INSTITUTE OF ECONOMIC STUDIES, FACULTY OF SOCIAL SCIENCES
CHARLES UNIVERSITY IN PRAGUE (*established 1348*)

ROBUST STATISTICS AND ECONOMETRICS

INSTITUTE OF ECONOMIC STUDIES
FACULTY OF SOCIAL SCIENCES
CHARLES UNIVERSITY IN PRAGUE

JAN ÁMOS VÍŠEK

Week 6

Content of lecture

1. Linear regression
   - Repetition - notations, history, goals, misconceptions, snags and reality
   - Outliers and leverage points
   - Estimating regression model by alternative methods

2. Feasible high breakdown point estimators
   - Deleting some observations

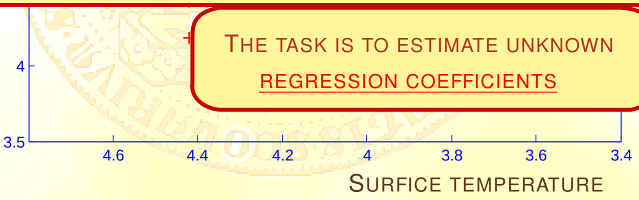
Recalling notations we have fixed in the the first lecture.

REGRESSION MODEL

$$\begin{aligned} Y_i &= X_i'\beta^0 + e_i \\ &= X_{i1}\beta_1^0 + X_{i2}\beta_2^0 + \ldots + X_{ip}\beta_p^0 + e_i, \end{aligned}$$

i = 1, 2, ...,n

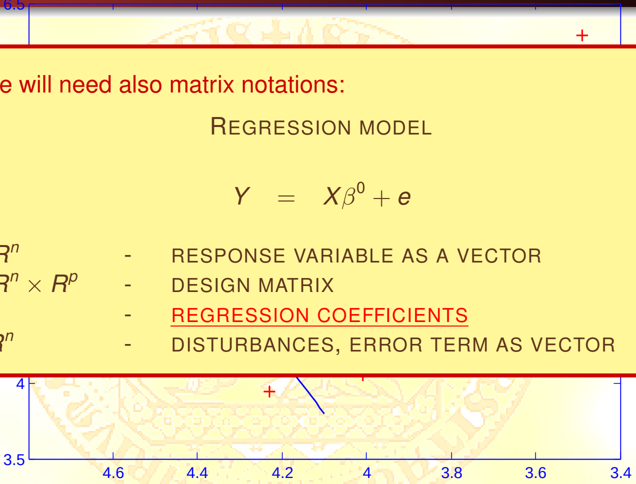
Galton, F. (1886): Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute vol. 15,. 246–263.*



THE TASK IS TO ESTIMATE UNKNOWN REGRESSION COEFFICIENTS

Enlarging a bit notations

LUMINIOUS



SURFICE TEMPERATURE

Today we will need also matrix notations:

REGRESSION MODEL

$$Y \quad = \quad X\beta^0 + e$$

$Y \in R^n$	-	RESPONSE VARIABLE AS A VECTOR
$X \in R^n \times R^p$	-	DESIGN MATRIX
β^0	-	REGRESSION COEFFICIENTS
$e \in R^n$	-	DISTURBANCES, ERROR TERM AS VECTOR

Recalling recommendable framework

We should use always the model with intercept, i. e. with desigh matrix

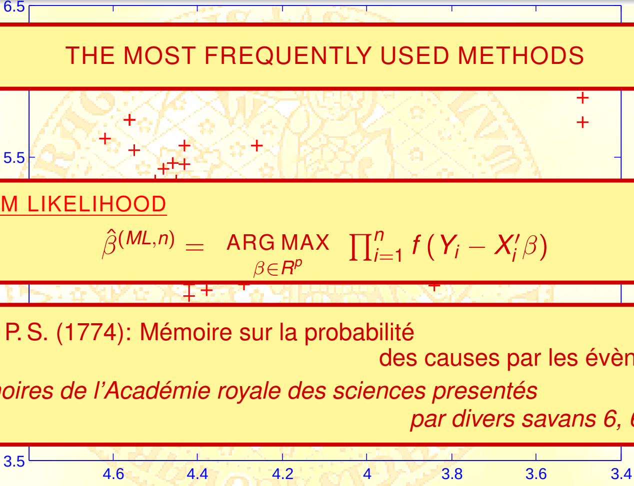
$$\begin{bmatrix} 1, & x_{1,2}, & \ldots, & x_{1,p} \\ 1, & x_{2,2}, & \ldots, & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1, & x_{n,2}, & \ldots, & x_{n,p} \end{bmatrix}.$$

with one exception - which one?

It force the estimator to do one important thing. Which one?

(We in fact impute an additional information into processing the data.)

A drop of history



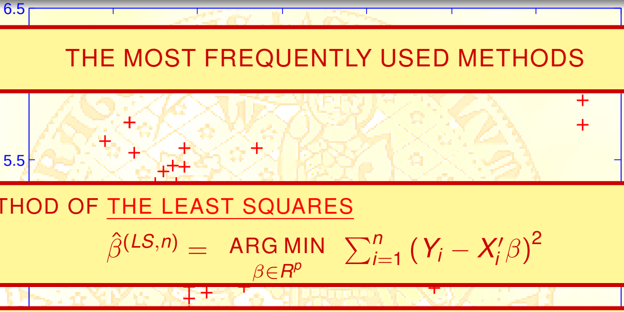
THE MOST FREQUENTLY USED METHODS

MAXIMUM LIKELIHOOD

$$\hat{\beta}^{(ML,n)} = \underset{\beta \in R^p}{\text{ARG MAX}} \prod_{i=1}^{n} f\left(Y_i - X_i'\beta\right)$$

Laplace, P. S. (1774): Mémoire sur la probabilité des causes par les évènemens. *Mémoires de l'Académie royale des sciences presentés par divers savans 6, 621 – 656.*

