

Introduction to Robust Statistics

Anthony Atkinson, London School of Economics, UK
Marco Riani, Univ. of Parma, Italy

Introduction to robust statistics

- Outliers are observations that are surprising in relation to the majority of the data:
- May be wrong - data gathering or recording errors - transcription? electronic if not manual
- May be correct and informative for example about departures from the assumed model. Ex. identifiable subsets in medical studies
- Should always be checked

Outline

- Simple sample
 - Introduction to theoretical concepts
 - M, S, MM, Tau estimators
- Regression
 - Transformations
 - Model choice
- Multivariate Analysis
- Clustering

Literature

- Hampel F.R., Ronchetti E.M., Rousseeuw P.J and Stahel W.A. (1986), Robust Statistics the Approach based on the Influence Function, Wiley, NY.
- Rousseeuw P.J. and Leroy A. (1987), Robust Regression and Outlier Detection, Wiley, NY
- Maronna R.A., Martin R.D. and Yohai V. (2006), Robust Statistics, Theory and Methods, Wiley, NY
- Atkinson A.C. and Riani M. (2000), Robust Diagnostic Regression Analysis, Springer NY.
- Huber, P. J. and Ronchetti, E. M. (2009). Robust Statistics, Second Edition. Wiley, New York

History of Robust statistics

- Awareness of the importance of immunizing against outliers / gross errors is as old as the experimental approach to science
- **Thucydides** (History of The Peloponnesian War): “in 428 B.C. the Plataeans, besieged by the Spartans, **excluded extreme measurements** when estimating the height of the walls and managed to break the siege”

History of Robust statistics

- **Legendre**: “if among these errors are some which appear to be too large to be admissible, then [...] will be rejected ”.
- **Edgeworth**: “the method of LS is seen to be our best course when we have thrown overboard a certain portion of our data ”.

Why it is not sufficient to screen data and remove outliers

- Users, even expert statisticians, do not always screen the data.
- It can be difficult or impossible to spot outliers in multivariate or highly structured data. Becoming increasingly difficult with “Yet Bigger Data”
- Rejecting outliers affects distributions - variances can be underestimated if data are ‘cleaned’. We would like procedures with defined statistical properties, such as size of tests. Machine learning?

Notation

Observations y_1, \dots, y_n
belonging to some sample space Y

$y_{(1)}, \dots, y_{(n)}$ = order statistics

A parametric model F_θ on the sample space
 $\theta \in R^p$

$T_n = T(y_1, \dots, y_n)$ = estimator based on n observations

Classical and robust theory

Classical

- The observations are distributed according to F_θ
- Example
- $F_\theta = N(\mu, \sigma^2)$
- $\theta = (\mu, \sigma^2)$

Robust

- F_θ is considered as a mathematical abstraction which is only an ideal approximation to reality. The goal is to produce statistical procedures which still behave fairly well under deviations from the assumed model

The grand plan

- Andrews *et al.* (1972) (the Princeton Robustness Study), at which time it was expected that all statistical analyses would, by default, be robust
- **“any author of an applied article who did *not* use the robust alternative would be asked by the referee for an explanation”.**

Prediction from 1972 Princeton study

- “From the 1970s to 2000 we would see ... extensions to linear models, time series, and multivariate models, and widespread adoption to the point where every statistical package would take the robust method as the default ...”

Importance of Robust Statistics

- A tremendous growth for about two decades from 1964
- However still not routinely used in practical data analysis and standard software
- As we shall see, many sets of data contain numerous outliers so robustness is a crucial aspect
- Recent developments are easy to apply and interpret

Properties of estimators

- Consistency
- Equivariance
- Sensitivity curve
- Breakdown point
- Efficiency
- Influence function
- MaxBias
- Gross error sensitivity

Properties of estimators

$\hat{T}_\infty(G)$ = value of the estimator when
 $n \rightarrow \infty$ and data are from G

- Consistency: the results become more and more precise when the number of observations increases
- Fisher consistency: at the model the estimator is equal to the parameter or $\hat{T}_\infty(F_\theta) = \theta$

Location equivariant estimators

- A location estimator T should be equivariant in the sense when a constant is added to the data (location) and when they are multiplied by a constant (scale), you get:

$$T(cy_1 + d, \dots, cy_n + d) = cT(y_1, \dots, y_n) + d$$

Example of location estimators

- Mean (\bar{y})
 - Median (Me)
 - α - trimmed mean (trim a proportion α from both ends of the data set and then take the mean), (\bar{y}_α)
 - α - Winsorized mean: replace a proportion α from both ends of the data set by the next closest observation and then take the mean.
-
- Example: 2, 4, 5, 10, 200
 - Mean = 44.2 $Me = 5$
 - 20% trimmed mean = $(4 + 5 + 10) / 3 = 6.33$
 - 20% Winsorized mean = $(4 + 4 + 5 + 10 + 10) / 5 = 6.6$

L-statistics

- **REMARK:** mean, α -trimmed mean and α - Winsorized mean, median are particular cases of L-statistics
- L-statistics: linear combination of order statistics. For example

$$T(y_1, \dots, y_n) = \sum_{i=1}^n a_i Y_{(i)} \quad a_{i,n} = \frac{1}{n} \Rightarrow \bar{y}$$

$$a_{i,n} = \begin{cases} \text{for } n \text{ odd} \\ 1 & i = \frac{n+1}{2} \\ 0 & i \neq \frac{n+1}{2} \end{cases} \quad a_{i,n} = \begin{cases} \text{for } n \text{ even} \\ \frac{1}{2} & i = \frac{n}{2}, \frac{n}{2} + 1 \\ 0 & \text{otherwise} \end{cases}$$

$\Rightarrow Me$

Scale equivariant estimators

- A scale estimator S should be equivariant, in the sense that

$$S(cy_1 + d, \dots, cy_n + d) = |c|S(y_1, \dots, y_n)$$

- **Remark:** the absolute value is needed because a scale estimate is always positive

Examples of scale estimators

- Standard deviation
- Interquartile range

$$IQR_n = y_{(n-[n/4]+1)} - y_{([n/4])}$$

At $F_\sigma = N(0, \sigma^2)$, $IQR(F_\sigma) = 2\Phi^{-1}(0.75)\sigma \neq \sigma$

$$IQRN_n = \frac{1}{2\Phi^{-1}(0.75)} \{y_{(n-[n/4]+1)} - y_{([n/4])}\}$$

$1/2\Phi^{-1}(0.75) = 0.7413 = \text{consistency factor}$

Examples of scale estimators

- Median Absolute deviation (*MAD*)

$$MAD_n = Me(|y_i - Me(Y_n)|)$$

Normalized version

$$MADN_n = \frac{1}{\Phi^{-1}(0.75)} MAD_n = 1.4826 \times MAD_n$$

Example

- Location scale model $y_i \sim N(\mu, \sigma^2)$
- Data $Y_{10} = \{y_1, \dots, y_{10}\}$ are the natural logs of the annual incomes of 10 people.
- 9.52 9.68 10.16 9.96 10.08
- 9.99 10.47 9.91 9.92 15.21
- Remark: the income of person 10 is much larger than the other values.

Classical versus robust estimators

	The 9 regular observations	All 10 observations
\bar{y}	9.965	10.49
Me	9.960	9.975
$\bar{y}_{0.10}$	10.021	9.966
SD	0.27	1.68
IQRN	0.13	0.17

- Classical estimators are highly influenced by the outlier
- Robust estimate computed from all observations is comparable with the classical estimate applied to non-outlying data
- How to compare robust estimators?

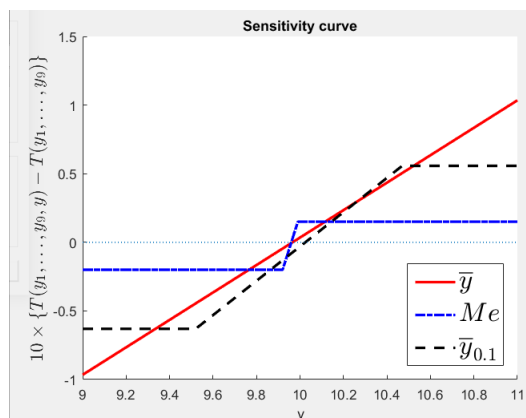
(Standardized) sensitivity curve (SC)

- Measure the effect of a single outlier on the estimator T_n .
- Assume we have $n - 1$ fixed obs. $Y_{n-1} = \{y_1, \dots, y_{n-1}\}$. Add an n -th observation at y , which can be any real number.

$$SC(y, \hat{T}_n, Y_{n-1}) = \{\hat{T}_n(Y_{n-1}, y) - \hat{T}_n(Y_{n-1})\}n$$
- For the arithmetic mean $SC(y, \hat{T}_n, Y_{n-1}) = y - \bar{y}_{n-1}$
- Note that SC depends strongly on the dataset Y_{n-1}

Sensitivity curve (example)

- Annual income data: let Y_9 consist of the 9 'regular' observations



(Finite sample) breakdown point

- Given data set with n obs.
- If replace m of obs. by *any* outliers and estimator stays in a bounded set, but doesn't when we replace $(m+1)$, the breakdown point of the estimator at that data set is m/n .
- breakdown point of the mean = 0

(Finite sample) breakdown point of the median

- n is even

$\underbrace{y(1), y(2), \dots, y_{(n/2-1)}}_{\text{Arbitrarily replaced}}$
 $\underbrace{y_{(n/2)}, y_{(n/2)+1}}_{Me=0.5(y_{(n/2)}+y_{(n/2+1)})}$
 $, \underbrace{y_{(n/2+2)}, \dots, y(n)}_{\text{Arbitrarily replaced}}$

- n is odd

$\underbrace{y(1), y(2), \dots, y_{(\frac{n-1}{2})}}_{\text{Arbitrarily replaced}}$
 $\underbrace{y_{(\frac{n+1}{2})}}_{Me=y_{(\frac{n+1}{2})}}$
 $, \underbrace{y_{(\frac{n+3}{2})}, \dots, y(n)}_{\text{Arbitrarily replaced}}$

$$bdp = \left\lfloor \frac{n-1}{2} \right\rfloor \text{ i.e. } \frac{n}{2} - 1 \text{ or } \frac{n-1}{2}$$

More formally

$$\theta \in \Theta$$

Y_m = the set of all datasets z of size n having $n-m$ elements in common with y

$$Y_m = \{z : \#(z) = n, \#(y \cap z) = n - m\}$$

$bdp(T_n, y)$ is the largest proportion of data points that can be arbitrarily replaced by outliers without the estimator leaving a set which is bounded and also bounded away from the boundary of $\theta \in \Theta$

$$bdp(T_n, y) = \frac{m^*}{n} \quad m^* = \max\{m \geq 0 : T_n \text{ bounded and also bounded away from } \partial\Theta \forall y \in Y_m\}$$

$\partial\Theta$ denotes the boundary of θ

Robust statistics deals with approximate models or model deviations

- We need to define a neighbourhood of the parametric model
- We consider the set of distributions

$$\{G_\epsilon = (1 - \epsilon)F_\theta + \epsilon W\}$$

- W is an arbitrary distribution function
- What happens to bdp when data are generated from $\{G_\epsilon\}$?

$$\hat{T}_\infty(G_\epsilon) = \text{value of the estimator when } n \rightarrow \infty \text{ and data are from } G_\epsilon$$

(Asymptotic) breakdown point (bdp)

- $bdp(T_\infty, F_\theta)$ is the largest $\epsilon^* \in (0, 1)$ such that for $\epsilon < \epsilon^*$, $T(G_\epsilon)$ as a function of W , remains bounded also bounded away from the boundary of Θ .

$$\{G_\epsilon = (1 - \epsilon)F_\theta + \epsilon W\}$$

- In symbols: there exists a bounded and closed set $K \subset \Theta$ such that $K \cap \partial\Theta = \emptyset$ and

$$\hat{T}_\infty(G_\epsilon) = \hat{T}_\infty((1 - \epsilon)F + \epsilon W) \in K \quad \{\forall \epsilon < \epsilon^* \text{ and } \forall W\}$$

(asymptotic) BDP

Location estimators

- $\bar{y}=0$
- median = $1/2$
- α -trimmed mean = α
- α - Winsorized mean = α

Scale estimators

- SD = 0
- IQRN = 0.25
- MADN = 0.5

(Asymptotic) relative efficiency (RE and ARE)

- For a fixed underlying distribution, the relative efficiency (RE) of an estimator \tilde{T}_n relative to that of \hat{T}_n is

$$\text{RE}(\tilde{T}_n; \hat{T}_n) = \frac{\text{variance of } \hat{T}_n}{\text{variance of } \tilde{T}_n}$$

- \hat{T}_n needs only RE times as many observations as \tilde{T}_n for the same variance
- Remark: use MSE for biased estimators
- ARE = limit of RE as $n \rightarrow \infty$

Examples of ARE

- Symmetric distribution μ = population mean = population median
- $\bar{y} \approx N(\mu, \sigma^2/n)$
- $Me \approx N\left(\mu, \frac{1}{n} \frac{1}{4f(\mu)^2}\right)$
- At normal distribution $\text{ARE}(Me; \bar{y}) = 2/\pi \approx 64\%$
- At t_5 $\text{ARE}(Me; \bar{y}) \approx 96\%$.
- At t_4 $\text{ARE}(Me; \bar{y}) \approx 112.5\%$
- At t_3 $\text{ARE}(Me; \bar{y}) \approx 162.5\%$
- At t_1 $\text{ARE}(Me; \bar{y}) = \infty$
- Is t_5 really a better model for the error distribution than the normal?

Robust statistics deals with approximate models or model deviations

- We need to define a neighbourhood of the parametric model

- We consider again the set of distributions

$$\{G_\epsilon = (1 - \epsilon)F_\theta + \epsilon W\}$$

- W is an arbitrary distribution function
- What happens to ARE when data are generated from $\{G_\epsilon\}$?

ARE with F_ϵ

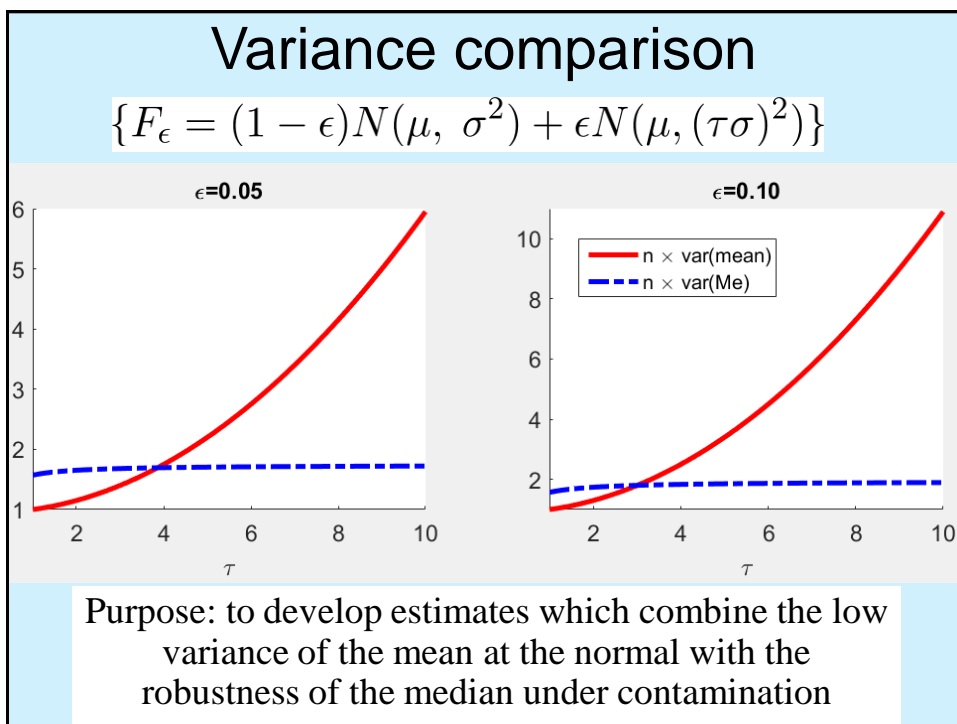
$$\{F_\epsilon = (1 - \epsilon)N(\mu, \sigma^2) + \epsilon N(\mu, (\tau\sigma)^2)\}$$

- i.e. not all measurements are equally precise

$$\text{var}(\bar{y}) = \frac{(1 - \epsilon) + \epsilon\tau^2}{n}$$

$$\text{var}(Me) = \frac{1}{n} \cdot \frac{\pi}{2(1 - \epsilon + \epsilon/\tau)^2}$$

- For $\tau=3$ and $\epsilon > 0.10 \Rightarrow \text{ARE}(Me; \bar{y}) > 1$



Contamination with point mass

$$\{G_\epsilon = (1 - \epsilon)F_\theta + \epsilon W\}$$

- One particular case is when W is the set of point mass distributions where the «point mass» δ_{y_0} is the distribution such that $P(y = y_0) = 1$

Interpretation

F_ϵ generates with probability $(1 - \epsilon)$ data from F_θ
and with probability ϵ data equal to y_0

$$\{F_\epsilon = (1 - \epsilon)F_\theta + \epsilon\delta_{y_0}\}$$

INFLUENCE FUNCTION (Hampel, 1974)

- Describes how the estimator reacts to a small amount of contamination at any point y_0
- Approximation to the relative change in the estimator caused by the addition of a small proportion of spurious observations at y_0 (small fraction ϵ of identical outliers)

$$\lim_{\epsilon \rightarrow 0} \frac{T(F_\epsilon) - T(F_\theta)}{\epsilon}$$

$T(F_\epsilon)$ Estimator in the contaminated model

$T(F_\theta)$ Estimator computed at true value

$\lim_{\epsilon \rightarrow 0}$ Infinitesimal amount of contamination

INFLUENCE FUNCTION (IF)

- Remark: the behaviour is referred to \hat{T}_∞

$$IF_{\hat{T}}(y_0, F_\theta) = \lim_{\epsilon \downarrow 0} \frac{\hat{T}_\infty((1 - \epsilon)F_\theta + \epsilon\delta_{y_0}) - \hat{T}_\infty(F_\theta)}{\epsilon}$$

$\epsilon \downarrow 0$ stands for “limit from the right”

$\hat{T}_\infty((1 - \epsilon)F_\theta + \epsilon\delta_{y_0})$ = asymptotic value of the estimate when the underlying distribution is F_θ and a fraction ϵ of outliers is equal to y_0

IF of \bar{y} , Me , \bar{y}_α at $N(\mu, 1)$

$$IF_{\bar{y}}(y, F_\theta) = y - \mu = \text{unbounded}$$

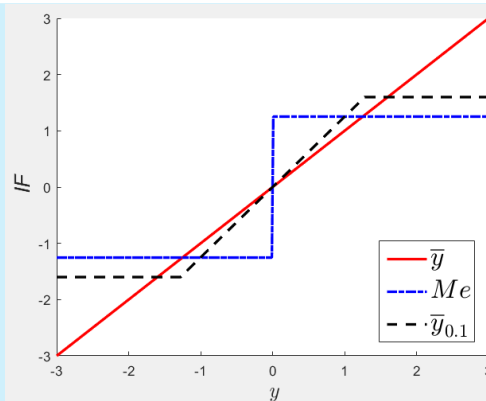
$$IF_{Me}(y, F_\theta) = \frac{\text{sign}(y - \mu)}{2\phi(0)}$$

$$IF_{\bar{y}_\alpha}(y, F_\theta) = \frac{\text{sign}(y - \mu)}{1 - 2\alpha} \min\{|y - \mu|, \Phi^{-1}(1 - \alpha)\}$$

PLOT OF THE INFLUENCE FUNCTION AT $N(0, 1)$ (\bar{y} , Me , \bar{y}_α)

$$IF_{Me}(y, N(0, 1)) = \frac{\sqrt{2\pi} \times \text{sign}(y)}{2}$$

$$IF_{\bar{y}_{0.10}}(y, N(0, 1)) = \frac{\text{sign}(y)}{1 - 2\alpha} \min\{|y - \mu|, 1.28\}$$



SC and IF

- *IF* : small fraction ϵ of identical outliers

$$IF_{\hat{T}}(y_0, F_\theta) = \lim_{\epsilon \downarrow 0} \frac{\hat{T}_\infty((1 - \epsilon)F_\theta + \epsilon\delta_{y_0}) - \hat{T}_\infty(F_\theta)}{\epsilon}$$

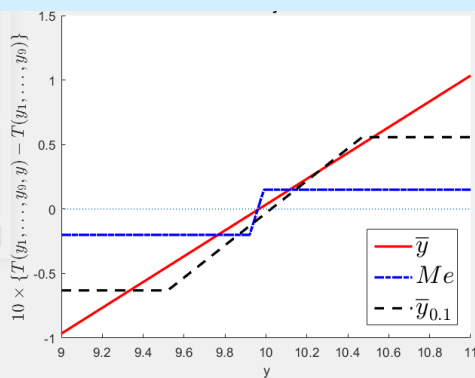
- *SC* : fraction of contamination is $1/n$

$$SC(y_0, \hat{T}_n, Y_{n-1}) = \{\hat{T}_n(Y_{n-1}, y_0) - \hat{T}_n(Y_{n-1})\}n$$

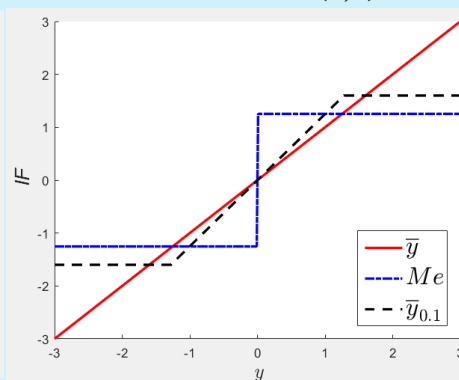
$$SC(y_0, \hat{T}_n, Y_{n-1}) \rightarrow_{a.s.} IF_{\hat{T}}(y_0, F_\theta)$$

SC and IF

Sensitivity curve of the
income data



Influence function at $N(0,1)$



IF and max. asymptotic bias

$$IF_{\hat{T}}(y_0, F_\theta) = \lim_{\epsilon \downarrow 0} \frac{\hat{T}_\infty((1 - \epsilon)F_\theta + \epsilon\delta_{y_0}) - \hat{T}_\infty(F_\theta)}{\epsilon}$$

- If ϵ is small the (asymptotic) bias

$$\hat{T}_\infty((1 - \epsilon)F_\theta + \epsilon\delta_{y_0}) - \hat{T}_\infty(F_\theta)$$

is approximated by

$$\epsilon \times IF_{\hat{T}}(y_0, F)$$

Remark: The *IF* (although it seems to be a particular measure of influence), is sufficient to describe the max. asymptotic bias of an estimator over a neighbourhood of the model because

$$\sup_W \|T(G_\epsilon) - T(F_\theta)\| \approx \epsilon \sup_{y_0} \|IF_{\hat{T}}(y_0, F_\theta)\|$$

$$\{G_\epsilon = (1 - \epsilon)F_\theta + \epsilon W\} \quad \{F_\epsilon = (1 - \epsilon)F_\theta + \epsilon\delta_{y_0}\}$$

IF and bdp

- *IF* = deals with infinitesimal values of ϵ
- bdp = largest ϵ an estimator can tolerate
- If an estimator has bdp = ϵ^* , $T_\infty(F)$ remains in a bounded set when F belongs to G_ϵ with $\epsilon \leq \epsilon^*$

$$\{G_\epsilon = (1 - \epsilon)F_\theta + \epsilon W\}$$

- What is the worst behaviour of the estimator for $\epsilon \leq \epsilon^*$?

MAXIMUM BIAS (MB) and BREAKDOWN POINT (*bdp*)

- The maximum (asymptotic bias) of T_n is

$$MB_{\hat{T}_\infty}(\epsilon, \theta) = \max \left\{ \left| \hat{T}_\infty(F) - \theta \right| : F \in G_\epsilon \right\}$$

$$\{G_\epsilon = (1 - \epsilon)F_\theta + \epsilon W\}$$

- MB gives the maximal possible effect on T due to any fixed fraction of contamination

$$bdp(T_\infty, F_\theta) = \max \left\{ \epsilon \geq 0 : MB_{\hat{T}_\infty}(\epsilon, \theta) < \infty \right\}$$

Summary values of IF: Gross error sensitivity (GES)

- The gross error sensitivity of T_n at F_θ

$$\gamma^*(T_n, F_\theta) = \sup_{y_0} |IF_{\hat{T}}(y_0, F_\theta)|$$

- GES measures the worst influence which a small amount of contamination of fixed size can have on the value of the estimator.
- Robust estimator = estimator with a bounded GES
- GES $< \infty \rightarrow$ B-robust (Bias robust)

Now for something constructive

Classes of estimators which have desirable properties

Class of M estimators

- Generalization of maximum likelihood estimators
- T_{MLE} for θ solve

$$\min_{\theta} \sum_{i=1}^n [-\log f(y_i; \theta)]$$

- M estimators are defined as the solution T_n for θ of the minimization problem

$$\min_{\theta} \sum_{i=1}^n \rho(y_i; \theta)$$

- $\rho =$ some convex function on $Y \times \theta$.
 ρ need not be related to any density

Estimating equations

- Suppose that ρ has a derivative

$$\psi(y; \theta) = \left[\frac{\partial}{\partial \theta_1} \rho(y; \theta), \dots, \frac{\partial}{\partial \theta_p} \rho(y; \theta), \right]$$

- then the estimate satisfies the implicit equation

$$\sum_{i=1}^n \psi(y_i; \theta) = 0$$

- Note that if

$$\psi(y; \theta) = \left[\frac{\partial}{\partial \theta_1} \log(y; \theta), \dots, \frac{\partial}{\partial \theta_p} \log(y; \theta), \right]$$

we obtain the MLE

M estimators of location

(y_1, y_2, \dots, y_n) iid with common cdf $F(y - \theta)$

- M estimators of location solve
- $$\min_{\theta} \sum_{i=1}^n \rho(y_i - \theta) \text{ or } \sum_{i=1}^n \psi(y_i - \theta) = 0 \text{ with } \psi = \rho'$$

Examples:

$$\rho(y) = \frac{y^2}{2} \Rightarrow \psi(y) = y \Rightarrow T_n = \bar{y} \text{ MLE for normal}$$

$$\rho(y) = |y| \Rightarrow \psi(y) = \text{sign}(y) \Rightarrow T_n = Me$$

MLE for double exponential

How to choose ρ or ψ ?

