# Introduction to Robust Statistics

**Elvezio Ronchetti**

Department of Econometrics
University of Geneva
Switzerland

*Elvezio.Ronchetti@metri.unige.ch*

*http://www.unige.ch/ses/metri/ronchetti/*



## ♦ Outline

♦ Introduction

♦ Basics

- Sensitivity Curve and Influence Function
- Breakdown Point
- Robust Inference

♦ Linear Models

♦ Generalized Linear Models

♦ Elements of Multivariate Analysis

♦ Conclusions

## ♦ Introduction

### Robust statistics

- deals with deviations from ideal models and their dangers for corresponding inference procedures

- primary goal is the development of procedures which are still reliable and reasonably efficient under small deviations from the model, i.e. when the underlying distribution lies in a neighborhood of the assumed model

Robust statistics is an extension of parametric statistics, taking into account that parametric models are at best only approximations to reality.



### Main aims of robust procedures

From a data-analytic point of view, robust statistical procedures will

(i) find the structure best fitting the majority of the data;

(ii) identify deviating points (outliers) and substructures for further treatment;

(iii) in unbalanced situations : identify and give a warning about highly influential data points (leverage points).

In addition to the classical concept of efficiency, new concepts are introduced to describe

- the local stability of a statistical procedure (the influence function and derived quantities)

- its global reliability or safety (the breakdown point).



The ancient, vaguely defined problem of robustness has been partly formalized into mathematical theories which yield optimal robust procedures and which provide illumination and guidance for the user of statistical methods.

## Robustness

- its purpose is to safeguard against deviations from the assumptions.

- It makes unnecessary getting the stochastic part of the model right.

## Diagnostics

- Its purpose is to find and identify deviations from the assumptions.

- It helps to make the functional part of the model right.



## ♦ Sensitivity Curve and Influence Function

### Sensitivity curve

- Observations $z_1, z_2, \ldots$ with underlying distribution (model) $F$.

- Statistic $T_n$ (function of the observations)

$$SC(z; z_1, \ldots, z_{n-1}, T_n)$$

$$= n\,[T_n \quad (z_1, \ldots, z_{n-1}, z) \quad - T_{n-1}(z_1, \ldots, z_{n-1})]$$

$$\downarrow \; n \to \infty$$

$$IF(z; T, F)$$

## Influence function of the mean

$$SC(z; z_1, \ldots, z_{n-1}, \text{mean}_n)$$

$$= \frac{\text{mean}_n(z_1, \ldots, z_{n-1}, z) - \text{mean}_{n-1}(z_1, \ldots, z_{n-1})}{\frac{1}{n}}$$

$$= \frac{\frac{1}{n}(z_1 + \ldots + z_{n-1} + z) - \frac{1}{n-1}(z_1 + \ldots + z_{n-1})}{\frac{1}{n}}$$

$$= \frac{\frac{1}{n}z - \left(\frac{1}{n-1} - \frac{1}{n}\right) \cdot (z_1 + \ldots + z_{n-1})}{\frac{1}{n}}$$

$$= z - \ \text{mean}_{n-1}(z_1, \ldots, z_{n-1})$$

$$\downarrow n \longrightarrow \infty$$

$$z - \ E_F Z = IF(z; \ \text{mean}\ , F)$$



## Influence function of the least squares estimator

Regression : $y_i = x_i^T \beta + u_i \quad i = 1, \ldots, n$

Least Squares Est : $\widehat{\beta}$

$$Q_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n-1} x_i x_i^T \longrightarrow Q, \ n \longrightarrow \infty.$$

$$SC\left((x, y); (x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), LS\right)$$

$$= \frac{n}{n-1} \ Q_{n-1}^{-1} \ x(y - x^T \ \widehat{\beta}_{n-1}) \ \frac{1}{1 + \frac{1}{n-1} x^T Q_{n-1}^{-1} x}$$

$$\begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & Q^{-1} & \beta & 1 \end{array}$$

$$\longrightarrow \ IF(x, y; LS, F) = Q^{-1} x (y - x^T \beta)$$

when $n \rightarrow \infty$.

- Can find $IF(z;T,F)$ for most est.

