

Robust Statistics, Robust Regression and Quantile Regression

By Keming Yu

This talk provides a brief review to *robust statistics*, *robust regression* and *quantile regression*. It is intended as a teaching tool, for people who are not already familiar with these statistics techniques. We begin with an introduction to and motivation for these techniques. We then outline various approaches to the techniques and discuss some typical application in finance and economics.

What does Robust mean?

Robust means stable or reliable.

Robust solution means that the best worst case.

Definitions differ in scope and content.

1. In the most general construction:

Robust models pertains to stable and reliable models.

Robust methods pertains to stable and reliable methods

2. Strictly speaking:

Threats to stability and reliability include

influential outliers in data.

(An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population.)

3. The most commonly used statistical methods in sciences and social sciences make assumptions about normality and variance homogeneity. However, often these assumptions are violated.

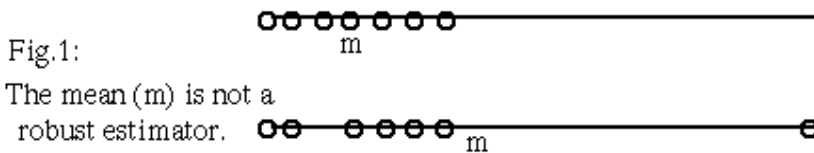
Part I, Robust Statistics

A fundamental task in many statistical analyses is to estimate a location parameter for the distribution; i.e., to find a typical or central value that best describes the data.

Mean, or the average value of a set of data or sample is often used as the estimate of population mean. The mean is the sum of the data points divided by the number of data points. That is,

$$\bar{Y} = \sum_{i=1}^N Y_i / N$$

However, as an example of a non-robust estimator, mean (or center of mass) of the set of points is not robust: if one point is moved "far" from the rest, the mean will follow (Fig.1). If we move the "corrupt" point to infinity, the mean will also go to infinity. This example indicates that robustness is particularly important when there is a possibility that our data set contains "contaminated" or "corrupt" data. It is also important when we simply do not want some data to have more importance than others. If we take the mean as an estimator, outlying points carry more "weight" than points near the mean. In Fig.1, deleting the outlying point would have a greater impact on the location of the mean than deleting a point in the dense region.



The 2nd example involves the real data given in Table 1 which are the results of an interlaboratory test (<http://www.quantlet.com/>). The boxplots are shown in Fig. 2 where the dotted line denotes the mean of the observations and the solid line the median.

Table 1: The results of an interlaboratory test involving 14 laboratories

1	2	3	4	5	6	7	9	9	10	11	12	13	14
1.4	5.7	2.64	5.5	5.2	5.5	6.1	5.54	6.0	5.1	5.5	5.9	5.5	5.3
1.5	5.8	2.88	5.4	5.7	5.8	6.3	5.47	5.9	5.1	5.5	5.6	5.4	5.3
1.4	5.8	2.42	5.1	5.9	5.3	6.2	5.48	6.1	5.1	5.5	5.7	5.5	5.4
0.9	5.7	2.62	5.3	5.6	5.3	6.1	5.51	5.9	5.3	5.3	5.6	5.6	

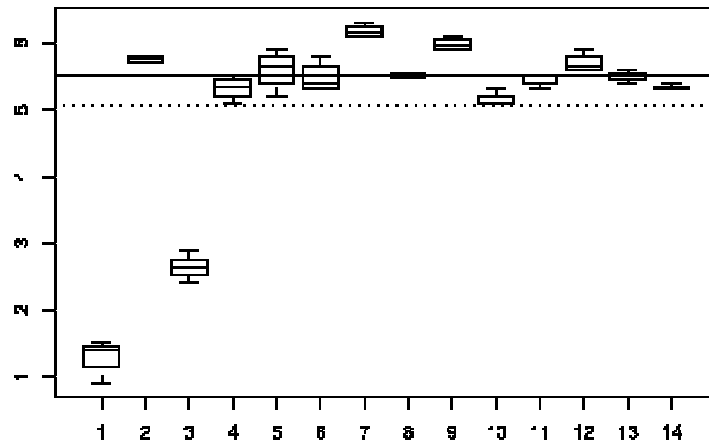


Fig. 2: A boxplot of the data of Table 1. The *dotted line* and the *solid line* denote respectively the mean and the median of the observations