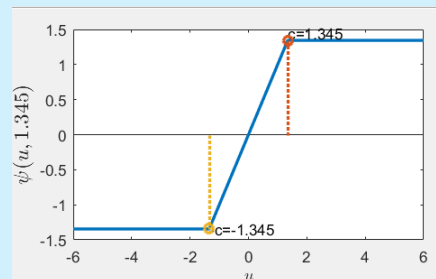
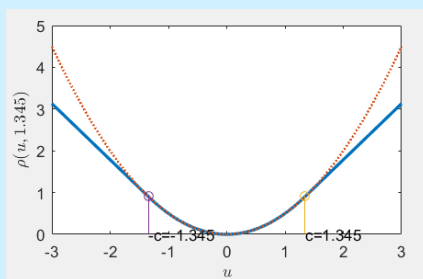
ρ and ψ functions

- A ρ function has the following characteristics
 - $\rho(u)$ is a non decreasing function of $|u|$
 - $\rho(0)=0$
 - $\rho(u)$ is increasing for $u >0$ such that $\rho(u) < \rho(\infty)$
- A ψ function denotes a function which is the derivative of a ρ function which implies
 - $\psi(u)$ is odd and $\psi(u) \geq 0$ for $u \geq 0$

Family of Huber functions

$$\rho(u) = \begin{cases} (u^2/2) & |u/c| \leq 1 \\ c|u| - c^2/2 & |u/c| > 1 \end{cases}$$

$$\psi(u) = \begin{cases} u & \text{if } |u/c| \leq 1 \\ c \times \text{sign}(u) & |u/c| > 1 \end{cases}$$



- the limit cases $c \rightarrow \infty$, $c \rightarrow 0$ are the mean and the median and we define $\psi(u,0)=\text{sign}(u)$. Monotonic ψ function.
- Brings in extreme observations to $\mu \pm c$.
- Corresponds to a density with normal centre and double-exponential tails.

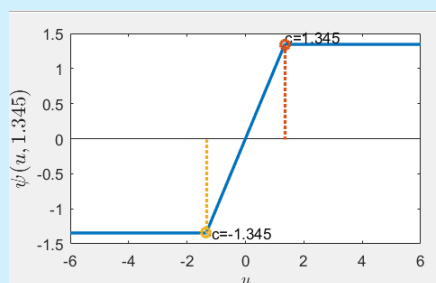
IF for (one dimensional location) M estimators

$$IF_{\hat{T}}(y_0, F_\theta) = \frac{\psi(y_0 - \hat{T}_\infty)}{E_F \psi'(y - \hat{T}_\infty)}$$

- Bounded influence if ψ is bounded
- The influence function and the ψ have the same shape
- IF for location estimation with a previously computed dispersion estimate $\hat{\sigma}$ is

$$IF_{\hat{T}}(y_0, F_\theta) = \hat{\sigma}_\infty \frac{\psi\left((y_0 - \hat{T}_\infty)/\hat{\sigma}_\infty\right)}{E_F \psi'\left((y - \hat{T}_\infty)/\hat{\sigma}_\infty\right)}$$

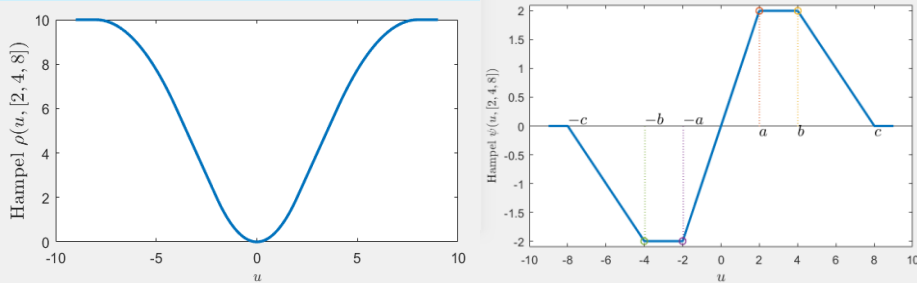
Redescending estimators



- The influence curve (proportional to psi function) is constant for all observations beyond a certain point.
- An M-estimator can be made more resistant by having the psi-function, (and hence the IF) return to 0.

Family of Hampel functions

$$\psi(u) = \begin{cases} u & |u| \leq a \\ a \times \text{sign}(u) & a \leq |u| < b \\ a \frac{c-|u|}{c-b} \times \text{sign}(u) & b \leq |u| < c \\ 0 & |u| \geq c \end{cases}$$



- Has a redescending psi function
- The 3 tuning constants provide flexibility for tuning the estimator.
- How to choose a , b and c ?

The rejection point

- The ψ function of the Hampel is 0 for $|u|$ larger than c . Therefore the IF is 0 for $|y| > c \hat{\sigma}$. We say that the rejection point is $r = c \hat{\sigma}$. Observations beyond the rejection point do not contribute to the value of the estimate (except possibly through the auxiliary scale estimate)
- The rejection point r is the least distance from the location estimate beyond which observations do not contribute to the value of the estimate (for a given auxiliary scale estimate)

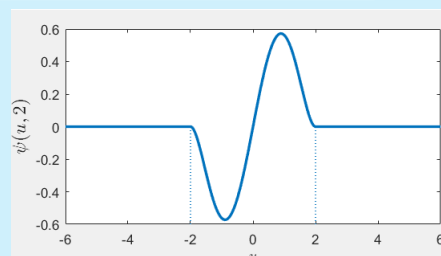
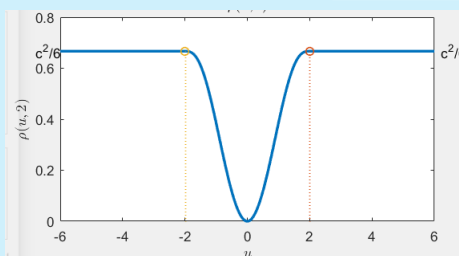
Winsor's principle

- Winsor's principle: «all distributions are normal in the middle» → we want to have a ψ function which resembles the one that is best for Gaussian data.
- The ψ function of MLE of μ at $N(\mu, \sigma^2)$ is linear therefore the ψ for M-estimators of location should be linear near the origin
- $\psi(u) \approx k u$ for small $|u|$ (where k is a nonzero constant usually standardized to $k=1$)
- Do you believe in Winsor's principle?

Family of Tukey's biweight functions

$$\rho(u) = \begin{cases} (c^2/6) \{1 - [1 - (u/c)^2]^3\} & |u/c| \leq 1 \\ (c^2/6) & |u/c| > 1 \end{cases}$$

$$\psi(u) = \begin{cases} (c^2/6)u[1 - (u/c)^2]^2 & |u/c| \leq 1 \\ 0 & |u/c| > 1 \end{cases}$$



- The constant c can be tuned for breakdown point (efficiency). Redescending psi function

Distribution of M estimators

- Let μ_0 be the solution of

$$E_F \psi(y - \mu_0) = 0.$$

- Centre of symmetry for symmetric F_0
- Then the distribution of M estimator is asymptotically normal

$$\mathcal{N}\left(\mu_0, \frac{v}{n}\right) \quad \text{with} \quad v = \frac{E_{F_0}\{\psi(y)^2\}}{\{E_{F_0}\psi'(y)\}^2} = \frac{A}{B^2}.$$

The asymptotic relative efficiency is

$$ARE(\hat{\mu}) = v_0/v$$

where v_0 is the asymptotic variance of the MLE

$$\mathcal{N}\left(\mu_0, \frac{v}{n}\right) \quad \text{with} \quad v = \frac{E_{F_0}\{\psi(y)^2\}}{\{E_{F_0}\psi'(y)\}^2} = \frac{A}{B^2}.$$

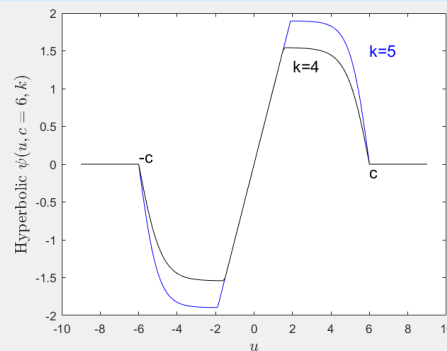
Goal: minimize variance subject to

- class of ψ function which have a finite rejection point
- control through a parameter k the change of variance sensitivity of the M-estimators (to investigate the infinitesimal stability of the asymptotic variance)
- Idea: define a psi functions with parameters A , B and k
- This has led to the hyperbolic tangent ψ function (Hampel Rousseeuw and Ronchetti, 1981)

Hyperbolic tangent ψ function

$$\psi(u) = \begin{cases} u & |u| \leq d \\ \sqrt{A(k-1)} \tanh\left(\sqrt{(k-1)B^2/A}(c-|u|)/2\right) \text{sign}(u) & d \leq |u| < c, \\ 0 & |u| \geq c. \end{cases}$$

$$0 < A < B < 2\Phi(c) - 1 - 2c \times \phi(c) < 1$$



- the central part of has to be linear in order to achieve a high asymptotic efficiency at the model
- Note that A , B and d are automatically determined after fixing k and c

Location M-estimate: computations

- Equation $\sum_{i=1}^n \psi(y_i - \hat{\mu}) = 0$ implies

$$\sum_{i=1}^n w_i(y_i - \hat{\mu}) = 0 \quad \text{with} \quad w_i = \psi(y_i - \hat{\mu}) / (y_i - \hat{\mu}).$$

- This suggests an iterative procedure
- Given some initial estimate (for example the median) or an estimate at step k ($\hat{\mu}_k$) compute

$$w_{i,k} = W(y_i - \hat{\mu}_k), \quad i = 1, 2, \dots, n$$

$$\hat{\mu}_{k+1} = \frac{\sum_{i=1}^n w_{i,k} y_i}{\sum_{i=1}^n w_{i,k}}$$

- Stop when $|\mu_{k+1} - \mu_k| < \epsilon$

Location M-estimate: computations

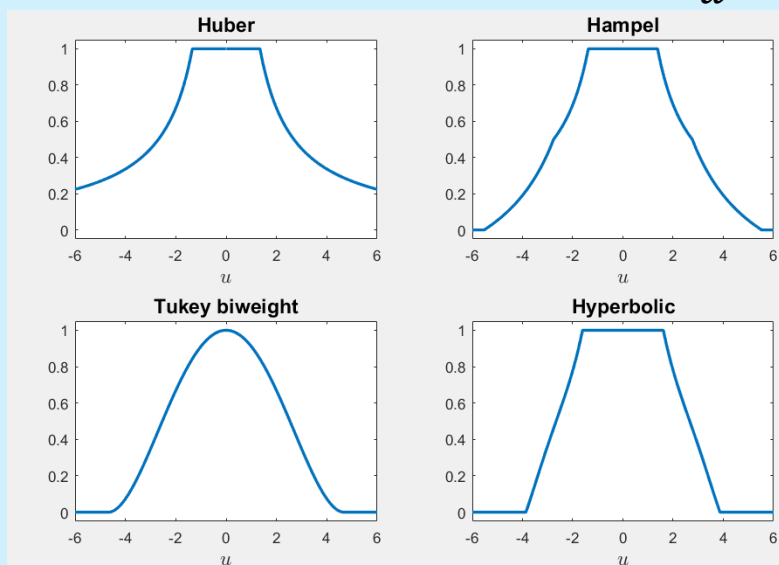
- $$\hat{\mu}_{k+1} = \frac{\sum_{i=1}^n w_{i,k} y_i}{\sum_{i=1}^n w_{i,k}}$$

- Idea: downweight outliers



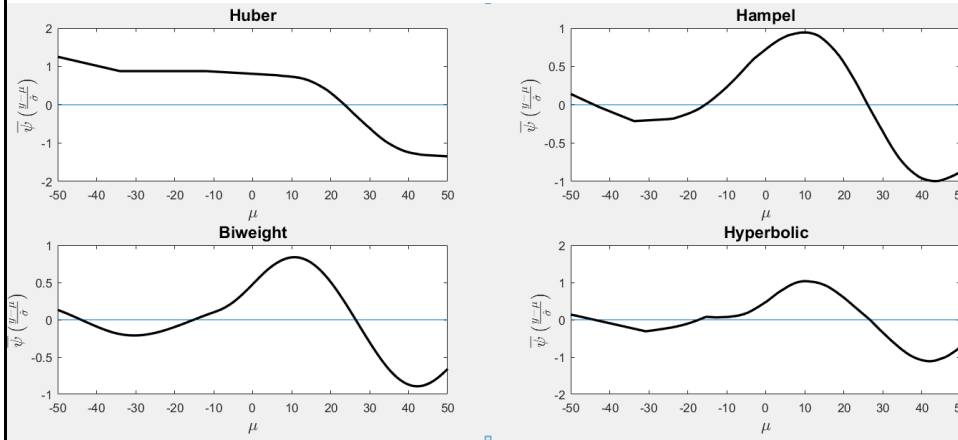
- If W is bounded and non increasing then the sequence converges to a solution
- If $\psi(u)$ is not monotone there may be multiple solutions

Comparison of $W(u) = \frac{\psi(u)}{u}$



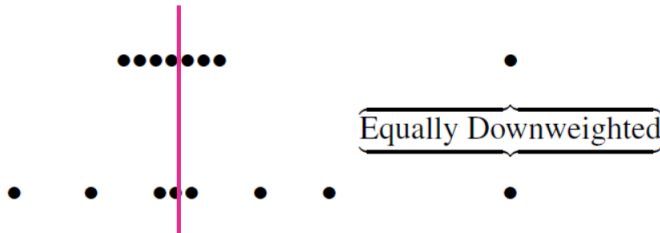
Example with multiple solutions

$y=[28\ 26\ 33\ 24\ 34\ 27\ 16\ 40\ -2\ 29\ 22\ 24\ 21\ 25\ 30\ 23\ 29\ 31\ 19\ -44\ -44\ -44]$; $\hat{\sigma} = \text{MADN}$, (efficiency set to 0.95)



M estimates of location are not scale equivariant

- The adaptive weights are not independent of the spread of the data (when the scale is not estimated)



- Exceptions include the mean and the median.

M estimate of location with auxiliary scale

- Use some (of course robust) estimate of scale, say $\hat{\sigma}_n$ and replace r in $\rho(r), \psi(r)$ by $(y - \mu)/\hat{\sigma}_n$

$$\hat{\mu} = \arg \min \sum_{i=1}^n \rho \left(\frac{y_i - \mu}{c \hat{\sigma}_n} \right) = 0$$

$$\hat{\mu} = \sum_{i=1}^n \psi \left(\frac{y_i - \mu}{c \hat{\sigma}_n} \right) = 0$$

- $c =$ tuning constant
- $\hat{\sigma}_n$ computed simultaneously?

M estimate of scale

- The MLE of σ for the scale family $\left(\frac{1}{\sigma}\right) f\left(\frac{y_i - \mu}{\sigma}\right)$ is:
 $\operatorname{argmax}_{\sigma} \prod_{i=1}^n \left(\frac{1}{\sigma}\right) f\left(\frac{y_i - \mu}{\sigma}\right)$
- Taking logs and differentiating with respect to σ we obtain
- $\frac{1}{n} \sum_{i=1}^n \left\{ -\frac{f'(\frac{y_i - \mu}{\sigma})}{f(\frac{y_i - \mu}{\sigma})} \frac{y_i - \mu}{\sigma} \right\} = 1$
- Idea: in order to bound the effect of large $(y_i - \mu)/\sigma$ replace what is in $\{ \}$ by a ρ function
- $\frac{1}{n} \sum_{i=1}^n \rho \left(\frac{y_i - \mu}{\sigma} \right) = \delta$
- $\delta = E\rho \left(\frac{y_i - \mu}{\sigma} \right)$ for consistency at the normal distribution

M-scale as a weighted RMS estimate

- Equation $\frac{1}{n} \sum_{i=1}^n \rho \left(\frac{y_i - \mu}{\sigma} \right) = \delta$

putting $W(y) = \rho(y)/y^2$

can be rewritten as

- $\frac{1}{n} \sum_{i=1}^n W \left(\frac{y_i - \mu}{\sigma} \right) \left(\frac{y_i - \mu}{\sigma} \right)^2 = \delta$

or as

- $\sigma^2 = \left(\frac{1}{n \delta} \right) \sum_{i=1}^n W \left(\frac{y_i - \mu}{\sigma} \right) (y_i - \mu)^2$

- this is a weighted mean square estimate

- Remark: μ is taken as known

Estimation of the M-scale

- Expression $\sigma^2 = \left(\frac{1}{n \delta} \right) \sum_{i=1}^n W \left(\frac{y_i - \mu}{\sigma} \right) (y_i - \mu)^2$ suggests an iterative procedure

- Start with some $\hat{\sigma}_0$ (for example MADN)

- In general, given $\hat{\sigma}_k$ (estimate of σ at step k) find the weights as $W \left(\frac{y_i - \mu}{\hat{\sigma}_k} \right)$

- $\hat{\sigma}_{k+1}^2 = \left(\frac{1}{n \delta} \right) \sum_{i=1}^n W \left(\frac{y_i - \mu}{\hat{\sigma}_k} \right) (y_i - \mu)^2$

- $\hat{\sigma}_{k+1}^2 = \hat{\sigma}_k^2 \left(\frac{1}{n \delta} \right) \sum_{i=1}^n W \left(\frac{y_i - \mu}{\hat{\sigma}_k} \right) \frac{(y_i - \mu)^2}{\hat{\sigma}_k^2}$

- Now given that $W(y) = \rho(y)/y^2$

- $\hat{\sigma}_{k+1}^2 = \hat{\sigma}_k^2 \left(\frac{1}{n \delta} \right) \sum_{i=1}^n \rho \left(\frac{y_i - \mu}{\hat{\sigma}_k} \right) = \hat{\sigma}_k^2 \frac{1}{\delta} \bar{\rho} \left(\frac{y_i - \mu}{\hat{\sigma}_k} \right)$

- ...

Simultaneous estimation of location and dispersion

- It is necessary to solve the system of equations
- $\sum_{i=1}^n \psi_{location} \left(\frac{y_i - \hat{\mu}}{\hat{\sigma}} \right) = 0$
- $\sum_{i=1}^n \frac{1}{n} \rho_{scale} \left(\frac{y_i - \hat{\mu}}{\hat{\sigma}} \right) = \delta$
- Remark: ρ_{scale} in order to distinguish it from the ρ function used for location.
- Given starting values $\hat{\mu}_0$ and $\hat{\sigma}_0$ (*Me* and *MADN*) or an estimate at step k , $\hat{\mu}_k$ and $\hat{\sigma}_k$ find the weights
- $w_{i,k} = W_{location} \left(\frac{y_i - \hat{\mu}_k}{\hat{\sigma}_k} \right)$
- $\hat{\mu}_{k+1} = \frac{\sum_{i=1}^n w_{i,k} y_i}{\sum_{i=1}^n w_{i,k}}$
- $\hat{\sigma}_{k+1}^2 = \hat{\sigma}_k^2 \left(\frac{1}{n \delta} \right) \sum_{i=1}^n \rho_{scale} \left(\frac{y_i - \hat{\mu}_k}{\hat{\sigma}_k} \right)$

Now regression

Regression setting

- Data (y_i, x_i) $i=1, 2, \dots, n$
- $y_i \in R$ Response
- $x_i \in R^p$ Predictors
- Regression model $y_i = x_i^T \beta + \sigma u_i$
- Predict y_i by $x_i^T \hat{\beta}$
- Residuals for given β : $r_i = r_i(\beta) = y_i - x_i^T \beta$

M estimates of regression

- They are defined as solution $\hat{\beta}$ to

$$\hat{\beta} = \arg \min \sum_{i=1}^n \rho \left(\frac{r_i(\hat{\beta})}{\hat{\sigma}} \right)$$

$$\sum_{i=1}^n \psi \left(\frac{r_i(\hat{\beta})}{\hat{\sigma}} \right) x_i = 0 \text{ where } \psi = \rho'$$

- For regression it is necessary to replace
- $y_i - \mu$ or $\frac{y_i - \mu}{\hat{\sigma}}$ in earlier expressions by r_i or $\frac{r_i}{\hat{\sigma}}$

Desired properties

- Scale equivariance

$$\tilde{y}_i = cy_i \rightarrow \beta(x_i, \tilde{y}_i) = c\beta(x_i, y_i)$$

- Affine equivariance

$$\tilde{x}_i = Ax_i \rightarrow \beta(\tilde{x}_i, y_i) = (A^T)^{-1}\beta(x_i, y_i)$$

- Regression equivariance

$$\tilde{y}_i = y_i + x_i^T \gamma \rightarrow \beta(x_i, \tilde{y}_i) = \beta(x_i, y_i) + \gamma$$

Why regression equivariance?

- If $y = X\beta + u$ and y is transformed as $\tilde{y} = y + X\gamma$ then $y = \tilde{y} - X\gamma \Rightarrow \tilde{y} - X\gamma = X\beta + u \Rightarrow$

$\tilde{y} = X\beta + X\gamma + u \Rightarrow \tilde{y} = X(\beta + \gamma) + u$ that is \tilde{y} satisfies the regression model with parameter vector $\tilde{\beta} = \beta + \gamma$

- Therefore if $\hat{\beta} = \hat{\beta}(X, y)$ is an estimate and data are transformed as $\tilde{y} = y + X\gamma$ we want that our new estimate of β is transformed as $\hat{\beta} + \gamma$

Computations (known scale)

- Recall LS regression: $\hat{\beta} = (X^T X)^{-1} X^T y$

- The constraint is (normal eqs)

$$\sum_{i=1}^n r_i(\beta) x_i = 0_{p \times 1}$$

- In robust regression the constraint is (weighted normal eqs)

$$\sum_{i=1}^n \psi\{r_i(\beta)\} x_i = 0 \quad \sum_{i=1}^n r_i(\beta) \frac{\psi\{r_i(\beta)\}}{r_i(\beta)} x_i = \sum_{i=1}^n r_i(\beta) w_i x_i = 0$$

with $w_i = \psi(r_i)/r_i$

- If w_i were known the above equation could be solved applying LS to $\sqrt{w_i} y_i$ and $\sqrt{w_i} x_i$

Adaptively weighted least squares

- Given some initial estimate of β say $\hat{\beta}_0$, first compute $\hat{\sigma}$ (for example MADN of the residuals)

- For $k=0, 1, 2, \dots$

- Given $\hat{\beta}_k$ compute residuals and weights as follows

$$r_{i,k} = y_i - x_i^T \hat{\beta}_k \quad i = 1, \dots, n \quad w_i = \psi(r_{i,k}/\hat{\sigma}) / (r_{i,k}/\hat{\sigma})$$

- Compute $\hat{\beta}_{k+1}$ solving

$$\sum_{i=1}^n w_{i,k} (y_i - x_i^T \hat{\beta}) = 0 \quad \hat{\beta}_{k+1} = \left(\sum_{i=1}^n w_{i,k} x_i x_i^T \right)^{-1} \sum_{i=1}^n w_{i,k} y_i x_i$$

- Stop when $\max_i (|r_{i,k} - r_{i,k+1}|) / \hat{\sigma} < \epsilon$

Remarks on the iterative procedure

- The algorithm converges if $W(x)$ is non increasing for $x > 0$
- If $\psi(u)$ is not monotone there may be multiple solutions
- For simultaneous estimation of β and σ the procedure is the same except that at each iteration $\hat{\sigma}$ is also updated (as in the location case)

Distribution of M estimates

- If X is fixed (or if x has a finite variance if it is random)

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N_p(0, \sigma^2 \gamma V_x^{-1})$$

$$V_x = E(xx') \text{ or } V_x = X'X \text{ if } X \text{ is fixed}$$

$$\gamma = \frac{E\psi(u/\sigma)^2}{(E\psi'(u/\sigma))^2}$$

- Remark: if x has a finite variance the efficiency of $\hat{\beta}$ does not depend on the distribution of x

COV matrix of estimated parameters

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N_p(0, \sigma^2 \gamma V_x^{-1})$$

- γ a correction factor depending on the ψ function which is used

$$\gamma = \frac{E\psi(u/\sigma)^2}{(E\psi'(u/\sigma))^2} \quad \hat{\gamma} = \frac{\frac{1}{n-p} \sum_{i=1}^n \psi\left(\frac{r_i}{\hat{\sigma}}\right)^2}{\left[\frac{1}{n} \sum_{i=1}^n \psi\left(\frac{r_i}{\hat{\sigma}}\right)\right]^2}$$

COV matrix of estimated parameters

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N_p(0, \sigma^2 \gamma V_x^{-1})$$

- Huber and Ronchetti suggest 3 expressions to estimate V_x

$$\hat{V}_x = \frac{1}{\frac{1}{n} \sum_{i=1}^n \hat{w}_i} X^T \hat{W} X,$$

where $w_i = w(r_i/\hat{\sigma}) = \hat{\sigma} \psi(r_i/\hat{\sigma})/r_i$

and $W = \text{diag}(\hat{w}_1, \dots, \hat{w}_n)$.

