University of Illinois | Department of Economics
Spring 2012 | Econ 574 | Roger Koenker

Lecture 10
"Maximum Likelihood Asymptotics under Non-standard Conditions: A Heuristic Introduction to Sandwiches"

Ref: Huber, 5th Berkeley Symp, 1967.
White, *Econometrica*, 1982, & 1994 monograph
Gourieroux & Monfort §8.4 and §24.

It is frequently the case that we would like like to investigate the limiting behavior of an M-estimator when the strict conditions of mle do not apply. For example, how does the LS estimator behavior when the errors are Student not Normal.

The general framework is the following: we observe an iid sequence $\{Y_i, \ldots, Y_n\}$ from G with a density g. Next we specify a model $F(y, \theta)$ for $\theta \in \Theta \in \Re^p$ *compact* and consider the quasi-MLE, or QMLE, $\hat{\theta}$, which maximizes

$$l(\theta) = \sum_{i=1}^{n} \log f(y_i, \theta)$$

Now we can (or should) ask what does $\hat{\theta}$ estimate? The answer is provided by the Kullback Leibler Information Criterion (KLIC)

$$\begin{aligned} K(g, f_\theta) &= \int \log(g(y)/f(y,\theta)) dG(y) \\ &= \int \log g(y) dG - \int \log f(y,\theta) dG \end{aligned}$$

Note that the first term is independent of θ so minimizing K wrt to θ is the same as maximizing expected log likelihood. It should be emphasized that the KLIC minimizer may or may not correspond to something close to what we anticipated interpreting as θ when we wrote down the original likelihood. In some simple cases interpretation carries forward nicely, but in others the minimizer may be some only some distorted shadow of its former self.

There is a large literature on the suitablity of K as a measure of discrepancy between distributions. We will assume that $K(g, f_\theta)$ has a unique minimum wrt to θ, and denote this minimizing value by θ^*. This may be regarded as a (fairly) harmless identifiability condition. We also need the fact that $E_G \log f(Y, \theta) < M < \infty$ for all $\theta \in \Theta$.

Theorem 1: Under the foregoing conditions, $\hat{\theta} \to \theta^*$ a.s.

Proof: Modified Wald proof. See Huber (1967) for a beautiful treatment under the weakest possible conditions.

The next question to arise is asymptotic normality. For this we need to define the following:

$$\begin{aligned} H_n(\theta) &= n^{-1}\sum_{i=1}^{n}\nabla^2 \log f(Y_i,\theta) \\ J_n(\theta) &= n^{-1}\sum \nabla \log f(Y_i,\theta)\cdot \nabla \log f(Y_i,\theta)' \end{aligned}$$

The latter matrix is usual called the outer product of the gradient and the former, the Hessian of the likelihood. They are both $p \times p$ matrices, and we need these limits:

$$\begin{aligned} H_0(\theta) &= E_G \nabla^2 \log f(Y,\theta) \\ J_0(\theta) &= E_G \nabla \log f \nabla \log f' \end{aligned}$$

Under the *crucial* proviso that the latter quantities exist, the KSLLN implies that $H_n(\theta) \to H_0(\theta)$ and $J_n(\theta) \to J_0(\theta)$. Now by the mean value theorem we have,

$$\begin{aligned} s_n(\theta^*) &\equiv n^{-1}\sum \nabla \log f(Y_i,\theta^*) \\ &= s_n(\hat{\theta_n}) + (\theta^* - \hat{\theta_n})'\nabla s_n(\tilde{\theta}_n) \\ &= s_n(\hat{\theta_n}) + (\theta^* - \hat{\theta_n})' H_n(\tilde{\theta}_n) \end{aligned}$$

but $s_n(\hat{\theta_n}) = 0$ [or at least $\to 0$] and $H_n(\tilde{\theta}_n) \to H_0(\theta^*)$ since $\hat{\theta_n} \to \theta^*$, so

$$\sqrt{n}(\theta^* - \hat{\theta_n}) = -\sqrt{n}[H_0(\theta^*)]^{-1} s_n(\theta^*) + o_p(1).$$

