Lecture 2
Some Useful Asymptotic Theory

As seen in the last lecture, linear least square has an analytical solution: $\hat{\beta}_{OLS} = (X'X)^{-1} X'y$. The consistency and asymptotic normality of $\hat{\beta}_n$ can be established using LLN, CLT and generalized Slutsky theorem. When it comes to nonlinear models/methods, the estimators typically do not have analytical solution. For example, a nonlinear regression model:

$$y = \alpha X^{\theta} + \varepsilon, E(\varepsilon|X) = 0. \tag{1}$$

A typical estimator is obtained by nonlinear least square:

$$(\hat{\alpha}_{NLS}, \hat{\theta}_{NLS}) = \arg \min_{(\alpha,\theta)\in A\times\Theta\subseteq R^2} \hat{Q}_n(\alpha, \theta), \tag{2}$$

where $\hat{Q}_n(\alpha, \theta) = n^{-1} \sum_{i=1}^{n} \left(y_i - aX_i^{\theta}\right)^2$. The nonlinear least square problem does not have analytical solution. In order to study the consistency of $(\hat{\alpha}_{NLS}, \hat{\theta}_{NLS})$, we need to know the asymptotic behavior of $\hat{Q}_n(\alpha, \theta)$ on the set $A \times \Theta \subseteq R^2$. In other words, we need the uniform (probability or almost sure) limit behavior of $\hat{Q}_n(\alpha, \theta)$ on the set $A \times \Theta \subseteq R^2$. Uniform laws of large numbers are tools serving that purpose.

Once the consistency is established, we then expand the objective function around the true value of the parameters, aiming to obtain the asymptotic dispersion of the estimators around the true value, namely the asymptotic distribution of the estimators. The expansion is made possible by the mean-value theorem. Details on consistency and asymptotic normality will be covered in the next few lectures. This lecture focuses on uniform laws of large numbers.

Let $\mathcal{X} \times \Theta$ be the Cartesian product of Euclidean sets $\mathcal{X}$ and Θ. Let $g(x, \theta)$ be a real-valued function defined on $\mathcal{X} \times \Theta$. Function $g(\cdot, \theta)$ is Lebesgue measurable for every $\theta \in \Theta$. Let $X_1, X_2, ...$ be a sequence of iid random variables on $\mathcal{X}$. A uniform (weak) law of large numbers defines a set of conditions under which[1]

$$\sup_{\theta\in\Theta} \left| n^{-1} \sum_{i=1}^{n} g(X_i, \theta) - Eg(X_i, \theta) \right| \to_p 0. \tag{3}$$

Trivial Case: $\Theta = \{\theta_1, ..., \theta_K\}$, $K < \infty$. That is, the set Θ is finite. In this case, pointwise LLN is enough. The condition for pointwise LLN is: $Eg(X_i, \theta) < \infty$ for all $\theta \in \Theta$. Suppose this

[1] It is a strong law of large number if the convergence holds almost surely instead of in probability. In this course, we only need weak law of large numbers, though some of the conditions we give today are strong enough to obtain strong law of large numbers.

condition hold. Then,

$$\sup_{\theta\in\Theta}\left|n^{-1}\sum_{i=1}^{n}g(X_i,\theta)-Eg(X_i,\theta)\right|\leq\sum_{k=1}^{K}\left|n^{-1}\sum_{i=1}^{n}g(X_i,\theta_k)-Eg(X_i,\theta_k)\right|\to_p 0. \tag{4}$$

When Θ is not finite, the idea is to approximate the supremum over Θ by the supremum over a finite subset of Θ. Loosely speaking, the two suprema are close if g is "not too volatile" on Θ. Various ULLNs restrict the "volatility" of g in different ways. Here we introduce an old and simple, but rather useful ULLN.

