Čížek, Pavel; Víšek, Jan Ámos

Working Paper

Least trimmed squares

SFB 373 Discussion Paper, No. 2000,53

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Čížek, Pavel; Víšek, Jan Ámos (2000) : Least trimmed squares, SFB 373 Discussion Paper, No. 2000,53, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:11-10047794

This Version is available at:
https://hdl.handle.net/10419/62211

Least Trimmed Squares

Pavel Čížek and Jan Ámos Víšek

Least trimmed squares (LTS) is a statistical technique for estimation of unknown parameters of a linear regression model and provides a "robust" alternative to the classical regression method based on minimizing the sum of squared residuals.

This chapter helps to understand the main ideas of robust statistics that stand behind the least trimmed squares estimator and to find out how to use XploRe for this type of robust estimation. As it is impossible to provide a profound introduction into this area here, we refer readers for further information to the bibliography.

Before proceeding to the next section, please type at the XploRe command line

```
library("metrics")
```

to load the necessary quantlibs (libraries). Quantlib `metrics` automatically loads `xplore`, `kernel`, `glm`, and `multi` quantlibs.

1 Robust Regression

1.1 Introduction

The classical least squares (LS) estimator is widely used in regression analysis both because of the ease of its computation and its tradition. Unfortunately, it is quite sensitive to higher amounts of data contamination, and this just adds together with the fact that outliers and other deviations from the standard linear regression model (for which the least squares method is best suited) appear quite frequently in real data. The danger of outlying observations, both in the direction of the dependent and explanatory variables, to the least

squares regression is that they can have a strong adverse effect on the estimate and they may remain unnoticed, especially when higher dimensional data are analyzed. Therefore, statistical techniques that are able to cope with or to detect outlying observations have been developed. One of them is the least trimmed squares estimator.

Figure 1: Least squares regression with outliers, `phonecal` data, `ls01.xpl`

The methods designed to treat contaminated data can be based on one of two principles. They can either detect highly influential observations first and then apply a classical estimation procedure on the "cleaned" data, or they can be

designed so that the resulting regression estimates are not easily influenced by contamination. Before we actually discuss them, especially the latter ones, let us exemplify the sensitivity of the least squares estimator to outlying observations.

The data set `phonecal` serves well this purpose. The data set, which comes from the Belgian Statistical Survey and was analyzed by Rousseeuw and Leroy (1987), describes the number of international phone calls from Belgium in years 1950–1973. The result of the least squares regression is depicted on Figure 1. Apparently, there is a heavy contamination caused by a different measurement system in years 1964–1969 and parts of year 1963 and 1970—instead of the number of phone calls, the total number of minutes of these calls was reported. Moreover, one can immediately see the effect of this contamination: the estimated regression line follow neither a mild upward trend in the rest of the data, nor any other pattern that can be recognized in the data. One could argue that the contamination was quite high and evident after a brief inspection of the data. However, such an effect might be caused even by a single observation, and in addition to that, the outlying observations do not have to be easily recognizable if analyzed data are multi-dimensional. To give an example, an artificial data set consisting of 10 observations and one outlier is used. We can see the effect of a single outlier on Figure 2—while the blue line represents the underlying model, the red thick line shows the least squares estimate. Moreover, the same figure shows that the residuals plot does not have to have any outlier-detection power (the blue thin lines represent interval $(-\sigma, \sigma)$ and the blue thick lines correspond to $\pm 3\sigma$).

As most statisticians are aware of the described threats caused by very influential observations for a long time, they have been trying to develop procedures that would help to identify these influential observations and provide "outlier-resistant" estimates. There are actually two ways how this goal can be achieved. First one relies on some kind of regression diagnostics to identify highly influential data points. Having identified suspicious data points, one can remove them, and subsequently, apply classical regression methods. These methods are not in the focus of this chapter. Another strategy, which will be discussed here, is to utilize estimation techniques based on the so-called robust statistics. These robust estimation methods are designed so that they are not easily endangered by contamination of data. Furthermore, a subsequent analysis of regression residuals coming from such a robust regression fit can then hint on outlying observations. Consequently, such robust regression methods can serve as diagnostic tools as well.

