5576	0.5421	0.5443	0.5353

The results may be, with a grain of salt, interpreted so that the *main* subpopulations behave like in the *market economy* while the others as in a *centrally planned economy* (we stress once again that it is somewhat exaggerated). Moreover, the results given in tables in previous text hint that there is a positive influence of the characteristic which is called *increasing return to scale* in the *main* subpopulation and negative in the *complementary.* It supports the same conclusion.

Conclusions

The conclusions are quite clear. Earlier, when the classic statistics studied the estimators as the maximum likelihood or the minimum χ^2, the evaluation of them seemed to be not very difficult task. The evaluation of the modern (robust) estimators is much more involving and a naive algorithm may betray us. So, the evaluation of them is to be taken as seriously as the proving plausible the theoretical features. Similarly, equipping the estimator by the accompanying tools, i. e. by test for the verification of the assumptions is unseparable part of establishing new estimator. Also searching for the consequences of under- or overfitting the model, presence of an influential point and/or collinearity etc. should be included into that process.

The *least trimmed squares* fulfill nearly all items of a modern paradigm of point estimation (the research on *the least weighted squares* which should rid us some problems with the least trimmed squares, is under process). Moreover, the heuristics of the least trimmed squares are so easy acceptable, that they may be apply even by the believers into a traditional paradigm of mathematical modeling. Finally, and it is also significant, the interpretation of results is not very far from the interpretation of *the ordinary least squares.* The advantage of the estimator is that nowadays several implementations are available

1. Appendix

LEMMA A.1. *Let us have* $\sum_{i=1}^{n} \|x_i\| = \mathcal{O}(n)$. *Then for any* $\Delta \in (0,1]$ *there is a* $K_\Delta < \infty$ *such that denoting for any* $n \in N$

$$m_n = \# \{ i : 1 \leq i \leq n, \|x_i\| > K_\Delta \}$$

we have $m_n < \Delta \cdot n$ *(where "*$\#A$*" denotes the number of elements of the set* A*).*

PROOF. Due to the assumptions of lemma there is C such that for all $n \in N$ we have $\frac{1}{n} \sum_{i=1}^{n} \|x_i\| < C$. Fix $\Delta \in (0,1]$ and put $K_\Delta = \frac{C}{\Delta} + 1$. Then

$$C > \frac{1}{n} \sum_{i=1}^{n} \|x_i\| = \frac{1}{n} \left\{ \sum_{\{i: \|x_i\| \leq K_\Delta\}} \|x_i\| + \sum_{\{i: \|x_i\| > K_\Delta\}} \|x_i\| \right\} > \frac{1}{n} m_n K_\Delta$$

and hence $m_n < n \cdot \frac{C}{K_\Delta} < n \cdot \Delta$.

Table A1

Data which were discussed in Figure 2

Datum given as a small circle is in the first column of the lower table.

x	0.2	0.39	0.6	0.87	1.2	1.94
y	-0.9	-0.459	-1.17	1.485	1.394	1.545

x	0.2	-0.2	-0.39	-0.6	-0.87	-1.2	-1.94
y	-0.3844	0.9	0.459	1.17	-1.485	-1.394	-1.545

Table A2

A pattern of data which were used in numerical illustration, i. e. data about the Czech economy

case	OKEC	X/S	VS/PH	SZ/PH	K/PH	R&D/PH	PH/W W)
1	101	0.49	2.85	3.79	5.21	51.2	2.16
2	102+103	0.13	1.36	2.59	3.98	51.2	3.67
3	111+112	0.18	1.51	2.87	3.98	1522	3.23
4	120+132	0.05	19.04	25.29	44.03	4417	0.37
5	141+142	0.22	1.32	3.16	5.04	472.9	3.4
6	143-145	0.31	2.46	4.68	4.88	472.9	2.32
7	151	0.02	3.44	5.66	4.1	10.1	2.2
8	152	0.04	0	5.01	2.26	10.1	2.91
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
83	352	0.39	2.81	4.32	4.2	312.4	2.49
84	353	0.03	6.46	2.42	2.54	2945.3	3.87
85	354	0.56	5.59	8.37	5.31	312.4	1.71
86	361	0.36	2.87	6.71	2.97	4.2	2.09
87	362	0.24	0.8	3.04	1.77	203.3	3.94
88	363-223	0.73	1.54	3.6	2.11	203.3	3.33
89	364-365	0.47	5.85	7.7	2.9	203.3	1.99
90	296-366	0.54	2.82	5.12	2.48	38.7	2.53
91	401	0.02	0.84	0.76	13.3	2.3	10.67

