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 7

Content of lecture

1. At the beginning of any lecture let us repeat .....
   - From basic econometrics
   - Repetition from the previous lecture
2. Feasible high breakdown point estimators
   - Deleting some observations
   - Frustrations and rebirths
   - Depressing the influence of some observations

One really devilish snag - ceteris paribus

Holding other factors fixed - Ceteris paribus:

Jeffrey Wooldridge discusses an employment of regression model as a tool for emulating the CETERIS PARIBUS, as follows:
Introductory Econometrics. A Modern Approach.
MIT Press, Cambridge, Massachusetts, second edition 2009

$$y_i = \beta_0^0 + \beta_1^0 \cdot x_{i1} + \beta_2^0 \cdot x_{i2} + ... + \beta_p^0 \cdot x_{ip} + u_i \quad i = 1, 2, ..., n$$

$$\Delta \hat{y} = \hat{\beta}_0^0 + \hat{\beta}_1^0 \cdot \Delta x_{i1} + \hat{\beta}_2^0 \cdot \Delta x_{i2} + ... + \hat{\beta}_p^0 \cdot \Delta x_{ip} + u_i \quad i = 1, 2, ..., n$$

usually $\Delta x_{i1} = 1$.

Warning - THIS INFORMATION IS OF LIMITED RELEVANCE AND IT MAY BE EVEN TOTALLY MISLEADING!

Misunderstanding the basic ideas can yield catastrophic conclusions

A regression for a “club of good health”

$$\begin{aligned} \mathit{Time\ Total} &= -3.62 + 1.27 \cdot \mathit{Weight} + 0.53 \cdot \mathit{Puls} \\ &\quad -0.51 \cdot \mathit{Strength} + 3.90 \cdot \mathit{Time\ per\ quarter\ of\ mile} + u_i. \end{aligned}$$

Then we can (frequently?) meet with a conclusion of type:

As the estimate of regression coefficient for Strength is negative,
the Strength has negative impact on Time Total.

or (even)

Although the coefficient of determination is small,
the polarities of the estimated coefficient corresponds to our ideas.

Naive interpretation can be misleading

As the estimate of regression coefficient for Strength is negative, the Strength has negative impact on Time Total.

This first assertion from the previous slide can be true,

under some circumstances,

but generally we cannot claim anything like that.

Naive interpretation can be misleading

Although the coefficient of determination is small,
the polarities of the estimated coefficient corresponds to our ideas.

This second assertion from the previous slide can have,

again under some circumstances,

a sense - but generally is false.

WHY ? ?

Naive interpretation can be misleading

Let me recall one of your homeworks on Econometrics I:
Find an example of pair of random variables. In this pair:

1. The first r. v. depends deterministically on the second one.
2. The r. v.'s are not correlated.

Let X_2 be a random variable with $\mathcal{L}(X_2) = N(0,1)$.
It will be our second r. v. .

Put $X_1 = -X_2^k$. This will be our first r. v..
Evidently - due to symmetry - $\mathbb{E}[X_1 \cdot X_2] = 0$. Then

$$cov(X1, X2) = \mathbb{E}(X_1 \dot{X}_2) - \mathbb{E}X_1 \cdot \mathbb{E}X_2 = 0$$

as $\mathbb{E}X_2 = 0$, i. e. X_1 and X_2 are not correlated.

Naive interpretation can be misleading

Consider regression model

$$Y = 1 + X_1 + X_2 + \varepsilon$$

for $k = 20$ (say), i. e. $X_1 = -X_2^{40}$. Then:

If $X_2 >> 1$, an increase of it X_2
yields a decrease of Y (despite of positive sign of X_2).

The First Estimator with 50% Breakdown Point

Repeated medians

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

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

(requiring approx. 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

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

(implementation will be discussed later).

Many advantages - maily

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 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

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

(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 $h = [\frac{n}{2}] + [\frac{p+1}{2}]$

   (Please, remember the optimal value of h).

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

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

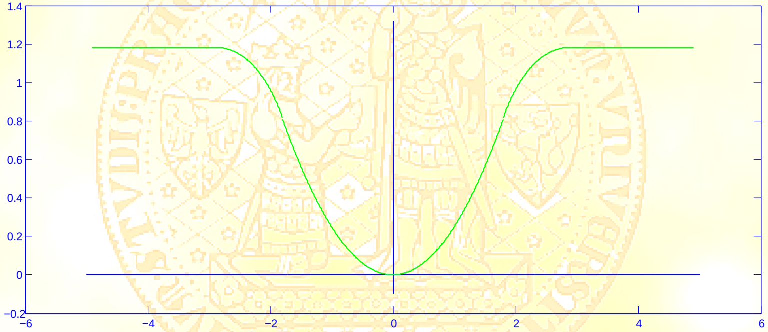
where $b = E\rho\left(\frac{e_i}{\sigma_0}\right)$ with $\sigma_0^2 = Ee_1^2$ (for ρ see next slide).

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

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



Finally - the victory

We have evidently reached something which is

"BOMB und IDIOTEN SICHER".

But maybe that it was only an illusion !!

It appeared that there is an "inborn" disadvantage
which is common to all robust estimator with high breakdown point.