- Huber derived another correction factor

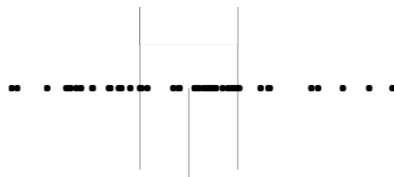
$$\hat{K}^2 = \left[1 + p \frac{\sum_{i=1}^n \left\{ \psi'\left(\frac{r_i}{\hat{\sigma}}\right) - \bar{\psi}'\left(\frac{r_i}{\hat{\sigma}}\right) \right\}^2}{\left\{ \sum_{i=1}^n \psi'\left(\frac{r_i}{\hat{\sigma}}\right) \right\}^2} \right]^2,$$

Least median of Squares (LMS)

$$\min_{\beta} Me\{r_i(\beta)^2\}$$

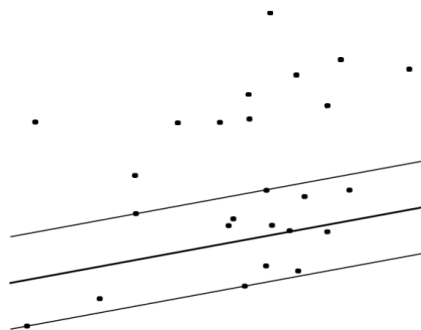
$$\min_{\beta} MAD\{r_i(\beta)\}$$

- In the univariate case LMS becomes the Midpoint of the SHORTest Half = SHORTH
- **SHORTH** = **shortest** interval that covers **half** of the values



Properties of LMS

- It is the centre-line of the shortest (narrowest) strip containing $\frac{1}{2}$ of the data



- It does not require a scale estimate
- Regression, scale and affine equivariant
- $bdp=0.5$, Fisher consistent and asymptotically normal

Drawbacks of LMS

- It displays marked sensitivity to central data values (not locally stable)
- It is very inefficient (converges at a rate $1/\sqrt[3]{n}$)

$$\sqrt{n} \|\hat{\beta} - \beta\| \rightarrow_p \infty$$

$$\sqrt[3]{n} \|\hat{\beta} - \beta\| = O_p(1)$$

Least Trimmed Squares regression (LTS)

- Least squares:
$$\min \sum_{i=1}^n r_i(\beta)^2$$
- Least trimmed squares:
$$\min \sum_{i=1}^h [r_i(\beta)^2]_{(i)}$$

$$[r(\beta)^2]_{(1)} \leq [r(\beta)^2]_{(2)} \leq \dots \leq [r(\beta)^2]_{(n)}$$

Characteristics of LTS

- A particular case of L-estimate of scale
- Regression, scale and affine equivariant
- Fisher consistent and asymptotically normal
- $\text{bdp} \approx \frac{\min(h, n-h)}{n}$
- converges at a rate $1/\sqrt{n}$
- Low efficiency (Ex. 7% at the normal distribution when $h=[n/2]$)

Regression S estimators

- LTS minimizes a robust residual scale estimate

$$\hat{\beta}_{LTS} = \arg \min_{\beta} \hat{\sigma}_{LTS}(\beta)$$

$$\hat{\sigma}_{LTS}(\beta) = \min \frac{1}{n-h} \sum_{i=1}^h [r_i(\beta)^2]_{(i)}$$

- Idea: minimize a more efficient robust scale estimator
- Regression S estimator minimizes an M estimate of scale

$$\hat{\beta}_S = \arg \min_{\beta} \hat{\sigma}_M(\beta)$$

Regression S estimators

- Least squares:
$$\min \sum_{i=1}^n r_i(\beta)^2$$
- S estimates
$$\min \hat{\sigma}_M(\beta)$$

where for each β , $\hat{\sigma}_M(\beta)$ solves

$$\frac{1}{n} \sum_{i=1}^n \rho_c \left(\frac{r_i(\beta)}{\sigma} \right) = K_c$$

bdp of S estimators

- If ρ satisfies the following conditions
 1. It is symmetric and continuously differentiable, and $\rho(0) = 0$;
 2. There exists a $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$;
 3. It is such that $K_c/\rho(c) = \text{bdp}$, with $0 < \text{bdp} \leq 0.5$,
- the asymptotic breakdown point of the S estimator tends to bdp when $n \rightarrow \infty$
- For consistency we require that

$$\frac{1}{n} \sum_{i=1}^n \rho_c \left(\frac{r_i(\beta)}{\sigma} \right) = K_c \quad E_{\Phi_{0,1}} \left[\rho \left(\frac{r_i}{s} \right) \right] = K_c$$

S estimates of regression. Tuning constant associated with a bdp

- From the equations

$$K_c = \text{bdp} \times \rho(c) \quad K_c = E_{\Phi_{0,1}} \left[\rho \left(\frac{r_i}{s} \right) \right]$$

- we can compute c . For example for Tukey's biweight ρ function we have

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if } |x| \leq c \\ \frac{c^2}{6} & \text{if } |x| > c, \end{cases}$$

$$\int_{-c}^c \left(\frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} \right) d\Phi_{0,1}(x) + \frac{c^2}{6} \Pr(|X| > c) = \text{bdp} \times \frac{c^2}{6}.$$

$$\left\{ \frac{\Pr(\chi_3^2 < c^2)}{2} - 3 \frac{\Pr(\chi_5^2 < c^2)}{2c^2} + 15 \frac{\Pr(\chi_7^2 < c^2)}{6c^4} + \frac{c^2}{3} (1 - \Phi(c)) \right\} = \text{bdp} \frac{c^2}{6}$$

S estimates of regression. Tuning constant associated with eff

- From the equations $\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N_p(0, \sigma^2 \gamma V_x^{-1})$

$$\gamma = \frac{E\psi(u/\sigma)^2}{(E\psi'(u/\sigma))^2}$$

- we can compute c . For example for Tukey biweight ρ function we have

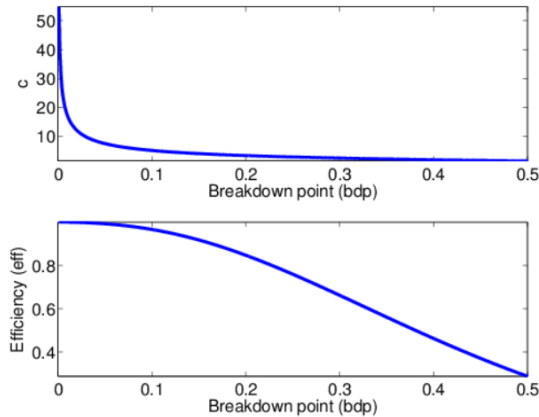
$$\int_{-c}^c \psi'(x) d\Phi(x) = 15 \frac{\Pr(\chi_5^2 < c^2)}{c^4} - 6 \frac{\Pr(\chi_3^2 < c^2)}{c^2} + \Pr(\chi_1^2 < c^2). \quad (19)$$

Similarly, we obtain for $\{\psi(x)\}^2$

$$\int_{-c}^c \{\psi(x)\}^2 d\Phi(x) = 9!! \frac{\Pr(\chi_{11}^2 < c^2)}{c^8} - 4 \times 7!! \frac{\Pr(\chi_9^2 < c^2)}{c^6} + 6 \times 5!! \frac{\Pr(\chi_7^2 < c^2)}{c^4} - 4 \times 3!! \frac{\Pr(\chi_5^2 < c^2)}{c^2} + \Pr(\chi_3^2 < c^2). \quad (20)$$

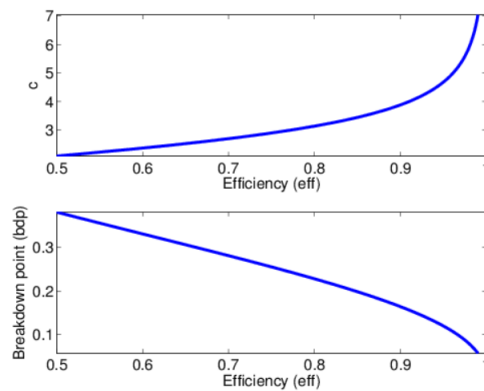
Efficiency and breakdown point

Consistency factor c (top panel) and efficiency (bottom panel) as a function of the breakdown point (bdp) for Tukey's Biweight.



Efficiency and breakdown point

Consistency factor c (top panel) and breakdown point (bottom panel) as a function of the efficiency (eff) for Tukey's Biweight.



TRADE OFF robustness-efficiency

- Hössjer (1992): an S-estimate with break down point equal to 0.5 has an asymptotic efficiency under normally distributed errors that is not larger than 0.33.

Breakdown point (bdp)	Consistency factor (c)	Asymptotic efficiency at the normal model (eff)
0.05	7.5453	0.9924
0.10	5.1824	0.9662
0.20	3.4207	0.8467
0.25	2.9370	0.7590
0.30	2.5608	0.6613
0.40	1.9880	0.4619
0.50	1.5476	0.2868

Table 1 Breakdown point, consistency factor and asymptotic efficiency at the normal model for Tukey's Biweight loss function in regression

S and LMS

- LMS is an S-estimate with a discontinuous ρ function
- Davies shows that estimates based on smooth ρ function have a convergence rate $n^{-0.5}$

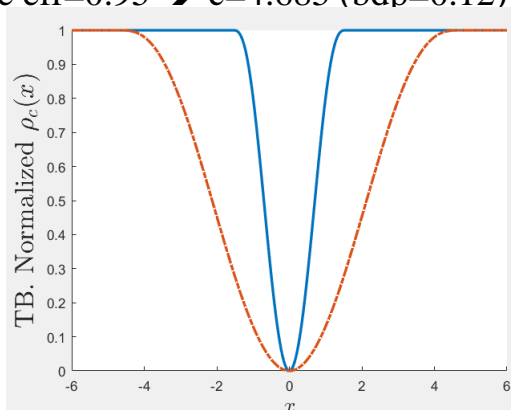
Regression MM estimators

$$\min \sum_{i=1}^n \rho_{eff=0.95} \left(\frac{r_i(\beta)}{\hat{\sigma}_S(\hat{\beta}_S)} \right)$$

- Idea: fix $bdp=0.5$ and using $\rho_{bdp=0.5}$ find $\hat{\beta}_S$ and $\hat{\sigma}_S$ using S estimators
- Fix $eff=0.95$, and using $\rho_{eff=0.95}$ using $\hat{\beta}_S$ and $\hat{\sigma}_S$ as starting values in the weighted least squares loop
- The estimate of the scale is kept fixed in the iterative procedure

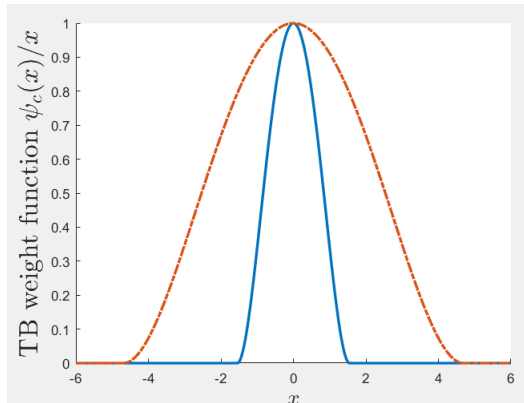
Claim of MM estimators

- HIGHLY ROBUST AND EFFICIENT
- Tukey's biweight rho (TB):
- blue line $bdp=0.5 \rightarrow c=1.548$ (eff=0.29)
- red line $eff=0.95 \rightarrow c=4.685$ (bdp=0.12)



Claim of MM estimators

- HIGHLY ROBUST AND EFFICIENT
- Tukey's biweight W (TB):
- blue line bdp=0.5 \rightarrow $c=1.548$ (eff=0.29)
- red line eff=0.95 \rightarrow $c=4.685$ (bdp=0.12)



Tau estimators (another attempt to break the link between bdp and eff)

- Unlike MM estimates do not require a preliminary scale estimate.
- If $\hat{\sigma}(\beta)$ solves the usual scale equation

$$\frac{1}{n} \sum_{i=1}^n \rho_{c_0} \left(\frac{r_i(\beta)}{\hat{\sigma}(\beta)} \right) = K_{c_0}$$

- define the scale tau as

$$\tau(\beta)^2 = \hat{\sigma}(\beta)^2 \frac{1}{n} \sum_{i=1}^n \rho_{c_1} \left(\frac{r_i(\beta)}{\hat{\sigma}(\beta)} \right)$$

- A regression tau estimate is defined by

$$\hat{\beta} = \min_{\beta} \tau(\beta)$$

Properties of tau estimators

$$\frac{1}{n} \sum_{i=1}^n \rho_{c_0} \left(\frac{r_i(\beta)}{\hat{\sigma}(\beta)} \right) = K_{c_0} \quad \tau(\beta)^2 = \hat{\sigma}(\beta)^2 \frac{1}{n} \sum_{i=1}^n \rho_{c_1} \left(\frac{r_i(\beta)}{\hat{\sigma}(\beta)} \right)$$

- Minimize a robust scale estimate (like S estimators) but (unlike S estimators) with a controllable efficiency
- Note that if $\rho_{c_0}(r) = \rho_{c_1}(r) = r^2 \rightarrow$ LS criterion
- In this case function ψ is a linear combination of ρ_{c_0} and ρ_{c_1}
- Claim: by an adequate choice of ρ the estimate can be made arbitrarily close to the LS estimate and therefore arbitrarily efficient at the normal distribution

Outlier detection

- We declare as an outlier any observation for which the absolute scaled residual

$$|r_i(\hat{\beta})| / \hat{\sigma} > \Phi^{-1}(1 - \alpha^*)$$

- Small sample correction factor?

Individual and simultaneous testing procedures

$$H_{0,i} : y_i \sim N(x_i'\beta, \sigma^2),$$

- which states that observation y_i comes from the postulated normal regression model.
- If the empirical test size is close to the nominal one, say α , we should thus expect a proportion of false outliers close to α for any uncontaminated data set (individual size)
- We can also use the whole set of n scaled residuals to test the hypothesis that no contamination is present in the data:

$$H_{0,\text{All}} : H_{0,1} \cap \dots \cap H_{0,n}.$$

- One expects to declare (at least one outlier) in a proportion α of the datasets

Individual and simultaneous threshold

- We use Bonferroni corrections for simultaneity, with level $\alpha^* = \frac{\alpha}{n}$, so taking the $1-\alpha^*$ cutoff value of the reference distribution.
- Reference distribution: it is customary to use the Chi squared reference distribution (although we are using robust estimation)

LTS (estimation of variance)

- Let
$$SST(\hat{\beta}_{LTS}) = \min \sum_{i=1}^h [r_i(\beta)^2]_{(i)}$$
- We base the estimator of σ^2 on this residual sum of squares. However, since the sum of squares contains only the central observations from a normal sample, the estimate needs scaling. The var of truncated normal is:

$$\sigma_T^2(h) = 1 - \frac{2n}{h} \Phi^{-1} \left(\frac{n+h}{2n} \right) \phi \left\{ \Phi^{-1} \left(\frac{n+h}{2n} \right) \right\}$$

ϕ and Φ are pdf and cdf of $N(0,1)$

- To estimate σ^2 we use
$$\hat{\sigma}_{LTS}^2(h) = \frac{SST(\hat{\beta}_{LTS})}{h \times \sigma_T^2(h)}$$

Small sample corr?

LMS and LTS reweighted (another attempt to break the link between bdp and eff)

- Giving weight 0 to observations for which

$$|r_{LTS,i}| = |r_i(\hat{\beta}_{LTS})| / \hat{\sigma}_{LTS}(h) > \Phi^{-1}(0.975)$$

- We then obtain a sample of reduced size $n-k$, possibly outlier free, to which OLS is applied.
- Let the parameter estimates be $\hat{\beta}_{LTSR}$ and $\hat{\sigma}_{LTSR}(h)$, the outliers are the k_1 observations rejected at the second stage

$$|r_{LTSR,i}| = |r_i(\hat{\beta}_{LTSR})| / \hat{\sigma}_{LTSR}(h) > \Phi^{-1}(1 - \alpha^*)$$

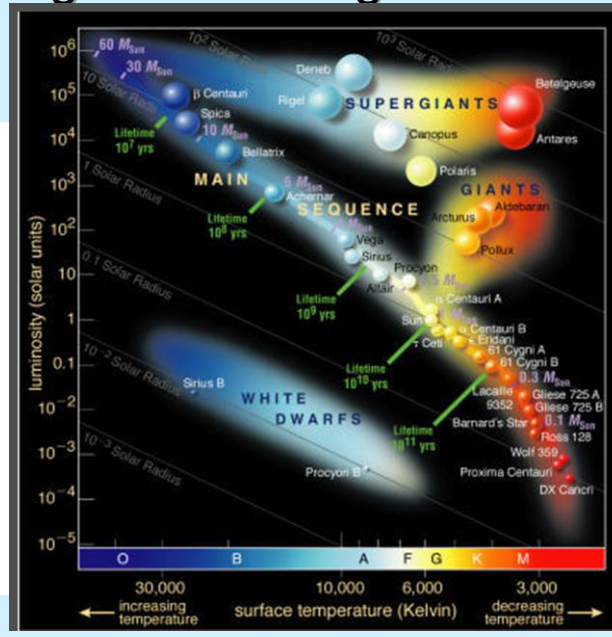
Huber and Ronchetti (2009)

- A plethora of alternative regression procedures have been devised whose goal is to improve the breakdown point ... Unfortunately, it seems that these alternative approaches have gone overboard with attempts to maximize the breakdown point, disregarding important other aspects, such as having reasonably high efficiency at the model. It is debatable whether any of these alternatives even deserve to be called robust, since they seem to fail the basic stability requirement of robustness. An approach through data analysis and diagnostics may be preferable.

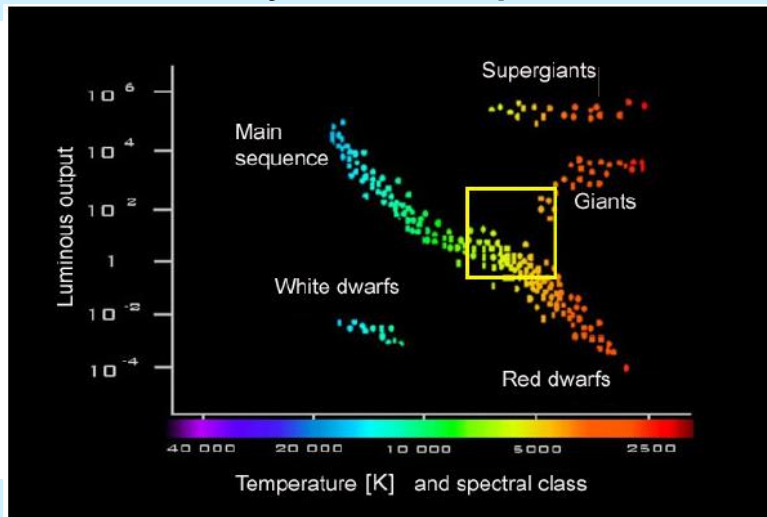
**Robust regression in
action**

Hertzspung Russell diagram.

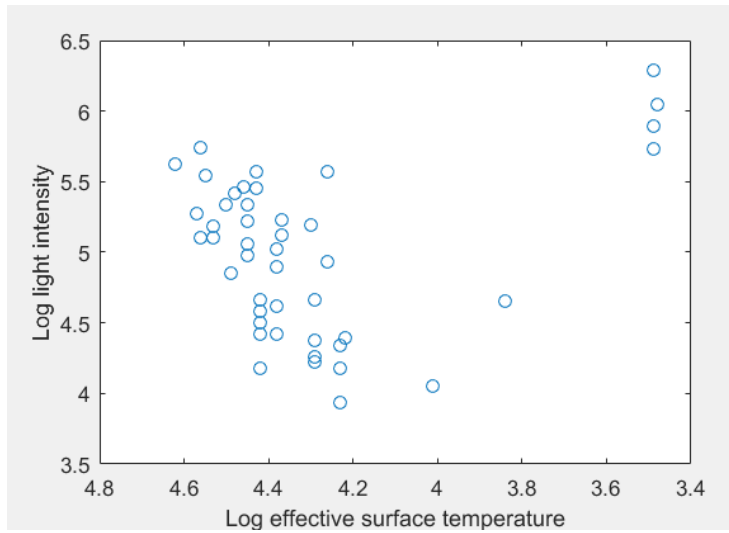
- Graph showing the luminosity of a star as a function of its surface temperature



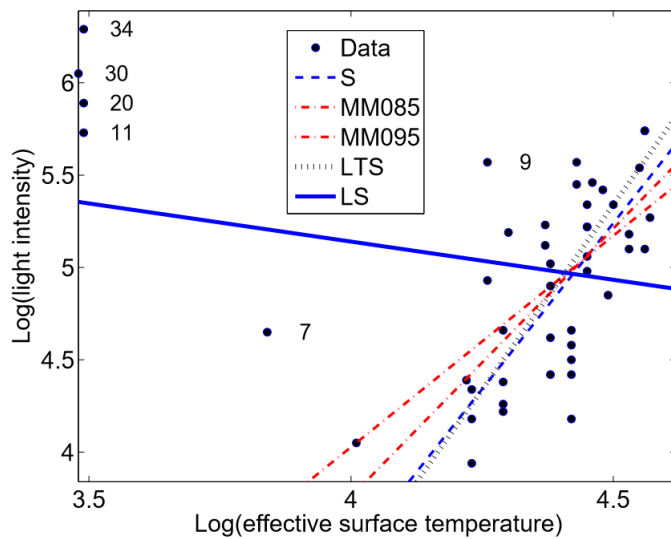
The extracted data come from the yellow square



Log light intensity vs Log effective surface temperature (reverse order)

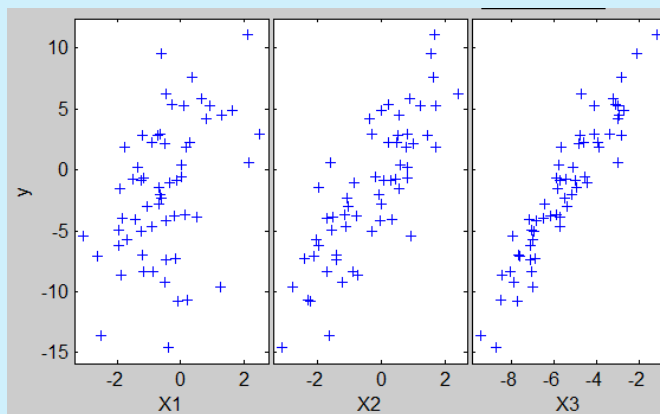


A comparison of different fits



Another example: a regression dataset with masked outliers (AR data)

- 60 observations, 3 explanatory variables



TRADITIONAL WAY OF DOING STATISTICS IN REGRESSION

```
>> mdlr = fitlm(X,y);
>> mdlr
```

```
Linear regression model:
y ~ 1 + x1 + x2 + x3
```

```
Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	11.174	0.67501	16.553	3.1288e-23
x1	-0.21796	0.17244	-1.264	0.21146
x2	1.4981	0.15534	9.6439	1.6733e-13
x3	2.2596	0.13668	16.531	3.3265e-23

```
Number of observations: 60, Error degrees of freedom: 56
```

```
Root Mean Squared Error: 1.09
```

```
R-squared: 0.965, Adjusted R-Squared 0.963
```

```
F-statistic vs. constant model: 510, p-value = 1.33e-40
```

Statistics toolbox: RobustOpts on