Table A2 (continued)

case	OKEC	CR3	TFPW	BAL	DP	Ln(FDI/W)
1	101	0.99	1.05	0.69	1.56	-4.6
2	102+103	0.94	1.64	1	1.86	-3.58
3	111+112	0.94	1.5	-0.98	1.66	-5.52
4	120+132	1	0.16	-0.74	1.97	-6.62
5	141+142	0.22	1.45	0.32	2.35	0.43
6	143-145	0.94	1.12	-0.14	1.8	-0.46
7	151	0.15	1.14	-0.08	1.38	-4.46
8	152	0.81	1.66	-0.42	1.97	-3.85
9	153	0.31	1.13	-0.46	1.39	-3.19
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
82	351	1	0.7	0.76	1.02	-3.21
83	352	0.59	1.23	0.86	4.12	-3.19
84	353	0.8	1.95	0.65	3.05	-0.72
85	354	0.6	0.88	0.09	3.24	0.28
86	361	0.27	1.21	0.29	2.02	-0.55
87	362	0.72	2.2	0.47	2.03	-1.84
88	363-223	0.39	1.86	-0.11	2.11	-3.42
89	364-365	0.31	1.18	-0.35	2.27	-0.1
90	296-366	0.31	1.46	0.45	2.15	-2.8
91	401	0.77	2.41	0.35	3.01	0.82

References

[1] 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.

[2] Antoch, J., Vorlíčková, D. (1992): *Vybrané metody statistické analýzy dat.* Academia, Praha, 1992.

[3] Arrow, K., H. Chenery, B. Minhas, R. Solow (1961): Capital-labor substitution and economic efficiency. *Review of Economic and Statistics, 45, 225 - 247.*

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

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

[6] Boscovisch, R. J. (1757): De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura eius ex exemplaria etiam sensorum impressa. *Bononiensi Scientiarum et Artium Instituto Atque Academia Commentarii 4, 353-396.*

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

[8] Douglas, P. H. (1948): Are there laws of production? *American Economic Review, 38, 1 - 41.*

[9] Galilei, G. (1632): *Dialogo dei masimi sistemi.*

[10] Galton F. (1886): Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute vol. 15, pp. 246–263.*

[11] Gauss F. C. (1809): Theoria molus corporum celestium. *Hamburg, Perthes et Besser.*

[12] Greene, W.H. (1993): *Econometric Analysis*, Macmillam Press, New York.

[13] Hampel, F. R. (1975): Beyond location parameters: Robust concepts and methods (with discussion). *Proceedings of the 40th Session of the ISI, vol. XLVI*

[14] 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 Sons.

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

[16] Hoerl, A. E., R. W. Kennard (1970 a): Ridge regression: Biased estimation for nonorthogonal problems *Technometrics 12, 55 - 68.*

[17] Hoerl, A. E., R. W. Kennard (1970 b): Ridge regression: Application to nonorthogonal problems *Technometrics 12, 69 - 82.*

[18] Huber, P. J. (1973): Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist. 1, 799–821.*

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

[20] Joss, J., A. Marazzi (1990): Probabilistic algorithms for *LMS* regression. *Computational Statistics & Data Analysis 9, 123–134.*

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

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

[23] Kmenta, J. (1986): *Elements of econometrics*, Macmillan Publishing Company, New York.