A shock and frustration - Engine Knock Data

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

Engine Knock Data $(n = 16, p = 4, h = 11)$

c	x_1	x_2	x_3	x_4	y
1	13.3	13.9	31	697	84.4
2	13.3	$\boxed{14.1}$	30	697	84.1
3	13.4	15.2 15.1	32	700	88.4
4	12.7	13.8	31	669	84.2
16	12.7	15.9	37	696	93.1

In fact they worked with two data sets.

Let's call these data “Correct”.

Let's call these data “Damaged”.

x_1 is spark timing x_2 air/fuel ratio
x_3 intake temperature x_4 exhaust temperature
y engine knock number

A shock and frustration

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

Engine Knock Data $(n = 16, p = 4, h = 11)$

The values of $\hat{\beta}^{(LMS,n,h)}$ by "elemental" algorithm !
(still included in some packages - see the next slide)

Data	Interc.	SPARK	AIR	INTK	EXHS.
Correct data ($x_{22} = 14.1$)	30.08	0.21	2.90	0.56	-0.01
Damaged data($x_{22} = 15.1$)	-86.50	4.59	1.21	1.47	0.07

x_1 is spark timing x_2 air/fuel ratio
x_3 intake temperature x_4 exhaust temperature
y engine knock number

What was the algorithm of computing the estimate?

1. Select randomly an elemental set of p points and fit a regression plane to them.
2. Compute all squared residuals and find the h-th smallest.
3. Repeat it “10 000” times and select that model (among these “10 000”) with smallest h-th squared residual.

An improvement of the algorithm - a geometric characterization

Joss, J., A. Marazzi (1990):
Probabilistic algorithms for LMS regression.
Computational Statistics & Data Analysis 9, 123-134.

The geometric characterization
of exact solution of LMS extremal problem:

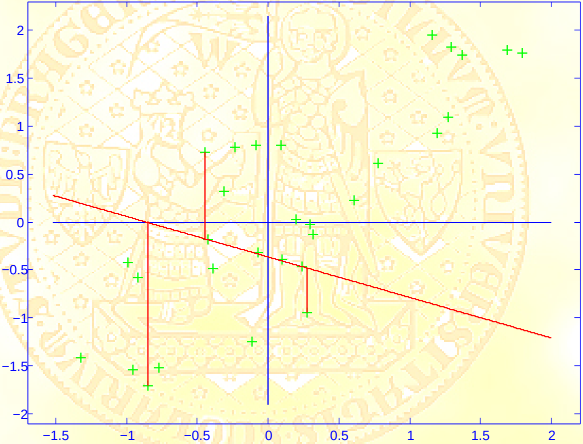
The exact solution has at least
$p + 1$ residuals of the same (absolute) value.

An improvement of the algorithm - a geometric characterization

1. Select randomly an elemental set of p points
   and fit a regression plane to them.
2. Perform (repeatedly) its shift and rotation
   to decrease the value of the h-th squared residual
   and to reach the geometric representation.
3. Repeat it “10 000” times
   and select that model (among these “10 000”)
   with smallest h-th squared residual.