```
>> mdlr = fitlm(X,y,'RobustOpts','on');
>> mdlr
```

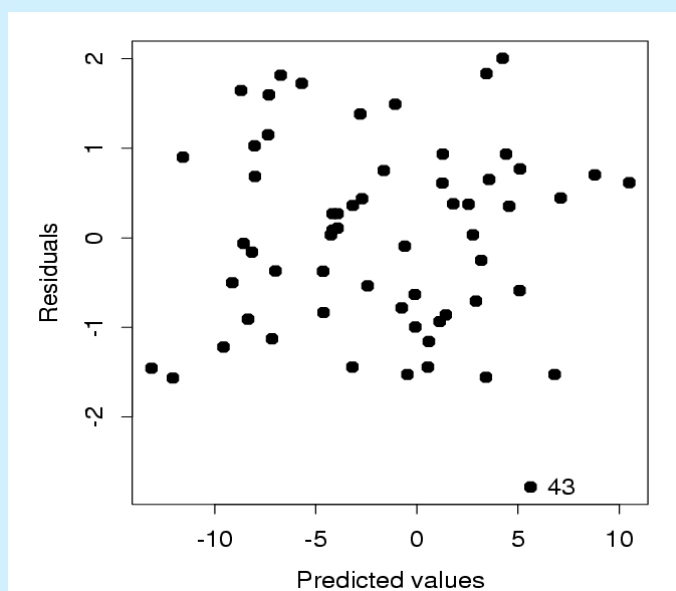
mdlr =
Linear regression model (robust fit):
y ~ 1 + x1 + x2 + x3

Estimated Coefficients:

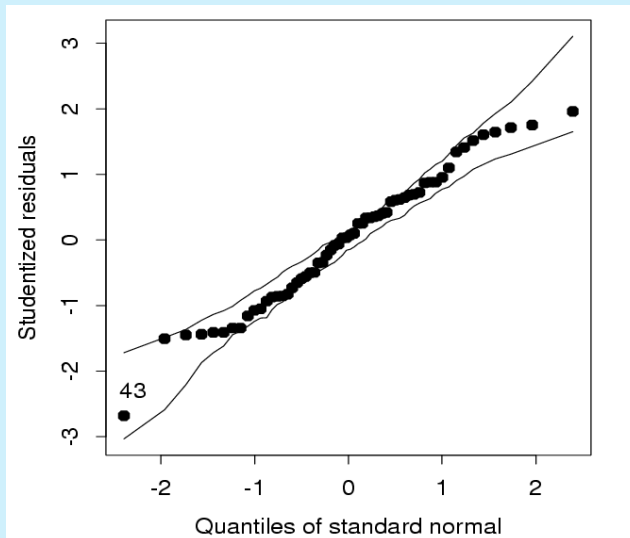
	Estimate	SE	tStat	pValue
(Intercept)	11.415	0.71721	15.915	1.9345e-22
x1	-0.25422	0.18322	-1.3875	0.17078
x2	1.4662	0.16506	8.8832	2.7871e-12
x3	2.3066	0.14523	15.883	2.1262e-22

Number of observations: 60, Error degrees of freedom: 56
Root Mean Squared Error: 1.16
R-squared: 0.961, Adjusted R-Squared 0.959
F-statistic vs. constant model: 456, p-value = 2.59e-39

LS residuals against predicted values



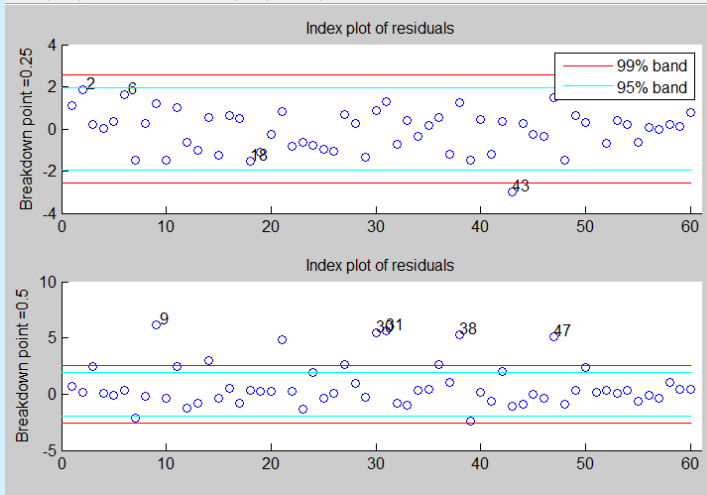
QQplot of studentized residuals



S and MM estimators

- Breakdown point (bdp)= percentage of outliers the estimator can cope with
- Efficiency (eff) = $\text{COV}(\beta_{\text{ROBUST}})/\text{COV}(\beta_{\text{LS}})$
- S \rightarrow fix breakdown point (efficiency depends on breakdown point). Ex. bdp=0.5 \rightarrow eff=0.29
- MM \rightarrow fix efficiency. Ex. eff=0.95 \rightarrow bdp=0.12

Analysis with robust S estimators:



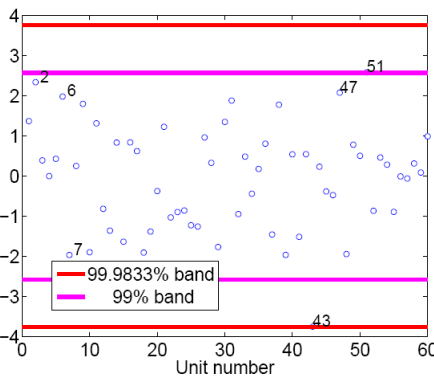
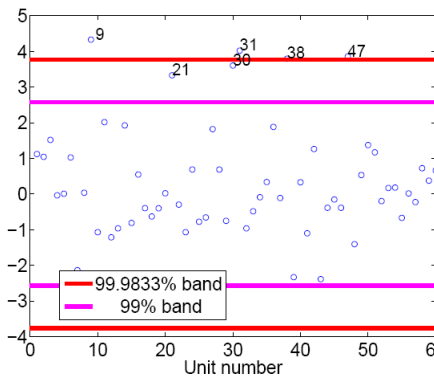
bdp=0.25

bdp=0.5

Analysis with robust estimators: MM

90% nominal efficiency

95% nominal efficiency



Individual and simultaneous confidence
99% bands

Traditional approach: compare robust and non robust fit

- Robust Inference as well as Classical Inference
- “... just which robust/resistant methods you use is not important – what is important is that you use some. It is perfectly proper to use both classical and robust/resistant methods routinely, and only worry when they differ enough to matter. But when they differ, you should think hard.”

– J. W. Tukey

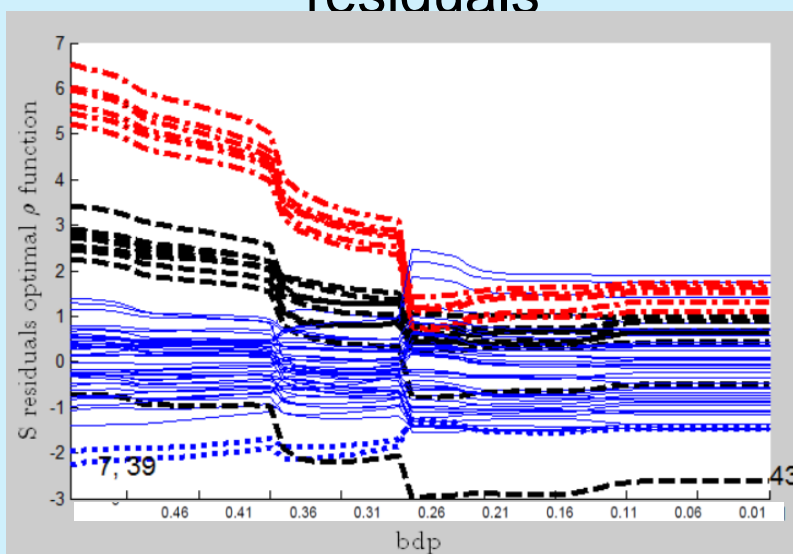
Consequences of the use of robust estimators

- Results obtained via a robust method are sometimes completely different
- Both in the use of traditional robust and non-robust statistical methods, researchers end up with a picture of the data.
- **WHY NOT TO WATCH A FILM OF THE DATA ANALYSIS?**

Consequences of the use of robust estimators

- Both in the use of traditional robust and non-robust statistical methods, researchers end up with a picture of the data analysis.
- The extension to more complex problems is difficult and requires ad hoc techniques
- The researcher loses the information that each unit, outlier or not, has on the final proposed estimate

Monitoring of scaled S residuals



How to summarize changes in fit?

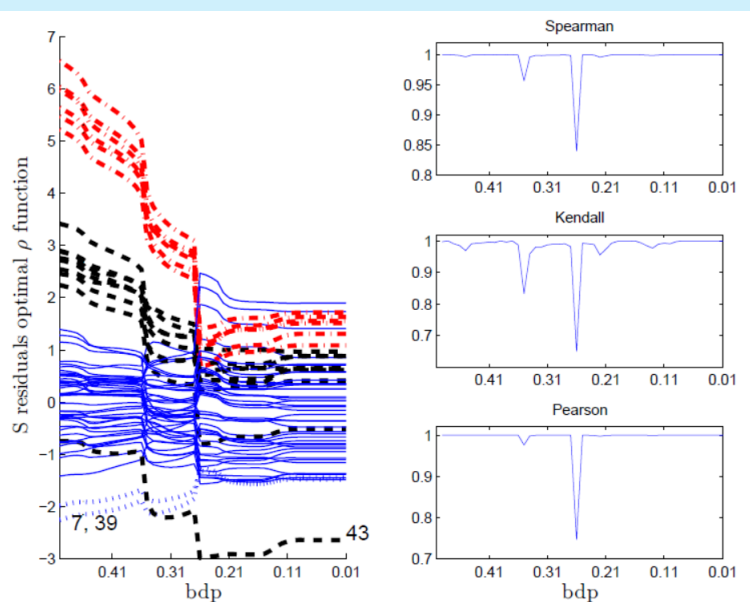
We consider three standard measures of correlation:

Spearman. The correlations between the ranks of the two sets of observations.

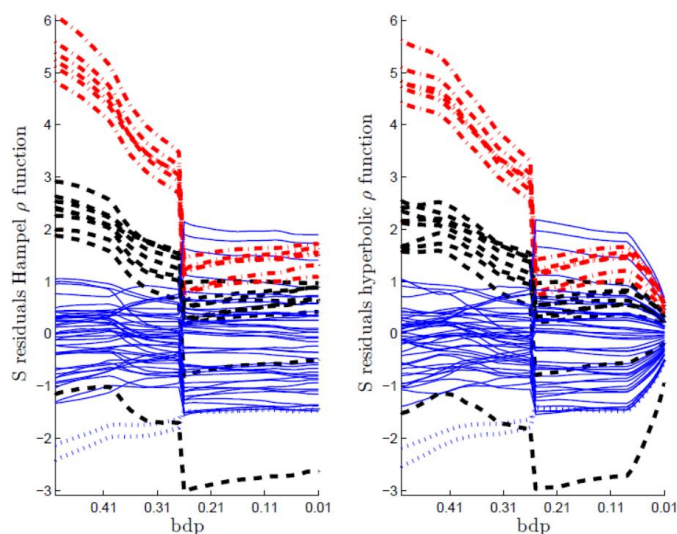
Kendall. Concordance of the pairs of ranks.

Pearson. Product-moment correlation coefficient

Monitoring of scaled S residuals



Hampel and hyperbolic rho function



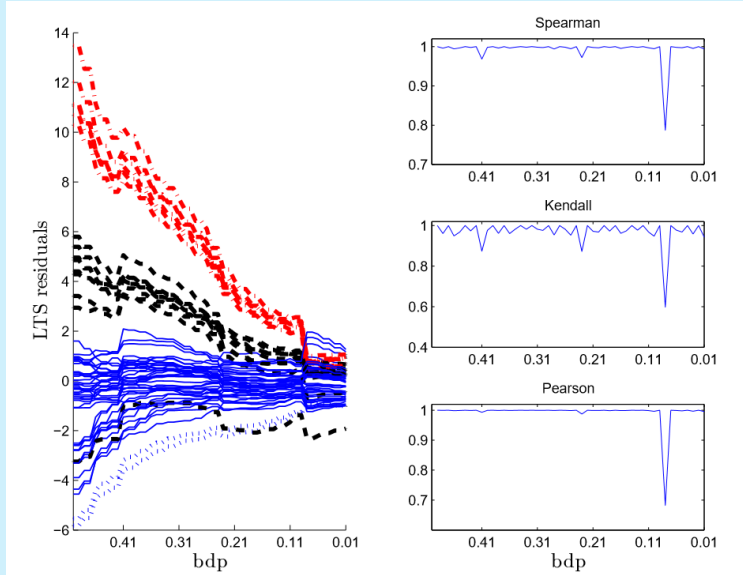
EMPIRICAL BDP AND EFFICIENCY

Table: Empirical breakdown point and efficiency during monitoring for the transition between very robust and least squares regression: five estimators and four ρ functions. The values are for the step before the switch to a non-robust fit

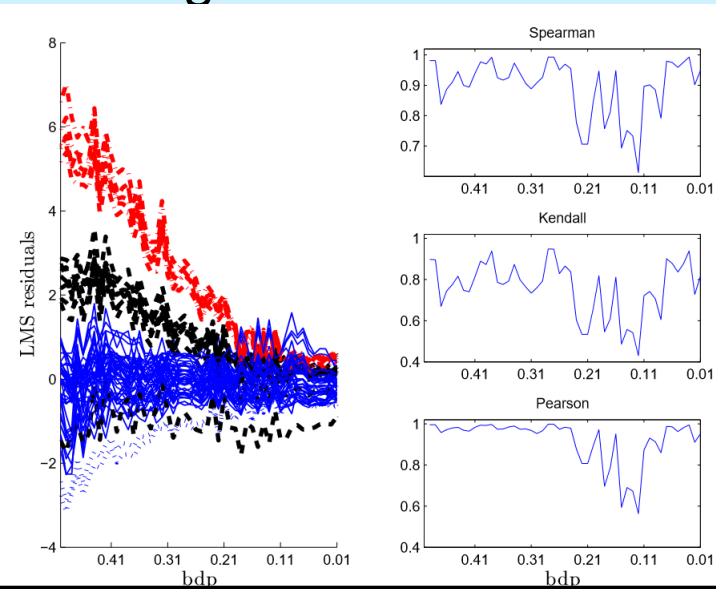
Estimator		Bisquare	Optimal	Hyperbolic	Hampel
S	bdp	0.27	0.27	0.26	0.27
$\tau = 0.85$	bdp	0.38	0.40	0.41	0.41
$\tau = 0.90$	bdp	0.45	0.48	— ^a	0.50
$\tau = 0.95$	bdp	— ^a	— ^a	— ^a	— ^a
MM	effic.	0.91	0.97	0.90	0.87

^a For these values of τ and ρ function, only non-robust solutions were obtained during monitoring.

Monitoring of scaled LTS residuals



Monitoring of scaled LMS residuals

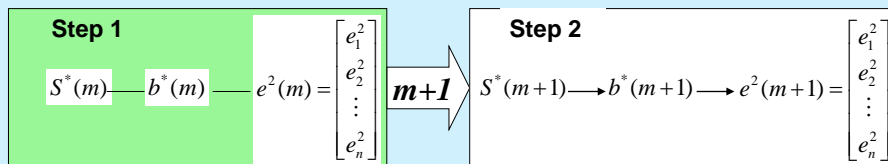


Forward Search in Linear Regression

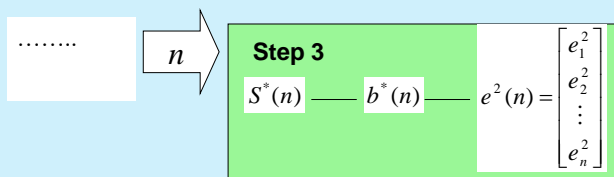
- STEP1: Start with very robust fit (LMS or LTS or S), then successively fit to larger subsets, found as those with the smallest residuals
- STEP2: Subset size increases until all the data are fitted. From LMS (LTS or S) to LS
- STEP3: MONITORING (scaled residuals, beta coefficients, ...)

Step 2: Adding observations during the Forward Search

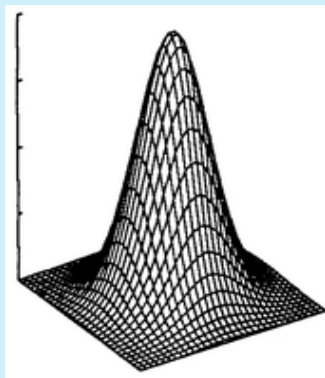
- Given $S^*(m)$, using $b^*(m)$, we compute the residuals for the n observations and select those which have the smallest squared $m+1$ residuals, $m=p, p+1, \dots, n$



- This step is repeated up to when all units are included into the subset



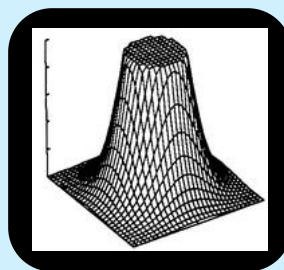
Characteristics of the FS



$m = p, p+1, p+2, \dots, n$



What is inside at step m



What is outside at step m

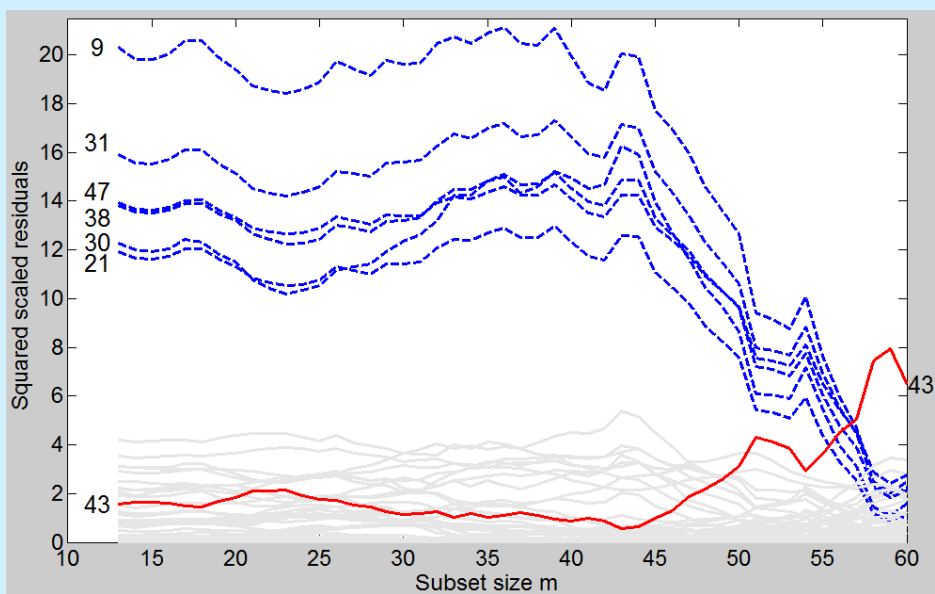
“New philosophy” of data analysis

- Our philosophy involves watching a film of data analysis rather than a snapshot.
- The crucial idea is to monitor how the fitted model changes as bdp decreases (S) or eff increases (MM) or, as in the “**forward search**”, whenever a new statistical unit is added to the subset.
- The slides which follow show the analysis of the AR data using the forward search

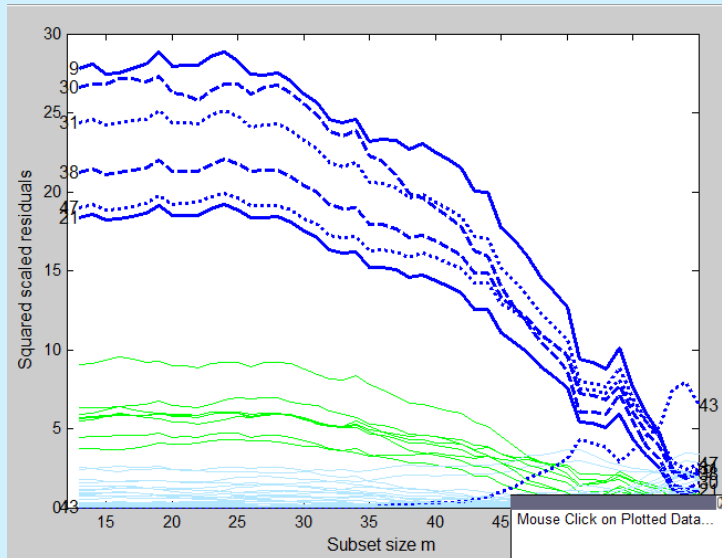
Target

- A tool that preserves the interpretative and computational simplicity of LS
- To develop a statistical approach that can attack relevant inferential issues in a unified way
- Italian expression “Botte piena e moglie ubriaca” (you can’t have your cake and eat it)

Monitoring of scaled residuals



Monitoring of scaled residuals



SIGNIFICANCE OF THE EXPLANATORY VARIABLES

Standard static approach

	All units	Without unit 43
t0	16.55	17.64
t1	-1.26	-1.93
t2	9.64	9.75
t3	16.53	17.66

Two alternative formulae for robust standard error in regression

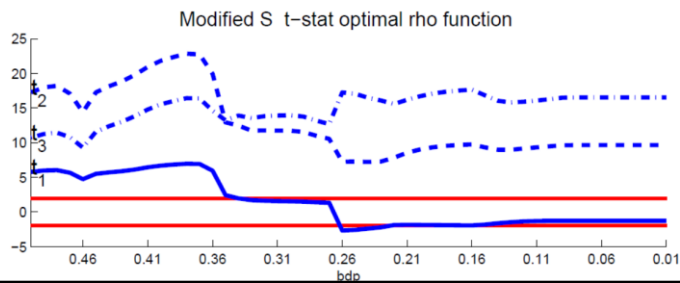
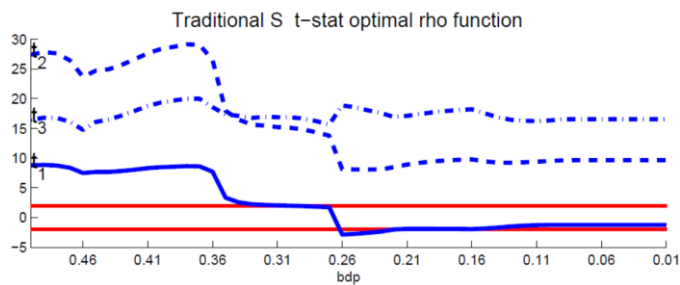
1

$$\hat{\sigma}^2 \hat{\gamma} (X^T X)^{-1}$$

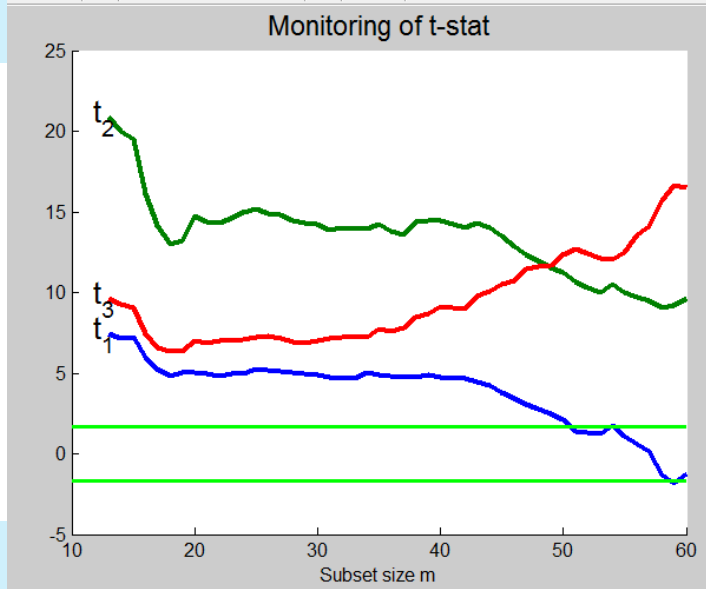
$$\hat{\sigma}^2 \hat{\gamma} \hat{K}^2 \frac{1}{\frac{1}{n} \sum_{i=1}^n \hat{w}_i} (X^T \hat{W} X)^{-1}$$

$$\hat{K}^2 = \left[1 + p \frac{\sum_{i=1}^n \left\{ \psi' \left(\frac{r_i}{\hat{\sigma}} \right) - \bar{\psi}' \left(\frac{r_i}{\hat{\sigma}} \right) \right\}^2}{\left\{ \sum_{i=1}^n \psi' \left(\frac{r_i}{\hat{\sigma}} \right) \right\}^2} \right]^2,$$

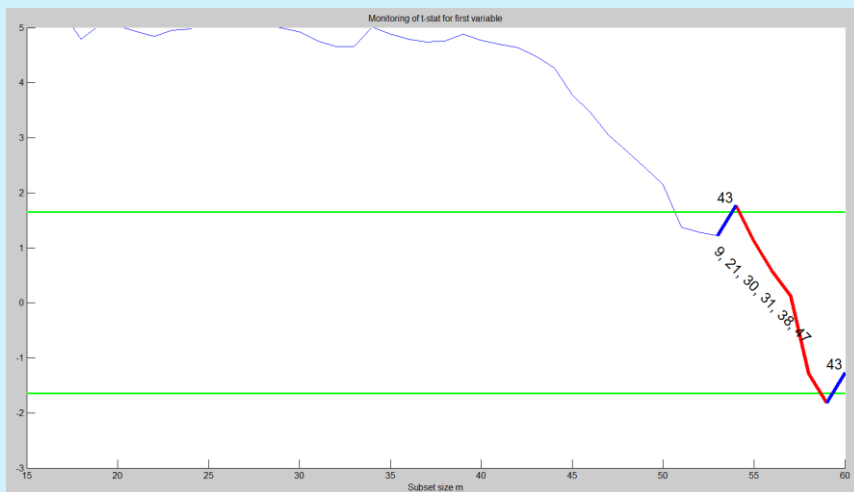
Robust t stats

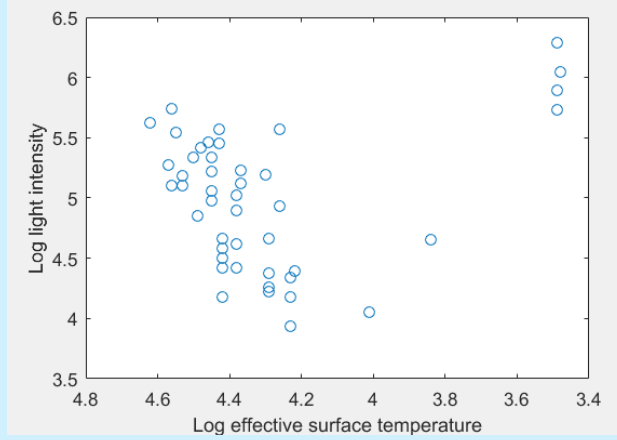


Monitoring of scaled t -statistics



Monitoring of scaled t -statistic for first variable

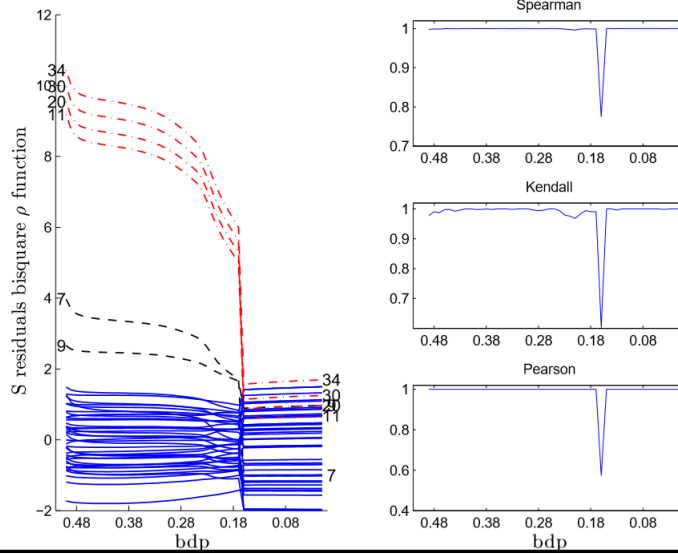




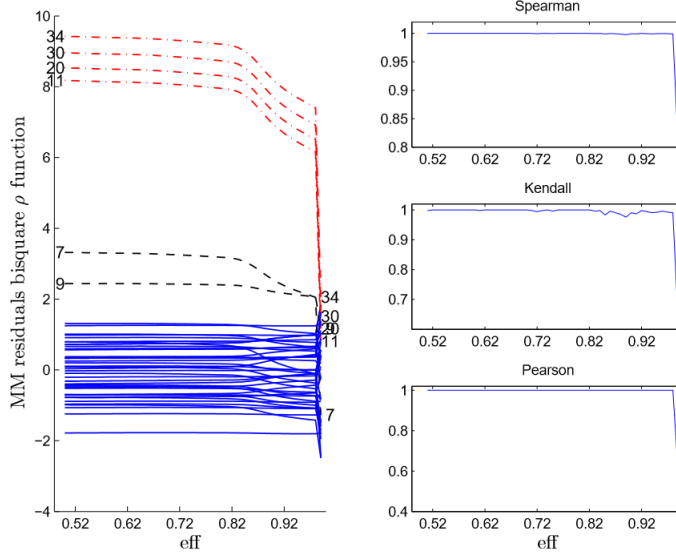
Stars data again

Analysis with the monitoring approach

Monitoring of scaled S residuals



Monitoring of MM residuals

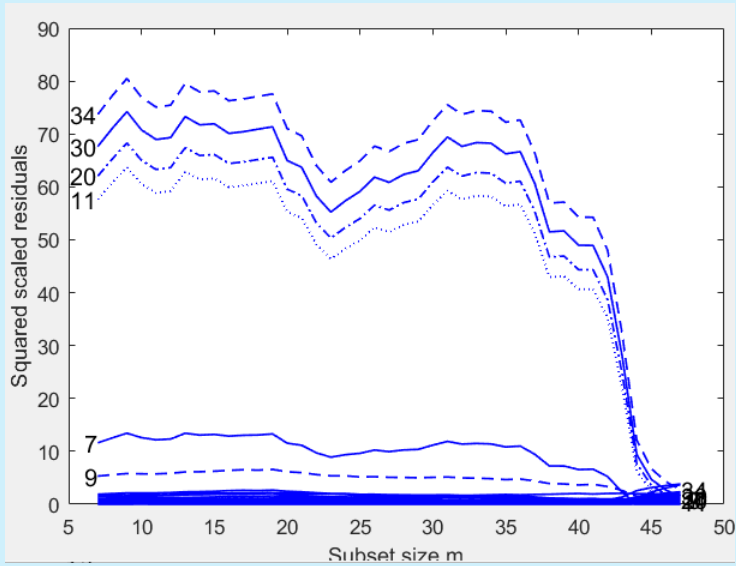


Stars data. Empirical breakdown point (bdp) or efficiency (eff) for MM: five estimators and four ρ functions. The values are for the step before the switch to a non-robust fit

Estimator		Bisquare	Optimal	Hyperbolic	Hampel
S	bdp	0.17	0.17	0.17	0.17
$\tau = 0.85$	bdp	0.14	0.14	0.14	0.16
$\tau = 0.90$	bdp	0.17	0.16	0.16	0.20
$\tau = 0.95$	bdp	0.26	0.21	0.24	— ^a
MM	eff	0.98	0.99	0.97	0.96

^a For this combination of τ and ρ function, only non-robust solutions were obtained during monitoring.

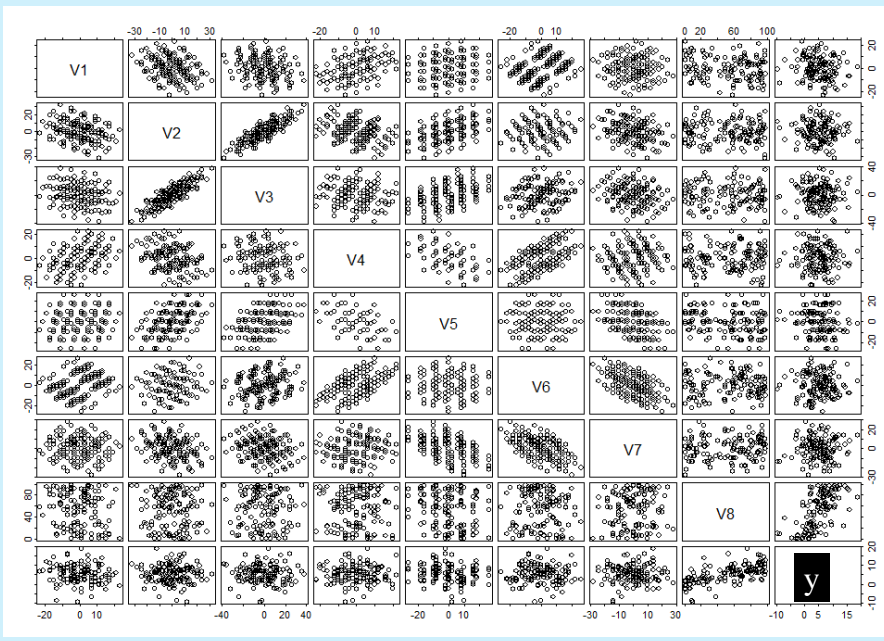
Monitoring of FS residuals



Hawkins data

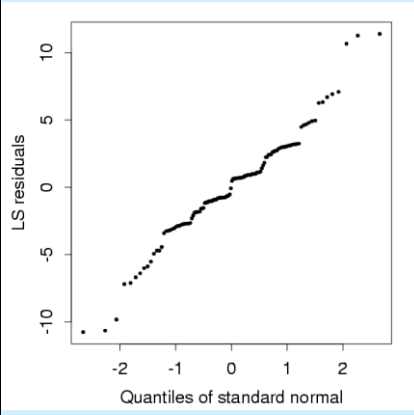
- 128 observations
- 8 explanatory variables

Hawkins data

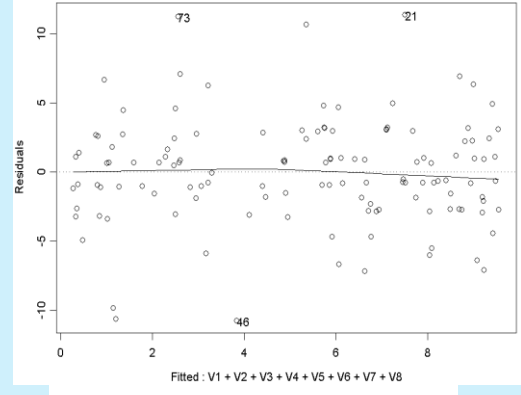


Example of static plots

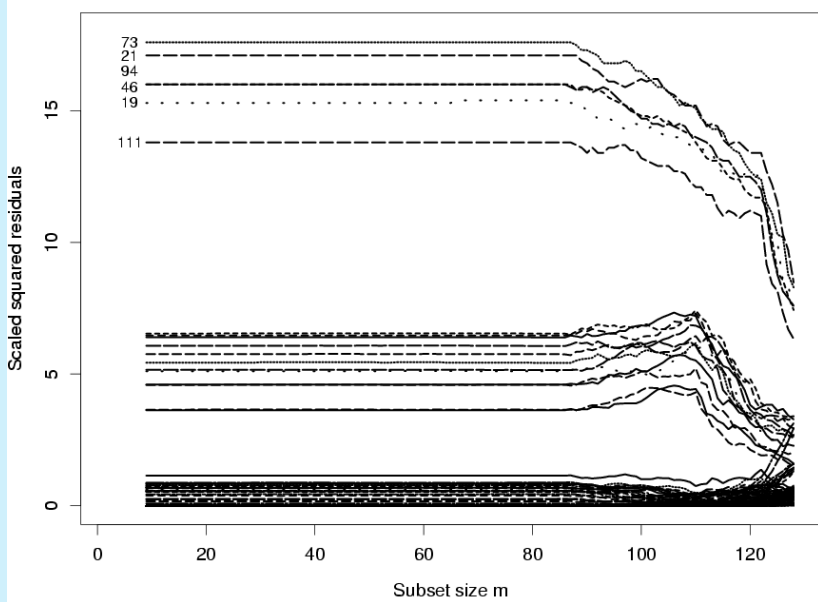
QQ plot of studentized res.



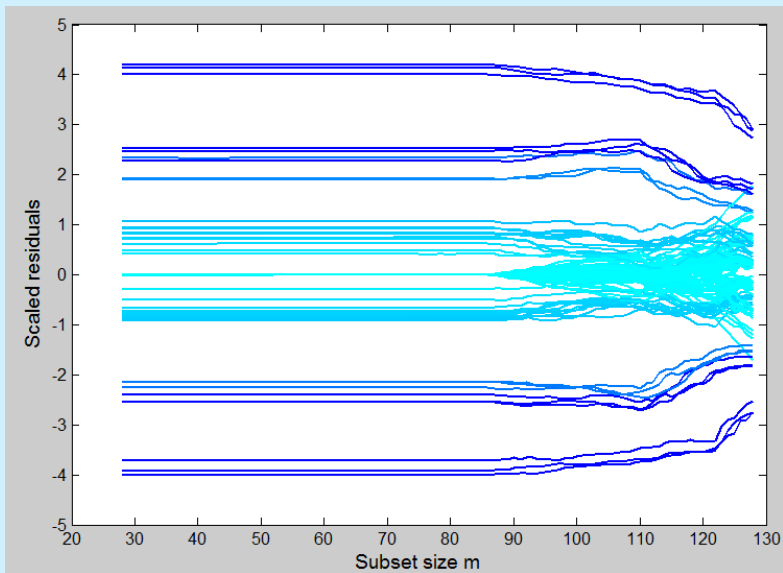
Residuals vs fitted values



HD: monitoring of squared scaled residuals (example of dynamic plot)



HD: monitoring of scaled residuals



- Example of dynamic plot

Other quantities to monitor

- **Maximum studentized residual** among the observations belonging to the subset
- **Minimum deletion residual** among the units not belonging to the subset

Studentized residuals

$$r_i = \frac{e_i}{s\sqrt{(1-h_i)}} = \frac{y_i - \hat{y}_i}{s\sqrt{(1-h_i)}}$$

We monitor: **maximum studentized residual** among the units belonging to the subset

$$r_{[m]} = \max \left| r_{i, S_*^{(m)}} \right| \quad \text{for } i \in S_*^{(m)}$$

$$m = p + 1, \dots, n.$$

Deletion residuals

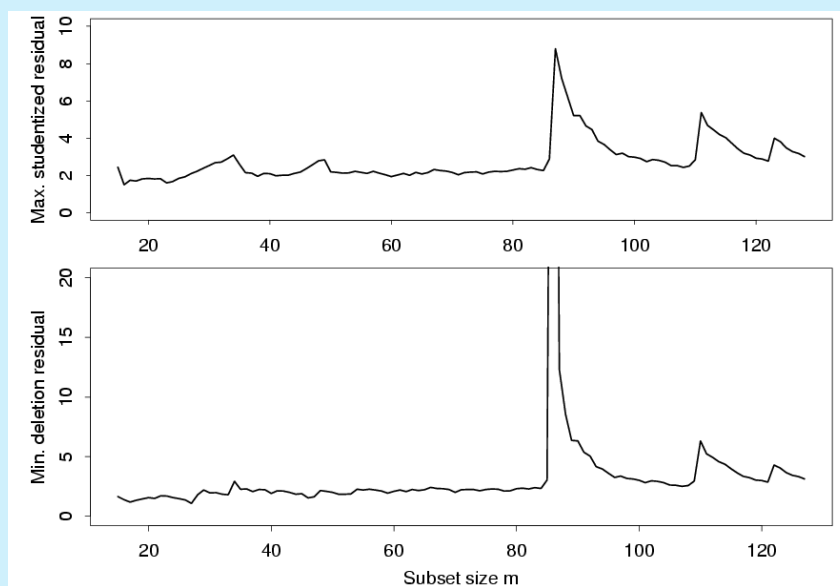
$$r_i^*(m^*) = \frac{y_i - x_i^T \hat{\beta}(m^*)}{\sqrt{s^2(m^*)\{1 + h_i(m^*)\}}}$$

$$i_{\min} = \arg \min |r_i^*(m^*)| \quad \text{for } i \notin S_*^{(m)}$$

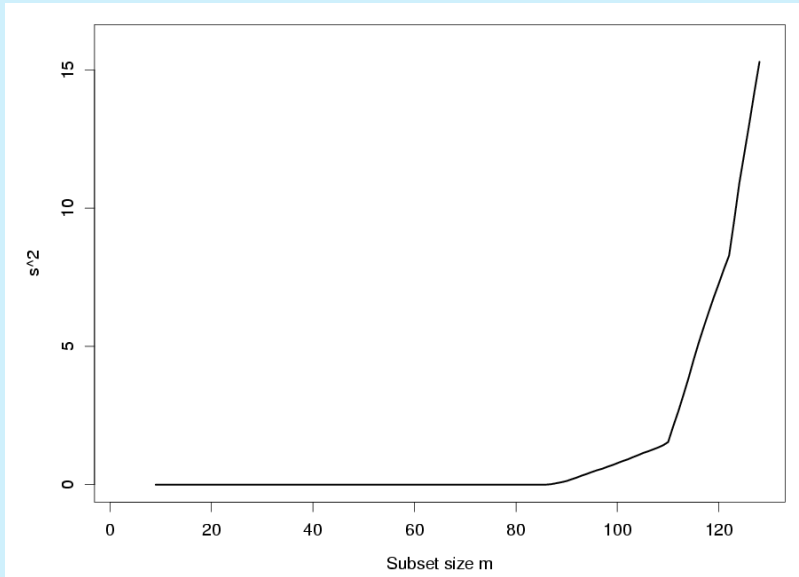
- **We monitor the minimum deletion residual (MDR)** among the units not belonging to the subset

$$r_{i_{\min}}^*(m^*) = \frac{e_{i_{\min}}(m^*)}{\sqrt{s^2(m^*)\{1 + h_{i_{\min}}(m^*)\}}}$$

HD: Monitoring max. stud. res and MDR



HD: Monitoring of s^2



Software

S, MM, LTS, LMS, MCD, MVE → all implemented in FSDA

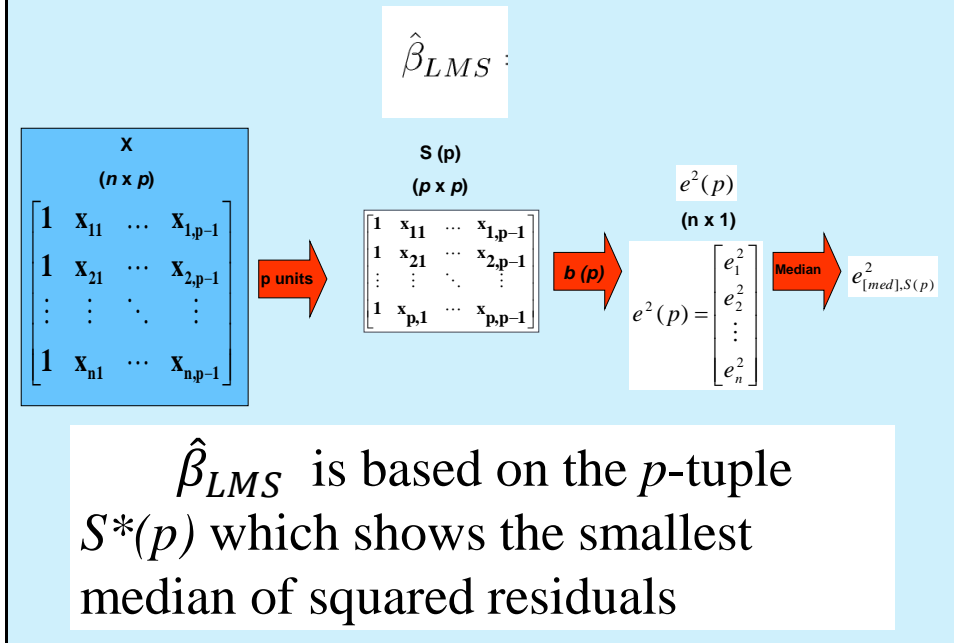
- S estimators
 - Sreg
 - Smult
- MM estimators
 - MMreg
 - MMmult
- LTS and LMS
 - LXS
- MCD, MVE

How to compute

$$\hat{\beta}_{LTS}, \hat{\beta}_{LMS}, \hat{\beta}_S$$

- Computational algorithms are based on subsampling
- They are implemented in function LXS.m
- and
- Sreg.m

Example: algorithms to find



The «heart» of the function LXS.m

