14.385

Nonlinear Econometrics

Lecture 3.

Theory: Consistency Continued.

Asymptotic Distribution of Extremum Estimators

1. MLE and M-Estimator Consistency

Here we apply extremum consistency theorem to establish consistency of MLE and other M-estimators.

Theorem 1 *(MLE Consistency): If z_i is i.i.d. with pdf $f(z|\theta)$ and (i) (Identification) $f(z_i|\theta) \neq f(z_i|\theta_0)$ with positive probability, for all $\theta \neq \theta_0$; (ii) (Compactness.) Θ is compact; (iii) (Continuity) $f(z|\theta)$ is continuous at all θ with probability one; (iv) (Dominance) $E[\sup_{\theta \in \Theta} |\ln f(z|\theta)|] < \infty$; then $\hat{\theta} \xrightarrow{p} \theta_0$.*

In MLE we have $\hat{\theta} \in \arg\inf_{\theta \in \Theta} \hat{Q}(\theta)$, where

$$\hat{Q}(\theta) = -E_n[\ln f(z_i, \theta)].$$

By the LLN, the limit of the MLE objective function will be $Q(\theta) = -E[\ln f(z|\theta)]$. The following result shows that the identification condition in condition (i) is sufficient for $E[\ln f(z|\theta)]$

to be uniquely maximized at the true parameter.

Lemma 1 (**Information inequality**): *If*

$$\int |\ln f(z|\theta)| f(z|\theta_0) dz < \infty$$

for each $\theta \in \Theta$ *then if* $f(z|\theta) \neq f(z|\theta_0)$ *we have*

$$E \ln[f(z|\theta)] < E \ln[f(z|\theta_0)]$$

Proof: By the strict version of Jensen's inequality and concavity of $\ln(v)$,

$$\int \ln[f(z|\theta)/f(z|\theta_0)] f(z|\theta_0) dz \quad (1)$$

$$< \ln\{\int [f(z|\theta)/f(z|\theta_0)]\, f(z|\theta_0) dz\} \quad (2)$$

$$= \ln \int f(z|\theta) dz = 0. \quad (3)$$

It is interesting to note that identification of θ_0, in the sense that changing the parameter changes the density, is sufficient condition for $Q(\theta) = E[\ln f(z|\theta)]$ to have a unique maximum at θ_0. For other extremum problems, it is often harder to give such simple sufficient condition for identifiability.

Proof of MLE consistency: It suffices to check conditions of Extremum Consistency Theorem. For identification condition (i) let $Q(\theta) = E[\ln f(z|\theta)]$. By the information inequality, for $\theta \neq \theta_0$,

$$Q(\theta) - Q(\theta_0) = E[\ln\{f(z|\theta)/f(z|\theta_0)\}] < 0,$$

giving identification. Compactness condition (ii) is assumed. Continuity condition (iii) and uniform convergence condition(iv) hold by the ULLN of L2 applied to $\hat{Q}(\theta) = E_n[\ln f(z_i|\theta)]$. Q.E.D.

Binary Choice Continued: The probit (conditional) likelihood for a single observation $z =$

(y, x) where $y \in \{0, 1\}$ is

$$f(z|\theta) = \Phi(x'\theta)^y [1 - \Phi(x'\theta)]^{1-y},$$

where $\Phi(v)$ is the standard normal CDF.

Theorem on Consistency of probit: *If $E[xx']$ is finite and nonsingular then the probit MLE $\hat{\theta}$ satisfies $\hat{\theta} \xrightarrow{p} \theta_0$.*

Proof: For the standard normal pdf $\phi(v)$, it is well known that $\partial \ln \Phi(v)/\partial v = \phi(v)/\Phi(v)$ is decreasing, so that $\ln \Phi(v)$ is concave. Also, $\ln[1 - \Phi(v)] = \ln \Phi(-v)$ is concave. Then, since a concave function of a linear function is concave, $\ln f(z|\theta) = y \ln \Phi(x'\beta) + (1 - y) \ln[\Phi(-x'\theta)]$ is concave. Therefore, to show consistency, it suffices to verify the conditions (i) - (iii) of Theorem on Consistency of Argmin Estimators with Convexity.

i) By nonsingularity of $E[xx']$, for $\theta \neq \theta_0$, $E[\{x'(\theta - \theta_0)\}^2] = (\theta - \theta_0)' E[xx'](\theta - \theta_0) > 0$, implying that $x'(\theta - \theta_0) \neq 0$ and hence $x'\theta \neq x'\theta_0$ with positive probability under θ_0. Both $\Phi(v)$ and $\Phi(-v)$ are strictly monotonic, so that $x'\theta \neq x'\theta_0$ implies both $\Phi(x'\theta) \neq \Phi(x'\theta_0)$ and $\Phi(-x'\theta) \neq \Phi(-x'\theta_0)$ Therefore,

$$f(z|\theta) = \Phi(x'\theta)^y \Phi(-x'\theta)^{1-y} \neq f(z \mid \theta_0),$$

with positive probability under θ_0.

ii) The set $\Theta = \mathbb{R}^p$, where p is the dimension of θ, is convex.

iii) By Feller inequality, there is a constant C such that $\phi(v)/\Phi(v) \leq C(1 + |v|)$, so that by integrating there is a constant C such that $|\ln \Phi(v)| \leq C(1 + |v|^2)$. Then $E[|\ln f(z|\theta)|] \leq E[2C(1 + |x'\theta|^2)] < \infty$ by existence of second moments of x.