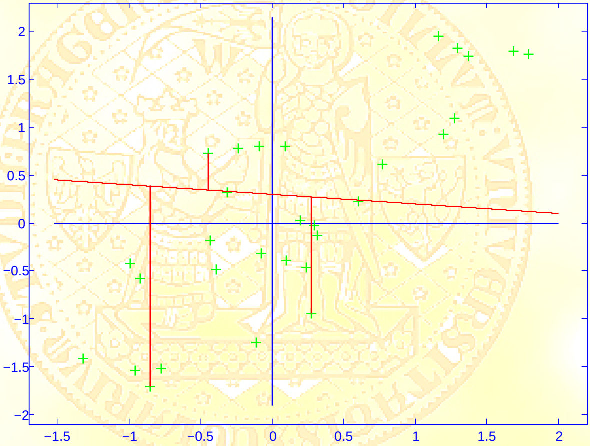
A geometric characterization

Unlucky selection of starting points



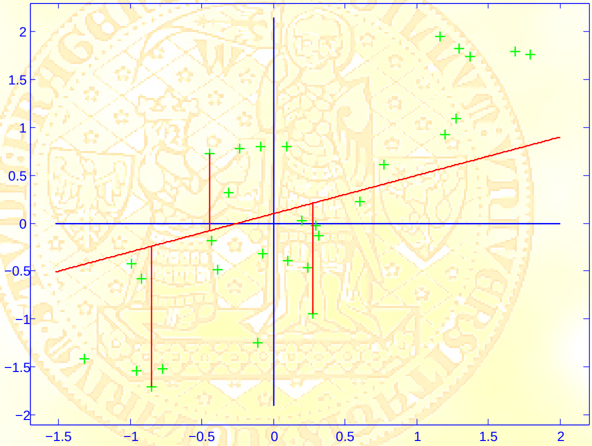
A geometric characterization

Starting shifting and spinning the line



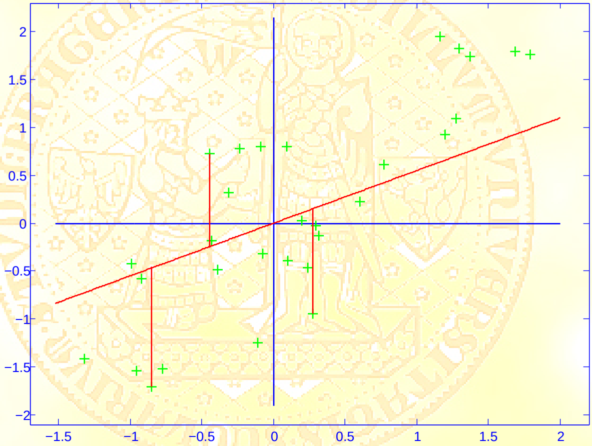
A geometric characterization

Continuing shifting and spinning the line



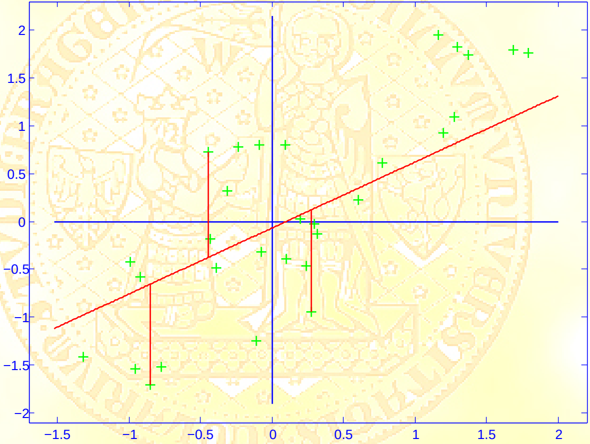
A geometric characterization

Nearly reaching the geometric characterization



A geometric characterization

Reaching the geometric characterization



A substantial improvement of the algorithm

- an employment of simplex method

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

A description is a bit complicated - it requires
to be familiar with a dual form of simplex method.

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

Engine Knock Data $(n = 16, p = 4, h = 11)$

c	x_1	x_2	x_3	x_4	y

The value of $\hat{\beta}^{(LMS,n,h)}$ by Boček-Lachout algorithm.

Data	Interc.	SPARK	AIR	INTK	EXHS.
Correct data ($x_{22} = 14.1$)	30.04	0.14	3.08	0.46	-0.01
Damaged data($x_{22} = 15.1$)	48.38	-0.73	3.36	0.23	-0.01

The difference between these two models is much lower.
So, the effect announced by H-S was a consequence of the bad algorithm.

x_1 is spark timing x_2 air/fuel ratio
x_3 intake temperature x_4 exhaust temperature
y engine knock number

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

Engine Knock Data $(n = 16, p = 4, h = 11)$

c	x_1	x_2	x_3	x_4	y
16	12.7	13.9	37	696	93.1

The value of $\hat{\beta}^{(LMS,n,h)}$ by Boček-Lachout algorithm.

Data	Interc.	SPARK	AIR	INTK	EXHS.
Correct data ($x_{22} = 14.1$)	30.04	0.14	3.08	0.46	-0.01

BUT THIS CONCLUSION - ALTHOUGH TRUE - WAS MISLEADING. WHY?

x_1 is spark timing x_2 air/fuel ratio
x_3 intake temperature x_4 exhaust temperature

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

Engine Knock Data $(n = 16, p = 4, h = 11)$

c	x_1	x_2	x_3	x_4	y

The value of $\hat{\beta}^{(LMS,n,h)}$ by Boček-Lachout algorithm.

Data	Interc.	SPARK	AIR	INTK	EXHS.
Correct data ($x_{22} = 14.1$)	30.04	0.14	3.08	0.46	-0.01

The correct conclusion is:

THE LARGE DIFFERENCE BETWEEN THE ESTIMATES WAS PARTIALLY DUE TO THE BAD ALGORITHM.

x_1 is spark timing x_2 air/fuel ratio
x_3 intake temperature x_4 exhaust temperature
y engine knock number

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

Engine Knock Data $(n = 16, p = 4, h = 11)$

Realize that $\binom{16}{11} = 4368$, so that we can compute $\hat{\beta}^{(LTS,16,11)}$ exactly, just computing $\hat{\beta}^{(OLS,11)}$ for all subsamples of size 11 and select the "best" one.

This is the exact value of $\hat{\beta}^{(LTS,n,h)}$!

Data	Interc.	SPARK	AIR	INTK	EXHS.
Correct data ($x_{22} = 14.1$)	35.11	-0.028	2.949	0.477	-0.009
Damaged data ($x_{22} = 15.1$)	-88.7	4.72	1.06	1.57	0.068

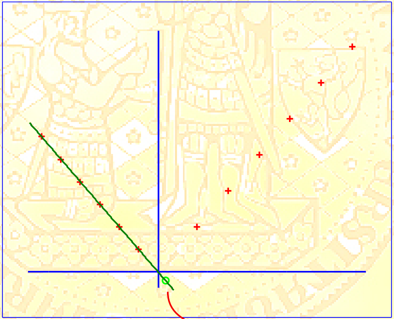
x_1 is spark timing x_2 air/fuel ratio

Víšek, J.Á (1994): A cautionary note on the method of Least Median of Squares reconsidered.
Transactions of the Twelfth Prague Conference 1994, 254 - 259.

