University of Illinois | Department of Economics
Spring 2017 | Econ 574 | Roger Koenker

Lecture 15
"Introduction to Robustness"

In econometrics the word "robustness" has no fixed meaning and seems to be used as a term of approval without any specific correlate. In statistics, since Hampel (1968), robustness means something quite specific. We will give a very brief exposition of Hampel's theory of qualitative robustness and then discuss some practical aspects of robust methods as they have developed over the last 30 years.[1]

Qualitative Robustness

It proves convenient to introduce a new way of representing estimators, rather than write

$$\hat{\theta} = \theta_n(y_1, \ldots, y_n)$$

we will write

$$\hat{\theta} = \theta_n(F_n)$$

where F_n is the usual empirical distribution function,

$$F_n(y) = n^{-1} \sum I(y_i \leq y)$$

E.g.,

$$\begin{aligned} \theta_n &= \int y dF_n(y) && \text{mean} \\ \theta_n &= F_n^{-1}(1/2) && \text{median} \\ \theta_n &= \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F_n^{-1}(u)du && \text{trimmed mean} \end{aligned}$$

The mapping $\theta_n(\cdot)$ induces a probability distribution for the estimator $\hat{\theta}_n$ under F which we may denote by $L_F(\theta_n)$. In general, $F_n \Rightarrow F$ and $\hat{\theta}_n \rightarrow \theta_\infty(F)$.

Def: Let $\mathcal{A}$ denote the Borel sets on $\Re$ for any $A \in \mathcal{A}$ and set $A^\varepsilon = \{x \in \Re| \ \inf_{y \in \mathcal{A}} |x-y| \leq \varepsilon\}$ the Prokhorov distance between F and G is given by,

$$\pi(F,G) = \ \inf \ \{\varepsilon | F\{A\} \leq G(A^\varepsilon) + \varepsilon \text{ for all } A \in \mathcal{A}\}$$

Robustness may now be defined as a continuity requirement on the mapping $\theta_n(\cdot)$. An Estimator, $\hat{\theta}$, is robust at F if small departures from F induce small departures in the distribution of $\hat{\theta}$ measured by Prokhorov distance.

[1] I learned only recently that Bahadur and Savage (1956) anticipates some of the central criticisms leveled by the robustness movement against conventional statistical practice based on moments. In particular, Bahadur and Savage point out that in any sufficiently rich class of non-parametric models means and variances *are not identified* and therefore conventional estimation and inference procedures about them are doomed to failure.

Def: (Hampel (1978) The sequence of estimators $\{\theta_n\}$ is robust at F iff for all $\varepsilon > 0$ there exists a $\delta > 0$ such that in Prokhorov metric, for all n,

$$\pi(F, G) < \delta \Rightarrow \pi(L_F(\theta_n), L_G(\theta_n)) < \varepsilon$$

This seems like a fairly innocuous requirement which should be satisfied by any reasonable estimator. Intuitively, it simply requires that if the model assumptions - here represented by F - don't change much, then the behavior of the estimator $\hat{\theta}_n$ won't change much either. Unfortunately, some reflection reveals that many common estimators in everyday use do not satisfy this requirement. Consider, for example the sample mean at $F = \Phi$, as we saw in Problem Set 1, the mixture density

$$F_\varepsilon = (1 - \varepsilon)\Phi + \varepsilon C$$

where C is the standard Cauchy df has the property that

$$\pi(\Phi, F_\varepsilon) = \varepsilon$$

But for any ε, the sample mean has, under F_ε, behavior radically different from its behavior under Φ. In particular, as $n \to \infty, L_\Phi(\theta_n) \to \delta_{\theta_0}$ and $L_{F_\varepsilon}(\theta_n) \to C$ so their P distance converges to 1.

A crucial tool in helping to understand qualitative robustness and one that provides some critical quantitative assessment of robustness is the influence function.

Def: The influence function, IF, of the estimator θ_n at F is

$$IF_{\theta_n,F}(x) = \lim_{\varepsilon \to 0}[\theta_n(F_\varepsilon) - \theta_n(F)]/\varepsilon$$

where $F_\varepsilon = (1 - \varepsilon)F + \varepsilon\delta_x$.

This is sometimes referred to as the Frechet derivative of the functional θ_n. We will not dwell on the technicalities which are well covered in Fernholz (1983) for example and Huber (1981).