- Gross-error sensitivity :

  maximum (over $z$) of $||IF||$

**WANTED**
PROCEDURES
WITH
**BOUNDED**
INFLUENCE FUNCTION

Reward : **ROBUSTNESS**



## ♦ Influence Function

$$z_1, \ldots, z_n \quad \text{iid} \qquad , z_i \sim F$$

$$T_n(z_1, \ldots, z_n)$$

$$T_n(z_1, \ldots, z_n) = T(F_n)$$

$T$ : functional on some subset of all distr.

$F_n$ : empirical distribution
(which assigns prob. $\frac{1}{n}$ to $z_1, \ldots, z_n$).

Influence Function of $T$ at $F$ :

$$IF(z;T,F) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon \triangle_z) - T(F)}{\varepsilon}$$

Hampel (1968), (1974), *J. Am. Stat. Ass.*

$\triangle_z$ : distr. which puts mass 1 at any point $z$.

Note : $IF(z;T,F) = \frac{\partial}{\partial \varepsilon} T\left((1-\varepsilon)F + \varepsilon \triangle_z\right)|_{\varepsilon=0}$

## Properties

- $IF$ describes the normalized influence on the estimate of an infinitesimal observation at $z$.
- $IF$ is the Gâteaux derivative of $T$ at $F$, or the integrand in the first term of the von Mises expansion

$$T(G) = T(F) + \int IF(z;T,F)\,d(G-F)(z) + O(||G-F||^2)$$

Math. treatment (e.g.) :
von Mises (1947), *Ann. Math. Stat.*
Fernholz (1983), Springer
Serfling (1980), Wiley



- $\varepsilon$-neighborhood $P_\varepsilon(F)$ of $F$ :

$$P_\varepsilon(F) = \{G | G = (1-\varepsilon)F + \varepsilon H, H \text{ arbitrary } \}$$

$$\begin{aligned} d(G,F) &= \sup_z ||G(z) - F(z)|| \\ &= \varepsilon \cdot \sup_z ||H(z) - F(z)|| \leq \varepsilon. \end{aligned}$$

For $G \in P_\varepsilon(F)$ :

$$T(G) = T(F) + \varepsilon \int IF(z;T,F)\,dH(z) + O(\varepsilon^2)$$

Bias curve: max bias over $\varepsilon$-neighborhood

$$b(\varepsilon;T,F) = \sup_{G \in P_\varepsilon(F)} ||T(G) - T(F)||$$

$$\underbrace{b(\varepsilon;T,F)}_{\text{max bias over neighborh.}} \approx \varepsilon \cdot \underbrace{\gamma^*(T,F)}_{\text{gr err sens}}$$

$$\gamma^*(T,F) = \sup_z ||IF(z;T,F)||$$

$IF$ describes the robustness (stability) properties of $T(\cdot)$

- For $G = F_n$ (empirical distr.)

$$T_n = T(F) + \frac{1}{n}\sum_{i=1}^{n} IF(z_i; T, F) + \ldots$$

$$\Longrightarrow \boxed{\sqrt{n}\,(T_n - T(F)) \sim_{as} N\,(0, V(T,F))}$$

$V(T,F) = E_F[IF(Z;T,F) \cdot IF^T(Z;T,F)]$
$E_F[IF(Z;T,F)] = 0$

$IF$ describes the efficiency properties of $T(\cdot)$.



- Connection to sensitivity curve $SC(z; z_1, \ldots, z_{n-1}, T_n)$

$$= n\left[T_n(z_1, \ldots, z_{n-1}, z) - T_{n-1}(z_1, \ldots, z_{n-1})\right]$$
$$= \frac{T\left((1-\frac{1}{n})F_{n-1} + \frac{1}{n}\triangle_z\right) - T(F_{n-1})}{\frac{1}{n}}$$

- Connection to jackknife

$$T_{(j)} = \quad T_{n-1}(z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_n)$$
$$j = 1, 2, \ldots n$$

Pseudo-values :

$$\begin{aligned} T_{*j} &= nT_n - (n-1)T_{(j)} \\ &= T_n + \underbrace{(n-1)\left[T_n - T_{(j)}\right]} \end{aligned}$$

$$\| $$

$$\tfrac{n-1}{n} SC(z_j; z_1, \ldots, z_{j-1}, z_{j+1}, \ldots z_n, T_n)$$

$$\|\wr$$

$$\tfrac{n-1}{n} IF(z_j; T, F)$$

Jackknife estimator :
Tukey (1958), *Ann. Math. Stat.*

$$\begin{aligned} T_{*\cdot} &= \tfrac{1}{n} \sum_{j=1}^{n} T_{*j} \\ &\approx T_n + \tfrac{1}{n} \sum_{j=1}^{n} IF(z_j; T, F) \end{aligned}$$

(von Mises expansion; one-step est.)