In M-estimation we have $\widehat{\theta} \in \arg\inf_{\theta \in \Theta} \widehat{Q}(\theta)$, with the criterion function taking a form of an average

$$\widehat{Q}(\theta) = E_n[m(z_i, \theta)].$$

The limit criterion function $Q(\theta)$ is assumed to be minimized at the true parameter value θ_0.

Theorem 2 *(M-Estimator Consistency): If z_i are i.i.d. or stationary and strongly mixing, and (i) (Identification) $Em(z_i, \theta) > Em(z_i, \theta_0)$ for for all $\theta \neq \theta_0$; (ii) (Compactness.) Θ is* *compact; (iii) (Continuity) for each* θ, $m(z,\theta)$ *is continuous at* θ *with probability one; (iv) (Dominance)* $E[\sup_{\theta\in\Theta}|m(z,\theta)|]<\infty$*; then* $\widehat{\theta}\xrightarrow{p}\theta_0$.

Proof: The argument is identical to the proof of MLE consistency, apart from verification of identifiability, which we assumed here directly.

2. Consistency of GMM

The consistency result for GMM is similar to that for MLE. The most important difference is in the identification hypotheses: here assume that

$$\bar{g}(\theta) = E[g(z,\theta)] = 0$$

must have a unique solution at the true parameter value θ_0.

Theorem 3 *(GMM Consistency): If z_i are i.i.d. or stationary and strongly mixing (i) $E[g(z,\theta_0)] = 0$ and $E[g(z,\theta)] \neq 0$ for $\theta \neq \theta_0$ and $\widehat{A} \xrightarrow{p} A$ with A positive definite; (ii) Θ is compact; (iii) for each θ, $g(z,\theta)$ is continuous at θ with probability one; (iv)*
$E[\sup_{\theta\in\Theta} \|g(z,\theta)\|] < \infty$, *then* $\widehat{\theta} \xrightarrow{p} \theta_0$.

It is often hard to specify primitive conditions for the identification condition (i). This condition amounts to assuming that there is a unique solution to the set of nonlinear equations $\bar{g}(\theta) = 0$. Global conditions for uniqueness are hard to specify. A simple local identification condition is that $\partial\bar{g}(\theta_0)/\partial\theta$ has full column rank p. Often, one can exchange the expectation and the derivative, so that this condition is

$$rank(E[\partial g(z_i,\theta_0)/\partial\theta]) = p.$$

Proof: We proceed by verifying the hypotheses of the Theorem on Consistency of Extremum Estimators (Theorem 2.1). Here we have

$$\widehat{Q}(\theta) = \widehat{g}(\theta)'\widehat{A}\widehat{g}(\theta) \text{ and } Q(\theta) = \bar{g}(\theta)'A\bar{g}(\theta).$$

Since $Q(\theta)$ has a unique minimum of zero at θ_0, the identification condition (i) of Theorem 2.1 is satisfied. Compactness condition (ii) holds by hypothesis.

Uniform convergence of $\widehat{Q}$ to Q (condition (iv)) and the continuity of Q (condition (iii)) "clearly follow" from (a) the uniform convergence of $\widehat{g}$ to a continuous function $\bar{g}$, where uniform convergence and continuity follows by the ULLN, (b) from the convergence of $\widehat{A}$ to A. $\square$

Remarks:

1. The details of the "clearly follow" step are as follows. By (iii) and (iv) of the GMM hypothesis, the ULLN lemma gives continuity of $\bar{g}(\theta)$ and $\sup_{\theta\in\Theta} \|\widehat{g}(\theta)-$

$\bar{g}(\theta)\| \xrightarrow{p} 0$ Therefore $Q(\theta)$ is continuous, so that condition (iii) of Extremum Consistency Theorem 2.1 holds. It follows from the triangle inequality that $\sup_{\theta\in\Theta} \|\hat{g}(\theta)\| \leq \sup_{\theta\in\Theta} \|\hat{g}(\theta) - \bar{g}(\theta)\| + \sup_{\theta\in\Theta} \|\bar{g}(\theta)\| = O_p(1)$. Let $\|A\|$ be the maximum eigenvalue of A. Then

$$
\begin{aligned}
|\hat{Q}(\theta) &- Q(\theta)| \leq |\hat{g}(\theta)'(\hat{A} - A)\hat{g}(\theta)| \\
&+ |[\hat{g}(\theta) - \bar{g}(\theta)]'A[\hat{g}(\theta) - \bar{g}(\theta)]| \\
&+ 2|\bar{g}(\theta)'A[\hat{g}(\theta) - \bar{g}(\theta)]| \\
&\leq \sup_{\theta\in\Theta} \|\hat{g}(\theta)\|^2 \|\hat{A} - A\| \\
&+ \sup_{\theta\in\Theta} \|\hat{g}(\theta) - \bar{g}(\theta)\|^2 \|A\| \\
&+ 2 \sup_{\theta\in\Theta} \|\bar{g}(\theta)\| \|A\| \sup_{\theta\in\Theta} \|\hat{g}(\theta) - \bar{g}(\theta)\| \xrightarrow{p} 0.
\end{aligned}
$$

Taking the supremum over Θ of the left-hand side, it follows that $\hat{Q}$ converges to Q uniformly.