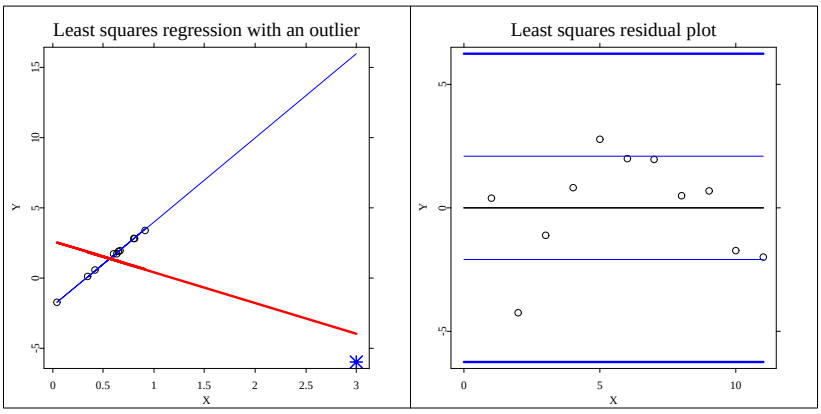

Figure 2: Least squares regression with one outlier and the corresponding residual plot, `ls02.xpl`

1.2 High Breakdown Point Estimators

Within the theory of robustness, several concepts exist. They range from the original minimax approach introduced in Huber (1964) and the approach based on the influence function (Hampel et al. 1986) to high breakdown point procedures (Hampel 1971), that is the procedures that are able to handle highly contaminated data. The last one will be of interest here as the least trimmed squares estimator belongs to and was developed as a high breakdown point method. To formalize the notion of the capability of an estimator to resist to some amount of contamination in the data, the **breakdown point** was introduced. For the simplicity of exposure, we present here one of its finite-sample versions suggested by Donoho and Huber (1983): *Take an arbitrary sample of n data points, $S_n = (x_1, \ldots, x_n)$, and let T_n be a regression estimator, i.e., applying T_n to the sample S_n produces an estimate of regression coefficients $T_n(S_n)$. Then the breakdown point of the estimator T_n at S_n is defined by*

$$\varepsilon_n^\star(T_n, S_n) = \frac{1}{n} \max \left\{ m \left| \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots,y_m} \|T_n(z_1, \ldots, z_n)\| < +\infty \right. \right\}, \qquad (1)$$

where sample $(z_1, \ldots, z_n)$ is created from the original sample S_n by replacing observations $x_{i_1}, \ldots, x_{i_m}$ by values $y_1, \ldots, y_m$. The breakdown point usually

does not depend on S_n. To give an example, it immediately follows from the definition that the finite-sample breakdown point of the arithmetic mean equals to 0 in a one-dimensional location model, while for the median it is 1/2. Actually, the breakdown point equal to 1/2 is the highest one that can be achieved at all—if the amount of contamination is higher, it is not possible to decide which part of the data is the correct one. Such a result is proved, for example, in Theorem 4, Chapter 3 of Rousseeuw and Leroy (1987) for the case of regression equivariance estimators (the upper bound on $\varepsilon_n^\star$ in this case is actually $([(n-p)/2]+1)/n$, where $[x]$ denotes the integer part of x).

There were quite a lot of estimators intended to have a high breakdown point, that is close to the upper bound, although some of them were not entirely successful in achieving this point because of their sensitivity to a specific kind of data contamination. One of truly high breakdown point estimators that reached the above mentioned upper bound of the breakdown point were the **least median of squares** (LMS) estimator (Rousseeuw 1984), which minimizes the median of squared residuals, and the **least trimmed squares** (LTS) estimator (Rousseeuw 1985), which takes as its objective function the sum of h smallest squared residuals and was indeed proposed as a remedy to the low asymptotic efficiency of LMS.

Before proceeding to the definition and a more detailed discussion of the least trimmed squares estimator, let us show the behavior of this estimator when applied to `phonecal` data used in the previous section. On Figure 3 we can see two estimated regression lines: the red thick line that corresponds to the LTS estimate, and for comparison purposes, the blue thin line that depicts the least squares regression result. While the least squares estimate is spoilt by outliers coming from years 1963–1970, the least trimmed squares regression line is not affected and outlines the trend one would consider as the right one.