An (academic) explanation by a shift of “inlier”

SENSITIVITY OF ANY HIGH-BREAKDOWN-POINT ESTIMATOR TO A SMALL CHANGE OF DATA

Model for the majority of data

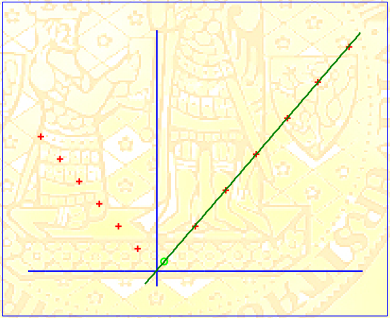


We are going to shift up this point “ o ”.

An (academic) explanation by a shift of "inlier"

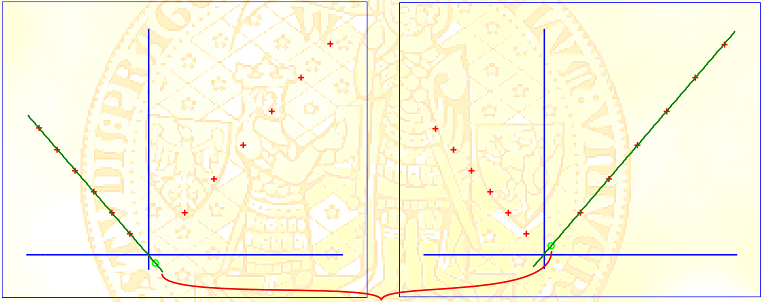
SENSITIVITY OF ANY HIGH-BREAKDOWN-POINT ESTIMATOR TO A SMALL CHANGE OF DATA

Again model for the majority of data



An (academic) explanation by a shift of "inlier"

In both cases the model is for the majority of data



Notice: *The closer* the point (" o ") is to the *y*-axe,

the smaller shift causes the "switch" of the model.

Final conclusion - frustration

We have built up the theory on the sand not on a solid rock base.

Is it really so ?

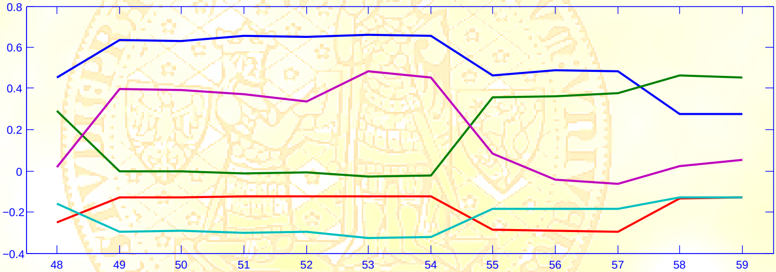
ANALYSIS OF THE EXPORT FROM THE CZECH REPUBLIC TO EU IN 1994

Number of industries 91

X_ℓ	-	*export from i-th industry,*
US_ℓ	-	*number of university-passed employees in the i-th industry,*
HS_ℓ	-	*nuber of high school-passed employees in the i-th industry,*
VA_ℓ	-	*value added in the i-th industry,*
K_ℓ	-	*capital in the i-th industry,*
CR_ℓ	-	*percentage of market occupied by 3 largest producers,*
$TFPW_\ell$	-	*by wages normed productvity in the i-th industry,*
Bal_ℓ	-	*Balasa index in the i-th industry„*
DP_ℓ	-	*cost discontinuity in 1993 in the i-th industry*
	-	*etc., about 20 explanatory variables*

NO REASONABLE MODEL BY OLS - COEFFICIENT OF DETERMINATION 0.28

ANALYSIS OF THE EXPORT FROM THE CZECH REPUBLIC TO EU IN 1994 BY MEANS OF THE *least trimmed squares*.



The development of the estimates of regression coefficients. The blue line represents $\hat{\beta}_1^{(LTS,n,h)}$ (down-scaledby $\frac{1}{10}$), the purple is $\hat{\beta}_8^{(LTS,n,h)}$, the green is $\hat{\beta}_3^{(LTS,n,h)}$, the red is $\hat{\beta}_4^{(LTS,n,h)}$ and light blue (the lowest curve) is $\hat{\beta}_6^{(LTS,n,h)}$ (down-scaled again by $\frac{1}{10}$). There is an evidnet break at 54.

