Introduction to Econometrics II: Recitation 1*

Seung-hun Lee†

January 22nd, 2020

1 Logistics

1.1 Recitation

- Location: IAB 1027
- Time: Wednesdays 10:10AM - 12:00PM
  - Note that not all of the 110 minutes of the time will be used as a recitation. I intend to run the recitation for 80 - 90 minutes, depending on the rigor of the material. The rest of the recitation will be used as an informal office hours where you can ask questions about problem sets, concepts covered, and etc.
- I will focus on reviewing the concepts covered on the classes before the recitation. In particular, I will attempt to give you an intuition on what various econometric methods aim to achieve, discuss key results and proofs, and mention how such methods are applied in various literatures in Economics. If there is demand, I am willing to incorporate different methods into the recitation.
- As you will notice in this semester, new methods arise in an attempt to fix drawbacks of the previous methods. I will attempt to establish the relationship among econometric methods by pointing out how one method makes up for flaws in the previous methods and so on.

*This is based on the lecture notes of Professors Jushan Bai and Bernard Salanie. I was also greatly helped by previous recitation notes from Paul Sungwook Koh and Dong Woo Hahm. The remaining errors are mine.

†Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

1.2 Office Hours

- Location: IAB 329A
- Time: Mondays 1:00PM - 2:00PM

1.3 Objective as a TA

- Post recitation notes by 10:00PM on Tuesdays and suggested problem set solutions
- Help you get through Econometrics sequence. This could mean helping you achieve high grades to avoid certs (for Economics Ph.D. students) or passing the course with a sufficient grade (other Ph.D. students).
- While I am the one teaching the course, it will not be complete without you. Do not hesitate to ask questions, make suggestions at any time. I am here to help you all in any way I can.

1.4 References

- Primary resources are the lecture notes of the professors
- I also make use of Angrist and Pischke(2009), Arellano(2003), Cameron and Trivedi(2005), Hansen(2019), Imbens and Rubin(2015), and Wooldridge(2010) for additional references.
- For Statistics, I usually refer to Casella and Berger(2002) and Hogg et al.(2014).
- For Linear Algebra, I rely on Gockenbach(2010), Lang(1987), and Strang(2009).
- I will also cite other papers as I go by whenever necessary.

2 Review on Asymptotic Theories

Econometrics is about drawing an empirical conclusion from the given data. In particular, we are interested in the features of the estimators that we obtain from the data. While we cannot be sure of the distribution of such estimators in a finite sample context, we can still use asymptotic theories to derive the approximate distribution. In particular, we are interested in the consistency and the asymptotic variance of an estimator. This is where various modes of convergence and related theorems come in handy.

Before moving on, here are the Markov Inequality and the Chevyshev Inequality that would serve useful in proving some key theorems on this note and throughout the semester.

Theorem 2.1 (Markov Inequality). *For any non-negative random variable X and $\epsilon > 0$,*

$$\Pr(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}$$

Theorem 2.2 (Chevyshev Inequality). *For any random variable X and $\epsilon > 0$,*

$$\Pr(|X - E(X)| \geq \epsilon) \leq \frac{var(X)}{\epsilon^2}$$

Another useful tool is the asymptotic notations, or what we know as "Big O and small o". The definitions and properties of each notations are as follows

Definition 2.1 (Asymptotic Notations). A sequence of non-random vectors X_n is said to be $O(1)$ if it is bounded and $o(1)$ if it converges to 0. In addition, if a_n is a sequence of non-random positive scalars, we say X_n is $O(a_n)$ if $\frac{X_n}{a_n}$ is bounded and $o(a_n)$ if $\frac{X_n}{a_n} \to 0$

Definition 2.2 (Asymptotic Notations (Stochastic)). A sequence of random vectors X_n is $O_p(1)$ if it is bounded in probability[a] and $o_p(1)$ if it converges to zero in probability[b]. If a_n is a random variable, then X_n is $O_p(a_n)$ if $\frac{X_n}{a_n}$ is $O_p(1)$. X_n is $o_p(a_n)$ if $\frac{X_n}{a_n} \xrightarrow{p} 0$.

Property 2.1. *The following holds true for O_p and o_p*

1. $o_p(1) \implies O_p(1)$
2. *If $X_n = O_p(a_n)$, then $X_n = o_p(b_n)$ for any b_n s.t. $a_n/b_n \to 0$*
3. *If random vector X_n converges in distrubtion to X, Then $X_n = O_p(1)$*
4. $o_p(1) + o_p(1) = o_p(1), O_p(1) + O_p(1) = O_p(1), o_p(1) + O_p(1) = O_p(1)$
5. $o_p(1)o_p(1) = o_p(1), O_p(1)O_p(1) = O_p(1), o_p(1)O_p(1) = o_p(1)$

[a] It means that for every $\epsilon > 0$, there is a constant $M_\epsilon > 0$ s.t. $\Pr(|X_n| > M_\epsilon) < \epsilon$ for all $n \in \mathbb{N}$

[b] Or for every $\epsilon > 0$, $\Pr(|X_n| > \epsilon) \to 0$

2.1 Various Modes of Convergence

A minimal criterion of a good estimator is consistency, which tells us that some estimator $\hat{\theta}_n$ converges to a true value θ for sufficiently large n. Convergence in probability is used to analyze whether the estimator is consistent or not.

Definition 2.3 (Convergence in Probability)**.** A sequence random variable $X_n \in \mathbb{R}$ **converges in probability** to a random variable X if for any $\epsilon > 0$, we have either one of

$$\lim_{n\to\infty} \Pr(|X_n - X| \leq \epsilon) = 1 \text{ or } \lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0$$

We denote this as $X_n \xrightarrow{p} X$.