A drop of history



THE MOST FREQUENTLY USED METHODS

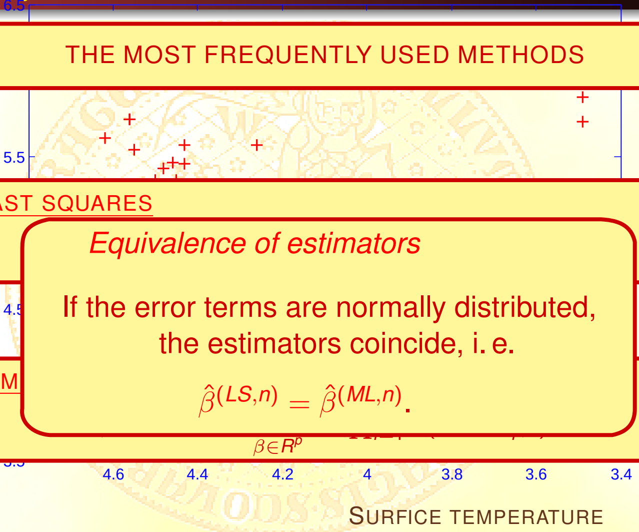
THE METHOD OF THE LEAST SQUARES

$$\hat{\beta}^{(LS,n)} = \underset{\beta \in R^p}{\text{ARG MIN}} \sum_{i=1}^{n} (Y_i - X_i' \beta)^2$$

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

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

A bit of theory



THE MOST FREQUENTLY USED METHODS

THE LEAST SQUARES

Equivalence of estimators

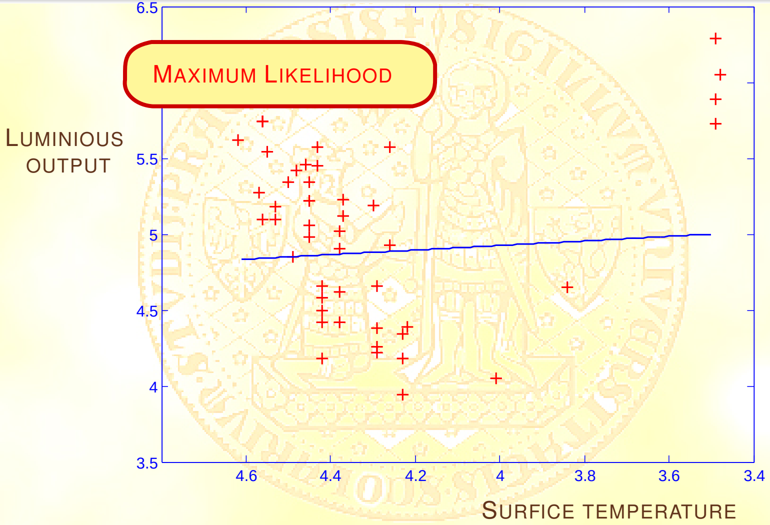
If the error terms are normally distributed, the estimators coincide, i. e.

$$\hat{\beta}^{(LS,n)} = \hat{\beta}^{(ML,n)}.$$

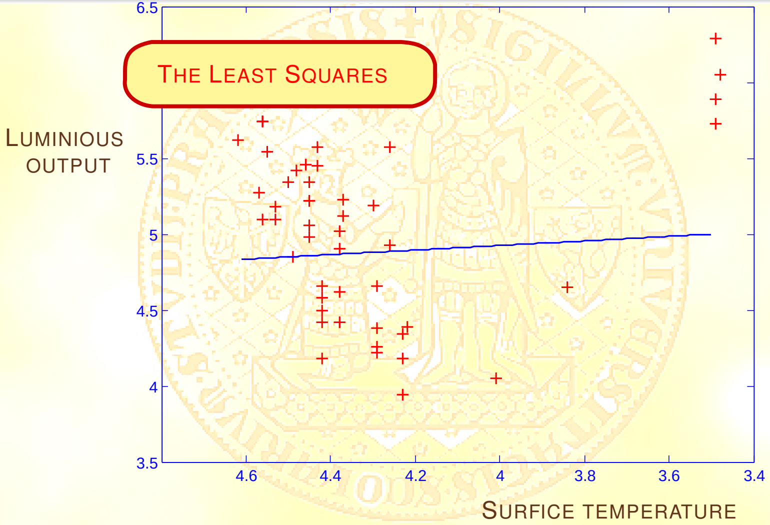
MAXIMUM

$\beta \in R^p$

A bit of theory



A bit of theory



Estimating by means of L_1 metric

$$\hat{\beta}^{(L_1,n)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} |Y_i - X_i'\beta|$$

Galilei, G. (1632): *Dialogo dei massimi sistemi.* Pisa.

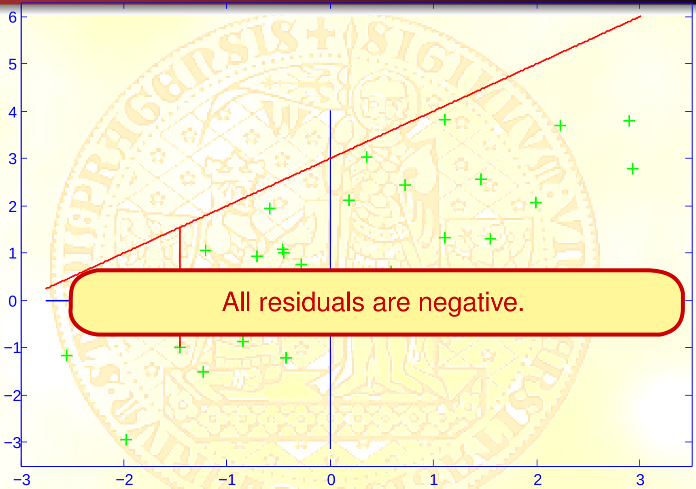
Boscovisch, R. J. (1757): De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura eius ex exemplaria etiam sensorum impressa.

But how did they solve this extremal problem?

