

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
- ◆ Sensitivity Curve and Influence Function
- ◆ Statistical Functionals
- ◆ Influence Function
- ◆ Breakdown Point
- ◆ M -estimators
- ◆ Optimal Robust Estimators
- ◆ Robust Inference
- ◆ Leading Example (Finance)
- ◆ Messages

◆ 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.

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.

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, \dots with underlying distribution (model) F .
- Statistic T_n (function of the observations)

$$\begin{aligned} & SC(z; z_1, \dots, z_{n-1}, T_n) \\ &= n [T_n(z_1, \dots, z_{n-1}, z) - T_{n-1}(z_1, \dots, z_{n-1})] \\ &\quad \downarrow n \rightarrow \infty \\ & IF(z; T, F) \end{aligned}$$

Influence function of the mean

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

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

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

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

$$\downarrow n \longrightarrow \infty$$

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

Influence function of the least squares estimator

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

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

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

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

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

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

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

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**

◆ Statistical Functionals

$$S : P \longrightarrow \mathcal{R}^P \quad , \quad F \longmapsto S(F)$$

Ex. $Z \sim F$

- Characteristic of a distr. F

$$\begin{aligned} T : F \longmapsto T(F) &= E_F Z = \int z dF(z) \\ &= \text{var}_F Z \\ &= F(a) = P_F [Z < a] \\ &= F^{-1}\left(\frac{1}{2}\right) \\ &= F^{-1}\left(\frac{3}{4}\right) - F^{-1}\left(\frac{1}{4}\right) \\ &\dots \end{aligned}$$

- Theil's index

$$\begin{aligned} I : F \longmapsto I(F) &= E_F \left[\left(\frac{Z}{\mu_F} \right) \cdot \log \left(\frac{Z}{\mu_F} \right) \right], \\ \mu_F &= E_F Z \end{aligned}$$

- Maximum likelihood estimator

$$T_{MLE} : E_F \left[\frac{\partial}{\partial \theta} \log g(Z|T_{MLE}(F)) \right] = 0$$

Model $g(\cdot|\theta)$, $Z \sim F$

T_{MLE} : ML functional

$$T_{MLE}(G(\cdot|\theta)) = \theta$$

- Level's functional

$\{F_\theta | \theta \in \Theta\}$

$H_0 : \theta = \theta_0$

Test statistic T_n

$$\alpha_0 = P_{F_{\theta_0}} [T_n > k_{1-\alpha_0}]$$

$$\implies \alpha : F \mapsto \alpha(F) = P_F [T_n > k_{1-\alpha_0}]$$

$$\alpha(F_{\theta_0}) = \alpha_0$$

◆ Influence Function

z_1, \dots, z_n iid, $z_i \sim F$

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

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

T : functional on some subset of all distr.

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

Influence Function of T at F :

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

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

Δ_z : distr. which puts mass 1 at any point z .

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

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

- ε -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 ε -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.)

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

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

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

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

- Connection to sensitivity curve

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

$$\begin{aligned} &= n \left[T_n(z_1, \dots, z_{n-1}, z) - T_{n-1}(z_1, \dots, z_{n-1}) \right] \\ &= \frac{T\left(\left(1 - \frac{1}{n}\right)F_{n-1} + \frac{1}{n}\Delta z\right) - T(F_{n-1})}{\frac{1}{n}} \end{aligned}$$

- Connection to jackknife

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

$$j = 1, 2, \dots, n$$

Pseudo-values :

$$T_{*j} = nT_n - (n-1)T_{(j)}$$

$$= T_n + \underbrace{(n-1) [T_n - T_{(j)}]}_{\parallel}$$

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

$$\parallel \parallel$$

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

Jackknife estimator : Tukey (1958)

$$T_{* \cdot} = \frac{1}{n} \sum_{j=1}^n T_{*j}$$

$$\approx T_n + \frac{1}{n} \sum_{j=1}^n IF(z_j; T, F)$$

(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) $\text{var}_F T_n$

(as) level of a test = $P_F [T_n > k_\alpha]$

...