ANALYSIS OF THE EXPORT FROM THE CZECH REPUBLIC TO EU IN 1994 BY MEANS OF THE *least trimmed squares*

has found:

MAIN SUBGROUP

with number of industries 54 and model

$$\frac{X_\ell}{S_\ell} = 4.64 - 0.032 \cdot \frac{US_\ell}{VA_\ell} - 0.022 \cdot \frac{HS_\ell}{VA_\ell} - 0.124 \cdot \frac{K_\ell}{VA_\ell} + 1.035 \cdot CR_\ell$$
$$-3.199 \cdot TFPW_\ell + 1.048 \cdot BAL_\ell + 0.452 \cdot DP_\ell + \varepsilon_\ell$$

X_ℓ	-	*export from i-th industry,*
US_ℓ	-	*number of university-passed employees in the i-th industry,*
HS_ℓ	-	*nuber of high school-passed employees in the i-th industry,*
VA_ℓ	-	*value added in the i-th industry,*
K_ℓ	-	*capital in the i-th industry,*
CR_ℓ	-	*percentage of market occupied by 3 largest producers,*
$TFPW_\ell$	-	*by wages normed productvity in the i-th industry,*
Bal_ℓ	-	*Balasa index in the i-th industry„*
DP_ℓ	-	*cost discontinuity in 1993 in the i-th industry*

with coefficient of determination 0.97 and stable submodels

ANALYSIS OF THE EXPORT FROM THE CZECH REPUBLIC TO EU IN 1994 BY MEANS OF THE *least trimmed squares*

has found:

COMPLEMENTARY SUBGROUP

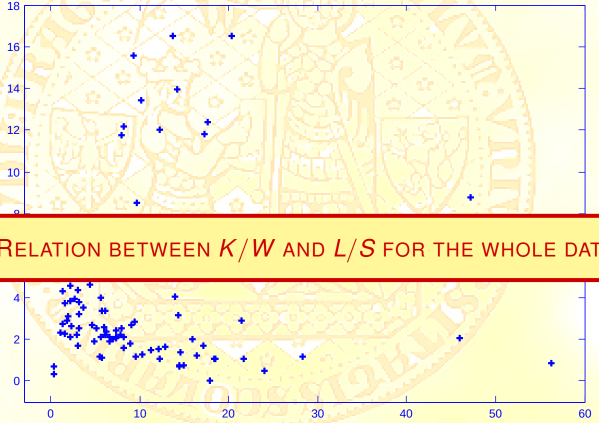
with number of industries 33 and model

$$\frac{X_\ell}{S_\ell} = -0.634 + 0.089 \cdot \frac{US_\ell}{VA_\ell} + 0.235 \cdot \frac{HS_\ell}{VA_\ell} + 0.249 \cdot \frac{K_\ell}{VA_\ell} + 1.174 \cdot CR_\ell$$
$$+0.690 \cdot TFPW_\ell + 2.691 \cdot BAL_\ell - 0.051 \cdot DP_\ell + \varepsilon_\ell$$

X_ℓ	-	*export from i-th industry,*
US_ℓ	-	*number of university-passed employees in the i-th industry,*
HS_ℓ	-	*nuber of high school-passed employees in the i-th industry,*
VA_ℓ	-	*value added in the i-th industry,*
K_ℓ	-	*capital in the i-th industry,*
CR_ℓ	-	*percentage of market occupied by 3 largest producers,*
$TFPW_\ell$	-	*by wages normed productvity in the i-th industry,*
Bal_ℓ	-	*Balasa index in the i-th industry„*
DP_ℓ	-	*cost discontinuity in 1993 in the i-th industry*

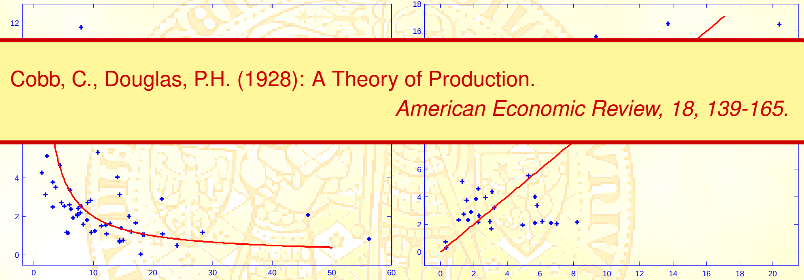
with coefficient of determination 0.93 and stable submodels

ANALYSIS OF THE EXPORT FROM THE CZECH REPUBLIC TO EU IN 1994 BY MEANS OF THE *least trimmed squares*.



RELATION BETWEEN K/W AND L/S FOR THE WHOLE DATA.

ANALYSIS OF THE EXPORT FROM THE CZECH REPUBLIC TO EU IN 1994 BY MEANS OF THE *least trimmed squares*.



Cobb, C., Douglas, P.H. (1928): A Theory of Production.
American Economic Review, 18, 139-165.

RELATION BETWEEN K/W AND L/S FOR THE *Main subpopulation* (LEFT PICTURE)
AND FOR THE *Complementary subpopulation* (RIGHT PICTURE).

Finally - all after - the victory

We haven't reached something which is

"BOMB und IDIOTEN SICHER"

but which is the powerful tool, if it is used with a care.

The least weighted squares

Residuals $\forall \beta \in R \rightarrow r_i(\beta) = Y_i - X_i'\beta$
Order statistics of squared residuals, i. e.

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

Definition

Let $w(u): [0,1] \rightarrow [0,1], w(0) = 1$, (nonincreasing). Then

$$\hat{\beta}^{(LWS,n,w)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r^2_{(i)}(\beta)$$

will be called *the least weighted squares (LWS)*.

Notice the order of words - there is also the Weigted Least Squares

Víšek, J. Á. (2000): Regression with high breakdown point.
Robust 2000 (eds. Antoch, J. Dohnal, G.), 324 - 356.