1 A Simple ULLN

The following theorem dates back to Jennrich (1969, Theorem 2) or even earlier:

Theorem (ULLN1): *Suppose* (a) Θ *is compact,* (b) $g(X_i,\theta)$ *is continuous at each* $\theta\in\Theta$ *with probability one,* (c) $g(X_i,\theta)$ *is dominated by a function* $G(X_i)$, i.e. $|g(X_i,\theta)|\leq G(X_i)$, *and* (d) $EG(X_i)<\infty$. *Then* (3) *holds.*

Proof. First, define $\Delta_\delta(X_i,\theta_0)=\sup_{\theta\in B(\theta_0,\delta)}g(X_i,\theta)-\inf_{\theta\in B(\theta_0,\delta)}g(X_i,\theta)$. Now $E\Delta_\delta(X_i,\theta_0)\downarrow 0$ as $\delta\downarrow 0$ because (i) $\Delta_\delta(X_i,\theta_0)\downarrow 0$ a.s. by condition (b), (ii) $\Delta_\delta(X_i,\theta_0)\leq 2\sup_{\theta\in\Theta}|g(X_i,\theta)|\leq 2G(X_i)$ by condition (c), and condition (d).

So, for all $\theta\in\Theta$ and $\varepsilon>0$, there exists $\delta_\varepsilon(\theta)$ such that $E\left[\Delta_{\delta_\varepsilon(\theta)}(X_i,\theta)\right]<\varepsilon$.

Obviously, we can cover the whole parameter space Θ by $\{B(\theta,\delta_\varepsilon(\theta)):\theta\in\Theta\}$. In fact, because Θ is compact, we can find a finite subcover, such that Θ is covered by $\cup_{k=1}^{K}B(\theta_k,\delta_\varepsilon(\theta_k))$.

Note that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left[n^{-1}\sum_{i=1}^{n}g(X_i,\theta)-Eg(X_i,\theta)\right]\\
=&\max_k\sup_{\theta\in B(\theta_k,\delta_\varepsilon(\theta_k))}\left[n^{-1}\sum_{i=1}^{n}g(X_i,\theta)-Eg(X_i,\theta)\right]\\
\leq&\max_k\left[n^{-1}\sum_{i=1}^{n}\sup_{\theta\in B(\theta_k,\delta_\varepsilon(\theta_k))}g(X_i,\theta)-E\inf_{\theta\in B(\theta_k,\delta_\varepsilon(\theta_k))}g(X_i,\theta)\right]\\
=&\,o_p(1)+\max_k\left[E\sup_{\theta\in B(\theta_k,\delta_\varepsilon(\theta_k))}g(X_i,\theta)-E\inf_{\theta\in B(\theta_k,\delta_\varepsilon(\theta_k))}g(X_i,\theta)\right]\\
=&\,o_p(1)+\max_k E\Delta_{\delta_\varepsilon(\theta_k)}(X_i,\theta_k)\\
\leq&\,o_p(1)+\varepsilon,
\end{aligned} \tag{5}$$

where the first equality holds by the WLLN, which applies because $E\left|\sup_{\theta\in B(\theta_k,\delta_\varepsilon(\theta_k))} g(X_i,\theta)\right| \leq EG(X_i) < \infty$, and the last inequality holds by the way we define $\delta_\varepsilon(\theta_k)$.

By analogous argument,

$$\inf_{\theta\in\Theta}\left[n^{-1}\sum_{i=1}^{n} g(X_i,\theta) - Eg(X_i,\theta)\right] \geq o_p(1) - \varepsilon. \tag{6}$$

The desired result follows from (5) and (6) by the fact that ε is chosen arbitrarily. ■

Comments on ULLN1: 1. Condition (a) is not a strong assumption and is often assumed in extremum estimation problems. It is effectively imposed in practice because computers don't deal with real infinity. Compactness can be replaced with boundedness without affecting the result or the proof.

2. Condition (b) does not require $g(x,\theta)$ to be continuous at all $\theta \in \Theta$ for a given x. Thus the theorem applies to the cases when the g functions are non-smooth. Condition (b) can usually be verified by visual inspection.