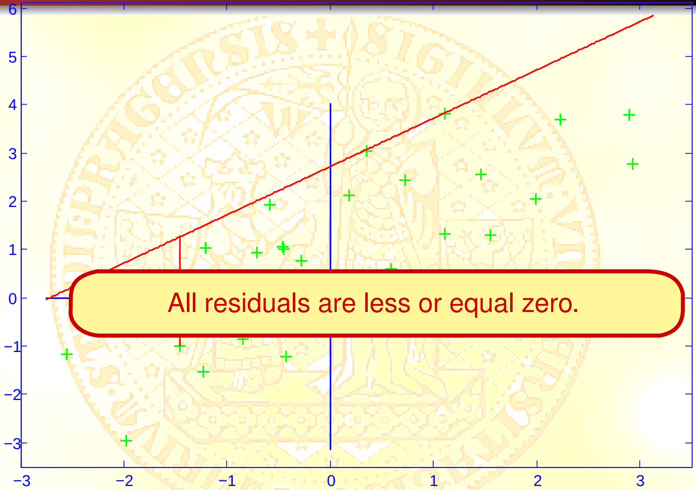
Laplace, P. S. (1793): Sur quelques points du systeme du mode. *Memoires de l'Academic Royale des Sciences de Paris, 1-87.*

(A hint on the next slide!!)

The solution can be found by the rule and pencil.

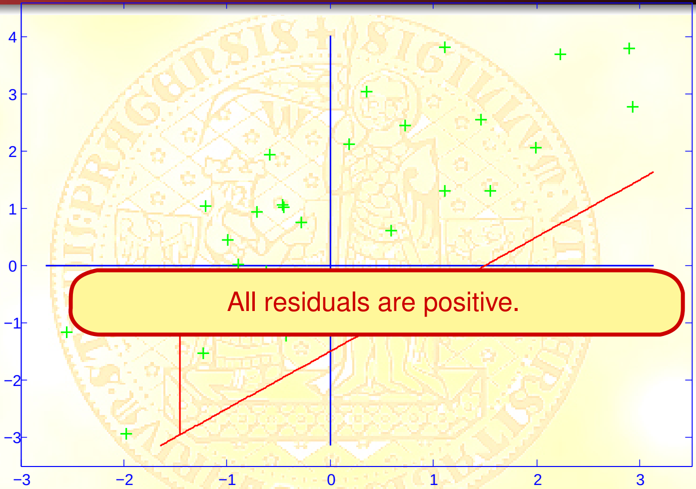


The solution can be found by the rule and pencil.

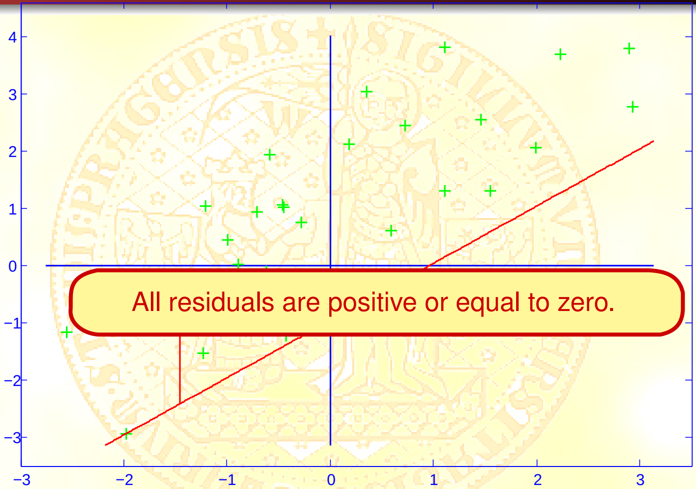


All residuals are less or equal zero.

The solution can be found by the rule and pencil.

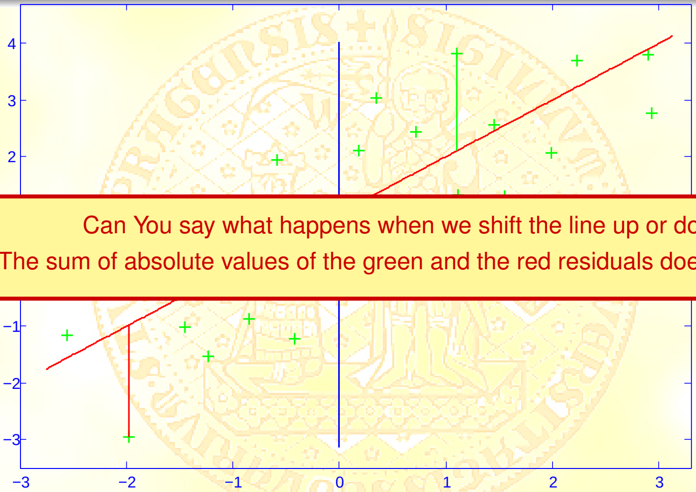


The solution can be found by the rule and pencil.



All residuals are positive or equal to zero.

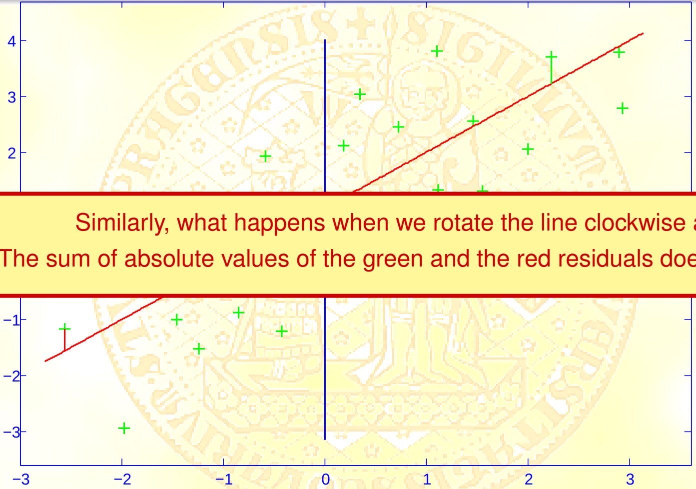
The solution can be found by the rule and pencil.



Can You say what happens when we shift the line up or down?

The sum of absolute values of the green and the red residuals doesn't change.