Recall that θ^* minimizes $E_G \log f(Y,\theta_n)$ so

$$\mathcal{L}(\sqrt{n} s_n(\theta^*)) \rightsquigarrow \mathcal{N}(0, J_0(\theta^*))$$

so it follows immediately that

$$\mathcal{L}(\sqrt{n}(\hat{\theta}_n - \theta^*)) \rightsquigarrow \mathcal{N}(0, [H_0(\theta^*)]^{-1} J(\theta^*) [H_0(\theta^*)]^{-1})$$

This is our first encounter with the Huber sandwich. Recall that in the special case that $F = G$, we would have $-H_0 = J_0$ and the sandwich would collapse to $I(\theta_0)^{-1}$.

Examples and Remarks

(1) Linear Model with Heteroscedasticity of unknown form,

$$\begin{aligned} y_i &= x_i\beta_0 + u_i \\ u_i &\sim \mathcal{N}(0, \sigma_i^2) \end{aligned}$$

We mistakenly assume $Euu' = \sigma^2 I$ i.e., $\sigma_i^2 \equiv \sigma^2$. So the QMLE is $\hat{\beta} = (X'X)^{-1}X'y$ and $\beta^* = \beta_0$ and there is no bias due to misspecification. Further, ignoring a constant,

$$\begin{aligned} l_n(\beta) &= (y - X\beta)'(y - X\beta) \\ \\ \text{so} \qquad & \nabla l_n(\beta) = X'(y - X\beta) \\ & E\nabla l_n \nabla l_n' = X'\Omega X \\ & \nabla^2 l_n(\beta) = -X'X \end{aligned}$$

so we have applying the theorem,

$$(\hat{\beta} - \beta) \rightsquigarrow \mathcal{N}(0, (X'X)^{-1}X'\Omega X(X'X)^{-1})$$

which is the well known Eicker-White heteroscedasticity-consistent covariance matrix result.

(2) There are innumerable other examples; perhaps the most famous is the case of robust estimation of location in which,

$$\begin{aligned} \rho(y - \theta) &= -\log f(y - \theta) \\ s_n(\theta) &= n^{-1} \sum_{i=1}^{n} \psi(y_i - \theta) \qquad\qquad \psi = \rho' \end{aligned}$$

If $\{\theta_n\}$ is a sequence of solutions of $s_n(\theta) = 0$, then, by the foregoing,

$$\mathcal{L}(\sqrt{n}(\hat{\theta}_n - \theta^*)) \to \mathcal{N}(0, E\psi^2(Y - \theta^*)/(E\psi'(Y - \theta^*))^2)$$

Here, $H_0 = E\psi'$ and $J_0 = E\psi^2$ and everything is nice and simple: iid, scalar.

Two examples within the example

For the choice $f = \phi$, i.e., standard normal we have $\psi(u) = u$ so $E\psi^2/(E\psi')^2 = \sigma_G^2$ so the variance of $\sqrt{n}(\hat{\mu} - \mu)$ tends to the variance of G for $\hat{\mu}$ chosen to be the sample mean. For f Laplace, i.e., $f(u) = \frac{1}{2}e^{-|u|}$ we have $\psi(u) =$ sgn (u) and $E\psi^2/(E\psi')^2 = 1/(2g(0))^2$ for any symmetric G. Thus, comparing the mean vs. median for $G = \Phi$ yields

$$\text{ARE } = \sigma_G^2/(1/(2g(0)))^2 = \frac{2}{\pi}$$

so the median is about 36% less efficient asymptotically than the mean at the normal model. In the Laplace case it is easy to see show that ARE = 2, so the median only needs half as many observations as the mean to acheive the same precision. This case is the subject of Kolmogorov's first statistics paper in 1931 and was also investigated by Laplace. In a dramatically different case, with G Cauchy, $\sigma_G^2 = \infty$ while $1/(2g(0))^2 = \pi^2/4 \approx 2.46$. The Fisher Information for the standard Cauchy is $1/2$ so the MLE, which achieves the CRLB of 2, is again quite a bit better than the median, but both are "infinitely better" than the mean.