We note that only the results of the Laboratories 1 and 3 lie below the mean whereas all the remaining laboratories return larger values. In the case of the median, 7 of the readings coincide with the median, 24 readings are smaller and 24 are larger. A glance at Fig. 2 suggests that in the absence of further information the Laboratories 1 and 3 should be treated as outliers. For the moment we note simply that the median is a robust statistic whereas the mean is not.

Similarly, *sample variance is often used to measure the variation of observations.*

That is, an example concerns quantifying the scatter of real valued observations x_1, \dots, x_n

Typically, there is a dispute about the relative merits of the following two statistics:

$$s_n = \left(\frac{1}{n} \sum (x_i - \bar{x})^2 \right)^{\frac{1}{2}} \quad \text{and} \quad d_n = \frac{1}{n} \sum |x_i - \bar{x}|.$$

Fisher (<http://www-groups.dcs.st-and.ac.uk/~history/PictDisplay/Fisher.html>) argued that for normal observations the standard deviation s_n is about 12% more efficient than the mean absolute deviation d_n .

In contrast Eddington (<http://www.usd.edu/phys/courses/phys300/gallery/clark/edd.html>) claimed that his experience with real data indicates that d_n is better than s_n .

People find a resolution of this apparent contradiction.

Consider the model

$$\mathcal{N}_\epsilon = (1 - \epsilon)N(\mu, \sigma^2) + \epsilon N(\mu, 9\sigma^2), \quad (1.1)$$

where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 $0 \leq \epsilon \leq 1$

(<http://mathworld.wolfram.com/NormalDistribution.html>) and . For data distributed according to (1.1) or the data is contaminated one can calculate

the asymptotic relative efficiency ARE of d_n with respect to s_n ,

$$ARE(\epsilon) = \lim_{n \rightarrow \infty} RE_n(\epsilon) = \lim_{n \rightarrow \infty} \frac{\text{Var}(s_n)/E(s_n)^2}{\text{Var}(d_n)/E(d_n)^2}$$

As Huber (Huber, P.J., 1981, *Robust Statistics*, Wiley, New York) states, the result is disquieting. Already for $\epsilon \geq 0.002$

ARE exceeds 1 and the effect is apparent for $ARE(\epsilon) = 2.035$

samples of size 1000. For $\epsilon = 0.05$ we have and simulations show that for samples of size 20 the relative efficiency exceeds 1.5 and increases to 2.0 for

samples of size 100. This is a severe deficiency of s_n as models such as \mathcal{N}_ϵ with ϵ between 0.01 and 0.1 often give better descriptions of real data than the normal distribution itself. Thus it becomes painfully clear that the naturally occurring deviations from the idealized model are large enough to render meaningless the traditional asymptotic optimality theory.

Remark: to illustrate a robust method in statistics, the most popular contaminated distribution family is the Tukey supermodel based on the Gaussian law:

$$F = \{ F: F(x) = (1 - \epsilon) \Phi(x) + \epsilon \Phi\left(\frac{x - \theta}{k}\right), 0 < \epsilon < 1, 1 < k \}.$$

Then how to measure the location of a distribution in a robust way if there are **outlier** in the observations. Besides mean, we have