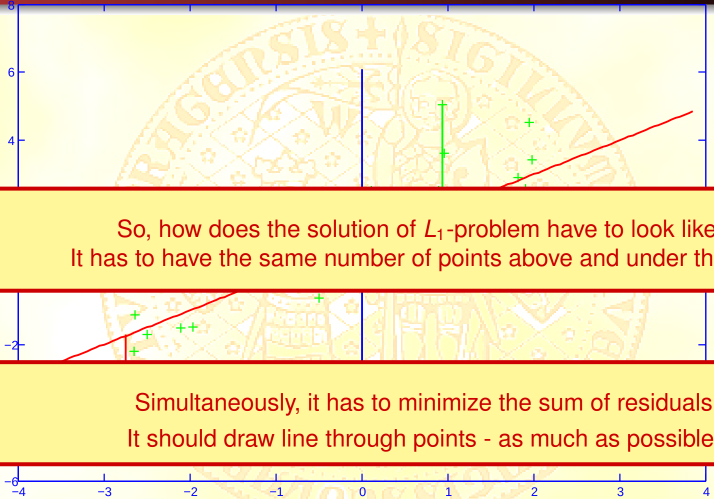
The solution can be found by the rule and pencil.



Similarly, what happens when we rotate the line clockwise a bit?

The sum of absolute values of the green and the red residuals doesn't change.

The solution can be found by the rule and pencil.



So, how does the solution of L_1-problem have to look like?
It has to have the same number of points above and under the line.

Simultaneously, it has to minimize the sum of residuals.
It should draw line through points - as much as possible.

In the 5th lecture *M*-estimators for general parameter were considered.

We have considered a general parameter family:

Let $\{F(x,\theta)\}_{\theta\in\Theta}$ and $\{f(x,\theta)\}_{\theta\in\Theta}$ be families of d. f.'s and densities, respectively.

Then we have put:

The solution of the extremal problem

$$\hat{\theta}^{(M,n)} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \rho\,(x_i,\theta)$$

is called *Maximum likelihood-like estimators of the parameter* θ or *M*-estimators of θ, for short.

(We are going to specify it for the regression framework but prior to it let's define outliers and leverage points.)

Influential observations - outliers

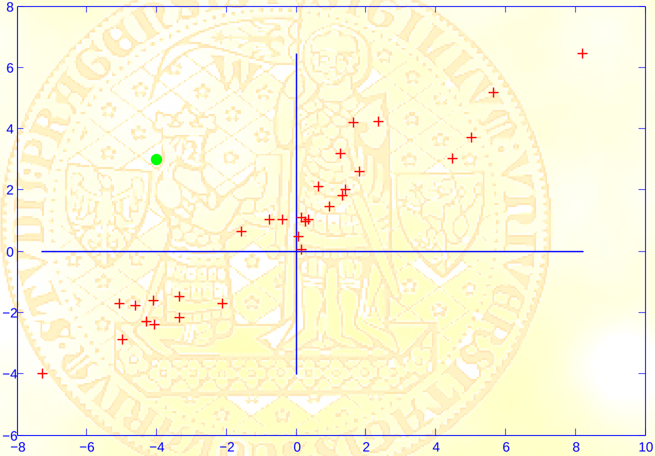
We speak about *outlier* if:

There is an observation
which has values of the explanatory variables "inside" the "cloud of data",

the value of the response variable is however
"far away" from the expected value of response variable.

From possible influential points this is less dangerous
- the figure on the next slide says much more.

Influential observations - outliers



Influential observations - leverage point

We speak about *good leverage point* if:

There is an observation which has

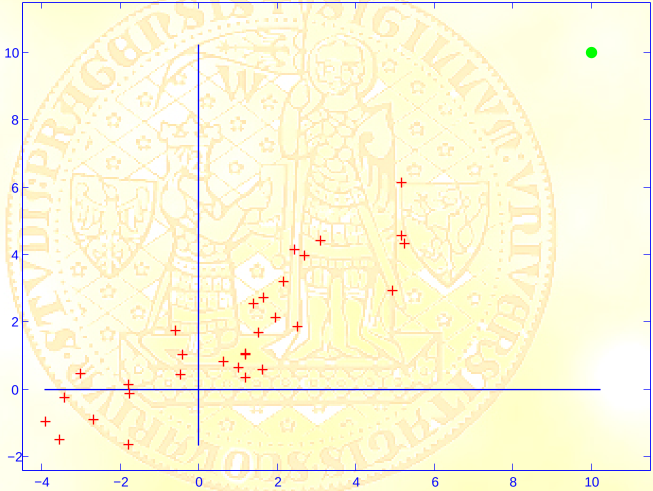
values of the explanatory variables “far away” from the “cloud of data”,

the value of the response variable is however the expected one.

From possible influential points this has a positive influence

- the figure on the next slide says much more.

Influential observations - leverage point



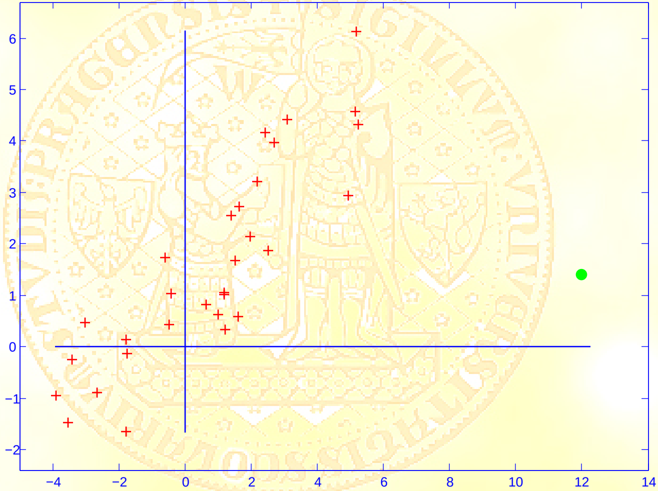
Influential observations - leverage point

We speak about *bad leverage point* if:

There is an observation which has
values of the explanatory variables "(far) away" from the "cloud of data"
and the value of the response variable is also
"(far) away" from the expected value of response variable.

From possible influential points this is the most dangerous
- the figure on the next slide says much more.

Influential observations - leverage point



Basic diagnostic tool

Hat matrix

$$X\,(X'X)^{-1}\,X'$$

$$\begin{bmatrix} \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix} \left(\begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \cdot \begin{bmatrix} \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix} \right)^{-1} \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} = \begin{bmatrix} \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \end{bmatrix}$$

and its diagonal - see the next several slides.