```
for i=1:nsamp
```

```
% Extraction of a subset
s=randsample(n,p);

% X and y based on subset|
Xb=X(s,:);
yb=y(s);

% Compute the vector of coefficients using matrice Xb and yb
b=Xb\yb;

% Residuals for all observations using b based on subset
r=y-X*b;

% Squared residuals for all the observations
r2=r.^2;

% Ordering of squared residuals
r2s=sort(r2);
```

The «core» of the function LXS.m

```

% Ordering of squared residuals
r2s=sort(r2);

if lms==1;
    rrob=r2s(h);
else
    rrob=sum(r2s(1:h));
end

if rrob<rmin

    % rmin = smallest ordered quantile or smallest truncated sum.
    rmin=rrob;

    % brob = \beta_lms or \beta_lts
    brob=b;

    % bs = units forming best subset according to lms or lts
    bs=s;
end

```

FAST LTS

- An alternative algorithm to find β_{LTS} is based on the so called concentration steps
- Given a candidate b_1 , let b_2 be the LS estimate based on the data corresponding to the h smallest absolute residuals. The scale estimate based on b_2 is not greater than the estimate based on b_1

$$\hat{\sigma}_2^2 = \sum_{i=1}^h r_{(i)2}^2 \leq \sum_{i \in I} r_{i2}^2 \leq \sum_{i \in I} r_{i1}^2 = \sum_{i=1}^h r_{(i)1}^2 = \hat{\sigma}_1^2.$$

I = the set of indexes corresponding to the smallest h squared residuals based on b_1

FAST LTS (1/2)

1. Find an initial \mathbf{b} by LS using a random subset of p observations
2. Calculate the residuals for the whole data set from a model with the \mathbf{b} just calculated
3. Use the subset of h observations with the lowest squared residuals to estimate a new \mathbf{b} via OLS.
4. Repeat from step 2. Each repeat of steps 2+3 is called a concentration step, or a C-step.

1. Find an initial \mathbf{b} by LS using a random subset of p observations

2. Calculate the residuals for the whole data set from a model with the \mathbf{b} just calculated

3. Use the subset of h observations with the lowest squared residuals to estimate a new \mathbf{b} via LS.

4. Repeat from step 2. Each repeat of steps 2+3 is called a concentration step, or a C-step.

```

%% Step 1
% subsample of p elements
s=randsample(n,p);
Xb=X(s,:);
yb=y(s);
% Regression using just p observations
b=regress(yb,Xb);

%% Step 2
% Residuals for all the observations
r=y-X*b;
% Sort squared residuals
[r2,IX]=sort(r.^2);
% Sum of smallest (n/2) squared residuals
r2LTS=sum(r2(1:(n/2)));
disp(r2LTS)

for i=1:6
    %% Step 3
    % Find the indexes of the units with the smallest n/2 squared residuals
    IX1=IX(1:n/2);
    % Find subset of y
    y1=y(IX1);
    % Find subset of X
    X1=X(IX1,:);
    % Find estimate of beta just using subset of y and X
    bet=regress(y1,X1);