Let's realize what the definition really asks for.

Definition

Let $w(u): [0,1] \rightarrow [0,1], w(0) = 1$, (nonincreasing). Then

$$\hat{\beta}^{(LWS,n,w)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r_{(i)}^2(\beta)$$

will be called *the least weighted squares (LWS).*

Notice:
The smallest residual obtains the largest weight

and vice versa

the largest residual obtains the smallest weight.

Does LWS exist at all ?

Is there always - for fixed $n \in N$ -

a solution of the extremal problem

$$\hat{\beta}^{(LWS,n,w)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r_{(i)}^2(\beta) \quad ?$$

To be able to answer it, let's make some preparatory steps.

An excursion to the history

First of all, let's recall that there is the classical:

The weighted least squares

Definition

Let $w_i \in [0,1], i = 1,2,...,n$ be weights and $W = \mathrm{diag}(w_1, w_2, ..., w_n)$ a diagonal matrix. Then

$$\hat{\beta}^{(WLS,n,w)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} w_i r_i^2(\beta) = \left(X'WX\right)^{-1} X'WY$$

will be called *the Weighted Least Squares (WLS)*.

Notice the order of words - it hints that

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

An excursion to the history

Notice also that we have the formula for the estimator -
- hence it can be easy implemented and computed.

$$\hat{\beta}^{(WLS,n,w)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} w_i r_i^2(\beta) = \left(X'WX\right)^{-1} X'WY.$$

How did we find the formula $\hat{\beta}^{(WLS,n,w)} = \left(X'WX\right)^{-1} X'WY$?

An excursion to the history

We can rewrite the definition

$$\hat{\beta}^{(WLS,n,w)} = \underset{\beta \in R^p}{\arg\min} \; \sum_{i=1}^{n} w_i r_i^2(\beta) = \sum_{i=1}^{n} \left(\sqrt{w}_i r_i(\beta)\right)^2.$$

Hence, considering transformed variables

$$\tilde{Y}_i = \sqrt{w}_i \cdot Y_i \qquad \text{and} \qquad \tilde{X}_i = \sqrt{w}_i \cdot X_i,$$

we have

$$\left(\tilde{Y}_i - \tilde{X}_i'\beta\right)^2 = (\sqrt{w}_i \cdot Y_i - \sqrt{w}_i \cdot X_i\beta)^2 \left[\sqrt{w}_i\,(Y_i - X_i\beta)\right]^2 = w_i r_i^2(\beta),$$

i. e. we look for
$$\min_{\beta \in R^p} \sum_{i=1}^{n} w_i r_i^2(\beta) = \min_{\beta \in R^p} \sum_{i=1}^{n} \left(\tilde{Y}_i - \tilde{X}_i'\beta\right)^2.$$

Finally, $\underline{\hat{\beta}^{(WLS,n,w)}(Y, X) = \hat{\beta}^{(OLS,n)}(\tilde{Y}, \tilde{X})}$.

Recalling the background of our model

Remenber that we consider the regression model

$$Y_i = X_i' \beta^0 + \varepsilon_i \quad \text{for} \quad i = 1, 2, ...$$

where $(X_i, \varepsilon_i)_{i=1}^{\infty}$ is an i.i.d. sequence of r.v.'s,

so X_i as well as ε_i (and hence also Y_i) are measurable mappings from the probability space $(\Omega, \mathcal{A}, P)$ to the real line (for simplicity).

So, if being consequential (důsledný), we should write

Finally, $\underline{\hat{\beta}^{(WLS,n,w)}(Y(\omega), X(\omega)) = \hat{\beta}^{(OLS,n)}(\tilde{Y}(\omega), \tilde{X}(\omega))}$.

We will need it a bit later.

Continuing preparatory considerations

1. Denote by Π the set of all permutations of numbers $\{1, 2, ..., n\}$ and its elements by $\boldsymbol{\pi}$, i. e. $\boldsymbol{\pi} \in \Pi \to \boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_n), \pi_i \in \{1, 2, ..., n\}$ and $\pi_i \neq \pi_j$ for $i \neq j$.
2. Put for any $\beta \in R^p$, $\boldsymbol{\pi} \in \Pi$ and $\omega \in \Omega$
$$S_n(\beta, \boldsymbol{\pi}, \omega) = \sum_{i=1}^n w_i r^2_{\pi_i}(\beta, \omega).$$
3. Fix $\omega \in \Omega$ and define for any $\boldsymbol{\pi} \in \Pi$
$$\hat{\beta}^{(WLS,n,w,\boldsymbol{\pi})}(\omega) = \arg\min_{\beta \in R^p} S_n(\beta, \boldsymbol{\pi}, \omega).$$
4. Realize that $S_n(\hat{\beta}^{(WLS,n,w,\boldsymbol{\pi})}, \boldsymbol{\pi}, \omega) = \min_{\beta \in R^p} S_n(\beta, \boldsymbol{\pi}, \omega)$. (1)
5. Finally, find $\hat{\boldsymbol{\pi}}$ so that
$$S_n(\hat{\beta}^{(WLS,n,w,\hat{\boldsymbol{\pi}})}, \hat{\boldsymbol{\pi}}, \omega) = \min_{\boldsymbol{\pi} \in \Pi} S_n(\hat{\beta}^{(WLS,n,w,\boldsymbol{\pi})}, \boldsymbol{\pi}, \omega).$$
6. Then (1) implies
$$S_n(\hat{\beta}^{(WLS,n,w,\hat{\boldsymbol{\pi}})}, \hat{\boldsymbol{\pi}}, \omega) = \min_{\boldsymbol{\pi} \in \Pi} \min_{\beta \in R^p} S_n(\beta, \boldsymbol{\pi}, \omega).$$