The stability analysis by means of the influence function can be performed on any statistical functional e.g.

(as)var$_F T_n$
(as) level of a test $= P_F\left[T_n > k_\alpha\right]$
$\ldots$

## ♦ Breakdown Point

The $IF$ shows how an estimator reacts to a small proportion of outliers.
Note that the sample mean cannot resist even one outlier !
Other estimators can, because their $IF$ is bounded.

What is the maximum amount of "perturbation" they can resist?

### Breakdown Point

Sample $Z = (z_1, \ldots, z_n)$
Statistic $T_n(Z)$

$$\text{bias } (m; T_n, Z) = \sup_{Z'} ||T_n(Z') - T_n(Z)||$$

$Z'$ : "corrupted" sample obtained by replacing any $m$ of the original $n$ data points by arbitrary values.

Breakdown point of $T_n$ (at $Z$) :

$$\varepsilon^*(T_n, Z) = \min \left\{ \frac{m}{n} | \text{ bias } (m; T_n, Z) = \infty \right\}$$



Examples:

Breakdown point of the

- mean: $\frac{1}{n}$
- $\alpha$-trimmed mean: $\alpha$

Robustness notions as
elementary calculus properties

of a function of one argument, namely its continuity, differentiability, and vertical asymptote.

The breakdown point tells us up to which distance the "linear approximation" provided by the influence function is likely to be of value.



## ♦ $M$-estimators

$z_1, \ldots, z_n$ iid
Huber(1964), *Ann. Math. Stat.*
Parametric model $\{F_\theta | \theta \in \Theta\}$
$M$-estimator $T_n : \sum_{i=1}^{n} \psi(z_i, T_n) = 0$

- $M$- estimators generalize $MLE$
  (for which $\psi(z, \theta) =$ score $= \frac{\partial}{\partial\theta} \log f_\theta(z)$)

- To any asymptotically normal estimator, there exists an asymptotically equivalent $M$-estimator.

- Properties :

$$IF(z; \psi, F) = M(\psi, F)^{-1} \psi(z, T(F))$$

$$\sqrt{n}(T_n - T(F)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V(\psi, F))$$

$$\begin{aligned} V(\psi, F) &= M(\psi, F)^{-1} Q(\psi, F) M(\psi, F)^{-T} \\ M(\psi, F) &= E_F[-\tfrac{\partial}{\partial\theta} \psi(Z, T(F))] \\ Q(\psi, F) &= E_F[\psi(Z, T(F)) \cdot \psi(Z, T(F))^T] \end{aligned}$$

**How do we construct (optimal) robust estimators?**

Example: location

$z_1, \ldots, z_n$ ind. observations from a distribution with location parameter $\mu$,
e.g. $z_i \sim \mathcal{N}(\mu, 1)$.

Two estimators of $\mu$:
the mean and the median.
Both are $M-$estimators with score functions:

$$\begin{aligned} \psi(z,\mu) &= z-\mu & \text{(mean)} \\ \psi(z,\mu) &= \text{sign}(z-\mu) & \text{(median)} \end{aligned}$$



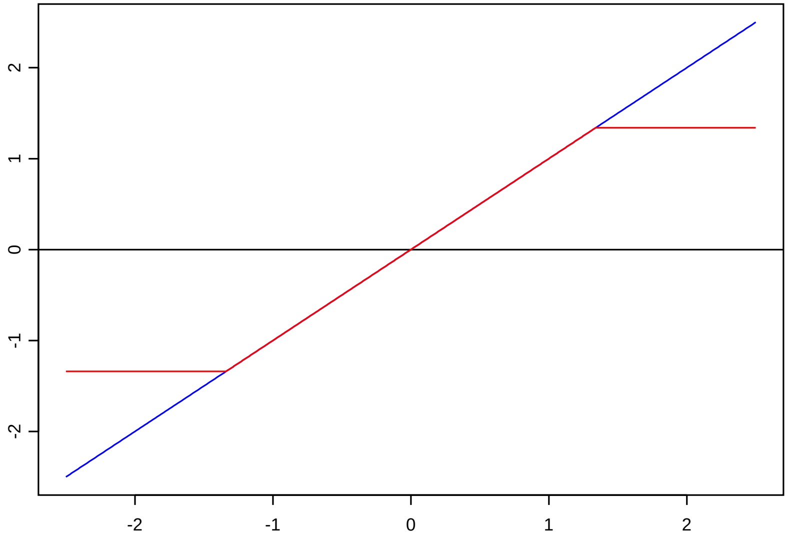
- Efficiency
  Under normality the mean is the most efficient estimator for $\mu$, while the median has efficiency $2/\pi = 64\%$.