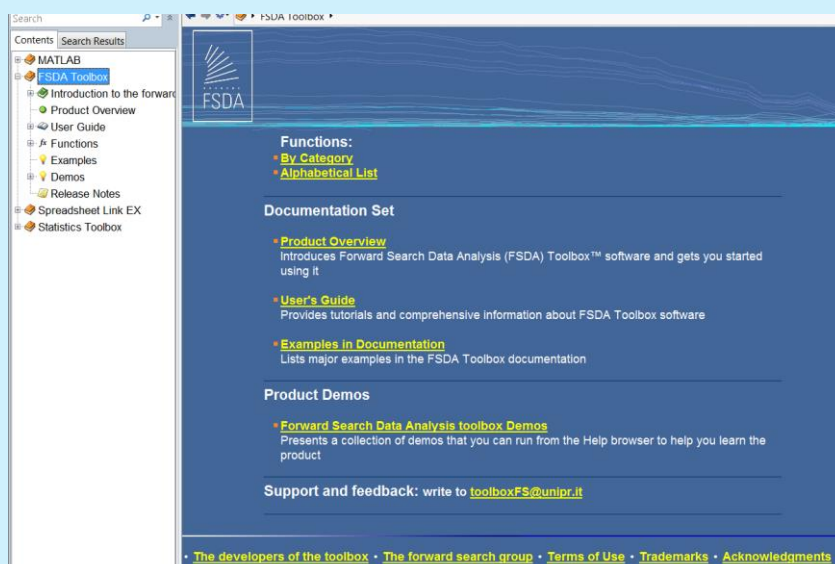
    %% Repeat step 2
    % r = residuals for all observations
    r=y-X*bet;
    % sort squared residuals
    [r2,IX]=sort(r.^2);
    % sum of the smallest squared residuals (truncated sum)
    r2LTS=sum(r2(1:n/2));
    % show value of truncated sum at each iteration
    disp(r2LTS);

```

FAST LTS (2/2)

- The algorithm then repeats the entire process (steps 1-4) a number of times (or until convergence)
- Each of these last repetitions yields an estimate of \mathbf{b} .
- The final estimate is the best \mathbf{b} of these ‘best’, that is the \mathbf{b} with the lowest sum of h squared residuals.

Dynamic visualization: the FSDA toolbox. Downloadable from <http://www.riani.it>



The screenshot shows the MATLAB interface with the FSDA toolbox documentation open. The left sidebar contains a navigation menu with the following items: MATLAB, FSDA Toolbox, Introduction to the forward search, Product Overview, User Guide, Functions, Examples, Demos, Release Notes, Spreadsheet Link EX, and Statistics Toolbox. The main content area displays the FSDA logo and the following sections:

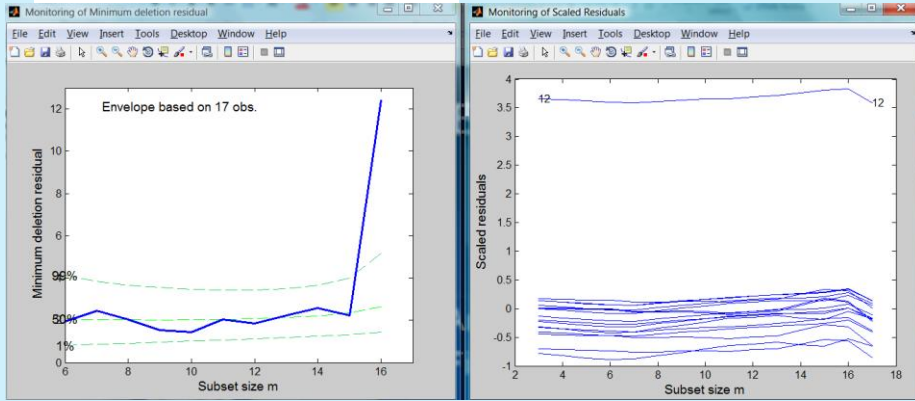
- Functions:**
 - [By Category](#)
 - [Alphabetical List](#)
- Documentation Set**
 - [Product Overview](#)
Introduces Forward Search Data Analysis (FSDA) Toolbox™ software and gets you started using it
 - [User's Guide](#)
Provides tutorials and comprehensive information about FSDA Toolbox software
 - [Examples in Documentation](#)
Lists major examples in the FSDA Toolbox documentation
- Product Demos**
 - [Forward Search Data Analysis toolbox Demos](#)
Presents a collection of demos that you can run from the Help browser to help you learn the product
- Support and feedback:** write to toolboxFS@uniipr.it

At the bottom, there are links for: [The developers of the toolbox](#) • [The forward search group](#) • [Terms of Use](#) • [Trademarks](#) • [Acknowledgments](#)

Forbes data

- Minimum deletion residual (mdrplot)

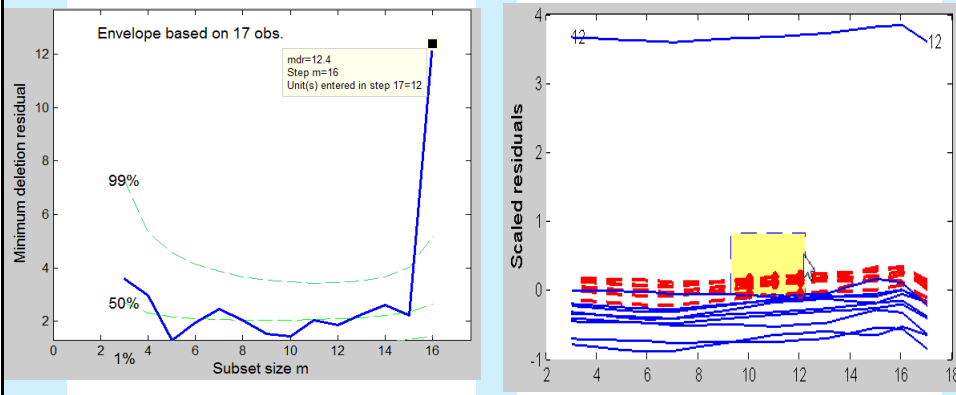
- Scaled residuals (resplot)



Brushing and linking

- Option datatooltip

- Option databrush



Dynamic visualization: the FSDA toolbox

Search

Contents Search Results

- MATLAB
- FSDA Toolbox
 - Introduction to the forward search philosophy of data
 - Product Overview
 - User Guide
 - Functions
 - Examples
 - Demos**
 - Release Notes
 - Spreadsheet Link EX
 - Statistics Toolbox

FSDA Toolbox DEMOS

FSDA Toolbox™ provides statisticians, engineers, scientists, researchers, financial analysts with a comprehensive set of tools to assess and understand Forward Search Data Analysis Toolbox™ software includes functions and interactive tools for analyzing and modeling data, learning and teaching statistics.

The Forward Search Data Analysis Toolbox™ supports a set of routines to develop robust and efficient regression analysis. In addition, it offers a rich set of graphical tools which enable us to explore the connection in the various features of the different forward plots.

All Forward Search Data Analysis Toolbox™ functions are written in the open MATLAB® language. This means that you can inspect the algorithms, code, and create your own custom functions.

Product |

Regression

- Hawkins data (4 min, 0 sec)
- Multiple regression data (6 min, 17 sec)
- Fishery data (5 min, 52 sec)

Multivariate Analysis

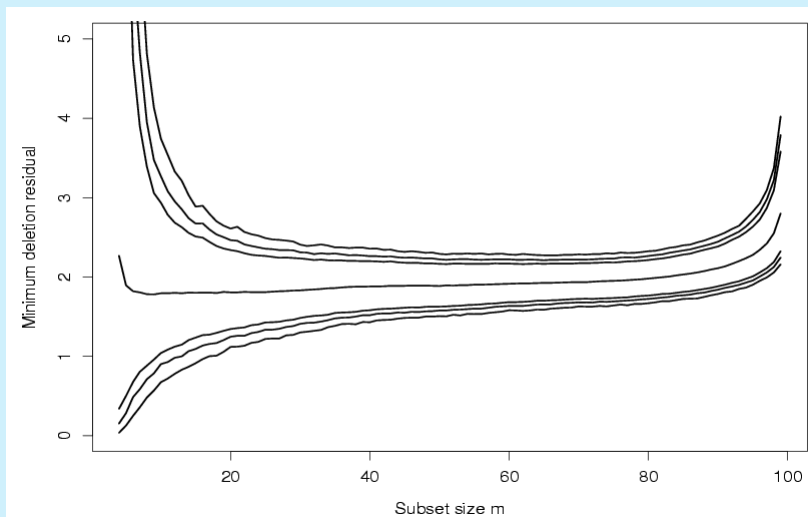
To be implemented in future releases of the product

Two approaches to the Forward Search

- Descriptive point of view 1998-2004
- Inferential point of view 2005-now
 - Forward confidence bands for minimum deletion residual (MDR)
 - Strategy to keep into account multiple testing

Example of the empirical distribution of MDR ($n=100$, $p=3$)

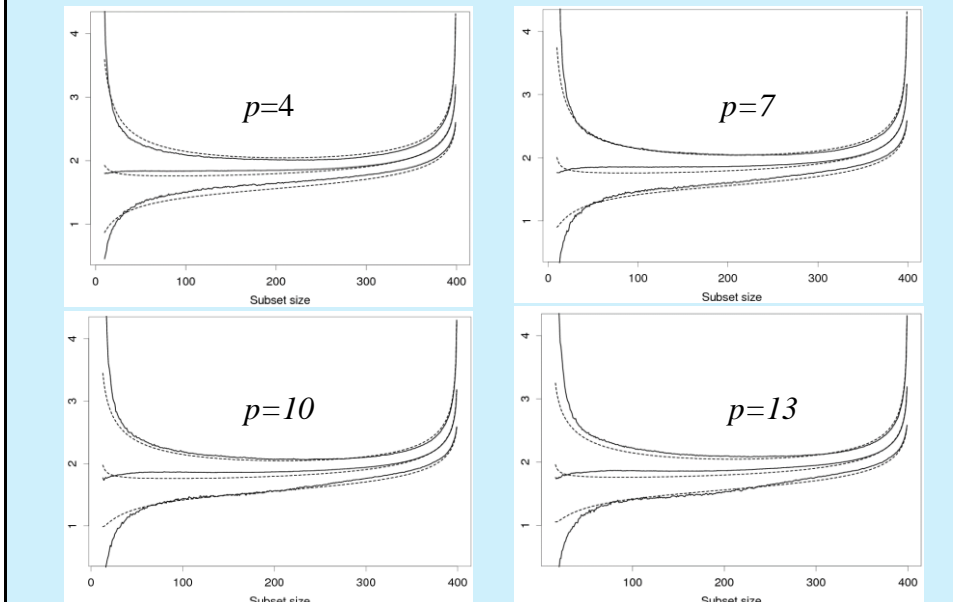
(1%, 2.5%, 5%, 50%, 95%, 97.5% and 99% bands)



How to approximate the forward distribution of the MDR

- Method 1: truncated samples
- Method 2: quantiles
- Method 3: exact order statistics (Riani Atkinson and Cerioli, 2009; JRSSb)

Forward MDR: comparison between theoretical and empirical. Quantiles 1, 50 and 99%. $n=400$

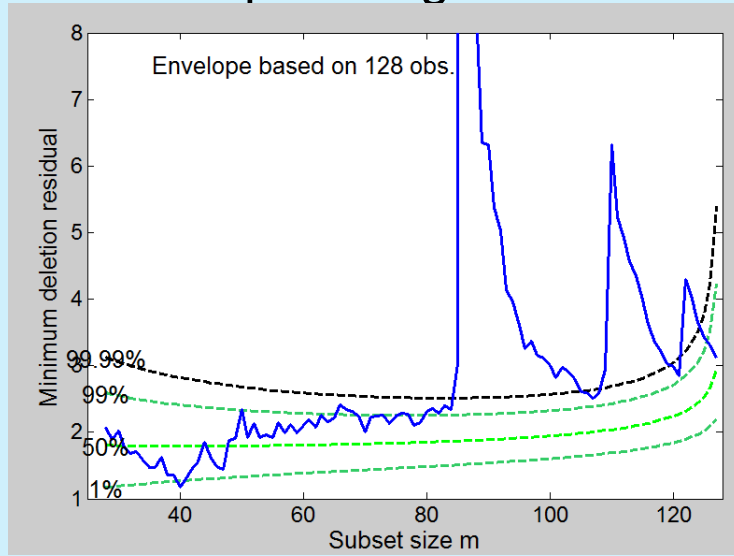


Automatic outlier detection procedure (file FSR.m)

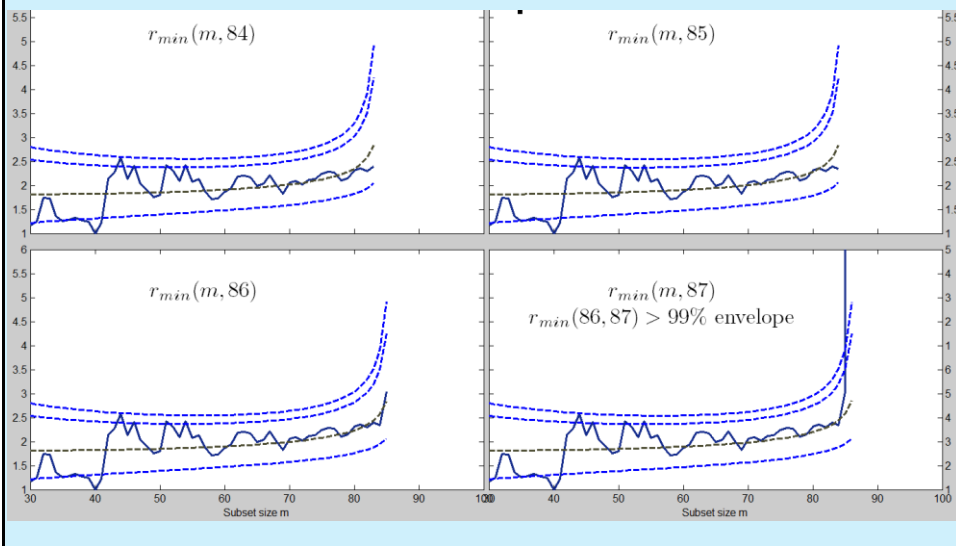
- Part I: signal detection
 - upper envelope exceedance
- Part II: signal validation
 - envelope superimposition

- Riani, Atkinson and Cerioli (2009), *JRSSB*

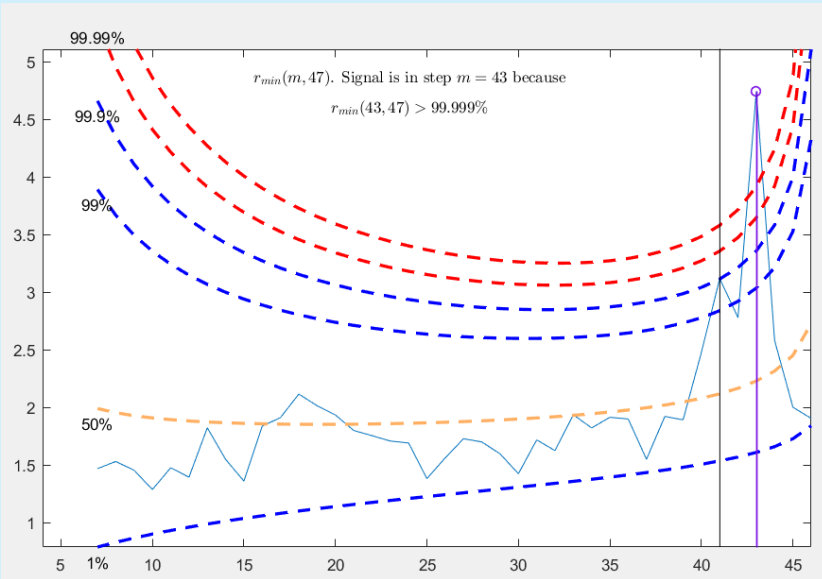
Automatic procedure for outlier detection: part I signal detection



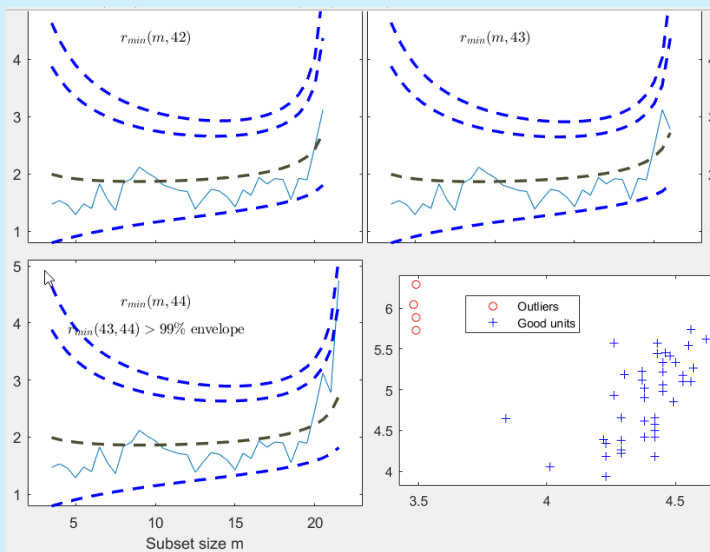
Automatic procedure for outlier detection: part II signal validation



Stars data. automatic FS



Stars data: envelope resuperimposition



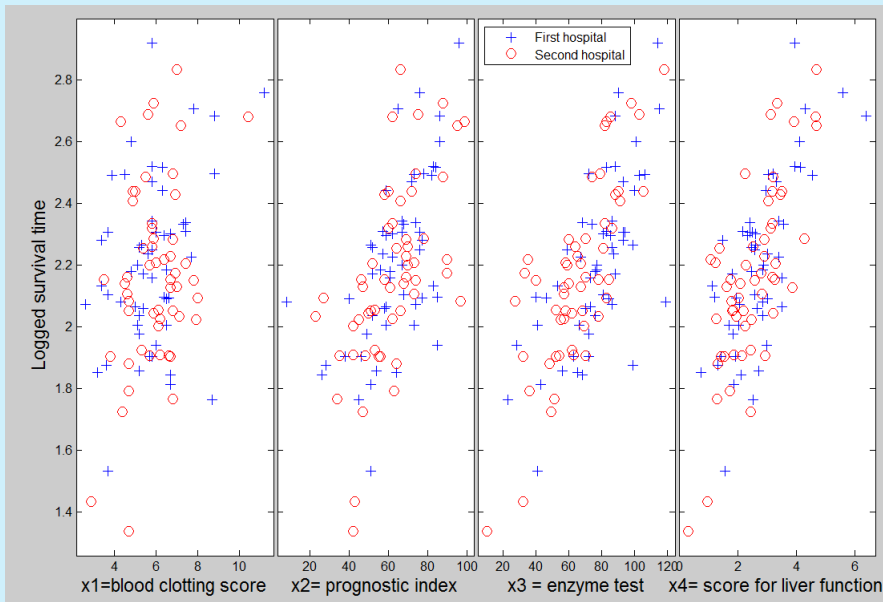
An example with two overlapping groups: hospital data (Neter et al. 1996)

- y = logged survival time of 54 patients undergoing liver surgery
- Other 54 observations are introduced to check the fitted model
- Neter *et al.*: “there is no systematic difference between the two sets”.

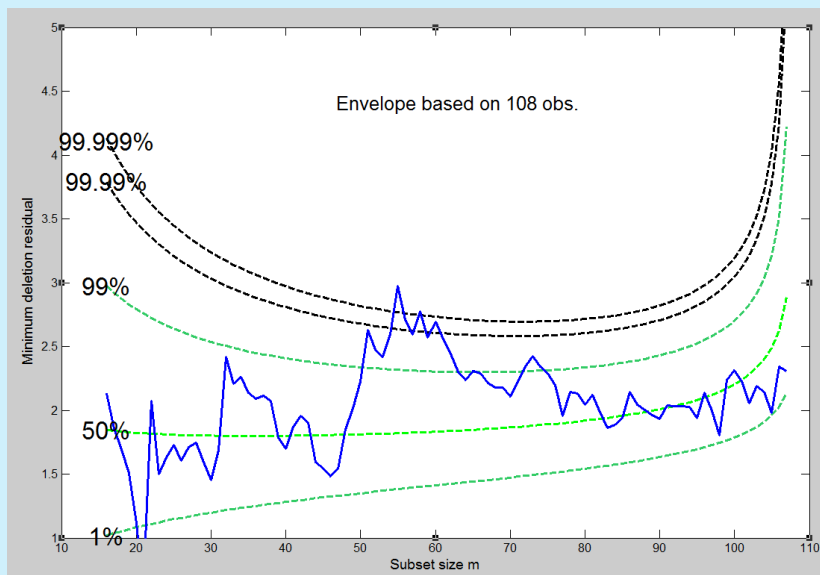
Data description

- y = logged survival time of 108 patients undergoing liver surgery
- x_1 = blood clotting score
- x_2 = prognostic index
- x_3 = enzyme test
- x_4 = score for liver function

Hospital data: yX plot

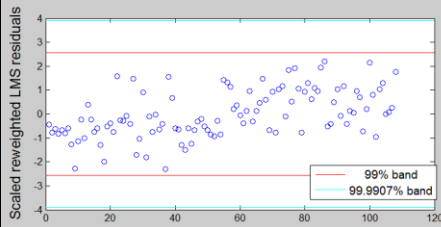
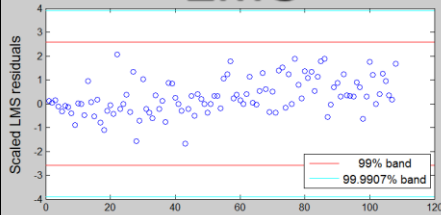


Hospital Data: forward plot of MDR with envelopes
 The difference in the two groups of observations is highly significant

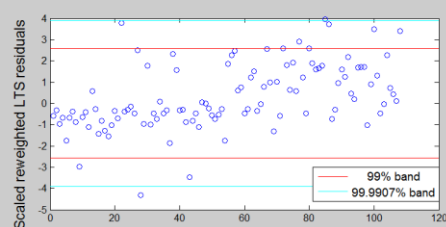
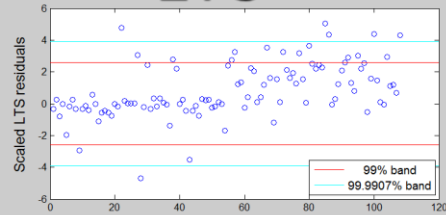


Hospital data: traditional robust analysis

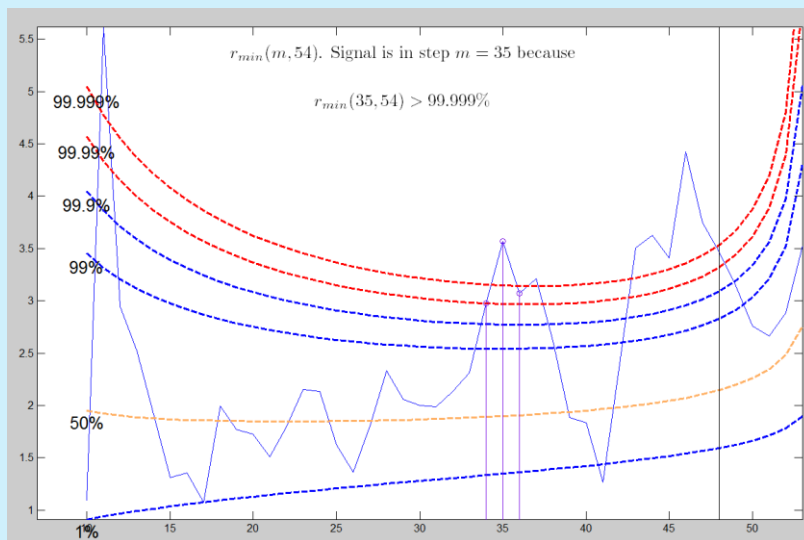
LMS



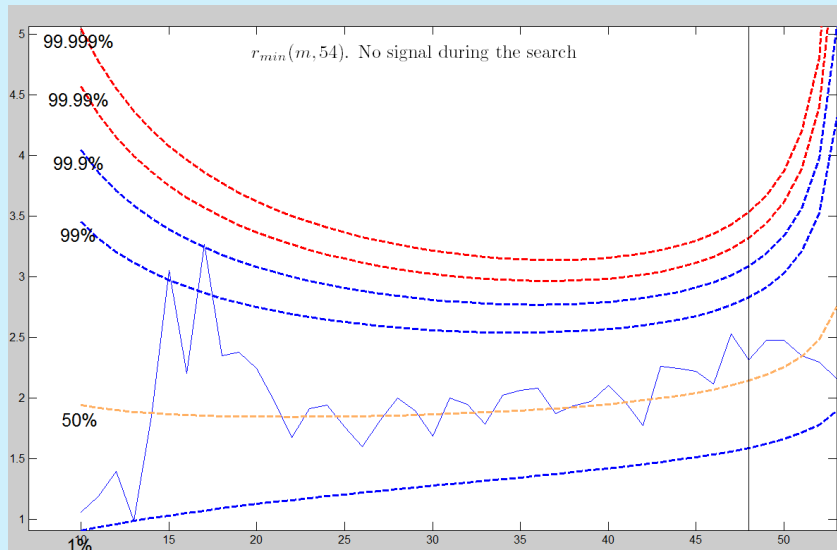
LTS



Monitoring MDR for the two groups separately: hospital 1



Monitoring MDR for the two groups separately: hospital 2



Bank Data