[24] Laplace, P. S. (1793): Sur quelques points du systeme du mode. *Memoires de l'Academic Royale des Sciences de Paris, 1-87.*

[25] Lax,D. A. (1975): An interim report of a Monte Carlo study of robust estimators of width. *Technical Report 93, Series 2, Dept. of Statistics, Princeton University, Princeton.*

[26] Legendre A. M. (1805): Nouvelles méthodes pour la détermination des orbites des comètes. *Paris, Courcier.*

[27] Maronna, R. A. (1976): Robust M-estimators of multivariate location and scatter. *Annals of Statistcs 4,51 - 67.*

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

[29] Martin, R. D., V. J. Yohai, R. H. Zamar (1989): Min-max bias robust regression. *Ann Statist. 17, 1608 - 1630.*

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

[31] Prigogine, I., I. Stengers (1977): La Nouvelle Alliance. *SCIENTIA, 1977, issues 5-12.*

[32] Prigogine, I., I. Stengers (1984): *Out of Chaos.* William Heinemann Ltd 1984.

[33] Rousseeuw, P.J. (1984): Least median of square regression. *Journal of Amer. Statist. Association 79, pp. 871-880.*

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

[35] Ruppert, D., R. J. Carroll (1980): Trimmed least squares estimation in linear model. *J. American Statist. Ass., 75 (372), pp. 828–838.*

[36] Schweingruber, M. (1980): Das Monte Carlo Verhalten einiger Verwerfungsregeln. *Diploma thesis. Fachgruppe fur Statistik. ETH. Zurich.*

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

[38] 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, Prague, 1994, 254 - 259.*

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

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

[41] Víšek J. Á. (1997 a): *Statistická analźa dat.* Vydavatelství Českh́o vysokého učení technického v Praze, 1997, ISBN 80-01-01735-4.

[42] Víšek J. Á. (1997 b): *Ekonometrie I.* Nakladatelství Univerzity Karlovy, 1997, ISBN 80-7184-483-7.

[43] Víšek J. Á. (1997 c): Contamination level and sensitivity of robust tests. *Handbook of Statistics, volume 15, 633 - 642, eds. G. S. Maddala & C. R. Rao, 1997, Amsterdam: Elsevier Science B. V., ISBN 0-444-82172-4*

[44] Víšek, J.Á.(1998 a): 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, pp. 581 - 586.*

[45] Víšek, J.Á.(1998 b): Robust instruments. *Robust'98 (ed. Jaromír Antoch & Gejza Dohnal, published by Union of Czechoslovak Mathematicians and Physicists), 1998, pp. 195 - 224.*

[46] Víšek J. Á. (1998 c): What is characterized by gross error sensitivity ? *Bulletin of the Czech Econometric Society, Volume 5 (1998), Issue 7, 111 - 124.*

[47] Víšek J. Á. (1999 a): Robust estimation of regression model. *Bulletin of the Czech Econometric Society, Volume 9/1999, 57 - 79.*

[48] Víšek J. Á. (1999 b): The least trimmed squares - random carriers. *Bulletin of the Czech Econometric Society, Volume 10/1999, 1 - 30.*

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

[50] Víšek, J.Á.(2000 b): Sensitivity analysis of M-estimates of nonlinear regression model: Influence of data subsets. To appear in *Annals of the Institute of Statistical Mathematics.*

[51] Víšek, J.Á.(2000 c): The least trimmed squares. Consistency, asymptotic normality and sensitivity study. *Preprint.*

[52] Víšek, J.Á.(2000 d): A new paradigm of (high breakdown point) estimation. Submitted to the proceedings the seminar "Data Analysis".

[53] Zvára, K. (1989): *Regresní analýza* (Regression Analysis – in Czech). Prague: Academia.

Charles University, Department of Macroeconomics and Econometrics, Institute of Economic Studies, Faculty of Social Sciences, Opletalova 26, CZ – 110 00 Prague 1, Czech Republic

E-mail address: visek@mbox.fsv.cuni.cz

Department of Stochastic Informatics, Institute of Information Theory and Automation, Czech Academy of Sciences, Pod Vodárenskou Věží 4, CZ – 182 08 Prague 8, Czech Republic

E-mail address: visek@utia.cas.cz