- Robustness
  Their influence function is proportional to their score function. The mean is not robust (unbounded IF), while the median is robust (bounded IF).

$\longrightarrow$ Best compromise between efficiency and robustness?

At the normal model, the Huber estimator, an $M-$estimator defined by the score function $\psi_c(\cdot)$, is the most efficient estimator for $\mu$ with a bounded influence function.

For $c = 1.345$ its efficiency at the normal model is 95%.




Huber's function

$$\psi_c(r) = \begin{cases} r & |r| \le c \\ c \cdot sign(r) & |r| > c. \end{cases}$$



## Huber's estimator of location

Huber(1964), *Ann. Math. Stat.*

The Huber estimator $T_n$ is an $M-$estimator defined as the solution of the implicit eqn.

$$\sum_{i=1}^{n} \psi_c(z_i - T_n) = 0$$

Rewrite as:

$$\sum_{i=1}^{n} w_c(r_i)(z_i - T_n) = 0$$

i.e.

$$T_n = \frac{\sum_{i=1}^n w_c(r_i) z_i}{\sum_{i=1}^n w_c(r_i)} \ ,$$

where

$r_i = z_i - T_n$ are the residuals and

$$\begin{aligned} w_c(r) &= \psi_c(r)/r \\ &= 1 \qquad |r| \le c \\ &= \frac{c}{|r|} \qquad |r| > c. \end{aligned}$$

## ♦ Robust Inference

- robustness of validity
  The level of the test should be stable under small, arbitrary departures from the distribution under the null hypothesis.

- robustness of efficiency
  The test should still have a good power under small, arbitrary departures from the distribution under specified alternatives.

Heritier & Ronchetti (1994), *J. Am. Stat. Ass.*



## Example: Bartlett's test

F-test for comparing two variances
Investigate the stability of the level of this test and its generalization to k samples (Bartlett's test)

| Distribution | $k = 2$ | $k = 5$ | $k = 10$ |
|---|---|---|---|
| Normal | 5.0 | 5.0 | 5.0 |
| $t_{10}$ | 11.0 | 17.6 | 25.7 |
| $t_7$ | 16.6 | 31.5 | 48.9 |

Actual level in % in large samples of Bartlett's test when the observations come from a slightly nonnormal distribution; from Box(1953), *Biometrika*

In view of its behavior this test would be more useful as a test for normality rather than as a test for equality of variances!

## Example:
## Two sample t-test and Wilcoxon test

Generate two samples of size 10 from $\mathcal{N}(0,1)$ and $\mathcal{N}(1.5,1)$ respectively:

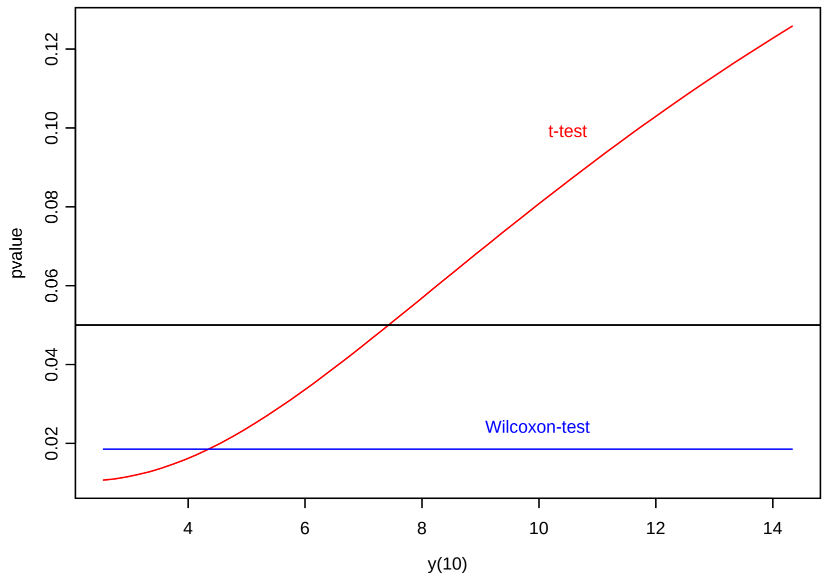
x
-1.7234313 -1.1028391 -0.8915296 -0.5941126 -0.4669093 -0.4511696 -0.3411728 0.3126089 1.1478631 1.2476020