- A (much) more complicated example - there is no simple model for all the data
- The aberrant observations do not seem to form a simple cluster
- 1,949 observations on the amount of money made from personal banking customers.
- The 13 explanatory variables describe the services used; all are discrete, one binary.
- Which activities are profitable?

Bank data: the 13 explanatory variables

Bank data: the thirteen explanatory variables

Variable number	Description	Number of zeroes
1	Personal loans	1666
2	Financing and hire-purchase	1529
3	Mortgages	1734
4	Life insurance	1503
5	Share account	435
6	Bond account	987
7	Current account	27
8	Salary deposits	742
9	Debit cards	1030
10	Credit cards	1003
11	Telephone banking	1459
12	Domestic direct debits	426
13	Money transfers	1596

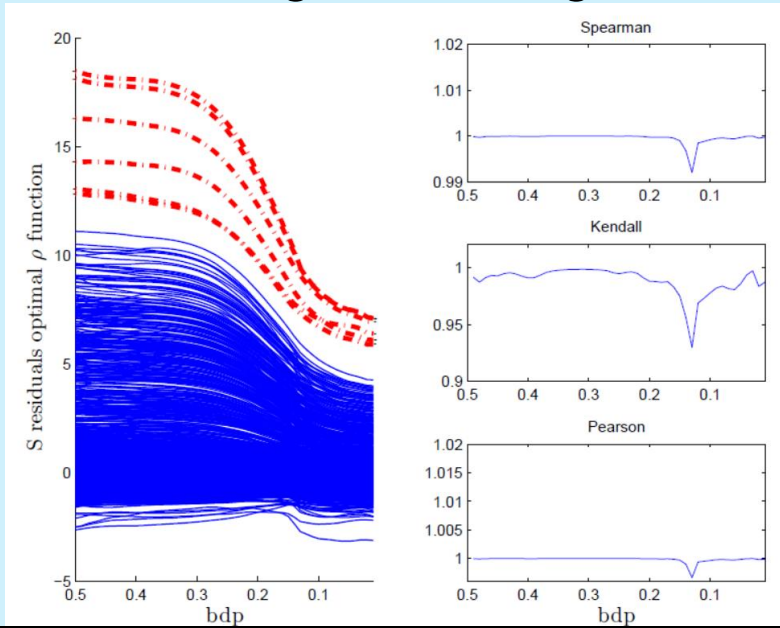
First 10 lines of the dataset

- All explanatory variables are discrete, taking values 0, 1, 2,...

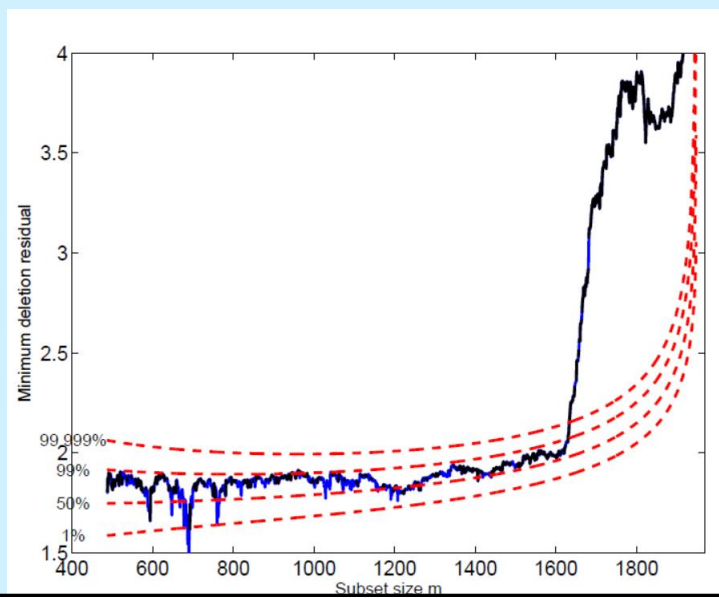
x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	y
0	0	0	0	0	0	0	3	1	0	0	0	3	0 506.0833
2	0	0	0	1	0	2	2	1	1	0	0	2	0 165.5833
0	0	1	0	2	4	4	2	1	2	0	0	10	0 262.25
0	0	0	0	1	2	2	2	2	0	0	0	7	0 92.33333
0	0	0	0	0	0	2	0	0	0	0	0	1	0 613.05
2	0	0	0	1	2	2	2	1	0	0	0	7	0 339.3333
0	1	0	0	2	3	2	0	1	1	1	1	7	0 247.5833
2	1	0	0	2	2	2	0	0	0	0	0	1	0 294.8333
0	1	0	0	0	0	2	0	0	1	2	1	8	1 315.5

- x1 = personal loans
- ...
- x7 = current account
- ...

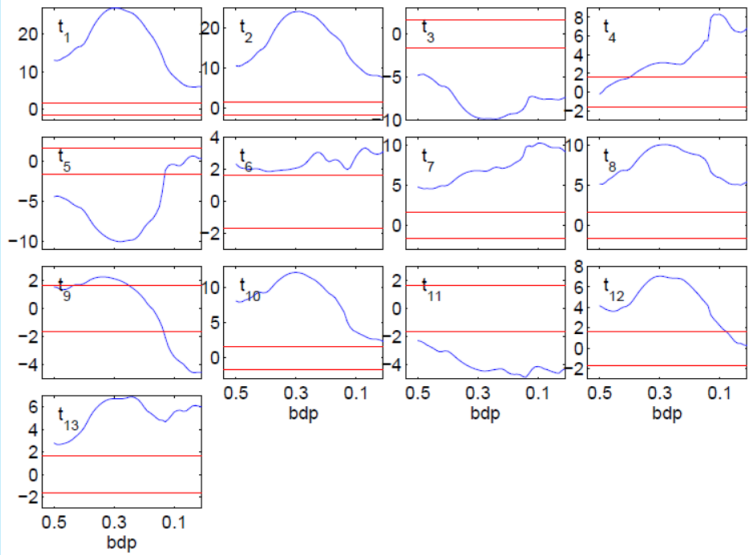
Monitoring robust regression



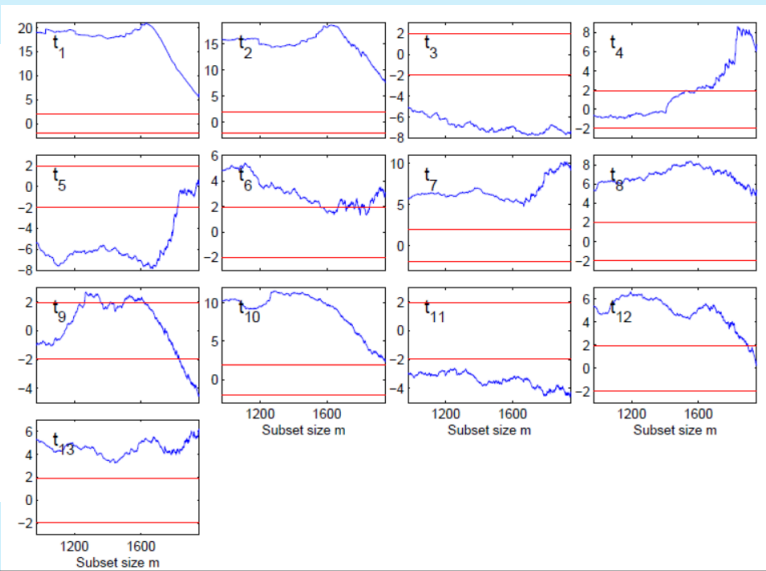
Forward plot of minimum deletion residual



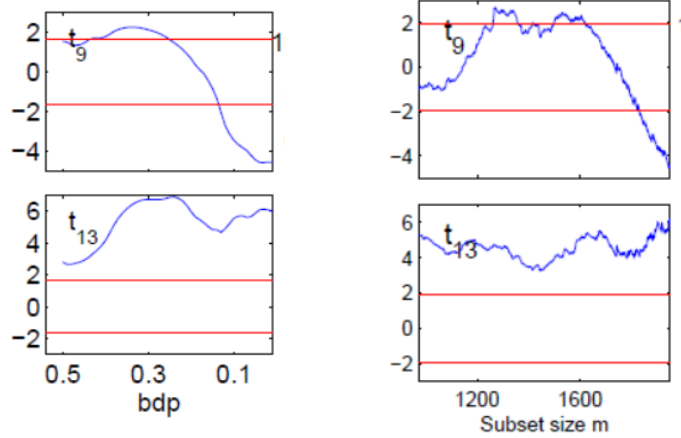
Monitoring t stat as a function of bdp



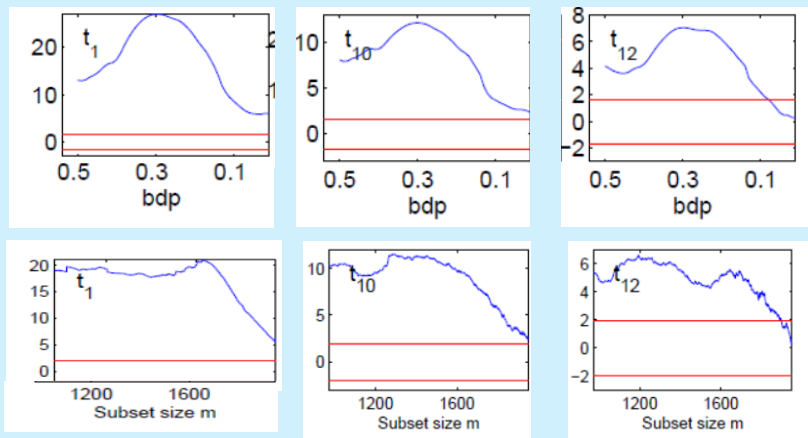
Monitoring t stat as a function of subset size



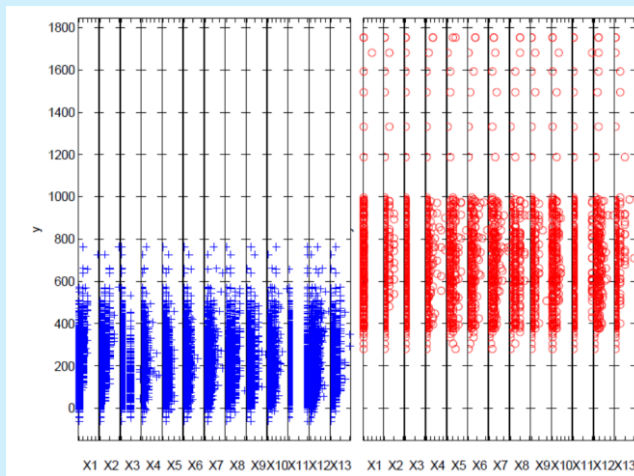
Monitoring of t stat as function of bdp (S estimator) or subset size (FS estimator)



Monitoring of t stat as function of bdp (S estimator) or subset size (FS estimator)



Bank data. Scatterplots of y against individual explanatory variables for the two parts of the data. Left-hand panel, the main body of the data. Right-hand panel, the remaining, somewhat different, 255 observations.

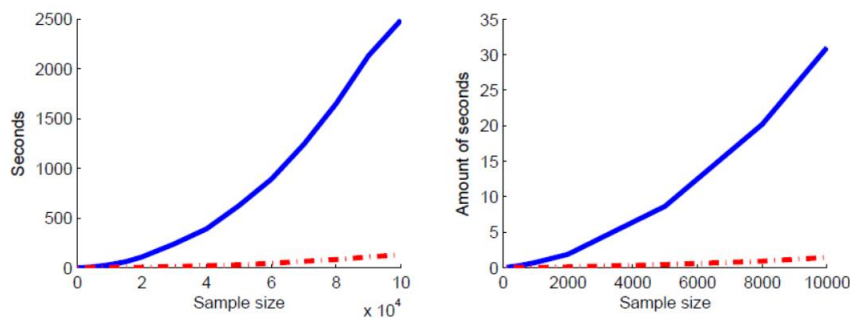


Computational time of
the forward search?

Fast efficient updating

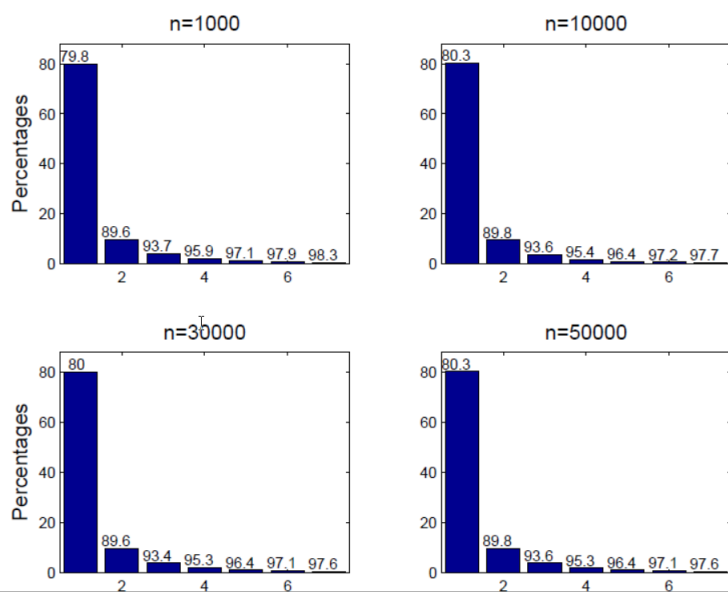
- Theorem: Assume that just one unit joins the subset from step m to step $m+1$. Then, S_{m+1} can be found with $2n+1$ logical operations and the computation of a sum.
- On the other hand, if $k > 1$ new units join the subset, it is necessary to compute k additional minima and k additional logical operations to find S_{m+1}

The time is now a roughly linear function of n .



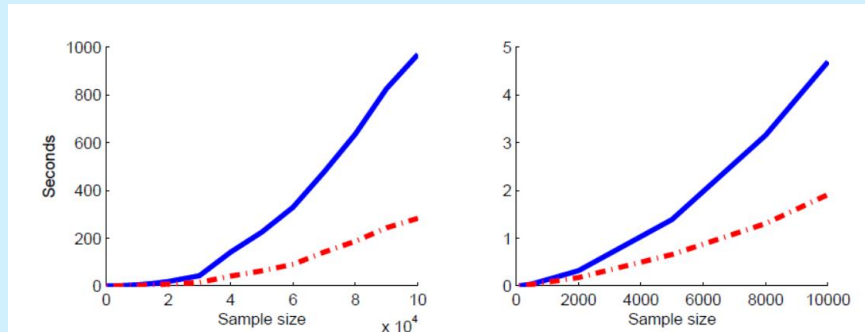
In passing from a subset of size m to a subset of size $m+1$, on average, how many new units join the subset?

X axis = number of new units which join subset in passing from m to $m+1$



New routines for efficient recursive updating

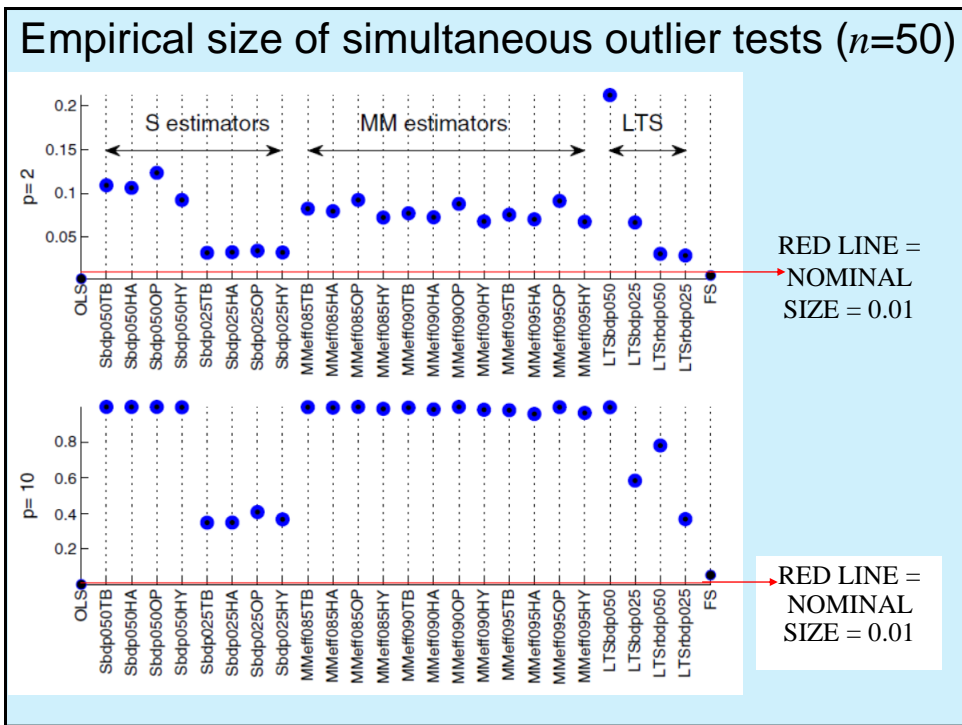
Comparison of computation time between the currently available version of the FSDA software (solid line) and the new routines



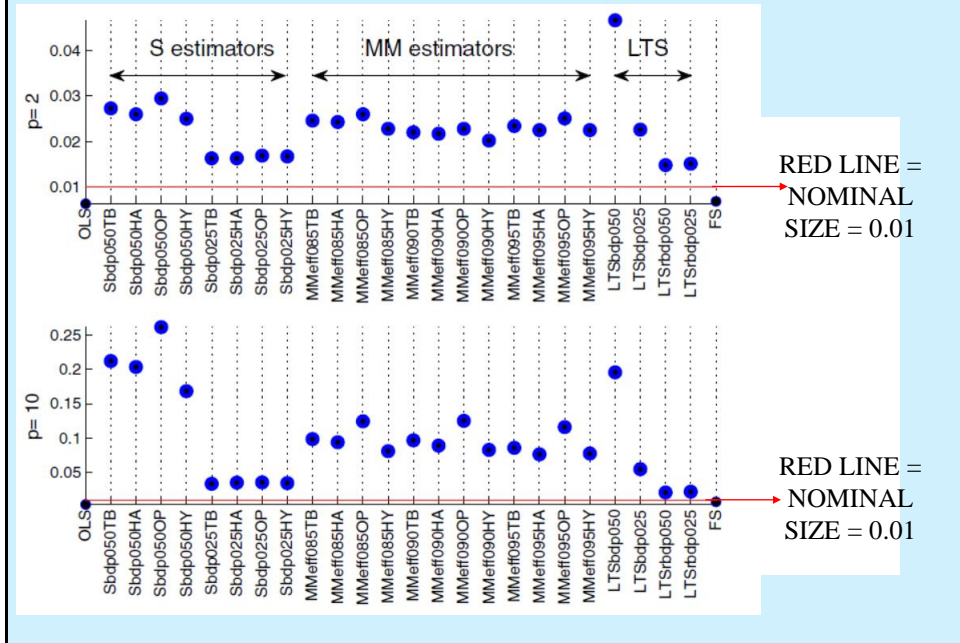
Theoretical properties of the FS

- Cerioli Farcomeni, and Riani (2014) show that the estimates obtained at step m and are **strongly consistent** under the null model and have **breakdown point** $1 - m/n$ under contamination: **the FS yields consistent high-breakdown estimators, but with adaptive breakdown point**

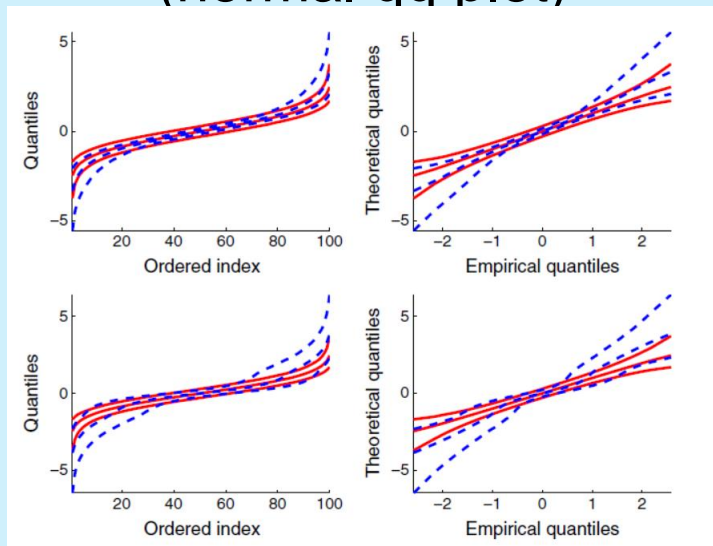
Nominal versus empirical size of robust estimators



Empirical size of simultaneous outlier tests ($n=200$)



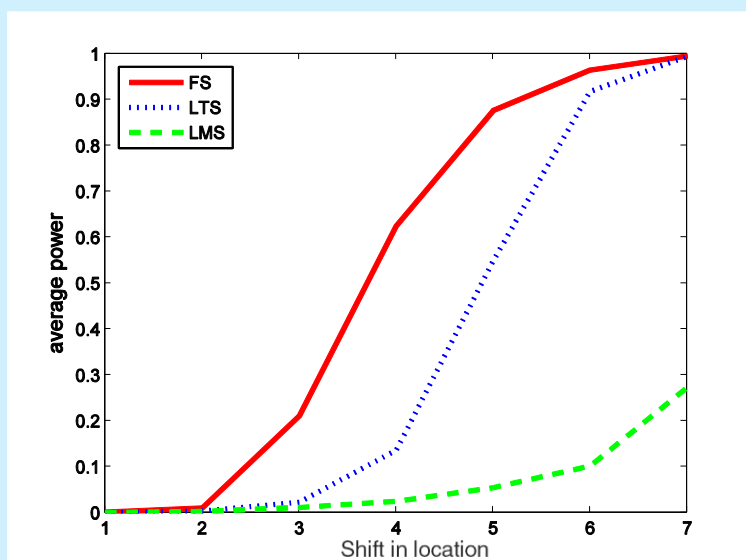
Distribution of robust residuals (normal qq plot)



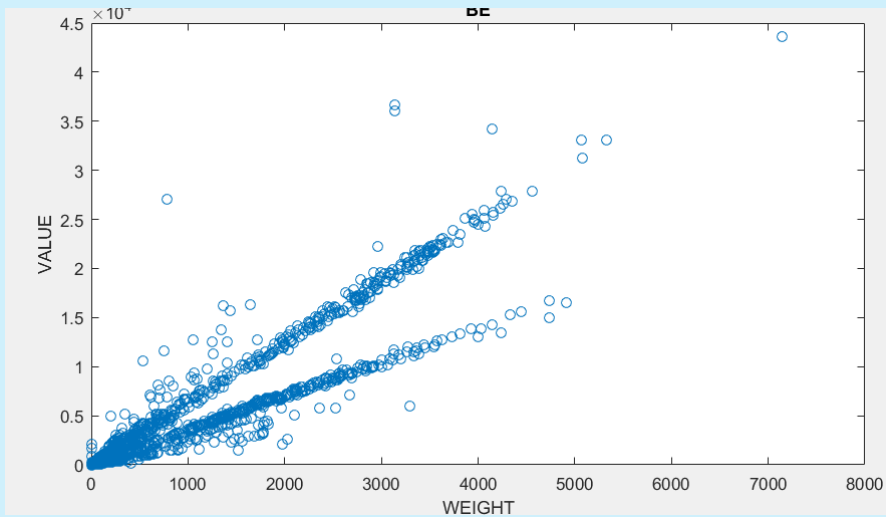
How to compare robust estimators

- **Average power:** average percentage of true outliers relative to the contaminated observations taken over all iterations.
- **Simultaneous power:** average number of outliers (both true and false) taken over all iterations.
- **Family wise error rate:** average number of iterations where at least one false outlier has been detected.
- **False discovery rate:** average percentage of false outliers relative to all outliers (both true and false) taken over all iterations.
- **Proportion of declared outliers in good data:** average percentage of false outliers relative to all iterations, divided by the number of non-contaminated observations.

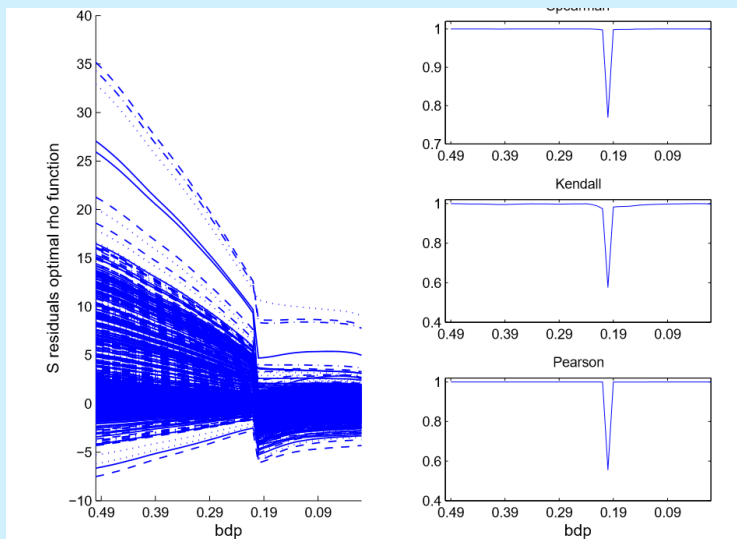
Regression: size and power comparison



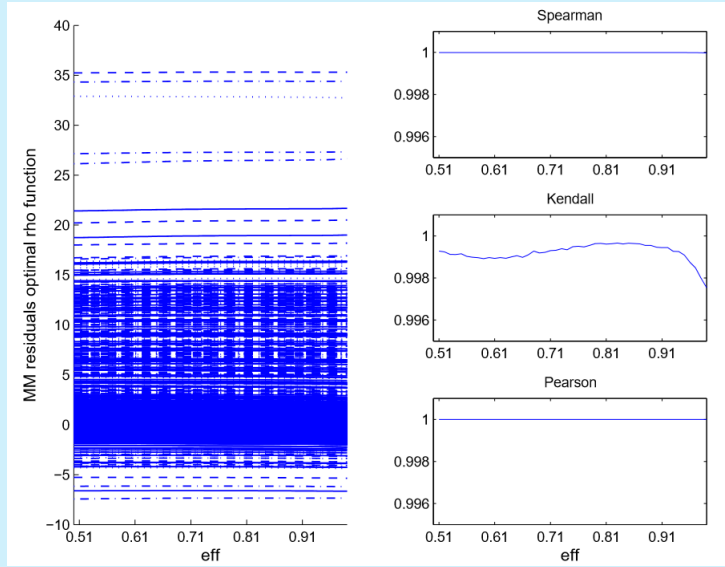
An example with international trade data



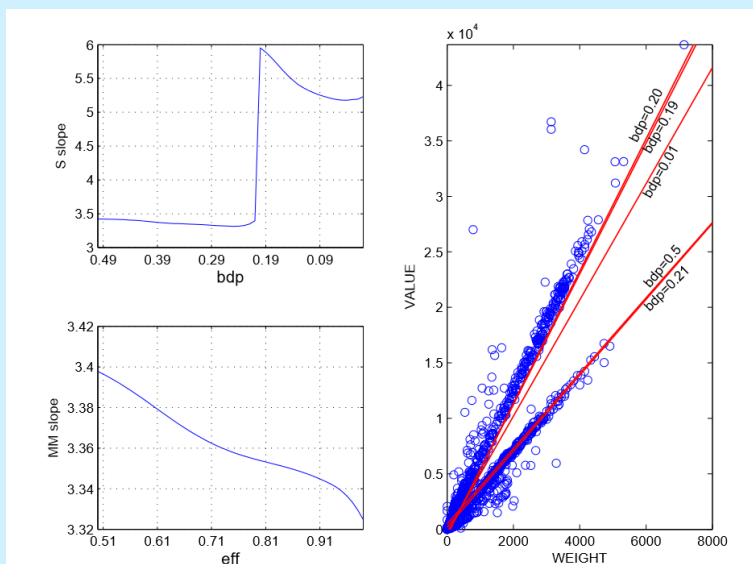
Monitoring of S residuals



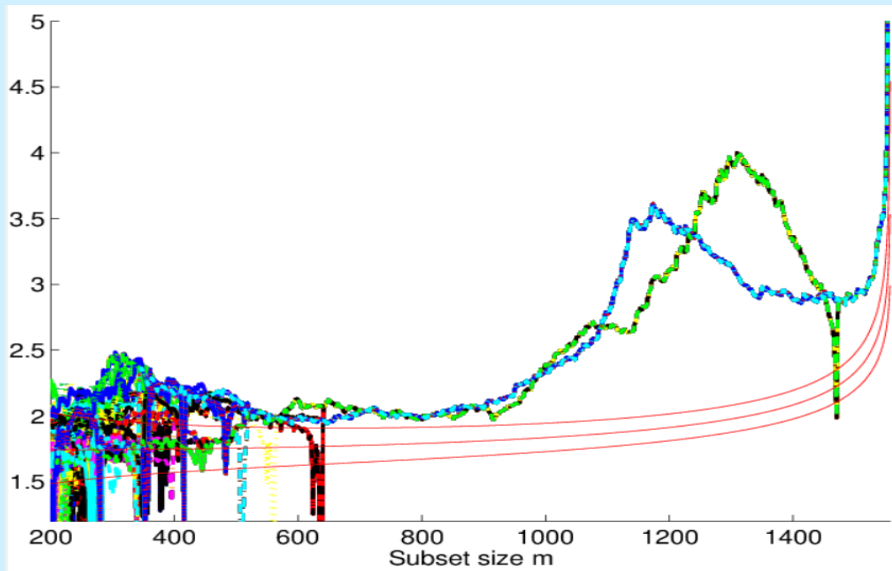
Monitoring of MM residuals



Monitoring of S and MM slope

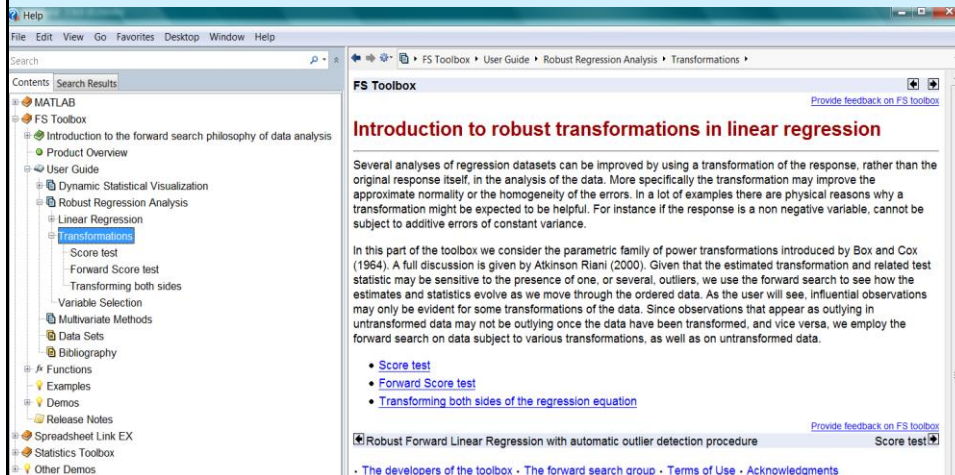


Random starts monitoring



Robust Transformations in Regression

Robust transformations in regression



Transformations of the Response

- Simple power transformation y^λ
- Continuous power transformation (Box Cox)

$$(y^\lambda - 1) / \lambda$$

Gives $\log y$ at $\lambda = 0$

The most widely used transformations are

$y(\lambda) = (y^\lambda - 1) / \lambda$	λ	transformation
	1	none
	0.5	square root
	0	logarithmic
	-1	reciprocal

Box Cox transformation

$$y(\lambda) = (y^\lambda - 1) / \lambda$$

Normalized Box Cox transformation

$$z(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda G^{\lambda-1}} & \lambda \neq 0 \\ G \log y & \lambda = 0 \end{cases}$$

G is the geometric mean of the observations

Likelihood of transformed observations

$$(2\pi\sigma^2)^{-n/2} \exp\{-(y(\lambda) - X\beta)^T(y(\lambda) - X\beta)/2\sigma^2\}J,$$

$$J = \prod_{i=1}^n \left| \frac{\partial y_i(\lambda)}{\partial y_i} \right|$$

The Jacobian allows for the change of scale of the response due to the transformation