3. ULLN1 can be readily extended to stationary data.

Example 1: (NLS continued) In the NLS example introduced at the beginning of the lecture, the NLS objective function is

$$\hat{Q}_n(\alpha,\theta) = n^{-1}\sum_{i=1}^{n}\left(y_i - aX_i^{\theta}\right)^2. \tag{7}$$

Suppose the parameter space is $[-C,C]\times[0,C]$ for some positive constant C.

Here $g(y_i,X_i;a,\theta) = \left(y_i - aX_i^{\theta}\right)^2$. It is continuous in a and θ by visual inspection. It is dominated by $G(y,x) = 2y^2 + 2C^2\left(|x|\vee 1\right)^{2C}$. The moment condition (d) is verified if y_i has finite second moment and $|X_i|$ has finite $2C'$th moment.

Example 2: (Maximum Likelihood – Probit) $Y_i = 1(X_i'\theta + \varepsilon_i \geq 0)$, $\varepsilon_i \sim N(0,1)$. Log-likelihood function for this model is

$$\ln L_n(\theta)/n = n^{-1}\sum_{i=1}^{n}[Y_i \ln\Phi(X_i'\theta) + (1-Y_i)\ln\Phi(-X_i'\theta)], \tag{8}$$

where $\Phi()$ is the cdf of the standard normal distribution. Here $g(Y_i,X_i;\theta) = Y_i\ln\Phi(X_i'\theta) + (1-Y_i)\ln\Phi(-X_i'\theta)$. It is continuous in θ by visual inspection. To find the dominating function, observe

that

$$\begin{aligned} |\ln \Phi(X_i^{'}\theta)| &= |\ln \Phi(0) + \lambda(X_i^{'}\tilde{\theta})X_i^{'}\theta| \\ &\leq |\ln \Phi(0)| + \lambda(X_i^{'}\tilde{\theta})|X_i^{'}\theta| \\ &\leq |\ln 2| + C \cdot |1 + X_i^{'}\tilde{\theta}||X_i^{'}\theta| \\ &\leq |\ln 2| + C \cdot (1 + \|X_i\| \ \|\theta\|)\|X_i\|\|\theta\|, \end{aligned} \tag{9}$$

where $\lambda(v) = \phi(v)/\Phi(v)$, the equality holds by a mean-value expansion, the first inequality holds by the triangular inequality, the second inequality holds by the well-known fact that $\lambda(v)$ is continuous, approaches 0 as v approaches positive infinity and approaches the negative 45 degree line as v approaches negative infinity, and the last inequality holds by the Cauchy-Schwartz inequality. Therefore, the dominating function for g is a multiple of $(1 + \|X_i\| \ \|\theta\|)\|X_i\|\|\theta\|$ plus a constant. Thus the moment condition (d) is verified as long as the space of θ is bounded and X has finite second moment.

Example 3: (Maximum Score, Manski, 1995) $Y_i = 1(X_i^{'}\theta + \varepsilon_i >= 0)$, where ε has conditional median (given x) of zero. The maximum score estimator is the maximizer of

$$\hat{Q}_n(\theta) = -n^{-1} \sum_{i=1}^{n} |Y_i - 1(X_i^{'}\theta > 0)|. \tag{10}$$

Here $g(y, x; \theta) = |y - 1(x^{'}\theta > 0)|$. It is continuous with probability one at each θ as long as $X_i^{'}\theta$ is continuously distributed for all θ. For example, if the coefficient of a component of X_i is nonzero for all $\theta \in \Theta$ and that component is a continuous random variable, then $X_i^{'}\theta$ is continuously distributed for all θ. The g function is dominated by the constant 2. Thus, conditions (c) and (d) of ULLN1 also hold.