1. median - the median is the value of the point which has half the data smaller than that point and half the data larger than that point. That is, if X_1, X_2, \dots, X_N is a random sample sorted from smallest value to largest value, then the median is defined as:

$$\tilde{Y} = Y_{(N+1)/2} \quad \text{if } N \text{ is odd}$$

$$\bar{Y} = (Y_{N/2} + Y_{(N/2)+1})/2 \quad \text{if } N \text{ is even}$$

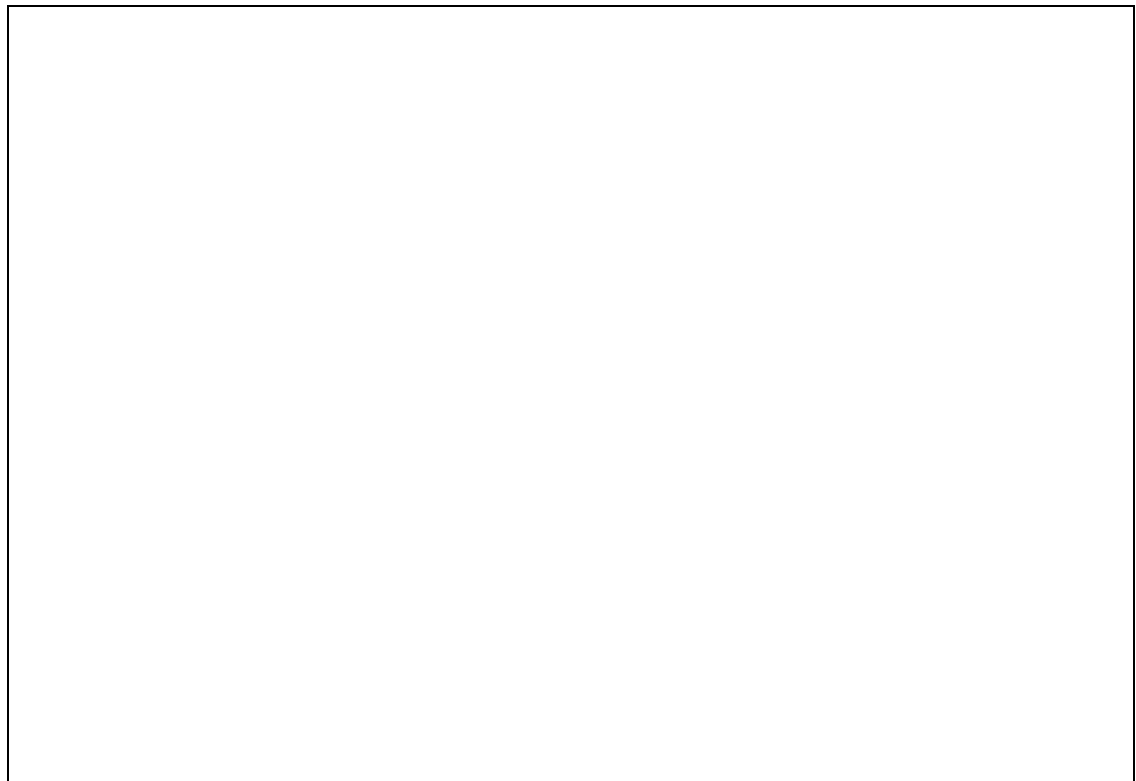
As mentioned above, a single point is enough to greatly influence the mean of a data set. By contrast, at least 50% of a data set must be moved to infinity in order to force the [median](#) to do the same.

The *BREAKDOWN POINT* is the proportion of data which must be moved to infinity so that the estimator will do the same.

Thus in univariate data set the median has a breakdown point of $1/2$ and the mean has a breakdown point of $1/n$, where n is the number of data. It has been shown that the maximum breakdown point for an estimator is $1/2$, so in 1D, the median excels according to this robustness criterion.

2. mode - the mode is the value of the random sample that occurs with the greatest frequency. It is not necessarily unique. The mode is typically used in a qualitative fashion. For example, there may be a single dominant hump in the data perhaps two or more smaller humps in the data. This is usually evident from a histogram of the data. The mode is the value of the peak of the distribution. The mode is biased the least by outliers and contaminants.