Examples:

1. Mean $T(F) = \int x dF(x)$

$$T(F_\varepsilon) = (1 - \varepsilon)\mu(F) + \varepsilon x$$

so

$$IF_{(F)}(x) = x - \mu(F)$$

2. Sample Variance

$$T(F) = \int (x - \mu)^2 dF(x)$$

$$\begin{aligned} IF_{T,F}(x) &= \lim_{t \to 0} \frac{(1 - \varepsilon)\sigma^2(F) + t(x - \mu)^2 - \sigma^2(F)}{\varepsilon} \\ &= (x - \mu)^2 - \sigma^2 \end{aligned}$$

in both cases since $IF(x) \to \infty$ as $x \to \infty$ we have the possibility of wrecking as much havoc as desired by only ε contamination.

Moral: Unbounded $IF \Rightarrow$ Qualitative nonrobustness.

3. M-estimators of location

   T_n defined implicitly by the estimating equation:

$$\begin{aligned} \sum \psi(x_i - T_n) &= 0 \\ \textit{or} \\ \int \psi(x - T_n) dF_n(x) &= 0 \end{aligned}$$

   Now to compute IF, write F_ε is above and write

$$T_n(F_\varepsilon) = T_n(F) + \Delta,$$

   so expanding we have

$$\begin{aligned} 0 &= \int [\psi(x-T) + \Delta\psi'(x-T)] dF_\varepsilon(x) \\ \Rightarrow \quad \Delta &= -\frac{\int \psi(x-T)dF_\varepsilon(x)}{\int \psi'(x-T)dF_\varepsilon(x)} \\ &= -\frac{\varepsilon \int \psi(x-T)d\delta_{x_0}}{\int \psi'(x-T)dF_\varepsilon(x)} \\ &= -\frac{\varepsilon\psi(x_0 - T)}{\int \psi'(x-T)dF_\varepsilon(x)} \end{aligned}$$

   so

$$IF_{T_n,F}(x) = \frac{\psi(x - T_n)}{\int \psi'(x - T_n)dF(x)}$$

   Since the denominator is just a constant we may interpret the numerator as containing the essential information about the shape of the IF.

   Some Comparative Anatomy of M Estimators and Their IF's

   Note that to the extent that F has narrow tails the influence function for the optimal estimator based on F increases in the tails rapidly, as in the Gaussian case. Whereas, when the tails are long the IF tends to redescend to zero as in the case of the Cauchy, for example.

4. Influence Functions for L-estimators

   L-estimators for the one-sample model are simply linear combinations of order statistics of the form

$$T_n = \sum_{i=1}^{n} w_i X_{(i)}$$

   To study the influence function of such estimators, we begin by developing the IF of a single sample quantile. Consider the identity (for $u \in (0, 1)$),

$$F_\varepsilon(F_\varepsilon^{-1}(u)) = u$$

   where as usual, $F_\varepsilon = ((1-\varepsilon)F + \varepsilon\delta_x)$, and F is some smooth df and δ_x is the df representing point mass one at x. Differentiating our identity with respect to ε, we have,

$$-F(F_\varepsilon^{-1}(u)) + \delta_x(F_\varepsilon^{-1}(u)) + f_\varepsilon(F_\varepsilon^{-1}(u)) \cdot \frac{d}{d\varepsilon} F_\varepsilon^{-1}(u) = 0$$

and evaluating at $\varepsilon = 0$, yields,

$$\begin{aligned} IF(x,T,F) &= \frac{d}{d\varepsilon} F_\varepsilon^{-1}(u) = \frac{u - \delta_x(F^{-1}(u))}{f(F^{-1}(u))} \\ &= \begin{cases} (u-1)/f(F^{-1}(u)) & \text{if} \quad x \le F^{-1}(u) \\ u/f(F^{-1}(u)) & \text{if} \quad x > F^{-1}(u) \end{cases} \end{aligned}$$

Example 1. The simplest form of L-estimators is the so called “systematic statistic”, linear functions of a finite number of order statistics (sample quantiles). So, for example, for fixed $m < n$, the estimator

$$T_n = \sum_{i=1}^{m} w_i F_n^{-1}(u_i)$$

would have

$$IF(x,T,F) = \sum_{i=1}^{m} w_k IF(x, F_n^{-1}(u_i), F)$$

If T_n is to be an estimator of *location*, then we would require that it be *location equivariant*, i.e.,

$$T_n(x_1, x_2, \ldots, x_n) + \theta = T_n(x_1 + \theta, x_2 + \theta, \ldots, x_n + \theta)$$