Example 4: (Quantile Regression) $Y_i = X_i'\theta + \varepsilon_i$, $i = 1, ..., n$; $Q_{\varepsilon_i}(q|X_i) = 0$ (the qth quantile of ε_i given X_i is zero). Then a quantile estimator sets $n^{-1} \sum_{i=1}^{n} \left(q - 1(Y_i - X_i'\hat{\theta}_n \leq 0)\right) X_i = 0$.

Here $g(y, x; \theta) = (q - 1(y - x'\theta))x$. It is continuous with probability one at each θ as long as $\varepsilon_i|X_i$ has continuous distribution (and thus $Y_i|X_i$ has continuous distribution). The moment condition holds if $E\,||X_i|| \leq \infty$ because $g(y, x; \theta)$ is dominated by $G(y, x) = 2\,|x|$.

2 Generic Uniform Convergence

The ULLN1 is enough for most of the uniform convergence results needed in the consistency proofs. But it has its limitations. It only applies to sample averages. Some criterion functions we use are not of the sample average form (e.g. pre-estimated parameters, U-statistics, etc.). Thus we also discuss a generic uniform convergence theorem. The central piece of the generic uniform convergence is

the stochastic equicontinuity concept, which will be useful in other aspects of extreme estimators besides consistency.

Let $G_n(\theta)$ be a generic sequence of random functions. In the special case of sample averages, $G_n(\theta) = n^{-1}\sum_{i=1}^n [g(X_i,\theta) - Eg(X_i,\theta)]$. In general, $G_n(\theta)$ is not necessarily of the sample average form.

Definition SE: $\{G_n(\theta) : n \geq 1\}$ *is stochastically equicontinuous on* Θ *if* $\forall \varepsilon > 0$, $\exists \delta > 0$ *such that*

$$\limsup_{n\to\infty} P\left(\sup_{\theta\in\Theta}\sup_{\theta'\in B(\theta,\delta)} |G_n(\theta) - G_n(\theta')| > \varepsilon\right) < \varepsilon. \tag{11}$$

Remark. In the definition of SE, we can replace "$< \varepsilon$" by "$= 0$". Doing so results in an equivalent definition. (Why?)

Another equivalent definition is: for any random sequences $\{\theta_n \in \Theta\}_{n\geq 1}$ and $\{\theta_n^* \in \Theta\}_{n\geq 1}$ such that $\|\theta_n - \theta_n^*\| \to_p 0, \|G_n(\theta_n) - G_n(\theta_n^*)\| \to_p 0$.

Theorem 2.1 (Generic Convergence) (a) If (i) Θ is a bounded Euclidean space, (ii) $G_n(\theta) \to_p 0$ $\forall\theta \in \Theta$ and (iii) $\{G_n(\theta) : n \geq 1\}$ is stochastically equicontinuous, then $\sup_{\theta\in\Theta} |G_n(\theta)| \to_p 0$.

(b) If $\sup_{\theta\in\Theta} |G_n(\theta)| \to_p 0$, then (ii) and (iii) hold.

Proof. (a) Because Θ is a bounded Euclidean space, for any $\delta > 0$, there exist a finite subset $\{\theta_k : k = 1, ..., K\}$ of Θ such that the open balls $\{B(\theta_k, \delta) : k = 1, .., K\}$ cover Θ. Consider an arbitrary $\varepsilon > 0$. Let δ be the positive number such that the inequality in Definition SE holds. Observe that