(3) In the strict mle case we 'know'(!) that $J_0(\theta^*) = -H_0(\theta^*)$ so $H^{-1}JH^{-1} = -H^{-1} = J^{-1}$ this is convenient, but undoubtedly overly utopian.

(4) If one is tempted to believe the mle fairy tale, then one can compute standard errors by several different asymptotic approximations. Either (i) approximations to H_0^{-1}, (ii) approximation to J_0^{-1} or the outer products matrix, or (iii) finally by approximation $H_0^{-1}J_0H_0^{-1}$. The latter is undoubtedly superior, but more difficult.

(5) Discrepancies between these versions of the covariance matrix of $\hat{\theta}$, suggest that since such discrepancies occur only under circumstances in which f, the model, and g, the process generating the data, differ it becomes interesting to test for specification using the size of the discrepancy. More interesting is the question: can we interpret the nature of the discrepancy applications? This is the focus of White's paper and subsequent work by Chesher, Lancaster and many others. More on this later in the course.

Likelihoods from exponential family

Ref: McCulloch, "KLIC and LEF's," *American Statistician*, 1988, GM also has a good discussion of some of these ideas based on a couple of *Econometrica* papers that they wrote with Trognon, in the 1980's.

In this section we explore briefly a special case of the foregoing general theory. Consider a model based on $X \sim P_\theta$ with exponential family density

$$p(x|\theta) = \exp\{x'\theta - c(\theta)\}g(x)d\mu(x)$$

Recall that $EX = \nabla c(\theta)$ and $VX = \nabla^2 c(\theta)$. Now suppose that we have iid observations on the random variable $Y \sim Q$, some other "data generating process". We would like to explore the consequences of analysing Y as if it were X. We will assume that Q is absolutely continuous with respect to μ so $dQ = qd\mu$, and will also assume $EY = m$ exists, i.e., $m < \infty$. Consider the KL divergence between P_θ and Q

$$K(Q, P) = \int q \log(q/p) d\mu$$

Note that this relation is not symmetric $K(Q, P) \neq K(P, Q)$, but $K(Q, P) \geq 0$ with equality iff $Q = P$, since by Jensen's inequality

$$-K(Q, P) = \int \log(p/q) q d\mu \leq \log(\int p d\mu) = 0.$$

The model P_θ provides a "nice" family of densities since
(i) The parameter space $\Theta = \{\theta \in \Re^P | p(x|\theta) < \infty\}$ is convex, and
(ii) The loglikelihood is concave, i.e., $\nabla^2 c(\theta)$ is positive definite.

Proposition 1: $K(Q, P_\theta)$ has a unique minimum, say θ^*, which is sometimes called the "pseudo-true" value of θ. (It plays the role of θ_0 is the strict MLE setup.) Further, if $\theta^* \in$ int Θ, then $\theta^* = \nabla c^{-1}(m)$.

Proof:

$$\begin{aligned} K(Q, P_\theta) &= \int q \log(q/p_\theta) d\mu \\ &= \int q(x) \log(q(x)/g(x)) d\mu + \int (c(\theta) - x'\theta) q d\mu \\ &= K(Q, P_0) + c(\theta) - m'\theta \end{aligned}$$

But $\nabla^2 c(\theta) >> 0$, so $c(\theta) - m'\theta$ is strictly convex on Θ. This establishes uniqueness of θ^*. If $\theta^* \in$ int Θ, then $\nabla c(\theta^*) = m$, so $\theta^* = (\nabla c)^{-1}(m)$.