y
-0.7651532 0.4464456 0.5107215 0.5611747 0.5929228 0.7118542 1.0405136 1.3153364 2.0116585 2.5419382

t-test: p-value = .011
Wilcoxon test: p-value = .018

Increase the largest value of the 2nd sample $y_{(10)} = 2.5419382$ by steps of size .2 and recompute the $p-$value of the t-test and Wilcoxon test.






$p-$value of the two sample $t-$test and Wilcoxon test as the value of the largest observation of the 2nd sample increases

Beyond some value of $y_{(10)}$, the $t-$test never rejects the null hypothesis.

In general:

- $t-$test: $\sim$ robustness of validity but no robustness of efficiency
- Wilcoxon test: robustness of validity but loses power in the presence of small deviations from normality



## ♦ Linear Models

$Y_1, ..., Y_n$ $\; n$ independent observations of a response variable:

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

$\beta \in \mathbb{R}^q$ is a vector of unknown parameters,
$x_i \in \mathbb{R}^q$ is a vector of explanatory variables,
and $E[u_i] = 0$, $var[u_i] = \sigma^2$.

The least squares estimator $\widehat{\beta}_{LS}$ of $\beta$ is a $M-$estimator defined by the estimating equation:

$$\sum_{i=1}^{n} (y_i - x_i^T \beta) \cdot x_i \;=\; 0.$$

Influence function of $\widehat{\beta}_{LS}$:

$$IF(x, y; \widehat{\beta}_{LS}, F) = Q^{-1} x \cdot u \ ,$$

where $u = y - x^T\beta$ and $Q = E[xx^T]$.

Unbounded both w.r. to $y$ and $x$.



## $M-$estimators for regression with bounded IF

$\longrightarrow$ Construct new robust $M-$estimators by bounding $u$ and $x$ through score function

$$\begin{aligned}
\psi((x, y), \beta) &= \psi_c(u/\sigma) \cdot x \qquad\qquad Huber \\
&= \psi_c(u/\sigma) \cdot w(x) \cdot x \ \ Mallows \\
&= \psi_{c/\|Ax\|}(u/\sigma) \cdot x \ \ \ Hampel - Krasker
\end{aligned}$$

where $\psi_c(\cdot)$ is the Huber function, $w(\cdot)$ is a weight function for the $x_i's$, and $A$ is a positive definite matrix defined in the space of the $x_i's$.

The Hampel-Krasker estimator is the optimal $B-$ robust estimator when the $IF$ is measured by the Euclidean norm.

## Robust tests for regression

The three classes of tests for general parametric models can be defined here with the score functions $\psi((x, y), \beta)$ given above.

In particular, the likelihood ratio type test for this model is the so-called $\tau-$ test; see Ronchetti(1982).

This test is a robust alternative of the classical $F-$test for regression.



## Looking for structures in the data: high breakdown point estimators

The breakdown point $\varepsilon^*$ of $M-$estimators depends on the design (the distribution of the $x's$).

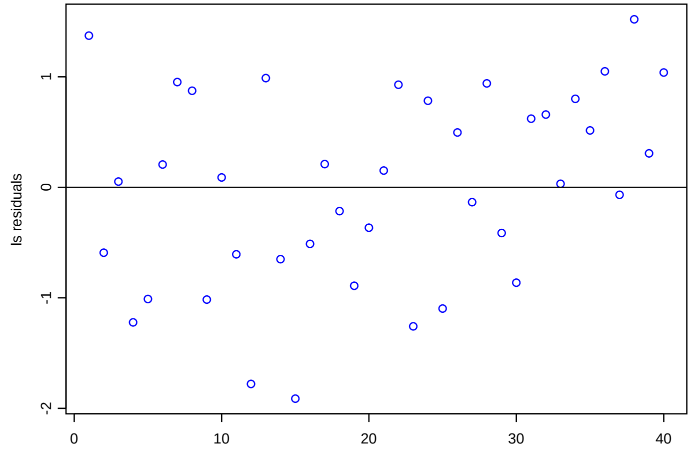
Example
$\varepsilon^*$ of $L_1-$estimator is 25% for uniform $x's$.