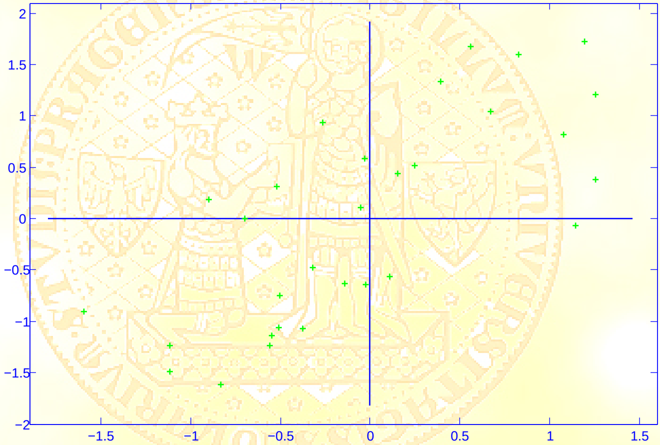
Recognizing the influential points

All these influential points can be easily recognized (for simplicity assume intercept in model).

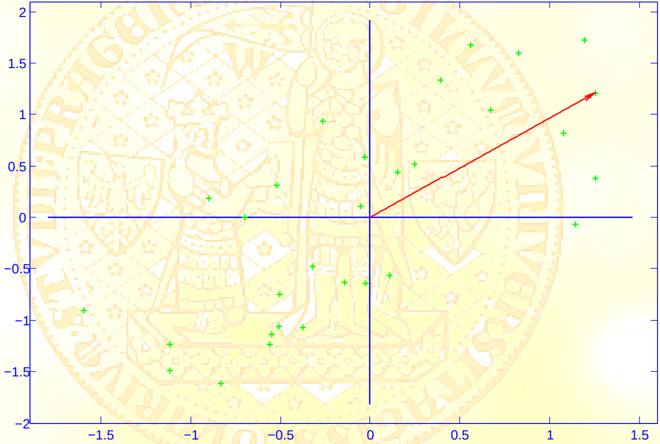
To see it, let's make some preliminary considerations. Realize that:

1. for any observation the vector of explanatory variables X_i specifies its location in the space of explanatory variables, i. e. in R^p,
2. $\|X_i\| = \sqrt{\sum_{j=1}^{p} X_{ij}^2}$ is the length of vector X_i, i. e. the distance of observation from the origin in R^p,
3. $\|X_i\|^2 = \sum_{j=1}^{p} X_{ij}^2 = X_i' I\!\!I X_i$ where $I\!\!I$ is the (diagonal) unit matrix,

Continuing in preliminary considerations



Continuing in preliminary considerations



Recognizing the influential points

All these influential points can be easily recognized (for simplicity assume intercept in model).

To see it, let's make some preliminary considerations. Realize that:

1. for any observation the vector of explanatory variables X_i specifies its location in the space of explanatory variables, i. e. in R^p,
2. $\|X_i\| = \sqrt{\sum_{j=1}^{p} X_{ij}^2}$ is the length of vector X_i, i. e. the distance of observation from the origin in R^p,
3. $\|X_i\|^2 = \sum_{j=1}^{p} X_{ij}^2 = X_i' I\!\!I X_i$ where $I\!\!I$ is the (diagonal) unit matrix,
4. substitute $I\!\!I$ by $(X'X)^{-1}$ and find what the value $d^2(X_i) = X_i' (X'X)^{-1} X_i$ represents.

Continuing in preliminary considerations

What is $d^2(X_i) = X_i' \left(X'X\right)^{-1} X_i$?

1. The first row (and the first column, of course) of $X'X$ is

$$n\bar{X}' = \left(n, \sum_{i=1}^{n} X_{i2}, \sum_{i=1}^{n} X_{i3}, \ldots, \sum_{i=1}^{n} X_{ip}\right).$$

Recalling the desigh matrix

$$\begin{bmatrix} 1, & x_{1,2}, & \ldots, & x_{1,p} \\ 1, & x_{2,2}, & \ldots, & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1, & x_{n,2}, & \ldots, & x_{n,p} \end{bmatrix}$$

and its transposition:

$$\begin{bmatrix} 1, & 1, & \ldots, & 1 \\ x_{1,2}, & x_{2,2}, & \ldots, & x_{n,2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1,p} & x_{2,p}, & \ldots, & x_{n,p} \end{bmatrix}.$$

Recalling the desigh matrix

Hence the first line of the matrix given by the product

$$\begin{bmatrix} 1, & 1, & \ldots, & 1 \\ x_{1,2}, & x_{2,2}, & \ldots, & x_{n,2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1,p} & x_{2,p}, & \ldots, & x_{n,p} \end{bmatrix} \times \begin{bmatrix} 1, & x_{1,2}, & \ldots, & x_{1,p} \\ 1, & x_{2,2}, & \ldots, & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1, & x_{n,2}, & \ldots, & x_{n,p} \end{bmatrix}$$

is $\left(n, \sum_{i=1}^{n} X_{i2}, \sum_{i=1}^{n} X_{i3}, \ldots, \sum_{i=1}^{n} X_{ip}\right) = n\overline{X}'$.

Continuing in preliminary considerations

What is $d^2(X_i) = X_i' \left(X'X\right)^{-1} X_i$?

1. The first row (and the first column, of course) of $X'X$ is
$$n\overline{X}' = \left(n, \sum_{i=1}^{n} X_{i2}, \sum_{i=1}^{n} X_{i3}, \ldots, \sum_{i=1}^{n} X_{ip}\right),$$
2. from $X'X\left(X'X\right)^{-1} = \mathbb{I}$ it follows that
$$n\overline{X}' \left(X'X\right)^{-1} = (1, 0, \ldots, 0), \text{ i. e. } \overline{X}' \left(X'X\right)^{-1} = (1/n, 0, \ldots, 0),$$
3. $\left(X_i - \overline{X}\right)' \left(X'X\right)^{-1} \left(X_i - \overline{X}\right)$
$$= X_i' \left(X'X\right)^{-1} X_i - \overline{X}' \left(X'X\right)^{-1} X_i - X_i' \left(X'X\right)^{-1} \overline{X} + \overline{X} \left(X'X\right)^{-1} \overline{X}$$
$$d^2(X_i) - 1/n - 1/n + 1/n = d^2(X_i) - 1/n\,.$$