This obviously requires that the weights used to define T_n sum to one. In addition, we would expect that if F were symmetric around zero, $T_n = T(F) = 0$, this requirement implies that the weights are “symmetric around 1/2.” Finally we would expect a reasonable location estimator to be *scale equivariant*, i.e.,

$$\sigma T_n(x_1, \ldots, x_n) = T_n(\sigma x_1, \ldots, \sigma x_n)$$

L-estimators may also be used to estimate *scale* or dispersion of a distribution, for example, the interquartle range is a simple robust alternative to the standard deviation for this purpose. For scale estimator we require *location invariance*, i.e.,

$$T_n(x_1, \ldots, x_n) = T_n(x_1 + \theta, \ldots, x_n + \theta)$$

and *scale equivariance*,

$$|\sigma| T_n(x_1, \ldots, x_n) = T_n(\sigma x_1, \ldots, \sigma x_n).$$

Note that the latter definition differs from its location counterpart in that σ is replaced by $|\sigma|$ to insure $T_n > 0$ for the scale estimation problem.

To investigate the large sample theory of such systematic statistics we obviously require that $f(F^{-1}(u_i)) > 0$ for $i = 1, \ldots, m$. Under this previso consider

$$\begin{aligned} Avar(T_n) &= EIF^2 \\ &= \sum_i \sum_j \frac{w_i w_j E(u_i - \delta_x(F^{-1}(u_i))(u_j - \delta_x F^{-1}(u_j))}{f(F^{-1}(u_i)) f(F^{-1}(u_j))} \end{aligned}$$

Note that for $u_i < u_j$

$$(u_i - \delta_x(\cdot))(u_j - \delta_x(\cdot)) = \begin{cases} (u_i-1)(u_j-1) & x < u_i \\ u_i(u_j-1) & u_i < x < u_j \\ u_i u_j & u_j < x \end{cases}$$

so taking expectations, we have (ugh!),

$$\begin{aligned} E(u_i - \delta_x(\cdot))(u_j - \delta_x(\cdot)) &= (u_i-1)(u_j-1)\int_{-\infty}^{F^{-1}(u_i)} dx + u_i(u_j-1)\int_{F^{-1}(u_i)}^{F^{-1}(u_j)} dx \\ &\quad + u_i u_j \int_{F^{-1}(u_j)}^{\infty} dx \\ &= (1-u_j)u_i \end{aligned}$$

So,

$$Avar(T_n) = w'\Omega w$$

where

$$\Omega = (w_{ij}) = \left(\frac{\min(u_i,u_j) - u_i u_j}{f(F^{-1}(u_i))f(F^{-1}(u_j))}\right)$$

An amusing sport, popular in the 1950's, is to compute "optimal" systematic statistics for particular distributions, e.g., take F to be Cauchy, and find the choice of $((w_i, u_i) \quad i = 1, \ldots, m)$ pairs for prespecified m to minimize the quadratic form $w'\Omega w$.

Another interesting exercise involves the comparison of the efficiency of the interquartile range and the standard deviation in the contaminated normal model, only slight amounts of contamination are required to make IQR preferable to s^2,

Example 2. A more interesting form for L-estimator are those with smooth weight functions,

$$T_n = \int_0^1 J(u) F_n^{-1}(u) du$$

Thus,

$$\begin{aligned} IF(x,T,F) &= \int_0^1 IF(x, F_n^{-1}(u), F) J(u) du \\ &= \int_0^1 J(u)\left[\frac{u - \delta_x(F^{-1}(u))}{f(F^{-1}(u))}\right] du \end{aligned}$$

let $u = F(y)$ so,

$$\begin{aligned} IF(x,T,F) &= \int J(F(y))(F(y) - \delta_x(y)) dy \\ &= \int_{-\infty}^{x} J(F(y)) dy - \int_{-\infty}^{\infty} (1 - F(y)) J(F(y)) dy \end{aligned}$$

2.1. Trimmed Mean

Perhaps the best, and oldest, of the "smooth" L-estimators are the trimmed means with

$$J(u) = \begin{cases} (1-2\alpha)^{-1} & \alpha < u < 1-\alpha \\ 0 & \text{otherwise} \end{cases}$$