Continuing preparatory considerations

We have from previous slide

$$S_n(\hat{\beta}^{(WLS,n,w,\hat{\boldsymbol{\pi}})}, \hat{\boldsymbol{\pi}}, \omega) = \min_{\boldsymbol{\pi}\in\Pi} \min_{\beta\in R^p} S_n(\beta, \boldsymbol{\pi}, \omega)$$

and let's change the order of minimization

$$S_n(\hat{\beta}^{(WLS,n,w,\hat{\boldsymbol{\pi}})}, \hat{\boldsymbol{\pi}}, \omega) = \min_{\beta\in R^p} \{\min_{\boldsymbol{\pi}\in\Pi} S_n(\beta, \boldsymbol{\pi}, \omega)\}.$$

Finally, keep in mind that

$$S_n(\hat{\beta}^{(WLS,n,w,\hat{\boldsymbol{\pi}})}, \hat{\boldsymbol{\pi}}, \omega) = \min_{\beta\in R^p} \left\{ \min_{\boldsymbol{\pi}\in\Pi} \sum_{i=1}^{n} w_i r_{\pi_i}^2(\beta, \omega) \right\}$$

(we will need it at the end of our considerations).

Continuing preparatory considerations

Now fix $\beta \in R^p$, put $w_i = w\left(\frac{i-1}{n}\right)$ and realize that

$$\begin{array}{lccccccccc} & w_1 & \geq & w_2 & \geq & \ldots & \geq & w_n & \\ \text{and} & & & & & & & & \\ & r^2_{(1)}(\beta) & \leq & r^2_{(2)}(\beta) & \leq & \ldots & \leq & r^2_{(n)}(\beta). & \quad (2) \end{array}$$

It implies that for any $\beta \in R^p$

$$\min_{\pi \in \Pi} \sum_{i=1}^{n} w_i r^2_{\pi_i}(\beta) = \sum_{i=1}^{n} w_i r^2_{(i)}(\beta). \qquad (3)$$

Assume that not, i.e. the minimum

$$\min_{\pi \in \Pi} \sum_{i=1}^{n} w_i r^2_{\pi_i}(\beta)$$

is realized for another permutation of observations than given by (2)
and try to show a contradiction - which would prove (3)!!

Continuing preparatory considerations

An exhibition of contradiction: If the assertion doesn't hold, there is a pair $1 \leq i \leq j \leq n$ such that

$$w_i \cdot r^2_{(i)}(\beta) + w_j \cdot r^2_{(j)}(\beta) > w_i \cdot r^2_{(j)}(\beta) + w_j \cdot r^2_{(i)}(\beta) \tag{4}$$

and

$$w_j = w_i - a, \quad a \geq 0 \quad \text{and} \quad r^2_{(j)}(\beta) = r^2_{(i)}(\beta) + b, \quad b \geq 0. \tag{5}$$

Plugging the values (5) into (4)

$$w_i \cdot r^2_{(i)}(\beta) + (w_i - a) \cdot (r2_{(i)}(\beta) + b) > w_i \cdot (r^2_{(i)}(\beta) + b) + (w_i - a) \cdot r^2_{(i)}(\beta),$$

gives evidently

$$2 \cdot w_i \cdot r^2_{(i)}(\beta) + w_i \cdot b - a \cdot r^2_{(i)}(\beta) - a \cdot b > 2 \cdot w_i \cdot r^2_{(i)}(\beta) + w_i \cdot b - a \cdot r^2_{(i)}(\beta)$$

a contradiction (except of $a = 0$ and $b = 0$).

Final conclusion

We had (we have proved it several slides backwards)

$$S_n(\hat{\beta}^{(WLS,n,w,\hat{\pi})}, \hat{\pi}) = \min_{\beta \in R^p} \left\{ \min_{\pi \in \Pi} \sum_{i=1}^{n} w_i r^2_{\pi_i}(\beta) \right\}$$

and we have just proved that for any $\beta \in R^p$

n n

Realize that $\hat{\pi} = \hat{\pi}(\omega)$.

Plugging the second row to the first one

$$S_n(\hat{\beta}^{(WLS,n,w,\hat{\pi})}, \hat{\pi}) = \min_{\beta \in R^p} \sum_{i=1}^{n} w_i r^2_{(i)}(\beta).$$

It implies

$$\hat{\beta}^{(WLS,n,w,\hat{\pi})} = \hat{\beta}^{(LWS,n,w)}.$$

Q.E.D.