2 Least Trimmed Squares

In this section the least trimmed squares estimator, its robustness and asymptotic properties, and computational aspects will be discussed.

2.1 Definition

First of all, we will precise the verbal description of the estimator given in the previous section. Let us consider a linear regression model for a sample (y_i, x_i) with a response variable y_i and a vector of p explanatory variables x_i:

$$y_i = \beta^T x_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

The least trimmed squares estimator $\hat{\beta}^{(LTS)}$ is defined as

$$\hat{\beta}^{(LTS)} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{h} r^2_{[i]}(\beta), \tag{2}$$

where $r^2_{[i]}(\beta)$ represents the i-th order statistic among $r^2_1(\beta), \ldots, r^2_n(\beta)$ with $r_i(\beta) = y_i - \beta^T x_i$ (we believe that the notation is self-explaining). The so-called trimming constant h have to satisfy $\frac{n}{2} < h \leq n$. This constant determines the breakdown point of the LTS estimator since the definition (2) implies that $n-h$ observations with the largest residuals will not affect the estimator (except of the fact that the squared residuals of excluded points have to be larger than the h-th order statistics among the squared residuals). The maximum breakdown point is attained for $h = [n/2]+[(p+1)/2]$ (see Rousseeuw and Leroy 1987, Theorem 6), whereas for $h = n$, which corresponds to the least squares estimator, the breakdown point equals to 0. More on the choice of the trimming constant can be found in Subsection 3.1.

Before proceeding to the description of how such an estimate can be evaluated in XploRe, several issues have to be discussed, namely, the existence of this estimator and its statistical properties (a discussion of its computational aspects is postponed to Subsection 2.2). First, the existence of the optimum in (2) under some reasonable assumptions can be justified in the following way: the minimization of the objective function in (2) can be viewed as a process in which we every time choose a subsample of h observations and find some β minimizing the sum of squared residuals for the selected subsample. Doing this for every subsample (there are $\binom{n}{h}$ of them) we get $\binom{n}{h}$ candidates for the LTS estimate and the one that commands the smallest value of the objective function is the final estimate. Therefore, the existence of the LTS estimator is basically equivalent to the existence of the least squares estimator for subsamples of size h.

Let us now briefly discuss various statistical properties of LTS. First, the least trimmed squares is regression, scale, and affine equivariant (see, for exam-

ple, Rousseeuw and Leroy 1987, Lemma 3, Chapter 3). We have also already remarked that the breakdown point of LTS reaches the upper bound $([(n-p)/2]+1)/n$ for regression equivariant estimators if the trimming constant h equals to $[n/2] + [(p+1)/2]$. Furthermore, the $\sqrt{n}$-consistency and asymptotic normality of LTS can be proved for a general linear regression model with continuously distributed disturbances (Víšek 1999b). Besides these important statistical properties, there are also some less practical aspects. The main one directly follows from the noncontinuity of the LTS objective function. Because of this, the sensitivity of the least trimmed squares estimator to a change of one or several observations might be sometimes rather high (Víšek 1999a). This property, often referred as high subsample sensitivity, is closely connected with the possibility that a change or omission of some observations may change considerably the subset of a sample that is treated as the set of "correct" data points. It does not have to be seen necessarily as disadvantageous, the point of view merely depends on the purpose we are using LTS for. See Víšek (1999b) and Section 3 for further information.

2.2 Computation

`b = lts(x, y{, h, all, mult})`
computes the least trimmed squares estimate of a linear regression model

The quantlet of quantlib `metrics` which serves for the least trimmed squares estimation is `lts`. To understand the function of its parameters, the algorithm used for the evaluation of LTS has to be described. Later, the description of the quantlet follows.