To get the IF for these estimators suppose, to begin, that $x < F^{-1}(\alpha)$, then the first term above is zero and,

$$\begin{aligned} \int F(y)J(F(y))dy &= \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} F(y)dy \\ &= \frac{1}{1-2\alpha}[yF(y)|_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} - \int yf(y)dy] \\ &= \frac{1}{1-2\alpha}[F^{-1}(1-\alpha)(1-\alpha) - F^{-1}(\alpha)\alpha - \int_{\alpha}^{1-\alpha} F^{-1}(u)du] \end{aligned}$$

and

$$\int J(F(y))dy = (F^{-1}(1-\alpha) - F^{-1}(\alpha))/(1-2\alpha).$$

so,

$$\int (1-F(y))J(F(y))dy = \frac{1}{1-2\alpha}[\mu(F) - F^{-1}(\alpha)]$$

where $\mu(F) = \int_{\alpha}^{1-\alpha} F^{-1}(u)du - \alpha(F^{-1}(1-\alpha) + F^{-1}(\alpha))$.

Now,

$$\int_{-\infty}^{x} J(F(y))dy = \frac{1}{1-2\alpha}\int dy = \begin{cases} 0 & x < F^{-1}(\alpha) \\ \frac{1}{1-2\alpha}[x - F^{-1}(\alpha)] & x \in (F^{-1}(\alpha), F^{-1}(1-\alpha)) \\ \frac{1}{1-2\alpha}[F^{-1}(1-\alpha) - F^{-1}(\alpha)] & x > F^{-1}(1-\alpha) \end{cases}$$

so we can write

$$IF(x,T,F) = \begin{cases} \frac{1}{1-2\alpha}[F^{-1}(\alpha) - \mu(F) & x < F^{-1}(\alpha) \\ \frac{1}{1-2\alpha}[x - \mu(F)] & x \in (F^{-1}(\alpha), F^{-1}(1-\alpha)) \\ \frac{1}{1-2\alpha}[F^{-1}(1-\alpha) - \mu(F)] & x > F^{-1}(1-\alpha) \end{cases}$$

Note that if F is symmetric around zero, then $\mu(F) = 0$ and we have the picture of the Huber influence function.

The large sample theory of the trimmed mean can be developed directly from the IF, $\sqrt{n}(T_n - \mu(F))$ is asymptotically normal provided $f(F^{-1}(u_i)) > 0, u_i \in \{\alpha, 1-\alpha\}$ and the asymptotic variance is

$$\begin{aligned} Avar(T_n) &= \left(\frac{1}{1-2\alpha}\right)^2 [\alpha(F^{-1}(\alpha) - \mu(F))^2 \\ &\quad + \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} (x-\mu(F))^2 dx + (1-\alpha)(F^{-1}(1-\alpha) - \mu(F))^2] \end{aligned}$$

2.2. Optimal Smooth L-estimators

It is natural to ask how should we choose J if we *knew* $F = F_0$. Assuming two continuous derivatives of $\log f_0$ the answer is,

$$J_0(F_0(y)) = -\frac{(\log f_0(y))''}{I(F_0)}$$

where $I(F_0)$ is Fisher's Information Number.

Note that,

$$\int J_0(F_0(y))dF(y) = 1$$

which implies that T_n is *location equivariant.* (Try proving this assertion.)

To verify that using J_0 hits the Crámer-Rao lower bound, note that

$$\begin{aligned} IF(x, T_{J_0}, F_0) &= -I(F_0)^{-1}[\int_{-\infty}^{x} \psi_0'(y)dy - \int_{-\infty}^{\infty}(1 - F_0(y))\psi_0'(y)dy] \\ &= -\frac{\psi_0(x)}{I(F_0)} \qquad [\text{use the fact that } \int F_0\psi_0' dy = \int \psi_0 f_0 dy = 0]. \end{aligned}$$

Thus,

$$EIF^2 = \int \frac{\psi_0^2(x)}{(I(F_0))^2} f_0(x)dx = \frac{1}{I(F_0)}.$$

5. Some Heuristic Asymptotics for M-estimators

Since $\sqrt{n}(F_n - F) = O_p(1)$ and if

$$T(F_n) = T(F) + \int IF(x, T, F)dF_n + o(\| F_n - F \|)$$

we have

$$\begin{aligned} \sqrt{n}(T(F_n) - T(F)) &= \sqrt{n}\int IF dF_n + \sqrt{n}\ \ o(O_p(1/\sqrt{n})) \\ &= \frac{1}{\sqrt{n}}\sum IF(X_i) + o_p(1) \\ &\rightsquigarrow \mathcal{N}(O, E\ \ IF^2(X_i)) \end{aligned}$$

where

$$E\ \ IF^2(x) = E\psi^2(X)/(E\psi'(X))^2$$

This may be interpreted as the prototypical Huber Sandwich formula.