Fig.4, the relationship of mode, median and mean in a distribution with outlier.



Mode is particularly adapted by the Bank of England who publishes the *Inflation Report*.

It is the responsibility of Monetary Policy Committee (MPC) to produce the forecast. The forecast is an estimate - probability distribution of possible outcomes for future inflation. Since 1997, the Bank of England quarterly *Inflation Report* has been published explicitly, in the form of a probability density.

Source:

<http://www.bankofengland.co.uk/education/targettwopointzero/inflation/whatsInflation.htm>

http://www.ims.nus.edu.sg/Programs/econometrics/files/kw_ref_13.pdf

<http://www.res.org.uk/economic/freearticles/october04.pdf>

A two-piece normal (2PN) distribution (2PN) is used by the Bank of England, 2PN has probability density functions with parameters μ , σ_1 , σ_2 as

$$f(x) = \begin{cases} A \exp[-(x - \mu)^2 / 2\sigma_1^2] & \text{for } x \leq \mu \\ A \exp[-(x - \mu)^2 / 2\sigma_2^2] & \text{for } x \geq \mu, \end{cases}$$

where $A = (\sqrt{2\pi}(\sigma_1 + \sigma_2)/2)^{-1}$ and μ is the mode.

It is formed by taking two halves of normal distributions with parameters (μ, σ_1) and (μ, σ_2) respectively and scaling them to give the common value $f(\mu) =$ at the mode, as above. The scaling factor applied to the left half of (μ, σ_1) pdf is $2\sigma_1/(\sigma_1 + \sigma_2)$ while to the right half of (μ, σ_2) is $2\sigma_2/(\sigma_1 + \sigma_2)$. If $\sigma_1 = \sigma_2$, then the 2PN distribution collapses to the symmetric normal distribution, but otherwise the density is skewed (to the right if $\sigma_1 < \sigma_2$ and to the left if $\sigma_1 > \sigma_2$) That is, $\sigma_1 < \sigma_2$, then the two-piece normal distribution has positive skewness with mean > median > mode and if $\sigma_1 > \sigma_2$, then the distribution is negatively skewed. The mean and variance are

$$E(x) = \mu + \sqrt{\frac{2}{\pi}}(\sigma_2 - \sigma_1),$$

$$V(x) = (1 - \frac{2}{\pi})(\sigma_2 - \sigma_1)^2 + \sigma_1\sigma_2,$$

The probability of outcomes between L_1 and L_2 for the two-piece normal distribution is derived in John, S. (1982) and is

$$pr[L_1 \leq x \leq L_2] = \int_{L_1}^{L_2} f(x) dx = \frac{2\sigma}{(\sigma_1 + \sigma_2)} \left[\Phi\left(\frac{L_2 - \mu}{\sigma}\right) - \Phi\left(\frac{L_1 - \mu}{\sigma}\right) \right],$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution and

$$\sigma = \begin{cases} \sigma_1, & L_1 \leq L_2 \leq \mu \\ \sigma_2, & \mu \leq L_1 \leq L_2. \end{cases}$$

The other issue is that the underlying return distribution is often skewed or contains outliers. In those cases, the mode is a better estimator of the overall time of divergence than the mean or median, as the mode is biased the least by outliers and contaminants (Dalenius, 1965). However, calculation of the mode is more difficult than the mean or median and this has limited its widespread application. But it has been pointed out that a typical relation between mode, mean and median such as *mode=3 median-2 mean* is true or approximately true for many unimodal probability distributions (Section 2.1 of Stuart and Ord, 1994, the appendix of Lee, 1994). We could impose the mode constraint on the construction of a unimodal return distribution. Moreover, as model is the most likely value of inflation forecasting or stock price or exchange rate, extending this relation into a linear forecasting for mode financial returns is very useful. For example, if we have historical times series of mean and median estimates and records of mode, then one-ahead mode prediction in terms of one step-ahead mean and median forecasts is

$\text{mod } e_{t+1} = \beta_0 + \beta_1 \text{median} + \beta_2 \text{mean}$, where β_0, β_1 and β_2 are parameters estimated by the least-squares rule.