There are two possible strategies how the least trimmed squares estimate can be determined. First one relies on the full search through all subsamples of size h and the consecutive LS estimation as described in the previous section, and thus, let us obtain the precise solution (neglecting ubiquitarian numerical errors). Unfortunately, it is hardly possible to examine the total of $\binom{n}{h}$ subsamples unless a very small sample is analyzed. Therefore, in most cases (when the number of cases is higher) only an approximation can be computed (note, please, that in the examples presented here we compute the exact LTS estimates as described above, and thus, the computation is relatively slow). The present algorithm does the approximation in the following way: having selected

randomly an $(p+1)$-tuple of observations we apply the least squares method on them, and for the estimated regression coefficients we evaluate residuals for all n observations. Then h-tuple of data points with the smallest squared residuals is selected and the LS estimation takes place again. This step is repeated so long until a decrease of the sum of the h smallest squared residuals is obtained. When no further improvement can be found this way, a new subsample of h observations is randomly generated and the whole process is repeated. The search is stopped either when we find s times the same estimate of model (where s is an a priori given positive integer) or when an a priori given number of randomly generated subsamples is accomplished. A more refined version of this algorithm suitable also for large data sets was proposed and described by Rousseeuw and Van Driessen (1999).

From now on, noninteractive quantlet `lts` is going to be described. The quantlet expects at least two input parameters: an $n \times p$ matrix `x` that contains n observations for each of p explanatory variables and an $n \times 1$ vector `y` of n observed responses. If the intercept is to be included in the regression model, the $n \times 1$ vector of ones can be concatenated to the matrix `x` in the following way:

```
x = matrix(rows(x))~x
```

Neither the matrix `x`, nor the vector `y` should contain missing (`NaN`) or infinite values (`Inf,-Inf`). Their presence can be identified by `isNaN` or `isNumber` and the invalid observations should be processed before running `lts`, e.g., omitted using `paf`. These two parameter are enough for the most basic use of the quantlet. Typing

```
b = lts(x,y)
```

results in the approximation of the LTS estimate for the most robust choice of $h = [n/2] + [(p+1)/2]$ using the default number of iterations. Though this might suffice for some purposes, in most cases we would like to specify also the third parameter—the trimming constant h—too. So probably the most common use takes the form

```
b = lts(x,y,h)
```

The last two parameters of the quantlet, particularly `all` and `mult`, provide a way to influence how the estimate is in fact computed. Parameter `all` allows

to switch from the approximation algorithm, which corresponds to `all` equal to 0 and is used by default, to the precise computation of LTS, which takes place if `all` is nonzero. As the precise calculation can take quite a long time if a given sample is not really small, a warning together with a possibility to cancel the evaluation is issued whenever the total number of iterations is too high. Finally, the last parameter `mult`, which equals to 1 by default, offers possibility to adjust the maximum number of randomly generated subsamples in the case of the approximation algorithm—this maximum is calculated from the size of a given sample and the trimming constant, and subsequently, it is multiplied by `mult`.

To have a real example, let us show how the time trend in `phonecal` data set was estimated in Section 1. The data set is two-dimensional, having only one explanatory variable `x`, year, in the first column and the response variable `y`, the number of international phone calls, in the second column. In order to obtain the LTS estimate for the linear regression of `y` on constant term and `x`, you have to type at the command line or in the editor window

```
z = read("phonecal")
x = matrix(rows(z)) ~ z[,2]
y = z[,3]
b = lts(x,y)
b
```

 `lts02.xpl`

The result of the above example should appear in the XploRe output window as follows:

```
Contents of coefs
[1,]  -5.6522
[2,]  0.11649
```

3 Supplementary Remarks

3.1 Choice of the Trimming Constant

As was already mentioned, the trimming constant h have to satisfy $\frac{n}{2} < h \leq n$ and indeed determines the breakdown point of LTS. The choice of this constant

depends mainly on the purpose for which we want to use LTS. There is, of course, a trade-off involved: lower values of h, which are close to the optimal breakdown point choice, lead to a higher breakdown point, while higher values improve efficiency (if the data are not too contaminated) since more information stored in data is utilized. The maximum breakdown point is attained for $h = [n/2] + [(p+1)/2]$. This choice is often employed when the LTS is used for diagnostic purposes (see Subsection 3.2). The most robust choice of h may be also favored when LTS is used for comparison with some less robust estimator, e.g., the least squares, since comparison of these two estimators can serve as a simple check of data and a model—if the estimates are not similar to each other, a special care should be taken throughout the subsequent analysis. On the other hand, it may be sensible to evaluate LTS for a wide range of values of the trimming constant and to observe how the estimate behaves with increasing h, because this can provide hints on the amount of contamination and possibly on suspicious structures of a given data set (for example, that the data set contains actually a mixture of two different populations).