Remark: Since $EX = \nabla c(\theta)$, we may interpret θ^* as the value of θ which makes P_θ have mean m, so we are really doing "moment matching," or GMM.

Proposition 2: Let $\hat{\theta}_n$ be the qmle based on a random sample $Y_1, \ldots, Y_n$ from Q using P_θ as the likelihood. Then, $\hat{\theta}_n \to \theta^*$ a.s.

Proof: The log likelihood is denoting a generic constant by K

$$l_n(\theta) = K + n(\bar{Y}_n'\theta - c(\theta)).$$

By the KSLLN, $\bar{Y}_n \to m$ a.s. so $n^{-1} l_n(\theta)$ tends, uniformly in θ, to $m'\theta - c(\theta)$ plus a term independent of θ. But $K(Q, P_\theta) = c(\theta) - m'\theta + K'$ so the maximizer of $l_n(\theta)$ tends to the minimizer of $K(Q, P_\theta)$.

Proposition 3: Suppose $V(Y) = \Omega$ exists and $\theta^* \in$ int Θ and denote $H = \nabla^2 c(\theta)$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \rightsquigarrow \mathcal{N}(0, H^{-1}\Omega H^{-1})$$

Proof: $\hat{\theta}_n = \nabla c^{-1}(\bar{Y}_n)$ for n sufficiently large, since $\theta^* \in$ int Θ. And $\theta^* = \nabla c^{-1}(m)$ by Proposition 1. Thus,

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta^*) &= \sqrt{n}(\nabla c^{-1}(\bar{Y}) - \nabla c^{-1}(m)) \\ &= \sqrt{n}(H^{-1}(\bar{Y}_n - m)) + o_p(1) \end{aligned}$$

Interpretation of KL-divergence via density estimation.

Given $Y_1, \ldots, Y_n$ we may construct the usual empirical *df*,

$$F_n(y) = n^{-1} \sum_{i=1}^{n} I(Y_i < y)$$

Now we can smooth this step function by adding a little noise to it, e.g.,

$$\begin{aligned} \hat{Y} &\sim F_n(y) \\ \tilde{Y} &= \hat{Y} + U \qquad \text{where } U \sim (G, g) \end{aligned}$$

then

$$\begin{aligned} f_{\tilde{Y}}(y) &= \int \hat{f}(x) g(y-x) dx \\ &= \int g(y-x) d\hat{F}(x) \\ &= n^{-1} \sum g(y - X_i) \end{aligned}$$

so this is the usual kernel density estimator in which we use $g(\cdot)$ as the kernel. And it provides a new interpretation of the kernel density estimation as smoothing the r.v. $\hat{Y} \sim F_n(y)$ by adding a little random noise thus replacing its step function df, by something smoother.

Now, let $\tilde{Q}_n$ denote the df of the convolution, $\tilde{Y}_n$. Note that the expectation.

$$\int x d\tilde{Q}_n(x) = \int \tilde{Q}_n^{-1} dx = E(\hat{Y}_n + U) = E\hat{Y} = \bar{Y}_n.$$

where we have made the change of variable $x = \tilde{Q}_n^{-1}$. And we have also assumed that the noise introduced by U has mean zero. We have also assumed that we haven't altered the support of the distribution by the smoothing operation. Now we can write

$$\begin{aligned} K(Q, P_\theta) &\approx K(\tilde{Q}_n, P_\theta) \\ &= K(\tilde{Q}_n, g d\mu) + \int (c(\theta) - x'\theta) \tilde{q}_n d\mu \\ &= K(\tilde{Q}_n, P_0) + c(\theta) - \bar{Y}_n' \theta \\ &= -n l_n(\theta) + K \end{aligned}$$

Thus the likelihood may be regarded as an estimate of the function mapping θ to $K(Q, P_\theta)$, since we don't know $K(Q, P_\theta)$ explicitly we compute θ^* by maximizing the likelihood rather than by minimizing $K(Q, P_\theta)$. The smoothing just facilitates the integration in a nice way.