◆ 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 $Z = (z_1, \dots, z_n)$

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) :

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

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, \dots, z_n iid Huber(1964)

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 $\psi(z, \theta) = \text{score} = \frac{\partial}{\partial \theta} \log f_\theta(z)$)
- To any asymptotically normal estimator, there exists an asymptotically equivalent M -estimator.
- Properties :

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

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

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

◆ Optimal Robust Estimators

z_1, \dots, z_n

Parametric model $\{F_\theta | \theta \in \Theta\}$

score $s(z, \theta) = \frac{\partial}{\partial \theta} \log f_\theta(z)$

$$\hat{\theta}_c : \sum_{i=1}^n \psi_c^{A,a}(z_i, \hat{\theta}_c) = 0$$

where

$$\psi_c^{A,a}(z, \theta) = h_c(A[s(z, \theta) - a])$$

$$h_c(z) = z \min \left\{ 1, \frac{c}{\|z\|} \right\} \text{ Huber function}$$

$$A, a : E_{F_\theta}[\psi_c^{A,a}(Z, \theta)] = 0$$

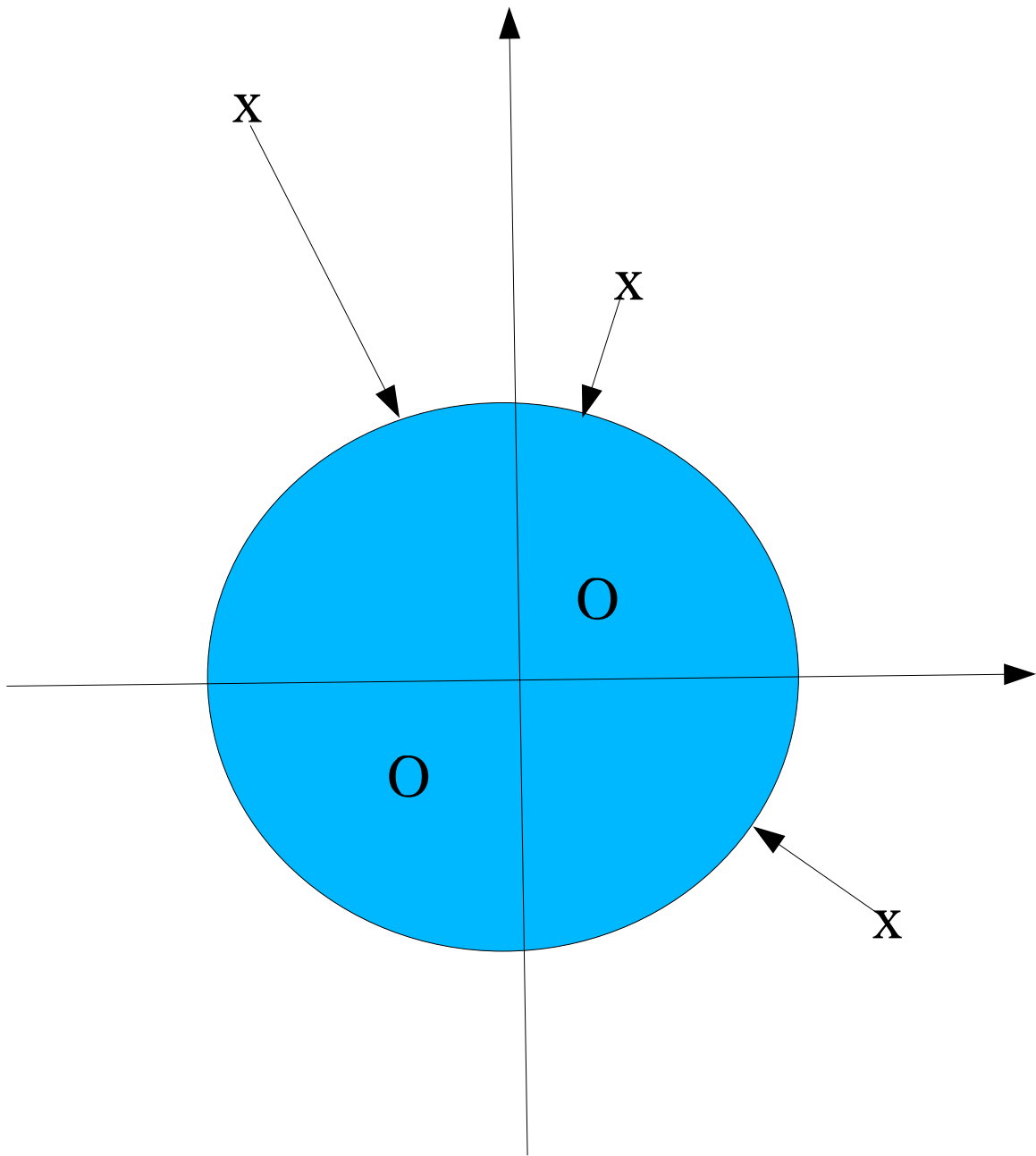
$$E_{F_\theta}[\psi_c^{A,a}(Z, \theta)s(Z, \theta)^T] = I$$

Then $\hat{\theta}_c$ is optimal B-robust in the sense that it minimizes trace $\{V(\psi, F_\theta)\}$, subject to a given upper bound c on the IF :

$$\sup_z \|IF(z, \psi, F_\theta)\| \leq c.$$

Hampel(1968)

Hampel, Ronchetti, Rousseeuw, Stahel (1986)



Sketch of the Huber function

$\psi_c^{A,a}$ can be interpreted as a *truncated version* of s . Because of the truncation, s must be shifted by a in order to satisfy Fisher consistency at the model. Moreover, the second condition ensures that c is an upper bound on the gross-error sensitivity.