Continuing in preliminary considerations

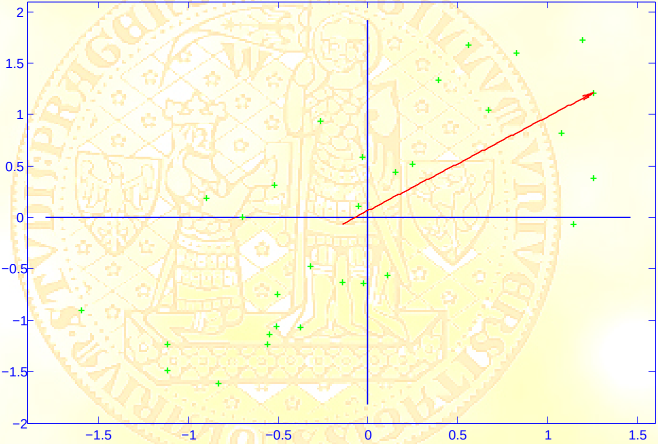
We have found:

$$d^2(X_i) = \left(X_i - \overline{X}\right)' \left(X'X\right)^{-1} \left(X_i - \overline{X}\right) + 1/n,$$

i. e. except of $1/n$, $d^2(X_i)$ is the squared distance of given observation from the "center of gravity" of the cloud of all observations.

Can we make an idea how large it is (typically)?

Continuing in preliminary considerations



Continuing in preliminary considerations

We easy verify that:

1. $d^2(X_i) = X_i' \left(X'X\right)^{-1} X_i = \left[X\left(X'X\right)^{-1} X'\right]_{ii}$,

A lot of information can be found in

Chatterjee, S., A. S. Hadi (1988):
Sensitivity Analysis in Linear Regression.
New York: J. Wiley & Sons.

$$trace\left(X\left(X'X\right)^{-1} X'\right) = trace\left(X'X\left(X'X\right)^{-1}\right) = trace\left(\mathbb{I}\right) = p,$$

4. the matrix $X\left(X'X\right)^{-1} X'$ has n diagonal elements,
hence each of them is approximately p/n large.

M-estimators for the regression framework.

The solution of the extremal problem

$$\hat{\beta}^{(M,n)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} \rho\left(Y_i - X_i'\beta\right)$$

is called

Maximum likelihood-like estimators of the regression coefficients

or *M*-estimators of β^0, for short.

(We can use the same ρ as for location and scale.)

We usually adopt some basic assumptions:

Let $F(x,r), x \in R^p, r \in R$ be a d.f. (with a density $f(x,r)$) governing the explanatory variables and disturbances in the regression model.

Evidently this form of definition inevitably implies that $\hat{\beta}^{(M,n)}$ is not scale- and regression-equivariant.

(possible solutions of the problem on the next but one slide). An advantage - on the other hand -

Computing *M*-estimate of regression coefficients

Consider the extremal problem (with $\rho(0) = 0$)

$$\hat{\beta}^{(M,n)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} \rho\left(Y_i - X_i'\beta\right) = \arg\min_{\beta \in R^p} \sum_{\left\{i: Y_i - X_i'\beta \neq 0\right\}} \rho\left(Y_i - X_i'\beta\right).$$

Write it as

$$\hat{\beta}^{(M,n)} = \arg\min_{\beta \in R^p} \sum_{\left\{i: Y_i - X_i'\beta \neq 0\right\}} \frac{\rho\left(Y_i - X_i'\beta\right)}{\left(Y_i - X_i'\beta\right)^2} \left(Y_i - X_i'\beta\right)^2$$

$$= \arg\min \sum^{n} w_i \cdot (Y_i - X_i'\beta)^2$$

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

where $W = \mathrm{diag}(w_1, w_2, \ldots, w_n)$.

And an iterative computation, starting with a preliminary "guess" of β^0 .

GM-estimators for the regression framework.

The solution of the extremal problem

$$\hat{\beta}^{(M,n)} = \arg\min_{\beta \in R^p, \sigma \in R^+} \sum_{i=1}^{n} \rho\left(\frac{Y_i - X_i'\beta}{\sigma}\right)$$

is called

General(ized) Maximum likelihood-like estimator of the regression coefficients or *GM*-estimator of β^0, for short.

(We can still use the same ρ as in previous.)

Evidently this estimator is scale- and regression-equivariant but the computation is not easy.

GM-estimators for the regression framework.

That is why we usually select a preliminary consistent (sufficiently robust) estimator of standard deviation of disturbances, say $\hat{\sigma}^{(n)}$ and put:.

The solution of the extremal problem

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

is also called *Generalized Maximum likelihood-like estimator of the regression coefficients* or *GM*-estimator of β^0, for short.

(We can still use the same ρ as in previous.)

This proposal is frequently used but even experienced statisticians are not aware that it has a drawback - see the next slide.

Repetition from the 3rd lecture

Equivariance of $\hat{\beta}^{(n)}$

$$\hat{\beta}(Y, X) : M(n, p+1) \to R^p$$

$$\textit{scale-equivariant} \ : \quad \forall c \in R^+ \quad \hat{\beta}(cY, X) = c\hat{\beta}(Y, X)$$

$$\textit{regression-equivariant} \ : \quad \forall b \in R^p \quad \hat{\beta}(Y + Xb, X) = \hat{\beta}(Y, X) + b$$

$$\text{Examples} : \hat{\beta}^{(OLS,n)} = (X'X)^{-1} X'Y$$

$$\hat{\beta}^{(L_1,n)} = \ ...$$

We have justified the requirement of equivariance

What is the equivariance of $\hat{\beta}^{(n)}$ good for ?

1. When the units of measurement have been changed,
   we don't need to recalculate the estimator
   - we just shift the decimal point
   (we are used to it from classical statistics).
2. The requirement of invariance and equivariance
   removed superefficiency.

Problems with studentization of residuals

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

To reach *scale-* and *regression-equivariance* of an *M*-estimator by

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

$\hat{\sigma}^{(n)}$ has to be *scale-equivariant* and *regression-invariant*.