3.2 LTS as a Diagnostic Tool

We have several times advocated the use of the least trimmed squares estimator for diagnostic purposes. Therefore, a brief guidance regarding diagnostics is provided in this subsection via an example. Let us look at `stacklos` data, which were already analyzed many times, for example by Drapper and Smith (1966), Daniel and Wood (1971), Carroll and Ruppert (1985), and Rousseeuw and Leroy (1987). The data consist of 21 four-dimensional observations characterizing the production of nitric acid by the oxidation of ammonia. The stackloss (y) is assumed to depend on the rate of operation (x_1), on the cooling water inlet temperature (x_2) and on the acid concentration (x_3). Most of the studies dealing with this data set found out among others that data points 1, 3, 4, 21, and maybe also 2 were outliers. First, the least square regression result

$$\hat{y} = -39.92 + 0.716x_1 + 1.295x_2 - 0.152x_3,$$

`ls03.xpl`, is reported for comparison with LTS, the corresponding residual plot is plotted on Figure 4 (once again, the blue thin lines represent $\pm\sigma$ and the blue thick lines correspond to $\pm 3\sigma$). There are no significantly large residuals with respect to the standard deviation, so without any other diagnostic statistics one would be tempted to believe that there are no outlying observations. On the contrary, if we inspect the least trimmed squares regression,

which produces

$$\hat{y} = -35.21 + 0.746x_1 + 0.338x_2 - 0.005x_3,$$

`lts03.xpl`, our conclusion will be different. To construct a residual plot for a robust estimator, it is necessary to use also a robust estimator of scale because the presence of outliers is presumed. Such a robust estimator of variance can be based in the case of LTS, for example, on the sum of the h smallest residuals or on the absolute median deviation $\mathrm{MAD}_i\, x_i = \mathrm{med}_i\, |x_i - \mathrm{med}_i\, x_i|$ as is the case on Figure 5. Inspecting the residual plot on Figure 5 (the blue lines represents again $\pm\sigma$ and $\pm 3\sigma$ levels, where $\sigma = 1.483\, \mathrm{MAD}_i\, r_i(\beta)$), observations 1, 2, 3, 4, and 21 become suspicious ones as their residuals are very large in the sense that they lie outside of the interval $(-3\sigma, 3\sigma)$. Thus, the LTS estimate provide us at the same time with a powerful diagnostic tool. One has naturally to decide which ratios $|r_i(\beta)/\sigma|$ are already doubtable, but value 2.5 is often used as a decisive point.

3.3 High Subsample Sensitivity

The final note on LTS concerns a broader issue that we should be aware of whenever such a robust estimator is employed. Already mentioned high subsample sensitivity is caused by the fact that high breakdown point estimators search for a "core" subset of data that follows best a certain model (with all its assumptions) without taking into account the rest of observations. A change of some observations may then lead to a large swing in composition of this core subset. This might happen, for instance, if the data are actually a mixture of two (or several) populations of data, i.e., a part of data can be explained by one regression line, another part of the same data by a quite different regression function, and in addition to that, some observations may suit both model relatively well (this can happen with a real data set too, see Benáček, Jarolím, and Víšek 1998). In such a situation, a small change of some observations or some parameters of the estimator can bring the estimate from one regression function to another. Moreover, application of several (robust) estimates is likely to introduce several rather different estimates in such a situation—see Víšek (1999b) for a detailed discussion. Still, it is necessary to have in mind that this is not shortcoming of the discussed estimators, but of the approach taken in this case—procedures designed to suit some theoretical models are applied to an unknown sample and the procedures in question just try to explain it by means of a prescribed model.