General Philosophy

The examples above illustrate a general phenomenon. An optimal statistical procedure based on a particular family of models \mathcal{M}_1 can differ considerably from an optimal procedure based on another family \mathcal{M}_2 even though the families \mathcal{M}_1 and \mathcal{M}_2 are very close. This may be expressed by saying that optimal procedures are often unstable in that small changes in the data or the model can lead to large changes in the analysis. The basic philosophy of robust statistics is to produce statistical procedures which are stable with respect to small changes in the data or model and even large changes should not cause a complete breakdown of the procedure.

PART II, ROBUST REGRESSION

2.1. Introduction

To illustrate, consider the straight-line model,

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

For this model the least squares estimates of the parameters would be computed by minimizing

$$Q = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Doing this by

1. taking partial derivatives of Q with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$,
2. setting each partial derivative equal to zero, and
3. solving the resulting system of two equations with two unknowns

yields the following estimators for the parameters:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

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.

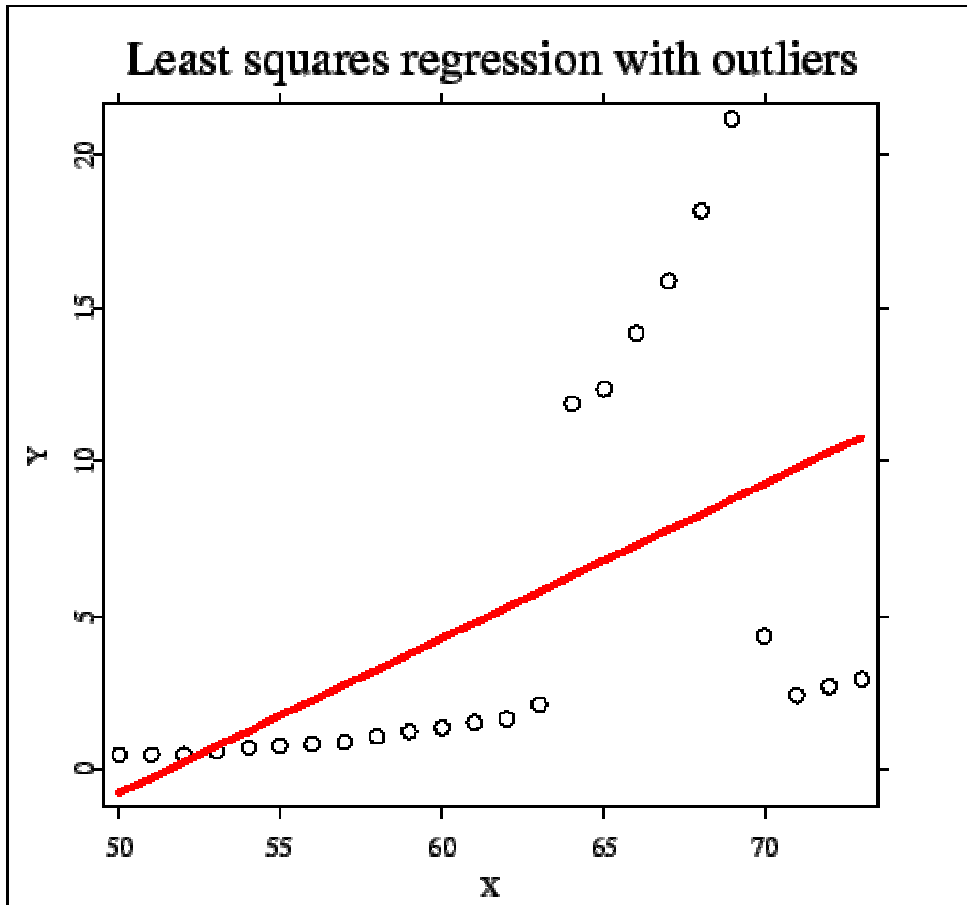


Figure 5: Least squares regression with outliers, [phonecal](#) data

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 Fig. 6. 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 Fig. 6--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 $\pm 3\sigma$).