Optimal robust estimator defined by :

$$\begin{aligned}
 \psi_c^{A,a}(z, \theta) &= h_c(A[s(z, \theta) - a]) \\
 &= A[s(z, \theta) - a] \cdot w_c(z, \theta) \\
 s(z, \theta) &= \frac{\partial}{\partial \theta} \log f_\theta(z) \\
 w_c(z, \theta) &= \min \left\{ 1, \frac{c}{\|A[s(z, \theta) - a]\|} \right\}
 \end{aligned}$$

Choice of the norm :

- Euclidean norm

$$c \geq \dim(\theta) / E_{\theta} \|s\|$$

- Self-standardized norm

$$\|IF\| = (IF^T V^{-1} IF)^{1/2}$$

(\longrightarrow 2nd equation defining A is different)

$$c \geq \sqrt{\dim(\theta)}$$

Sketch of the proof

- Class of M -est. $\{\psi\}$
 minimize trace $\{V(\psi, F_\theta)\}$
 subject to $\|IF(z; \psi, F_\theta)\| \leq c \quad \forall z$

$$\begin{aligned} V(\psi, F_\theta) &= M(\psi, F_\theta)^{-1} Q(\psi, F_\theta) M(\psi, F_\theta)^{-T} \\ IF(z; \psi, F_\theta) &= M(\psi, F_\theta)^{-1} \psi(z, \theta) \\ M(\psi, F_\theta) &= - \int \frac{\partial}{\partial \theta} [\psi(z, \theta)] dF_\theta(z) \\ &= \int \psi(z, \theta) s(z, \theta)^T dF_\theta(z) \end{aligned}$$

- ψ is determined up to the multipl. with a matrix

$$\longrightarrow \text{w.l.g. } M(\psi, F_\theta) = I$$

We have to solve the following problem :

$$\text{minimize trace } \int \psi(z, \theta) \psi(z, \theta)^T dF_\theta(z)$$

$$\text{subject to } \begin{aligned} \|\psi(z, \theta)\| &\leq c \quad \forall z \\ M(\psi, F_\theta) &= I \end{aligned}$$

- A, a defined by the given implicit equations. Then

$$\begin{aligned}
& \int \{\psi - A[s - a]\} \{\psi - A[s - a]\}^T \\
&= \int \psi\psi^T + \underbrace{A \int [s - a][s - a]^T A^T}_{\text{fixed}} \\
&\quad - \underbrace{A \int [s - a]\psi^T}_I - \underbrace{\int \psi[s - a]^T A^T}_I
\end{aligned}$$

→ minimize trace $\int \psi\psi^T$ equivalent to
min trace $\int \{\psi - A[s - a]\} \{\psi - A[s - a]\}^T$

- minimize $\int \|\psi - A[s - a]\|^2$
subject to $\|\psi\| \leq c$

$$\longrightarrow \psi = \psi_c^{A,a} = h_c(A[s - a])$$

◆ 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.*

z_1, \dots, z_n

Parametric model $\{F_\theta | \theta \in \Theta \subset \mathcal{R}^p\}$

Notation : $\theta = \begin{bmatrix} \theta_{(1)} \\ \theta_{(2)} \end{bmatrix} \begin{matrix} \} p - q \\ \} q \end{matrix}$

$A_{(ij)}$ $i, j = 1, 2$ corresp. partition of $p \times p$ matrix

Test : $H_0 : \theta_{(2)} = 0$.

Classical tests (as. equivalent):

Wald, score (Rao), likelihood ratio

Robust tests will be based on robust M -estimators T_n of θ .