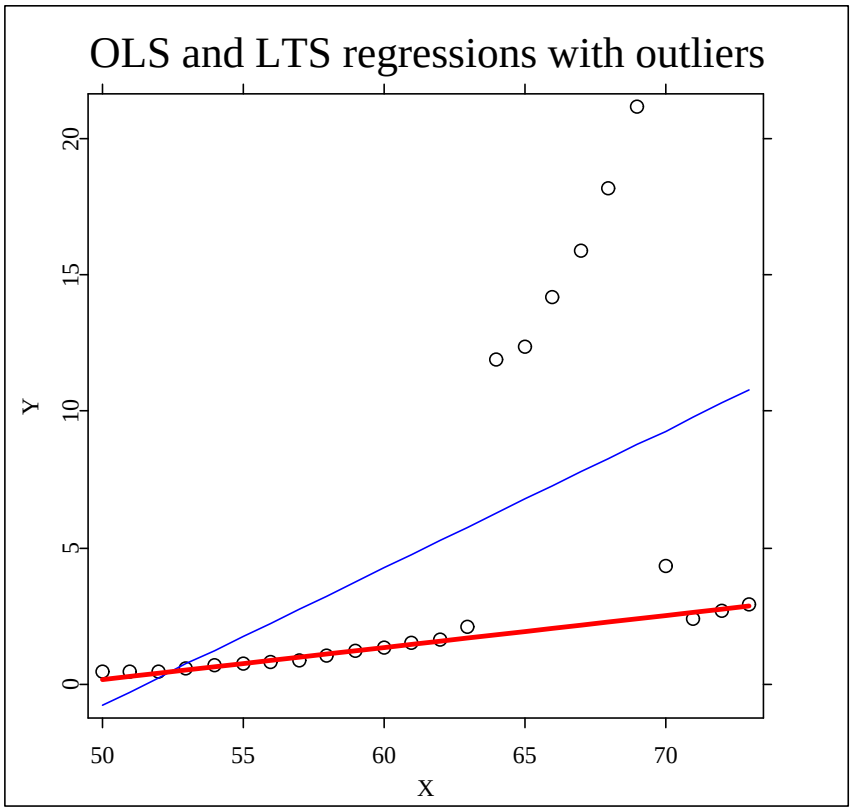

Figure 3: Least trimmed squares regression with outliers, `phonecal` data, `lts01.xpl`