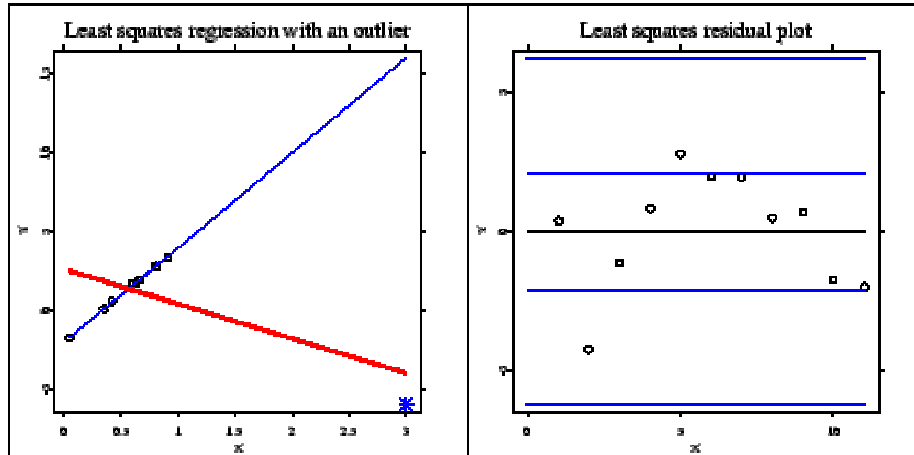


Fig. 6: Least squares regression with one outlier and the corresponding residual plot,

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.

2.2. LEAST TRIMMED SQUARES

The idea of LTS is to minimize the sum of squares using "smallest residuals" only.

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 :

$$y_i = \beta^T x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

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

$$\hat{\beta}^{(LTS)} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^h r_{[i]}^2(\beta), \quad (2.2)$$

where $r_{[i]}^2(\beta)$ represents the i -th order statistic among $r_1^2(\beta), \dots, r_n^2(\beta)$ 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 (2.2) 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.