Useful Exercises

(a) Show that if ψ is optimal, then $E\ \ IF^2$ specializes to the inverse of Fisher's information.

(b) Show that if ψ is not optimal $E\ \ IF^2$ is greater than the inverse of Fisher's information and interpret this result.

Asymptotically Minimax Estimators

Huber(1964) posed the following problem. Let F_ε denote the following family of Φ-contaminated df's:

$$F_\varepsilon = \{F | F = (1-\varepsilon)\Phi + \varepsilon H, \quad H \in \mathcal{H}\}$$

where $\mathcal{H}$ is the set of df's *symmetric* about 0.

Q. What is the least favorable member of F_ε, i.e., the member which makes it as difficult as possible to estimate the center of distribution, i.e., maximizes the asymptotic variance of the best possible estimator of the location parameter θ from a sample from $F_\varepsilon(x-\theta)$.

A. The answer requires us to minimize, over F_ϵ,

$$I(F) = \int (f'/f)^2 f dx$$

This is a decidedly nontrivial problem and has the following solution:

$$f^*(x) = \begin{cases} (1-\varepsilon)\phi(x) & x \in [-k,k], \\ c\exp\{-\lambda|x|\} & \text{otherwise.} \end{cases}$$

for some constants, c, λ and k depending on ϵ. Thus, the least favorable density is Gaussian in the center and exponential in the tails. It is not easy to find an elementary treatment of this solution, but recently Jiaying Gu has suggested to me a nice calculus of variations argument that provides the essentials of the argument. I've added this as a new example to the brief tutorial that I've provided for 574 on calculus of variations as Lecture 12a.

The least favorable nature of the exponential is quite interesting – thicker tails than the exponential are actually informative about θ in a "negative way". The optimal ψ function for this least favorable f is the form

$$\psi(u) = \min\{k, \max\{u, -k\}\}$$

and is the M-estimator form of the more commonly used L-estimator, the α-trimmed mean.

$$\hat{\theta}_n = (1-2\alpha)^{-1} \int_\alpha^{1-\alpha} x dF_n(x).$$

The correspondance between ϵ, k and the trimming proportion α is suggested by some examples in Table 1. Table 2 provides asymptotic variances for this estimator for various choices of the trimming proportion α for several scale mixtures of the normal distributions. Note that although there is a small efficiency loss at the normal model, there is a potentially large gain from trimming in longer tailed error conditions.

Optimal k's and α's for Huber's
Model of Contamination

ε	k	α
0.001	2.63	0.005
0.01	1.95	0.031
0.05	1.40	0.102
0.10	1.14	0.164
0.20	.86	0.256

Asymptotic Variances of Trimmed Least Squares Estimators
for Contaminated Gaussian Distributions

Proportion	Relative	Trimming Proportion				
Contamination	Scale	0.0	.05	.10	.25	.50
0.00	1.0	1.00	1.03	1.06	1.19	1.57
0.05	3.0	1.40	1.16	1.17	1.29	1.68
0.05	5.0	2.20	1.20	1.20	1.31	1.70
0.05	10.0	5.95	1.25	1.23	1.33	1.72
0.10	3.0	1.80	1.32	1.30	1.39	1.80
0.10	5.0	3.40	1.46	1.38	1.45	1.85
0.10	10.0	10.90	1.65	1.45	1.49	1.89
0.25	3.0	3.00	2.14	1.85	1.80	2.26
0.25	5.0	7.00	4.11	2.39	2.01	2.45
0.25	10.0	25.75	13.65	3.65	2.19	2.61

References

Hampel. F. (1968) Contributions to the theory of Robust Estimation, Ph.d. Thesis, UC Berkeley.

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

Fernholz, L.T. (1983) *von Mises Calculus for Statistical Functionals*, Springer-Verlag.

Huber, P.J. (1964) Robust estimation of a location parameter, *Annals of Statistics*, 35, 73-101.

Bahadur, R. and L. Savage (1956) The Nonexistence of Certain Statistical Procedures in Nonparametric Problems, *Annals of Statistics*, 27, 1115-1122.