Asymptotic Distribution of Extremum Estimators Assume conditions of Theorem 1, so that

$$\hat{\theta} = \arg\inf_{\theta \in \Theta} \hat{Q}(\theta)$$

is consistent for

$$\theta_0 = \arg\inf_{\theta \in \Theta} Q(\theta).$$

If $\hat{Q}$ is smooth, then $\hat{\theta}$ is a root of the first order condition:

$$\nabla \hat{Q}(\hat{\theta}) = 0, \tag{4}$$

which we can expand as*

$$\nabla \hat{Q}(\theta_0) + \nabla^2 \hat{Q}(\theta^*)(\hat{\theta} - \theta_0) = 0, \tag{5}$$

and solve for the estimate

$$\sqrt{n}(\hat{\theta} - \theta_0) = -[\nabla^2 \hat{Q}(\theta^*)]^{-1} \sqrt{n} \nabla \hat{Q}(\theta_0). \tag{6}$$

Then we look for conditions to give as a LLN:

$$\nabla^2 \hat{Q}(\theta^*) \to_p J \tag{7}$$

*$\nabla^2 \hat{Q}(\theta^*)$ stands for a Hessian matrix $\nabla^2 \hat{Q}$ with each row evaluated at different intermediate point θ^*.

and a CLT:

$$\sqrt{n}\nabla\hat{Q}(\theta_0) \to_d N(0, \Omega), \qquad (8)$$

so that by CMT

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d J^{-1}N(0, \Omega) = N(0, J^{-1}\Omega J^{-1}). (9)$$

In some case, like in the case of regular MLE, we will have a **generalized information matrix equality**:

$$-J = \Omega,$$

leading to the simplification of the asymptotic variance formula to J^{-1}.

Let us formalize this into a Theorem:

Theorem 4 *If* $\hat{\theta} \xrightarrow{p} \theta_0$ *and i)* $\theta_0 \in interior(\Theta)$; *ii)* $\hat{Q}(\theta)$ *is twice continuously differentiable in a neighborhood* N *of* θ_0*; iii)* $\sqrt{n}\nabla\hat{Q}(\theta_0) \xrightarrow{d} N(0, \Omega)$; *iv) there is* $J(\theta)$ *that is continuous at* θ_0 *and*

$\sup_{\theta\in\mathcal{N}} \left\|\nabla^2\widehat{Q}(\theta) - J(\theta)\right\| \xrightarrow{p} 0$; *v)* $J = J(\theta_0)$ *is nonsingular. Then*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, J^{-1}\Omega J^{-1}).$$

Proof: By consistency, $\widehat{\theta}$ is in a small open neigborhood $\mathcal{N}$ of θ_0 wp $\to 1$. Condition (ii) enables (4) and (5). Condition (iv) and the continuous mapping theorem imply (7): $\nabla^2\widehat{Q}(\theta^*) \to_p J$. (To see this note $\nabla^2\widehat{Q}(\theta^*) - J(\theta^*) \to_p 0$, and $J(\theta^*) - J(\theta_0) \to_p 0$ by $\theta^* \to_p \theta_0$ and continuous mapping theorem). This in turn implies by (v) and the continuous mapping theorem: $[\nabla^2\widehat{Q}(\theta^*)]^{-1} \to_p J^{-1}$, which enables (6). By Slutsky we have (9). Q.E.D.

Remark* **(Non-Smooth Case)**: Is $\hat{Q}$ is not smooth but has a unique gradient $\nabla \hat{Q}$ almost everywhere (for points of non-uniqueness, choose any subdifferential to be a gradient). Assume that we can define $\hat{\theta}$ as a root of the first order condition holding approximately:

$$\nabla \hat{Q}(\hat{\theta}) = o_p(1/\sqrt{n}). \tag{10}$$

We then impose the **stochastic differentiability** condition, replacing the previous Taylor expansion, namely:

$$\nabla \hat{Q}(\hat{\theta}) = \nabla \hat{Q}(\theta_0) + [J + o_p](\hat{\theta} - \theta_0), \tag{11}$$

and solve for the estimate

$$\sqrt{n}(\hat{\theta} - \theta_0) = -[J + o_p]^{-1}\sqrt{n}\nabla \hat{Q}(\theta_0). \tag{12}$$

Then as before, we require a CLT:

$$\sqrt{n}\nabla \hat{Q}(\theta_0) \to_d N(0, \Omega), \tag{13}$$

and conclude by CMT

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, G^{-1}\Omega G^{-1}). \tag{14}$$

Stochastic differentiability is somewhat more difficult to establish but is quite fruitful for quantile regression and related methods. We'll cover some methods for establishing this condition in theoretical exercises.