$$\begin{aligned}
P\left(\sup_{\theta\in\Theta} |G_n(\theta)| > 2\varepsilon\right) &= P\left(\max_k \sup_{\theta\in B(\theta_k,\delta_\varepsilon)} |G_n(\theta) - G_n(\theta_k) + G_n(\theta_k)| > 2\varepsilon\right) \\
&\leq P\left(\max_k \sup_{\theta\in B(\theta_k,\delta_\varepsilon)} |G_n(\theta) - G_n(\theta_k)| + \max_k |G_n(\theta_k)| > 2\varepsilon\right) \\
&\leq P\left(\max_k \sup_{\theta\in B(\theta_k,\delta_\varepsilon)} |G_n(\theta) - G_n(\theta_k)| > \varepsilon\right) + P\left(\max_k |G_n(\theta_k)| > \varepsilon\right) \\
&\leq P\left(\sup_{\theta\in\Theta}\sup_{\theta'\in B(\theta,\delta)} |G_n(\theta) - G_n(\theta_k)| > \varepsilon\right) + P\left(\max_k |G_n(\theta_k)| > \varepsilon\right). \qquad (12)
\end{aligned}$$

Thus,

$$\limsup_{n\to\infty} P\left(\sup_{\theta\in\Theta} |G_n(\theta)| > 2\varepsilon\right) \leq \varepsilon + 0 = \varepsilon, \tag{13}$$

which implies that $\sup_{\theta\in\Theta}|G_n(\theta)| \to_p 0$.

(b) The implication of (ii) is immediate. To show that the uniform convergence implies (iii), observe that

$$P\left(\sup_{\theta\in\Theta}\sup_{\theta'\in B(\theta,\delta)}|G_n(\theta)-G_n(\theta_k)|>\varepsilon\right)\leq P\left(2\sup_{\theta\in\Theta}|G_n(\theta)|>\varepsilon\right)\to 0. \tag{14}$$

■

Example. U-Statistic. Newey, 1991: Let $m(z,\tilde{z},\theta)$ be a function of a pair of data arguments that is symmetric in the data arguments, i.e., $m(z,\tilde{z},\theta) = m(\tilde{z},z,\theta)$. Consider a U-statistic, depending on θ and its population analog

$$\begin{aligned}\hat{Q}_n(\theta) &\equiv 2\sum_{t=1}^{n}\sum_{s>t} m(z_t,z_s,\theta)/(n(n-1))\\ Q(\theta) &= E[m(z_t,z_s,\theta)], \theta\in\Theta,\end{aligned} \tag{15}$$

where z_t is assumed i.i.d. Results on convergence of $\hat{Q}_n(\theta)$ is well known (see e.g. Serfling (1980)). We can turn it into uniform covergence result using the generic uniform convergence result.

Proposition 2.1. Suppose that Θ is a compact metric space, $E[|m(z_t,z_s,\theta_0)|] < \infty$ for some $\theta_0\in\Theta$, and there are $b(z,\tilde{z})$ such that $E[b(z_1,z_2)]<\infty$ and for $\tilde{\theta},\theta\in\Theta$, $|m(z,\tilde{z},\tilde{\theta})-m(z,\tilde{z},\theta)|\leq b(z,\tilde{z})\|\tilde{\theta}-\theta\|$. Then

$$\sup_{\theta\in\Theta}|\hat{Q}_n(\theta)-Q(\theta)| = o_p(1), \tag{16}$$

and $Q(\theta)$ is continuous.

Proof. We verify the three conditions in Theorem 2.1(a). Condition (i) holds automatically. Condition (ii) holds by Theorem 5.4.A of Serfling (1980), which applies because for any $\theta\in\Theta$,

$$\begin{aligned}E[|m(z_t,z_s,\theta)|] &\leq E[|m(z_t,z_s,\theta_0)|]+E[|m(z_t,z_s,\theta_0)-m(z_t,z_s,\theta)|]\\ &\leq E[|m(z_t,z_s,\theta_0)|]+E[b(z,\tilde{z})\|\tilde{\theta}-\theta\|]\\ &\leq E[|m(z_t,z_s,\theta_0)|]+E[b(z,\tilde{z})]diameter(\Theta)\\ &<\infty.\end{aligned} \tag{17}$$