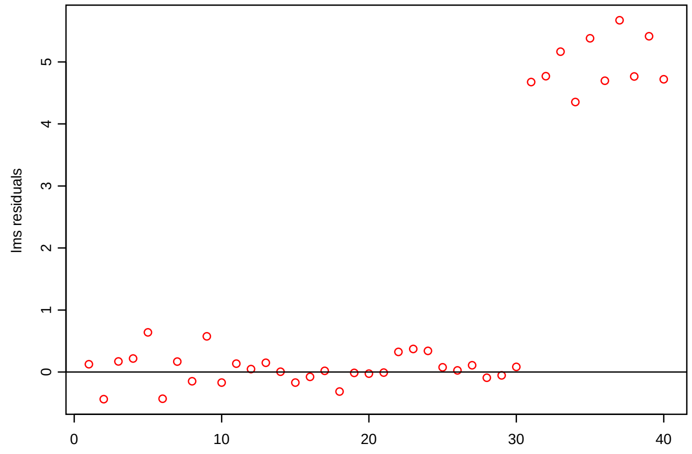
More generally for $M-$estimators:

- $\varepsilon^*$ is arbitrarily close to 0 for longer-tailed designs
- $\varepsilon^* \leq 1/dim(\beta)$;
  Maronna(1976), *Ann. Stat.*

$\longrightarrow$ Search for high (50%?) breakdown point equivariant estimators






LS and LMS residuals

$u_i = y_i - x_i^T \beta, \qquad i = 1, \ldots, n$

- Least Median of Squares (LMS)
  Hampel(1975), Rousseeuw(1984)

$$\min_{\beta} \ \text{med}_i \ u_i^2$$

- Least Trimmed Squares (LTS)
  Rousseeuw(1984), *JASA*

$$\min_{\beta} \sum_{i=1}^{h} u_{(i)}^2,$$

  where $u_{(1)}^2 \leq \ldots \leq u_{(n)}^2$

- $S-$estimators
  Rousseeuw and Yohai(1984)

$$\min_{\beta} \ s(u_1(\beta), \ldots, u_n(\beta)),$$

  where $s(\cdot)$ is an $M-$estimator of scale with a bounded $\rho-$fct., i.e. solution of

$$\frac{1}{n} \sum_{i=1}^{n} \rho(\frac{u_i}{s}) = K$$



## Properties

- An $S-$estimator with a smooth $\rho-$function is an $M-$estimator with score function $\psi(u/s)x$ and $\psi(\cdot) = \rho'(\cdot)$ redescending (multiple solutions!).
  Ex. Tukey's biweight
  $\longrightarrow$ asymptotic normality and corresponding tests

- LMS and LTS are $S-$estimators with discontinous $\rho-$functions

- High breakdown point ($\approx$ 50%)

- Low efficiency

- Computational aspects

To improve efficiency: $MM-$estimators
Yohai(1987), *Ann. Stat.*

(1) Compute a high breakdown point regression estimator, typically an $S$-estimator and its resulting estimator of scale $s$ based on a loss function $\rho_0(\cdot)$.

(2) Compute an $M-$estimator $\widehat{\beta}$ satisfying the equation

$$\sum_{i=1}^{n} \psi(\frac{y_i - x_i^T \widehat{\beta}}{s}) x_i = 0 \ ,$$

where $\psi(\cdot)$ is a smooth redescending score function, such as Tukey's biweight.



## ♦ Generalized Linear Models

$Y_1, ..., Y_n$ $n$ independent observations of a response variable.
The distribution of $Y_i \in$ exponential family, $E[Y_i] = \mu_i$, $var[Y_i] = V(\mu_i)$ and

$$g(\mu_i) = x_i^T \beta, \qquad i = 1, ..., n,$$

$\beta \in \mathbb{R}^q$ is a vector of unknown parameters,
$x_i \in \mathbb{R}^q$ is a vector of explanatory variables,
$g(.)$ is the link function.

- Estimation of $\beta$:

  maximum likelihood or quasi-likelihood (equivalent if g(.) is the canonical link, e.g.
  logistic regression: $\text{logit}(\mu) = \log(\frac{\mu}{1-\mu})$
  Poisson regression: $\log(\mu)$ ).