A technical trick - Jaroslav Hájek, rank tests

Ranks $\rho(\beta, j)$ of the squared residuals: We put

$$\boxed{\rho(\beta, j) = i \in \{1, 2, ..., n\} \quad \text{if} \quad r_j^2(\beta) = r_{(i)}^2(\beta)}$$

Firstly - read what it means, secondly - explain what it allows.

$$\hat{\beta}^{(LWS,n,w)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r_{(i)}^2(\beta)$$

$$= \underset{\beta \in R^p}{\arg\min} \sum_{j=1}^{n} w\left(\frac{\rho(\beta, j) - 1}{n}\right) r_j^2(\beta)$$

The last form of definition says:
The method decides itself which weight is assigned to which residual -
- we can speak about an implicit weighting.

Let's return once again to the question: Does LWS exist at all ?

There is always (for fixed $n \in N$) a solution of the extremal problem

$$\hat{\beta}^{(LWS,n,w)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r_{(i)}^2(\beta)$$
$$= \arg\min_{\beta \in R^p} \sum_{j=1}^{n} w\left(\frac{\rho(\beta,j)-1}{n}\right) r_j^2(\beta).$$

Notice that when we want to find LWS, we are looking for

the WLS with weights $w\left(\frac{\rho(\beta,1)-1}{n}\right), w\left(\frac{\rho(\beta,2)-1}{n}\right), \ldots, w\left(\frac{\rho(\beta,n)-1}{n}\right)$.

We will need it a bit later.

An excursion to the history - once again.

Definition of $\hat{\beta}^{(}WLS, n, w)$ in the matrix form -
- it'll give an answer how we have found a formula for it.

Let $w_i \in [0,1], i = 1, 2, ..., n$ be weights
and $W = \mathrm{diag}(w_1, w_2, ..., w_n)$ a diagonal matrix. Then

$$\hat{\beta}^{(WLS,n,w)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{n} w_i r_i^2(\beta) = (Y - X\beta)' \, W \, (Y - X\beta) \qquad (6)$$

will be called *the Weighted Least Squares (WLS)*.

Taking derivative of (6) (with respect to β),
we obtain normal equations:

$$X'W(Y - \beta) = 0.$$

Performing the multiplication $\rightarrow$ $X'WY = X'WX\beta$ and hence:

$$\hat{\beta}^{(WLS,n,w)} = (X'WX)^{-1} \, X'WY.$$

Remember the third row from bottom.

An excursion to the history - continued

So, we have the normal equations for WLS:

$$X'W(Y-\beta)=0.$$

They can be written in the vector form as:

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

Let us take it into account

together with the last conclusion on one from the previous slides -

- see the next slide.

Employing the results of several previous slides:

1. On some of the previous slides we had:

   Notice that when we want to find LWS, we are looking for

   the WLS with weights $w\left(\frac{\rho(\beta,1)-1}{n}\right), w\left(\frac{\rho(\beta,2)-1}{n}\right), \ldots, w\left(\frac{\rho(\beta,n)-1}{n}\right)$.

2. On the last slide we had:

   The normal equations for the WLS

   can be written in the vector form:

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

3. Hence:

   The estimator $\hat{\beta}^{(LWS,n,w)}$ is one of

   the solutions of the normal equations

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

Have we helped to ourselves?

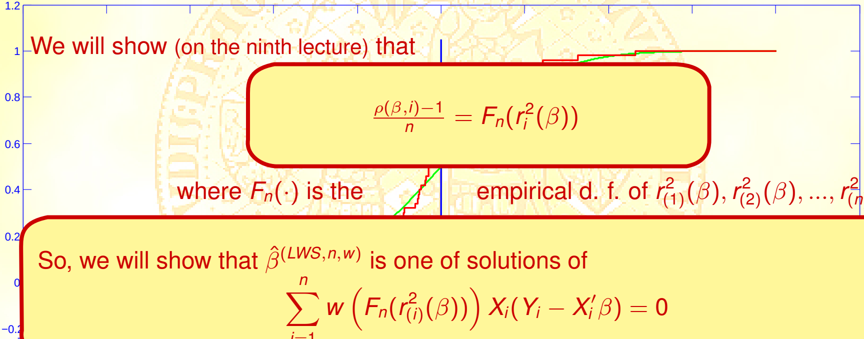
Have we obtained something which can be useful
for establishing the properties of $\hat{\beta}^{(LWS,n,w)}$?

It seems that no but the correct answer is: YES !!

Let's realize what is the rank of the squared residual $\rho(\beta, i)$.

For a hint see the next slide !

On the second lecture we saw the graph of empirical d.f.:



We will show (on the ninth lecture) that

$$\frac{\rho(\beta,i)-1}{n} = F_n(r_i^2(\beta))$$

where $F_n(\cdot)$ is the empirical d. f. of $r_{(1)}^2(\beta), r_{(2)}^2(\beta), ..., r_{(n)}^2(\beta)$.

So, we will show that $\hat{\beta}^{(LWS,n,w)}$ is one of solutions of

$$\sum_{i=1}^{n} w\left(F_n(r_{(i)}^2(\beta))\right) X_i(Y_i - X_i'\beta) = 0$$

and employing the fact that e. d. f. converges to the underlying d. f., we can ...

THANKS FOR ATTENTION