The studentization requires special estimator of scale

Equivariance - invariance of $\hat{\sigma}^2$

$$\hat{\sigma}^2(Y,X) : M(n,p+1) \to R^+$$

scale-equivariant : $\forall c \in R^+ \quad \hat{\sigma}^2(cY,X) = c^2\hat{\sigma}^2(Y,X)$

regression-invariant : $\forall b \in R^p \quad \hat{\sigma}^2(Y+Xb,X) = \hat{\sigma}^2(Y,X)$

Examples : $s_n^2 = \frac{1}{n-p}\sum_{i=1}^n r_i^2(\hat{\beta}^{(OLS,n)})$ if $\mathcal{L}(\varepsilon) = \mathcal{N}(\mu,\sigma^2)$

$\hat{\sigma}_{(L_1,n)} = MAD$ if $\mathcal{L}(\varepsilon) = DoubleExp(\lambda)$

$\hat{\sigma}_{(L_1,n)} = 1.483 \cdot MAD$ if $\mathcal{L}(\varepsilon) = \mathcal{N}(\mu,\sigma^2)$

$MAD = \operatorname*{med}_{1\le i\le n} \left| r_i(\hat{\beta}^{(L_1,n)}) - \operatorname*{med}_{1\le i\le n} r_i(\hat{\beta}^{(L_1,n)}) \right|$, $\quad E_{\mathcal{N}(0,1)}\, MAD = (1.483)^{-1}$

The studentization requires special estimator of scale

There are not too much estimators of scale of disturbances
which are consistent, scale-equivariant and regression-invariant:

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

Jurečková, J., P. K. Sen (1993): Regression rank scores scale statistics and

Their common feature - all these estimators are based
on the scale- and regression-equivariant estimator of β^0.

IMS Collections. Nonparametrics and Robustness in Modern Statistical Inference
and Time Series Analysis: Festschrift for Jana Jurečková, Vol. 7(2010), 254 - 267.

Let's remember for the next study:

Preliminary conclusion

We should prefer (robust) estimators of regression coefficients which are *"automatically"* scale- and regression-equivarint.

A pursuit for highly robust estimator of regression coefficients

Let's recall:

- *Breakdown point* - "finite" sample version

$$x_1, x_2, \dots, x_n \quad \Rightarrow \quad T_n(x_1, x_2, \dots, x_n)$$

- Find maximal m_n such that for any
$y_1, y_2, \dots, y_{m_n} \quad \Rightarrow \quad |T_n(x_1, x_2, \dots, x_{n-m_n}, y_1, y_2, \dots, y_{m_n})| < \infty$

$$(\ 0 < T_n(x_1, x_2, \dots, x_{n-m_n}, y_1, y_2, \dots, y_{m_n}) < \infty \text{ - for scale }).$$

- Put

$$\varepsilon^* = \lim_{n\to\infty} \frac{m_n}{n}.$$

A pursuit for highly robust estimator of regression coefficients

Hampel's approach - characteristics of the functional T at the d. f. F

- *Breakdown point* - "finite" sample version - examples

$$x_1, x_2, ..., x_n \quad \Rightarrow \quad T_n(x_1, x_2, ..., x_n) = \frac{1}{n} \sum_{i=1}^{\infty} x_i.$$

- Maximal m_n such that for any

$$y_1, y_2, ..., y_{m_n} \quad \Rightarrow \quad |T_n(x_1, x_2, ..., x_{n-m_n}, y_1, y_2, ..., y_{m_n})| < \infty$$

is zero,

- hence

$$\varepsilon^* = 0.$$

A pursuit for highly robust estimator of regression coefficients

Hampel's approach - characteristics of the functional T at the d. f. F

- *Breakdown point* - "finite" sample version - examples

$$x_1, x_2, ..., x_n \quad \Rightarrow \quad T_n(x_1, x_2, ..., x_n) = med\{x_1, x_2, ..., x_n\}.$$

- Maximal m_n such that for any

$$y_1, y_2, ..., y_{m_n} \quad \Rightarrow \quad |T_n(x_1, x_2, ..., x_{n-m_n}, y_1, y_2, ..., y_{m_n})| < \infty$$

is $\frac{n}{2}$,

- hence

$$\varepsilon^* = \frac{1}{2}.$$

A pursuit for highly robust estimator of regression coefficients

Hence, already in seventies, a question appeared:

CAN WE CONSTRUCT AN ESTIMATOR OF REGRESSION COEFFICIENTS

WITH $\varepsilon^* = \frac{1}{2}$?

see e. g.

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.

or

BICKEL, P. J. (1975): ONE-STEP HUBER ESTIMATES IN THE LINEAR MODEL. *J. Amer. Statist. Assoc. 70, 428–433.*

We had: Problems with studentization of residuals

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

To reach *scale-* and *regression-equivariance* of an *M*-estimator by

$$\hat{\beta}^{(M,n)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} \rho\left(\frac{Y_i - X_i'\beta}{\hat{\sigma}^{(n)}}\right)$$

$\hat{\sigma}^{(n)}$ has to be *scale-equivariant* and *regression-invariant*.

Assume we are able to find $\hat{\sigma}^{(n)}$ fulfilling the requirements
- we can have still problems.

Problems of *M*-estimators towards leverage points

M-estimator given by

$$\hat{\beta}^{(M,n)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} \rho\left(\frac{Y_i - X_i'\beta}{\hat{\sigma}^{(n)}}\right)$$

has to fulfill

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

If $\|X_i\|$ is large, the i-th observation has large impact on $\hat{\beta}^{(M,n)}$.
The influence of leverage points on *M*-estimators can be (very) harmful.

Possible remedy for *M*-estimators

What about to define *M*-estimator by

$$\hat{\beta}^{(M,n,w)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} w(X_i)\rho\left(\frac{Y_i - X_i'\beta}{\hat{\sigma}^{(n)}}\right)$$

where $w(.)$ is a weight function.

$\hat{\beta}^{(M,n,w)}$ is also called

Generalized Maximum likelihood-like estimator

of the regression coefficients or *GM*-estimator of β^0, for short.