- Inference and variable selection:

  Standard asymptotic inference based on likelihood ratio, Wald and score test is readily available for these models.

However ...



## Robustness

Given $n$ observations $x_1, ..., x_n$ of a set of $q$ explanatory variables ($x_i \in \mathbb{R}^q$), and when $g(\mu_i)$ is the canonical link, the maximum likelihood estimator and the quasi-likelihood estimator of $\beta$ are the solutions of the following system of equations

$$\begin{aligned} \sum_{i=1}^{n} r_i \frac{1}{V^{1/2}(\mu_i)} \mu_i' &= \\ \sum_{i=1}^{n} (y_i - \mu_i) \cdot x_i &= 0, \end{aligned} \quad (1)$$

where $r_i = \frac{y_i - \mu_i}{V^{1/2}(\mu_i)}$ are the Pearson residuals, $\mu_i = g^{-1}(x_i^T \beta)$, and $\mu_i' = \frac{\partial \mu_i}{\partial \beta}$.

The maximum likelihood and the quasi-likelihood estimator defined by (1) can be viewed as an M-estimator with score function

$$\psi(y_i; \beta) = (y_i - \mu_i) \cdot x_i. \quad (2)$$

Since $\psi(y;\beta)$ is unbounded in $x$ and $y$, the influence function of this estimator is unbounded and the estimator is not robust.

Several alternatives have been proposed. One of these alternative methods is the class of M-estimators of Mallows's type (Cantoni and Ronchetti 2001, *JASA*) defined by the score function

$$\psi(y_i;\beta) = \nu(y_i,\mu_i)w(x_i)\mu_i' - a(\beta), \qquad (3)$$

where $a(\beta) = \frac{1}{n}\sum_{i=1}^{n} E[\nu(y_i,\mu_i)]w(x_i)\mu_i'$, $\nu(y_i,\mu_i) = \psi_c(r_i)\frac{1}{V^{1/2}(\mu_i)}$, and $\psi_c$ is the Huber function defined by

$$\psi_c(r) = \begin{cases} r & , \quad |r| \le c \\ c \cdot sign(r) & \quad |r| > c. \end{cases}$$

When $w(x_i) = 1$, we obtain the so-called Huber quasi-likelihood estimator.



Standard inference based on robust quasi-deviances is available.

$\longrightarrow$ robust likelihood ratio test
is based on twice the difference between the robust quasi-likelihoods with and without restrictions

When the link function is the identity, this test becomes the $\tau-$test defined for linear regression.

## ♦ Elements of Multivariate Analysis



$x_1, \ldots, x_n$ iid

$x_i \in \mathbb{R}^p \quad (x_i \sim N(\mu, \Sigma))$

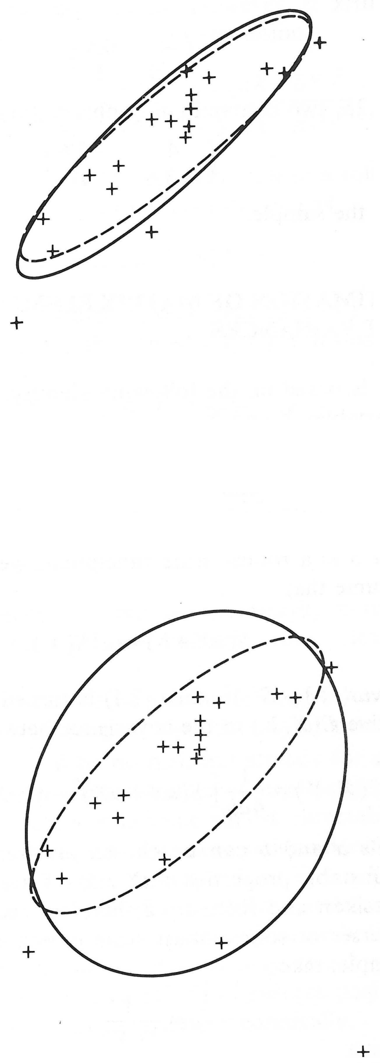
Classical estimators of location $\mu$ and scatter $\Sigma$ (MLE under normal model)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

Key for multivariate analysis:

- principal components analysis
- discriminant analysis
- factor analysis



Influence of outliers on classical and robust covariance estimates; from Huber(1981)



New class of estimators; which properties?