Let $B_n = 2\sum_{t=1}^{n}\sum_{s>t} b(z_s, z_t)/(n(n-1))$. Then $E[B_n] = E[b(z_1, z_2)] < \infty$. We have,

$$\begin{aligned}|\hat{Q}_n(\tilde{\theta}) - \hat{Q}_n(\theta)| &\leq 2\sum_{t=1}^{n}\sum_{s>t} |m(z_t, z_s, \tilde{\theta}) - m(z_s, z_t, \theta)|/(n(n-1)) \\ &\leq B_n\|\tilde{\theta} - \theta\|. \end{aligned} \tag{18}$$

Condition (iii) is implied by this display and $E(B_n) < \infty$. Now that we have verified all three conditions of Theorem 2.1(a), the theorem applies and gives us the desired result.

3 Some Probability Concepts

These probability concepts are used repeatedly. I put them here for easy reference.

Convergence in probability: A random sequence $\{X_n\}_{n=1}^{\infty}$ converges in probability to a nonstochastic scalar c if for all $\varepsilon > 0$, $\lim_{n\to\infty} \Pr(|X_n - c| > \varepsilon) = 0$. We usually write $X_n \to_p c$ or $\text{plim}_{n\to\infty} X_n = c$.

The o_p notation. Consider random sequences $\{X_n\}_{n=1}^{\infty}$ and $\{Y_n\}_{n=1}^{\infty}$. We say $X_n = o_p(Y_n)$ (in English: X_n is of order smaller than Y_n) if $\frac{X_n}{Y_n} \to_p 0$. For example, $X_n = o_p(1)$ means $X_n \to_p 0$ as $n \to \infty$.

The O_p notation. We say $X_n = O_p(Y_n)$ (in English: X_n is of order no larger than Y_n) if for any $\varepsilon > 0$, there exists C large enough such that $\lim_{n\to\infty} \Pr(|X_n/Y_n| > C) < \varepsilon$. If $X_n = O_p(1)$, we call X_n stochastically bounded. Any sequence of random variables that converges in distribution is stochastically bounded.

You may also encounter some variations of the o_p and O_p notation. The o notation is used in deterministic context: $a_n = o(b_n)$ means $\lim_{n\to\infty} a_n/b_n = 0$. The O notation is used also in deterministic context: $a_n = O(b_n)$ means a_n/b_n is a bounded sequence.

"Almost surely" or "with probability one". These phrases are added to statements about random objects (functions or variables), when we allow some exceptions to the statement. The exceptions are allowed for some values of the random objects. The exceptions cannot occur too often. They can only occur on a probabilistically negligible set of occasions. For example, we say $g(X_i, \theta_0)$ is continuous in θ at $\theta = \theta_0$ with probability one, if $\lim_{\theta\to\theta_0} g(x, \theta) = g(x, \theta_0)$ for all $x \in \mathcal{X}_1$ such that $\Pr(x \in \mathcal{X}_1) = 1$. The set $\mathcal{X}_1$ can be a proper subset of the support of X_i. To be more specific, suppose $X_i \sim N(0, 1)$. The support of X_i is R. The set $\mathcal{X}_1$ can be $R - \{0\}$, or even $R - \{\text{the set of all rational numbers}\}$.

Dominated Convergence Theorem. Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables and X is a random variable. If (a) $X_n \to X$ almost surely, (b) $|X_n| \leq Z$ almost surely, and (c) $E(Z) < \infty$, then $E(X_n) \to E(X)$.

References

Andrews, D. W. K., 1992, "Generic Uniform Convergence," *Econometric Theory*, 8, 241-257

Andrews, D. W. K., 1994, "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, Ch 37, 2247-2294

Jennrich, R. I., 1969, "Asymptotic Properties of Non-Linear Least Squares Estimators," *The Annals of Mathematical Statistics*, 40, 633-643

Newey, W. K., 1991, "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59(4), 1161-1167.

Serfling, R. J., 1980, *Approximation Theorems of Mathematical Statistics*, John Wiley and Sons, Inc.