In our example the Jacobian is the determinant of the matrix

$$J = \begin{vmatrix} \frac{\partial y_1(\lambda)}{\partial y_1} & \frac{\partial y_1(\lambda)}{\partial y_2} & \dots & \frac{\partial y_1(\lambda)}{\partial y_n} \\ \frac{\partial y_2(\lambda)}{\partial y_1} & \frac{\partial y_2(\lambda)}{\partial y_2} & \dots & \frac{\partial y_2(\lambda)}{\partial y_n} \\ \dots & \dots & \ddots & \vdots \\ \frac{\partial y_n(\lambda)}{\partial y_1} & \frac{\partial y_n(\lambda)}{\partial y_2} & \dots & \frac{\partial y_n(\lambda)}{\partial y_n} \end{vmatrix}$$

In our example the Jacobian is the determinant of the matrix

$$= \begin{vmatrix} y_1^{\lambda-1} & 0 & \dots & 0 \\ 0 & y_2^{\lambda-1} & \dots & 0 \\ \dots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & y_n^{\lambda-1} \end{vmatrix}$$

Expression for the Jacobian

$$J = \prod_{i=1}^n |y_i^{\lambda-1}|$$

$$J = G^{n(\lambda-1)}$$

A simpler but identical form for the likelihood is obtained with the normalized transformation defined as

$$z(\lambda) = y(\lambda) / J^{1/n},$$

$$z(\lambda) = \frac{y^\lambda - 1}{\lambda G^{\lambda-1}}$$

for which the Jacobian is 1

The likelihood becomes

$$(2\pi\sigma^2)^{-n/2} \exp\{-(z(\lambda) - X\beta)^T(z(\lambda) - X\beta)/2\sigma^2\}$$

For fixed λ the likelihood is maximized by the least squares estimate

$$\hat{\beta}(\lambda) = (X^T X)^{-1} X^T z(\lambda)$$

The residual sum of squares of the $z(\lambda)$ is

$$R(\lambda) = z(\lambda)^T (I - H)z(\lambda) = z(\lambda)^T Az(\lambda)$$

The maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2(\lambda) = R(\lambda)/n.$$

The maximized log likelihood is

$$L_{\max}(\lambda) = -(n/2) \log\{R(\lambda)/(n - p)\}$$

So that the estimate of λ minimizes $R(\lambda)$

$$R(\lambda) = z(\lambda)^T (I - H)z(\lambda) = z(\lambda)^T Az(\lambda)$$

- The model to be fitted is

$$z(\lambda) = X\beta + \varepsilon$$

- That is ordinary least squares with response $z(\lambda)$
- For fixed λ use LS
- Compare different λ by RSS $z(\lambda)$
- Or by likelihood ratio test
- Both require search over values of λ

Likelihood ratio test

$$T_{LR} = 2\{L_{\max}(\hat{\lambda}) - L_{\max}(\lambda_0)\}$$

$$= n \log\{R(\lambda_0)/R(\hat{\lambda})\}$$

How is it possible to avoid to compute the MLE of λ ?

→ *score test* for transformation

Score test for transformation

$$z(\lambda) = x^T \beta + \varepsilon$$

$$z(\lambda) \cong z(\lambda_0) + (\lambda - \lambda_0) \left. \frac{\partial z(\lambda)}{\partial \lambda} \right|_{\lambda=\lambda_0}$$

$$z(\lambda) \cong z(\lambda_0) + (\lambda - \lambda_0) w(\lambda_0)$$

$$z(\lambda_0) = x^T \beta - (\lambda - \lambda_0) w(\lambda_0) + \varepsilon$$

The score test ($H_0: \lambda = \lambda_0$) is the t -statistic on the constructed variable $w(\lambda_0)$

How to construct $w(\lambda)$ from y

$$z(\lambda_0) = x^T \beta - (\lambda - \lambda_0) w(\lambda_0) + \varepsilon$$

$$\begin{aligned} w(\lambda) = \frac{dz(\lambda)}{d\lambda} &= \frac{\lambda \dot{y}^{\lambda-1} y^\lambda \log y - (\dot{y}^{\lambda-1} + \lambda \dot{y}^{\lambda-1} \log \dot{y})(y^\lambda - 1)}{(\lambda \dot{y}^{\lambda-1})^2} \\ &= \frac{y^\lambda \log y}{\lambda \dot{y}^{\lambda-1}} - \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} (1/\lambda + \log \dot{y}). \end{aligned} \quad (4.29)$$

Score test is implemented in function `score.m`

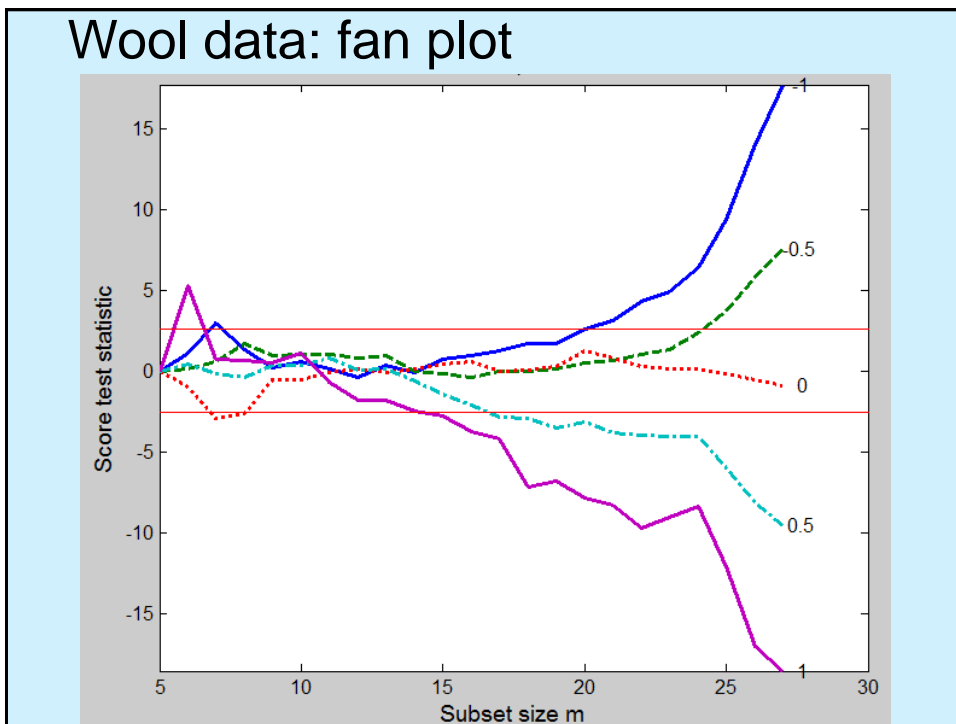
```
z=(y.^la(i)-1)/(la(i)*G^(la(i)-1));
w=(y.^la(i).*log(y)-(y.^la(i)-1)*(1/la(i)+log(G)))/(la(i)*G^(la(i)-1));
```

Disadvantage of the score test

- Does not allow identification of the individual observations
- Is not robust to the presence of atypical observations

FAN PLOT

- Fan plot: forward plot of t test for the constructed variable $w(\lambda)$ for five values of λ : - 1, -0.5, 0, 0.5 and +1. Usually enough
- FSRfan.m implements the monitoring of the score test
- fanplot.m produces the fan plot
- Ex. wool data: factorial experiment, 3 expl. variables

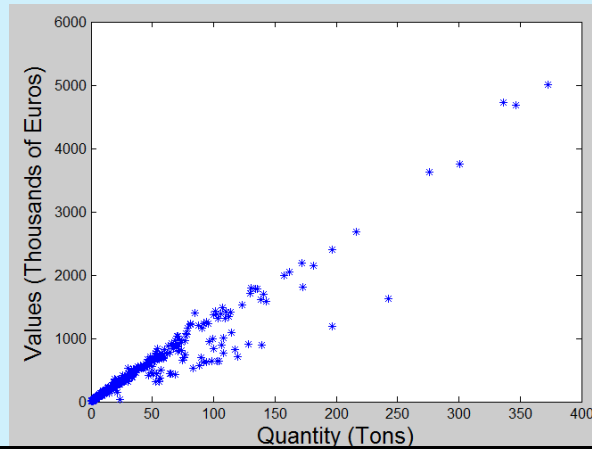


Fan Plot

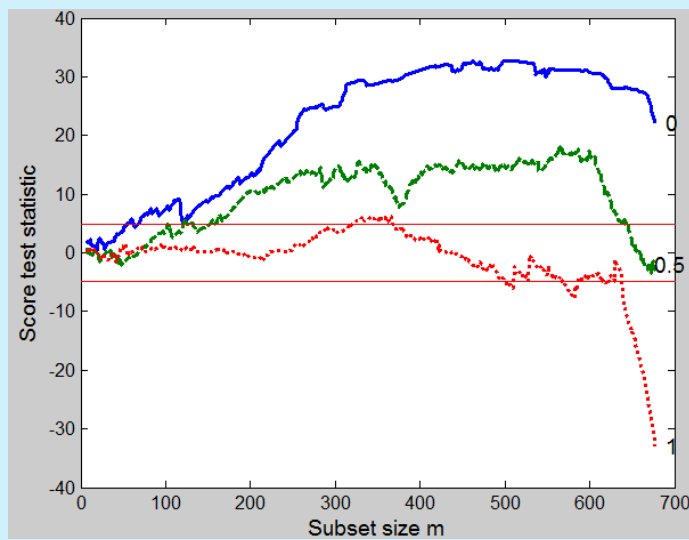
- Central horizontal bands at ± 2.58 , 1% (if normal approx. OK)
- $\lambda=0$ is supported by all the data
- For $\lambda = 1$ and 0.5 last cases are 19, 20, 21: 3 largest observations
- For $\lambda = -1$ and -0.5 last cases are 9, 8, 7: three smallest observations
- For correct transformation, no order

International trade data example

- 677 monthly aggregates of EU import flows of a fishery product (y =Values, X = quantity)

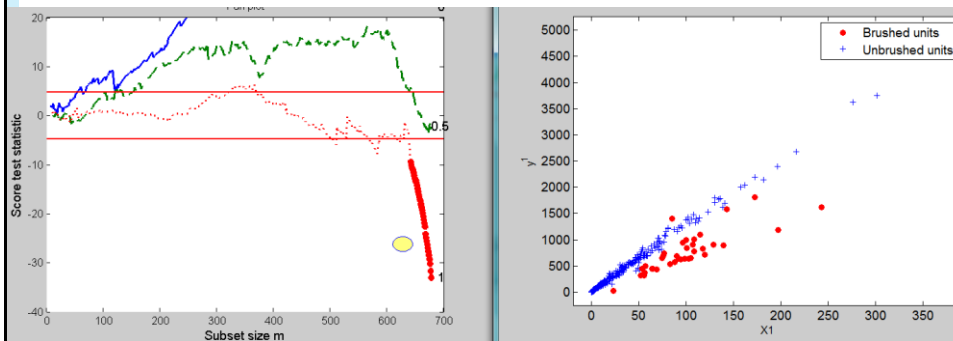


Fan plot



Dynamic visualization

Dynamic link from the fan plot to the yXplot



An example about regression with transformations and outliers

Source: Atkinson and Riani (2006) *JCGS*

Example: loyalty cards

- 509 observations on the behavior of customers with loyalty cards from a supermarket chain in Northern Italy

Variables

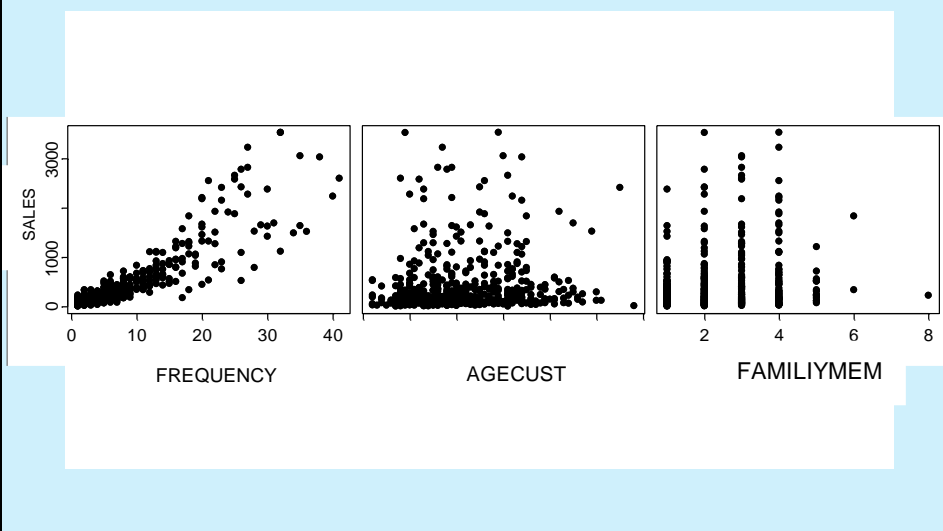
y : the amount, in euros, spent at the shop over six months

x_1 : the number of visits to the supermarket in the six month period

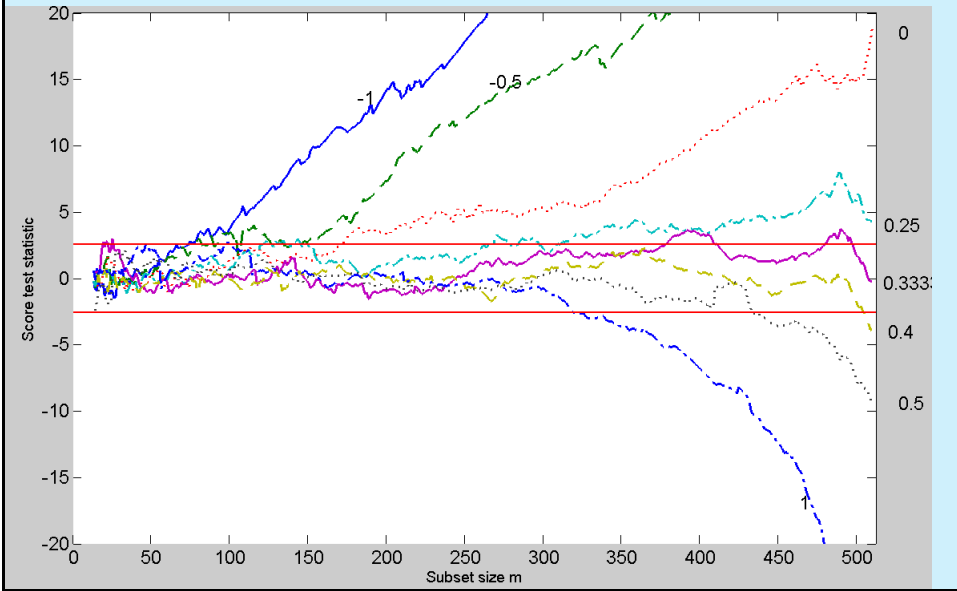
x_2 : the age of the customer

x_3 : the number of members of the customer's family.

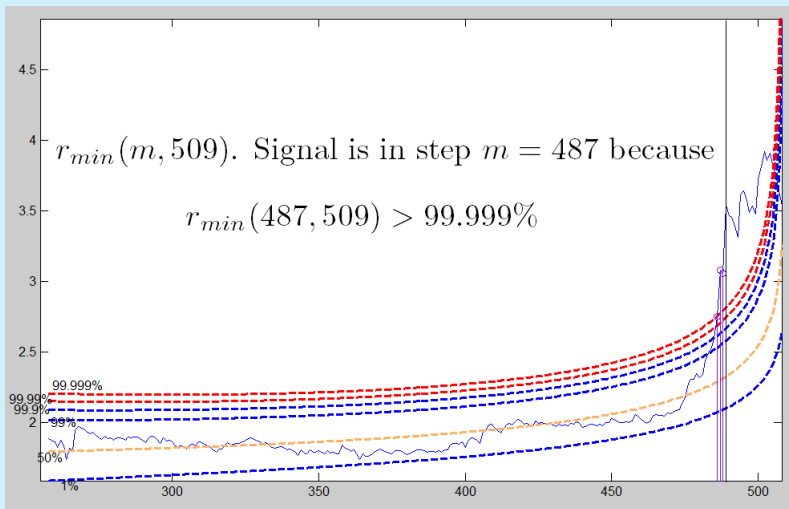
Loyalty cards: yX plot



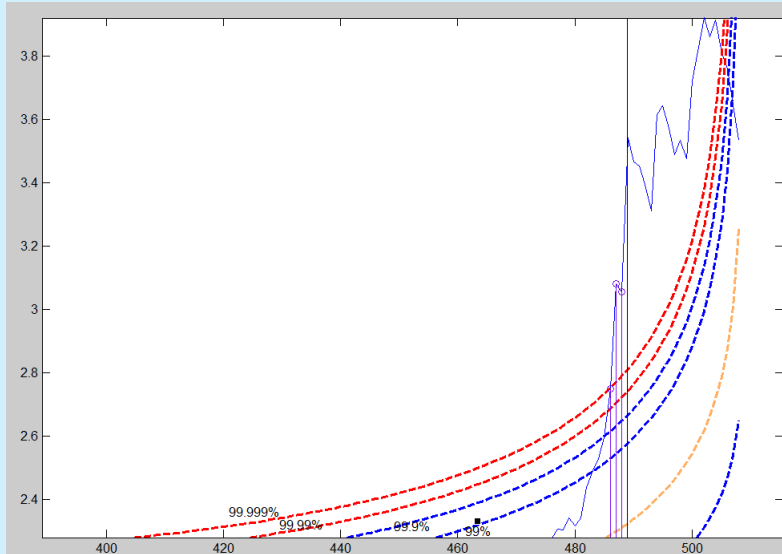
Fan plot



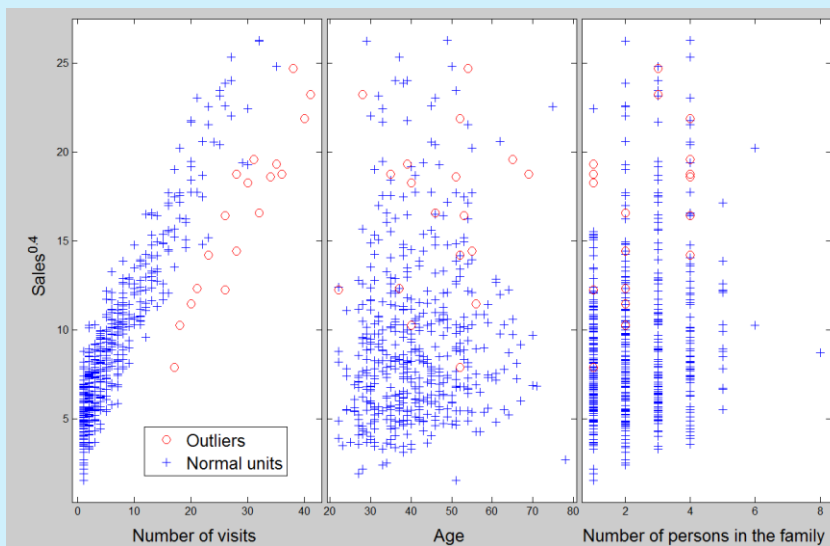
Transformed y: monitoring minimum deletion residual (MDR)



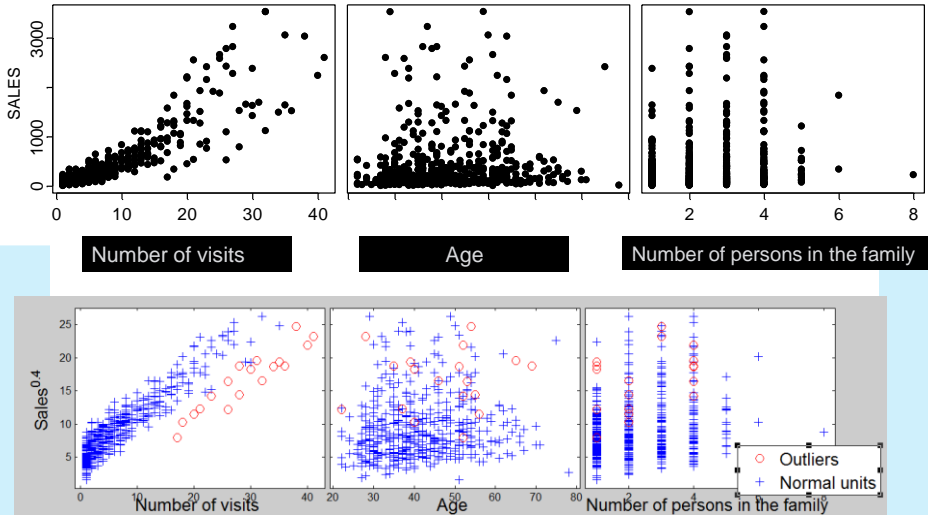
Monitoring minimum deletion residual (zoom of final part)



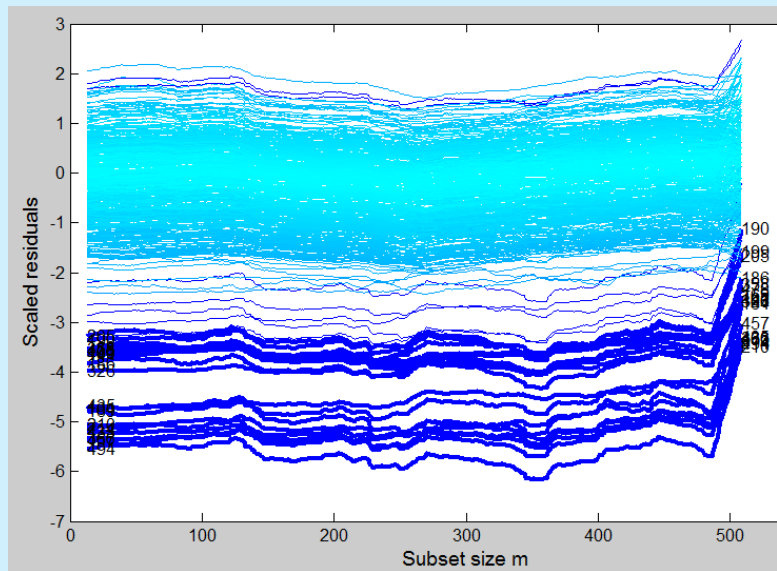
yX plot (using transformed observations) with outliers highlighted



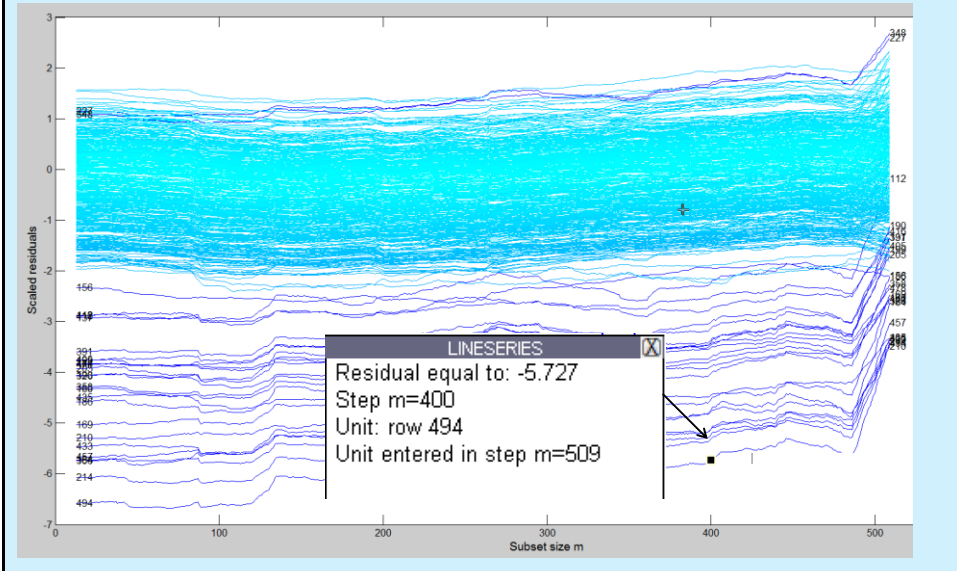
Loyalty cards: yX plot



Monitoring residuals



Monitoring residuals: dynamic visualization through datatooltips



Monitoring residuals: dynamic visualization through databrushing