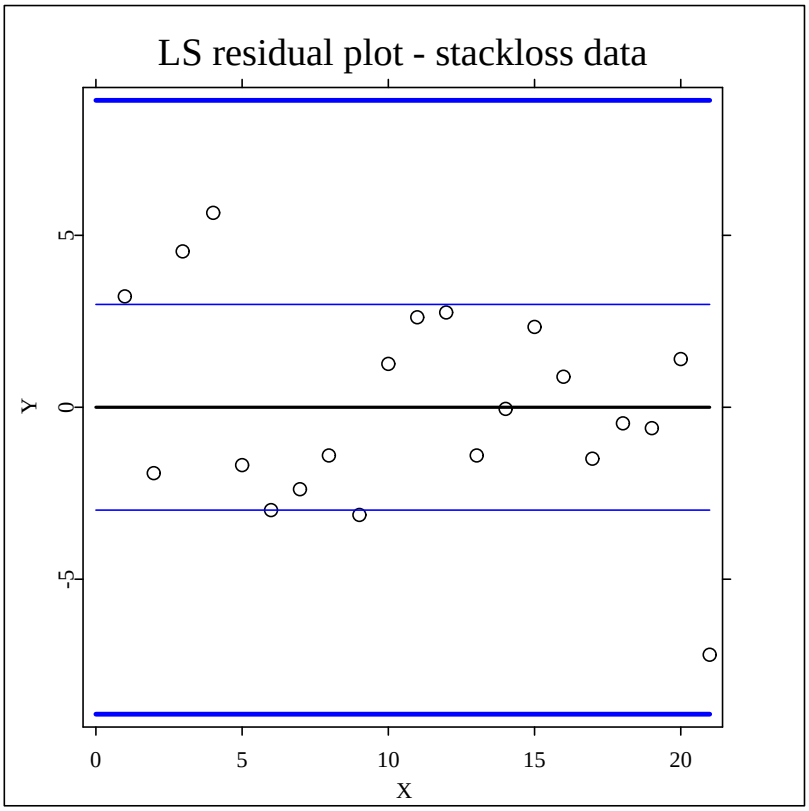

Figure 4: The LS residual plot for `stacklos` data, `ls04.xpl`

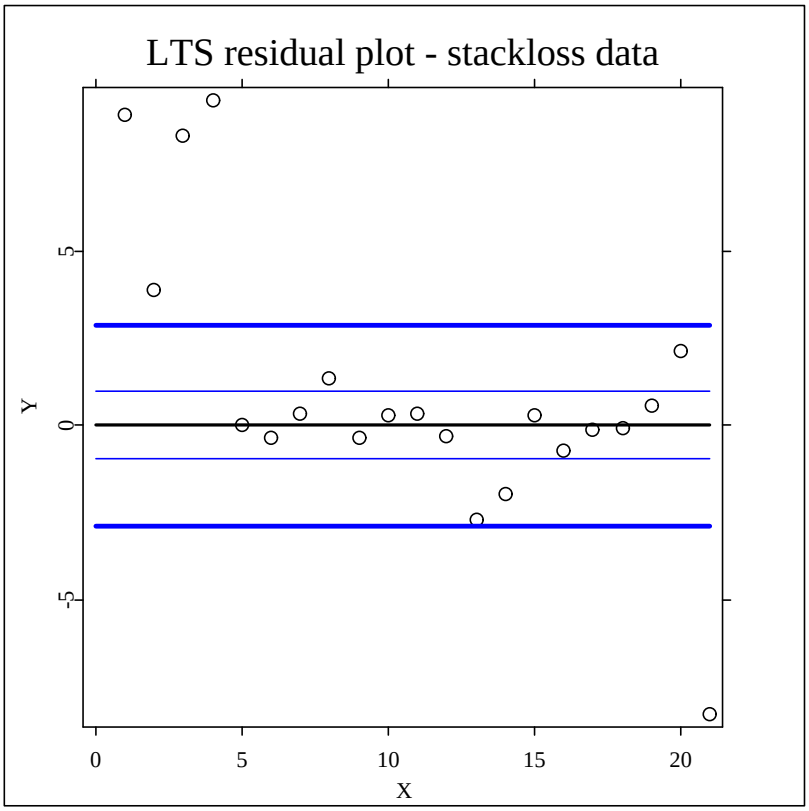

Figure 5: The LTS residual plot for `stacklos` data, `lts04.xpl`

References

Benáček, V., Jarolím, M., and Víšek, J. Á. (1998). Supply-side characteristics and the industrial structure of Czech foreign trade, *Proceedings of the conference Business and economic development in central and eastern Europe: Implications for economic integration into wider Europe, ISBN 80-214-1202-X*, Technical university in Brno together with University of Wisconsin, Whitewaters, and the Nottingham Trent university, 51–68.

Carroll, R. J. and Ruppert, D. (1985). Transformations in regression: A robust analysis, *Technometrics* **27**, 1–12.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, John Wiley & Sons, New York.

Drapper, N. R. and Smith, H. (1966). *Applied Regression Analysis*, John Wiley & Sons, New York.

Donoho, D. L. and Huber, P. J. (1983). The notion of the breakdown point, in *A Festschrift for Erich Lehmann*, edited by P. Bickel, K. Doksum, and J. L. Hodges, Jr., Wadsworth, Belmont, CA.

Hampel, F. R. (1971). A general qualitative definition of robustness, *Annals of Mathematical Statistics* **42**, 1887–1896.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics, The Approach Based on Influence Function*, Wiley series in Probability and Mathematical Statistics, New York.

Huber, P. J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics* **35**, 73–101.

Rousseeuw, P. J. (1984). Least Median of Squares Regression, *Journal of American Statistical Association* **79**, 871–880.

Rousseeuw, P. J. (1985). Multivariate Estimation With High Breakdown Point, *Mathematical Statistics and Applications, Vol. B*, edited by W. Grossmann, G. Pflug, I. Vincze, and W. Werty, Reidel, Dordrecht, Netherlands, 283–297.

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

Rousseeuw, P. J. and Van Driessen, K. (1999). Computing LTS Regression for Large Data Sets, *Technical Report, University of Antwerp.*

Víšek, J. Á. (1996). On high breakdown point estimation, *Computational Statistics* **11**, 137–146.

Víšek, J. Á. (1999a). The least trimmed squares—random carriers, *Bulletin of the Czech Econometric Society*, Volume **10/1999**, 1–30.

Víšek, J. Á. (1999b). On the diversity of estimates, to appear in *Computational Statistics and Data Analysis.*