- affine equivariance
- good robustness properties under local perturbations (bounded IF)
- good robustness properties under global perturbations (high breakdown point)
- good efficiency under a broad class of underlying distributions
- $n^{1/2}$ consistency, asymptotic normality
- computational simplicity

Not necessarily in order of priority and possibly conflicting requirements!

Affine equivariance:

location vector $t(x_1, \ldots, x_n) \in \mathbb{R}^p$
scatter matrix $V(x_1, \ldots, x_n)$ a $p \times p$ pos. def. symm. matrix

Then $\forall b \in \mathbb{R}^p$ , $B$ nonsingular $p \times p$ matrix:

$$\begin{aligned} t(Bx_1 + b, \ldots, Bx_n + b) &= B \cdot t(x_1, \ldots, x_n) + b \\ V(Bx_1 + b, \ldots, Bx_n + b) &= B \cdot V(x_1, \ldots, x_n) \cdot B^T \end{aligned}$$



## $M-$estimators for location and scatter

Maronna(1976), *Ann. Stat.*
Huber(1977)

$(t, V)$ solution of the implicit eqn.

$$t = \frac{\sum_{i=1}^n w_1(d_i) x_i}{\sum_{i=1}^n w_1(d_i)}$$

$$V = \frac{\sum_{i=1}^n w_2(d_i)(x_i - t)(x_i - t)^T}{\sum_{i=1}^n w_2(d_i)}$$

where

$$\begin{aligned} d_i &= d(x_i; t, V) \\ &= [(x_i - t)^T V^{-1} (x_i - t)]^{1/2} \end{aligned}$$

(robust Mahalanobis distance)

w.l.o.g. $t(F) = 0$ and $V(F) = I$.

$$\begin{aligned} IF(x; t, F) &\propto w_1(\|x\|)x \\ IF(x; V, F) &= -2\Gamma \end{aligned}$$

where $\frac{1}{p} tr(\Gamma) \propto w_2(\|x\|)(\frac{\|x\|^2}{p} - 1)$
$\Gamma - \frac{1}{p} tr(\Gamma) I \propto w_2(\|x\|)\|x\|^2 (\frac{xx^T}{\|x\|^2} - \frac{I}{p})$

$\longrightarrow$ To bound the IF choose e.g.:

$$\begin{aligned} w_1(d) &= \min(1, c/d) \\ w_2(d) &= \min(1, c/d^2) \end{aligned}$$

Breakdown point $\leq 1/p$



## High breakdown point estimators for location and scatter

- Minimum Volume Ellipsoid (MVE)
  Rousseeuw(1984), *JASA*
  Find the ellipsoid $\{x | d^2(x; t, V) \leq 1\}$ with minimum volume which covers at least 50% of the data $\longrightarrow t, V$

- Minimum Covariance Determinant (MCD)
  Rousseeuw(1984), *JASA*
  $t$ is the average of the $h$ points for which the determinant of the cov. matrix is minimal and $V$ is the corresponding cov. matrix

- $S-$estimators Rousseeuw & Yohai(1984)
  Lopuhaa(1989)

$$\min \ |V|$$

  under the constraint

$$\frac{1}{n} \sum_{i=1}^{n} \rho(d_i) = b_0,$$

  for a bounded function $\rho(\cdot)$.

## General references (books)


- Huber, P.J.(1981)
  *Robust Statistics*,
  Wiley (paperback 2004)

- Hampel,F.R., Ronchetti,E.M., Rousseeuw,P.J., Stahel, W.A. (1986)
  *Robust Statistics: The Approach Based on Influence Functions*,
  Wiley (paperback 2005)

- Maronna R. A., Martin, R.D., Yohai, V. J. (2006)
  *Robust Statistics: Theory, and Methods*, Wiley

- Heritier S., Cantoni E., Copt S., Victoria-Feser M.P. (2008)
  *Robust Methods in Biostatistics*,
  Wiley, to appear.




## Some common misunderstandings

- Robust statistics replaces classical statistics.

- The normality assumption is "guaranteed" by the central limit theorem.

- If the errors are non-normal, I change the specification of the errors.

- I use classical procedures after removing outliers. Therefore I do not need any robust procedures.

- Robust statistics cannot be used when the errors are asymmetric.

## ♦ Messages

- There exist robust statistical procedures which complement classical estimators and tests for general parametric models.

- Whenever you can do a likelihood analysis, you can do a robust analysis.