Classes of tests

(i) Wald type test statistic

$$W_n^2 = (T_n)_{(2)}^T \left[V_{(22)}(\psi, F_\theta) \right]^{-1} (T_n)_{(2)}$$

(ii) Score (Rao) type test statistic

$$R_n^2 = Z_n^T C^{-1} Z_n$$

$$Z_n = \frac{1}{n} \sum_{i=1}^n \psi(z_i, T_n^0)_{(2)}$$

T_n^0 is the M -est. in the reduced model

$$\sum_{i=1}^n \psi(z_i, T_n^0)_{(1)} = 0 \quad , \quad T_{n(2)}^0 = 0$$

$$\begin{aligned} C &= M_{(22 \cdot 1)} V_{(22)} M_{(22 \cdot 1)}^T \\ &= \text{as cov } (Z_n) \end{aligned}$$

$$M_{(22 \cdot 1)} = M_{(22)} - M_{(21)} M_{(11)}^{-1} M_{(12)}$$

(iii) Likelihood ratio type test statistic

$$S_n^2 = \frac{2}{n} \sum_{i=1}^n \left[\rho(z_i, T_n) - \rho(z_i, T_n^0) \right]$$

$\rho(z, 0) = 0, \frac{\partial}{\partial \theta} \rho(z, \theta) = \psi(z, \theta)$
 T_n, T_n^0 M -est. in full and reduced model
resp.

The three test statistics can be written as functionals of the empirical distribution function.

→ Functionals as quadratic forms $U(F)^T U(F)$:

$$U_W(F) = V(\psi, F_\theta)_{(22)}^{-1/2} T(F)_{(2)} \quad \text{Wald}$$

$$U_R(F) = C(\psi, F_\theta)^{-1/2} Z(F) \quad \text{Score}$$

When M is symmetric positive definite, the likelihood ratio test statistic is as equivalent to the quadratic form defined by

$$U_{LR}(F) = M(\psi, F_\theta)_{(22.1)}^{-1/2} T(F)_{(2)}$$

$T(F)$ (resp. $Z(F)$) are the functionals associated with T_n (resp. Z_n).

Asymptotic distribution

Under the sequence of alternatives

$$H_{1,n} : \theta_{(2)} = \frac{\Delta}{\sqrt{n}}$$

- nW_n^2, nR_n^2 are asymptotically $\chi_q^2(\delta)$,
 $\delta = \Delta^T V(\psi, F_{\theta_0})_{(22)}^{-1} \Delta$.
- nS_n^2 is asymptotically $\sum_{i=1}^q (\lambda_i^{1/2} N_i + \mu_i)^2$,
where N_1, \dots, N_q are iid $N(0, 1)$.

Robustness

- $H_0 : \theta = \theta_0, \theta_{0(2)} = 0,$
 $\theta_{0(1)}$ unspecified
- Contamination :
$$F_{\varepsilon,n} = \left(1 - \frac{\varepsilon}{\sqrt{n}}\right)F_{\theta_0} + \frac{\varepsilon}{\sqrt{n}}G$$
- Statistical functional $U_n = U(F^{(n)})$, where $F^{(n)}$ is the empir. distr. function, such that $U(F_{\theta_0}) = 0$, $IF(z; U, F_{\theta_0})$ bounded and

$$\sqrt{n}(U_n - U(F_{\varepsilon,n})) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_q) \quad .$$

- $\alpha(F)$ level of the test based on quadratic form $nU_n^T U_n$ when the underlying distr. is F and $\alpha(F_{\theta_0}) = \alpha_0$ the nominal level.

Then :

$$\lim_{n \rightarrow \infty} \alpha(F_{\varepsilon, n}) = \alpha_0 + \varepsilon^2 \mu \left\| \int IF(z; U, F_{\theta_0}) dG(z) \right\|^2 + o(\varepsilon^2)$$

where $\mu = -\frac{\partial}{\partial \delta} H_q(\eta_{1-\alpha_0}; \delta) |_{\delta=0}$, $H_q(\cdot; \delta)$ is the cumulative distr. fct. of $\chi_q^2(\delta)$, and $\eta_{1-\alpha_0}$ the $1 - \alpha_0$ quantile of a central χ_q^2 .

Similar result for the power.

To have a stable level in a neighborhood around the hypothesis, bound the influence function of the functional $U(F)$!

Special case :

$$\begin{aligned} G &= \Delta_z(\text{point mass at } z) \\ U &= U_W, U = U_R \end{aligned}$$

$$\lim_{n \rightarrow \infty} \alpha(F_{\varepsilon, n}) = \alpha_0 + \varepsilon^2 \cdot \underbrace{\mu \cdot IF_{(2)}^T V_{(22)}^{-1} IF_{(2)}}_{\text{self-stand. infl.}} + o(\varepsilon^2)$$

where

$$IF_{(2)} = IF(z; T_{(2)}, F_{\theta_0})$$

To obtain robust Wald and score type tests, bound the self-standardized influence function of the est. $T_{(2)}$.

Optimality theory.

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 Meth-
ods*, Wiley

◆ 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.