A pursuit for highly robust estimator of regression coefficients

Prior to continuing let us make an agreement:

For any $\beta \in R^p$

$$r_i(\beta) = Y_i - X_i'\beta \qquad \text{not only} \qquad \cancel{r_i(\hat{\beta}) = Y_i - X_i'\hat{\beta}}$$

Order statistics

$$r^2_{(1)}(\beta) \le r^2_{(2)}(\beta) \le \dots \le r^2_{(n)}(\beta),$$

some texts alternatively employ

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

A pursuit for highly robust estimator of regression coefficients

Regression quantiles

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

$$\hat{\beta}^{(\alpha)} = \underset{\beta \in R^p}{\arg\min} \left\{ \sum_{i=1}^{n} \left[\alpha \cdot |r_i(\beta)| \cdot I\{r_i(\beta) < 0\} + (1-\alpha) \cdot |r_i(\beta)| \cdot I\{r_i(\beta) > 0\} \right] \right\}$$

$$\hat{\beta}^{(L,n)} = \sum_{\ell=1}^{K} c_\ell \hat{\beta}^{(\alpha_\ell)}$$

$\hat{\beta}^{(\alpha)}$ is *M*- and simultaneously *L*-estimator

A pursuit for highly robust estimator of regression coefficients

The trimmed least squares (TLS)

Ruppert, D., R. J. Carroll (1980):

Trimmed least squares estimation in linear model.

J. Americal Statist. Ass., 75 (372), 828–838.

Trimming by $\left[x' \cdot \hat{\beta}^{(\alpha_1)}, x' \cdot \hat{\beta}^{(\alpha_2)}\right] \quad 0 \leq \alpha_1 < \alpha_2 \leq 1 \quad \rightarrow \quad \hat{\beta}^{(TLS,n)_{(\alpha_1,\alpha_2)}}$

A pursuit for highly robust estimator of regression coefficients

DISAPPOINTMENT

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

$$!!! \quad \varepsilon^* = \frac{1}{p} \quad !!!$$

(p - dimension of regression model)

The First Estimator with 50% Breakdown Point

Repeated medians

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

$$\hat{\beta}^{(j)} = \underset{i_1=1,2,\ldots,n}{\text{med}} \left(\ldots \left(\underset{i_{p-1}=1,2,\ldots,n}{\text{med}} \left(\underset{i_p=1,2,\ldots,n}{\text{med}} \left(\hat{\beta}_j \left(i_1, i_2, \ldots, i_p \right) \right) \right) \right) \right)$$

(requiring approx. $p \cdot n^p$ evaluations of model and orderings of estimates of coefficients
- nearly surely never implemented)

The first solution broke the mystery and implied a chain of others

Rousseeuw, P. J. (1983): Least median of square regression.
Journal of Amer. Statist. Association 79, pp. 871-880.

the Least Median of Squares

$$\hat{\beta}^{(LMS,n,h)} = \arg\min_{\beta \in R^p} r^2_{(h)}(\beta) \qquad \frac{n}{2} < h \le n,$$

(implementation will be discussed later).

Many advantages - mainly

1. *breakdown point equal to* $([\frac{n-p}{2}] + 1)n^{-1}$ *if* $h = [\frac{n}{2}] + [\frac{p+1}{2}]$
2. *scale- and regression equivariant* (without any studentization of residuals).

Main disadvantage $\sqrt[3]{n}\left(\hat{\beta}^{(LMS,n,h)} - \beta^0\right) = \mathcal{O}_p(1)$ (other will be discussed later).

Let's remove the deficiency of low rate of convergence of LMS

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

the Least Trimmed Squares

$$\hat{\beta}^{(LTS,n,h)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{h} r_{(i)}^2(\beta) \qquad \frac{n}{2} < h \leq n,$$

(Notice the order of words, remember there is also the Trimmed Least Squares.)

Many advantages - e. g.

1. the breakdown point equal to $([\frac{n-p}{2}] + 1)n^{-1}$ if $h = [\frac{n}{2}] + [\frac{p+1}{2}]$
2. *scale- and regression equivariant*
3. $\sqrt{n}\left(\hat{\beta}^{(LTS,n,h)} - \beta^0\right) = \mathcal{O}_p(1)$

Let's increase the efficiency with simultaneously keeping high breakdown point

Rousseeuw, P. J., V. Yohai (1984):
Robust regressiom by means of *S*-estimators.
Lecture Notes in Statistics No. 26 Springer Verlag, New York, 256-272.

S-estimators

$$\hat{\beta}^{(S,n,\rho)} = \arg\min_{\beta \in R^p} \left\{ \sigma \in R^+ : \sum_{i=1}^{n} \rho\left(\frac{r_i(\beta)}{\sigma}\right) = b \right\}$$

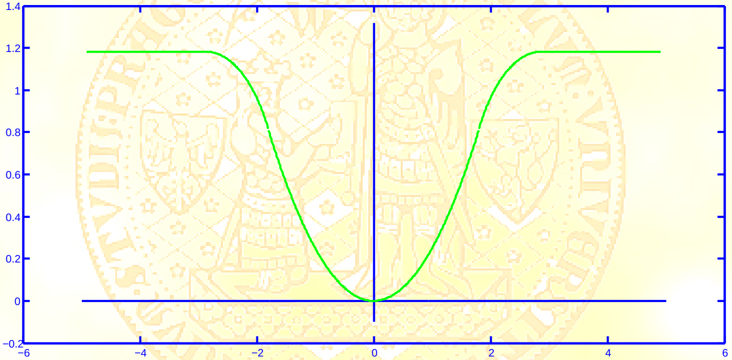
where $b = E\rho\left(\frac{e_i}{\sigma_0}\right)$ with $\sigma_0^2 = Ee_1^2$ (for ρ see next slide).

Many advantages - e. g.

1. the breakdown point equal to 50%,
2. *scale- and regression equivariant,*
3. $\sqrt{n}\left(\hat{\beta}^{(S,n,\rho)} - \beta^0\right) = \mathcal{O}_p(1)$,
4. much better utilization of information from data, i. e. higher efficiency than LTS.

Peter Rousseeuw's objective function ρ

$$\rho : (-\infty, \infty) \to (0, \infty),\ \rho(x) = \rho(-x),\ \rho(0) = 0, \rho(x) = c \text{ for } x > d.$$



THANKS FOR ATTENTION