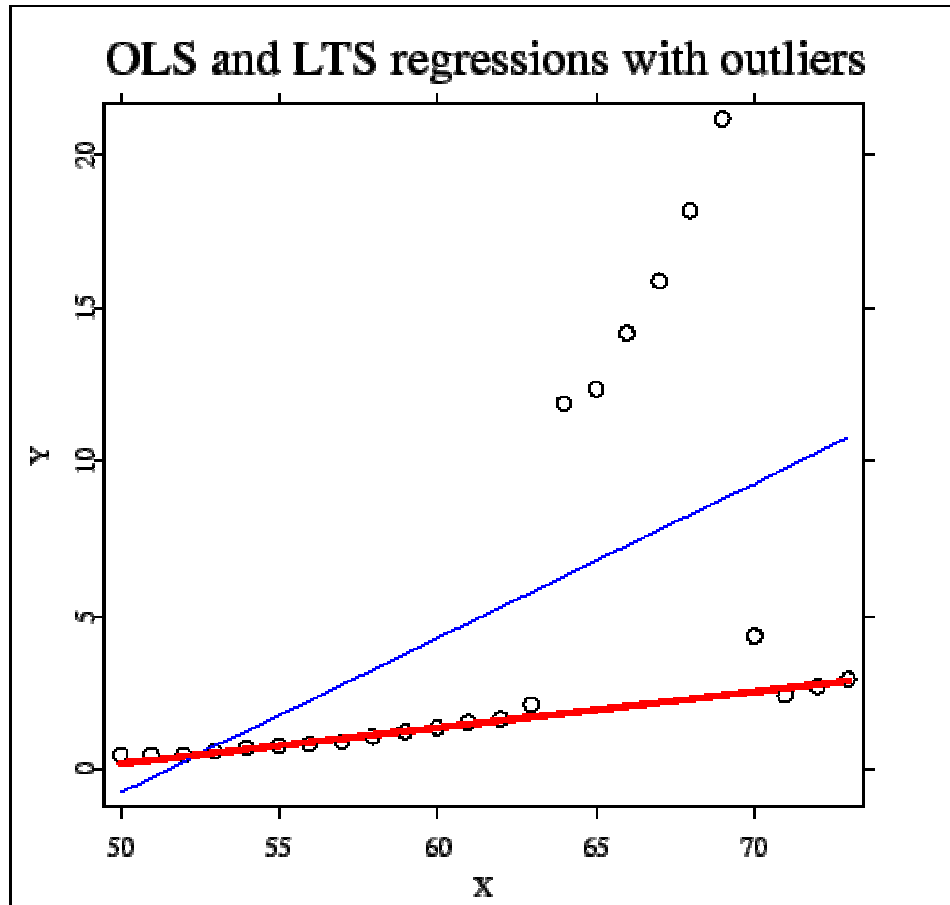


Fig. 7: Least trimmed squares regression with outliers, [phonocal](#) data

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 [phonocal](#) data used in the previous section. On Fig. 8 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 spoiled 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.

Part III, M-regression

Linear least-squares estimates can behave badly when the error distribution is not normal, particularly when the errors are heavy-tailed. One remedy is to remove influential observations from the least-squares fit. Another approach, termed robust regression, is to employ a fitting criterion that is not as vulnerable at least squares to unusual data.

The most common general method of robust regression is *M*-estimation, introduced by Huber(1964).

Consider the linear model

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i \\ &= x_i' \beta + \varepsilon_i \end{aligned}$$

For the *i* th of *n* observations. The fitted model is

$$\begin{aligned} y_i &= a + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik} + e_i \\ &= x_i' b + e_i \end{aligned}$$

The general *M*-estimator minimizes the *objective function*

$$\sum_{i=1}^n \rho(e_i) = \sum_{i=1}^n \rho(y_i - x_i' b)$$

Where the function ρ gives the contribution of each residual to the objective function.

A reasonable ρ should have the following properties:

- $\rho(e) \geq 0$
- $\rho(0) = 0$
- $\rho(e) = \rho(-e)$
- $\rho(e_i) \geq \rho(e_i')$ for $|e_i| > |e_i'|$

For example, for least-squares estimation, $\rho(e_i) = e_i^2$.

Let $\Psi = \rho'$ be the derivative of ρ . Differentiating of the objective function with respect to the coefficients, b , and setting the partial derivatives to 0, produces a system of $k = 1$ estimating equations for the coefficients:

$$\sum_{i=1}^n \Psi(y_i - x_i' b) x_i' = 0$$

Table 1

Method	Objective Function	Weight Function
Least-Squares	$\rho_{LS}(e) = e^2$	$w_{LS}(e) = 1$
Huber	$\rho_H(e) = \begin{cases} \frac{1}{2} e^2 & \text{for } e \leq k \\ k e - \frac{1}{2} k^2 & \text{for } e > k \end{cases}$	$w_H(e) = \begin{cases} 1 & \text{for } e \leq k \\ \frac{k}{ e } & \text{for } e > k \end{cases}$

Bisquare	$\rho_B(e) = \begin{cases} \frac{k^2}{6} \left\{ 1 - \left[1 - \left(\frac{e}{k} \right)^2 \right]^3 \right\} & \text{for } e \leq k \\ k^2 / 6 & \text{for } e > k \end{cases}$	$w_B(e) = \begin{cases} \left[1 - \left(\frac{e}{k} \right)^2 \right]^2 & \text{for } e \leq k \\ 0 & \text{for } e > k \end{cases}$
----------	--	--