The estimator $\hat{\theta}_n$ is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$. To show, consistency, we rely on the Weak Law of Large Numbers (hereafter WLLN).

Theorem 2.3 (WLLN)**.** *Let $\{X_n | n \geq 1\}$ be a sequence of IID random variables with $E|X| < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\mu = E[X_1]$. Then, $\bar{X}_n \xrightarrow{p} \mu$.*

Proof. For simplicity, I work with the case where the variance of the random variables are finite, namely $var(X_1) = \sigma^2 < \infty$. For any $\epsilon > 0$,

$$\begin{aligned}\Pr(|\bar{X}_n - \mu| > \epsilon) &\leq \frac{E(|\bar{X}_n - \mu|^2)}{\epsilon^2} \text{ } (\because \text{Chevyshev}) \\ &= \frac{\sigma^2}{n\epsilon^2} \text{ } (\because \text{Sample variance of } X_n = \sigma^2/n)\end{aligned}$$

By taking $n \to \infty$, $\frac{\sigma^2}{n\epsilon^2} \to 0$. Therefore, $\lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| > \epsilon) = 0$ □

We are also interested in the asymptotic distribution of the estimator. In particular, we are interested in checking whether the estimator is asymptotically normal.

Definition 2.4 (Convergence in Distribution)**.** Let X_n be a sequence of random variables with distribution $F_n(x) = \Pr(X_n \leq x)$. Let X be a random variable whose distribution is $F(x) = \Pr(X \leq x)$. A sequence random variable $X_n \in \mathbb{R}$ **converges in distribution** to a random variable X if at all x at which $F(x)$ is continuous,

$$\Pr(X_n \leq x) \to \Pr(X \leq x)$$

We denote this as $X_n \xrightarrow{d} X$.

In words, $X_n \xrightarrow{d} X$ does not necessarily mean that X_n and X are close - but their CDFs are close. If $X_n \xrightarrow{p} X$, X_n and X are close. Therefore, convergence in distribution has a weaker requirement than convergence in probability.

Convergence in distribution is applied when we use a central limit theorem.

Theorem 2.4 (Lindberg-Levy CLT). *If $\{X_n | n \geq 1\}$ is a sequence of IID random variables with $E(X_1) = \mu < \infty, var(X_1) = \sigma^2 < \infty$, then*

$$\sqrt{n}(\bar{X}_n - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} N(0, \sigma^2)$$

or in standardized notation,

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right) \xrightarrow{d} N(0, 1)$$

To answer whether the approximation provided by the CLT is accurate, we refer to the Berry-Esseen Theorem.

Theorem 2.5 (Berry-Esseen Theorem). *For all sample sizes n and for all $t \in \mathbb{R}$,*

$$\left| \Pr\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq t\right) - \Phi(t) \right| < \frac{C \times E(|X_i - \mu|^3)}{\sigma^3 \sqrt{n}}$$

where the estimate of C has changed over time: $C \in (0.4097, 0, 4748)$.

The key advancement of this theorem is that it contains some implication about the rate of a convergence of a probability distribution of a random sample to a normal distribution. It also implies that the accuracy depends on the standardized absolute third moment of the distribution of X_i, $\frac{E(|X_i - \mu|^3)}{\sigma^3}$, and sample size.

We can extend the CLT to a multi-variate setting using Cramer-Wold Device.

Theorem 2.6 (Cramer-Wold Device). *Let X_n and X be random vector in $\mathbb{R}^k$. Then,*

$$X_n \xrightarrow{d} X \iff \lambda' X_n \xrightarrow{d} \lambda' X, \ \forall \lambda \in \mathbb{R}^k$$

Theorem 2.7 (Multivariate Lindberg-Levy CLT). *Let $X_1, .., X_n \in \mathbb{R}^k$ be IID and $E||X_1||^2 < \infty$. Then, as $n \to \infty$,*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, E[(X-\mu)(X-\mu)'])$$

where $X \in \mathbb{R}^k$.

Proof. Since X_i is IID, and $\lambda \in \mathbb{R}^k$, $\lambda' X_i$ is also IID, where $E[\lambda' X_i] = \lambda' \mu$ and $var(\lambda' X_i) = \lambda' var(X_i) \lambda$. By Lindberg-Levy CLT (scalar version), we can get

$$\sqrt{n}(\lambda' \bar{X}_n - \lambda' \mu) \xrightarrow{d} N(0, \lambda' var(X_i) \lambda) \iff \lambda'(\sqrt{n}(\bar{X}_n - \mu)) \xrightarrow{d} \lambda'(N(0, var(X_i)))$$

Then, apply Cramer-Wold device to get $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, var(X_i))$ □

2.2 Some Useful Theorems

To get the distribution and check for consistency of a seemingly complicated estimators, we can use the following theorems.

Theorem 2.8 (Continuous Mapping Theorem). *If $X_n \xrightarrow{p,d} X$ as $n \to \infty$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ is continuous at every point of a set C s.t. $P(X \in C) = 1$ (or g has the set of discontinuity points D s.t. $\Pr(X \in D) = 0$), then*

$$g(X_n) \xrightarrow{p,d} g(X)$$

Theorem 2.9 (Slutsky's Theorem). *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ ($c \in \mathbb{R}$) as $n \to \infty$*

1. $X_n \pm Y_n \xrightarrow{d} X \pm c$
2. $X_n Y_n \xrightarrow{d} Xc$
3. $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$*, provided that* $c \neq 0$

If $X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y$, *then the following holds*

1. $X_n \pm Y_n \xrightarrow{p} X \pm Y$
2. $X_n Y_n \xrightarrow{p} XY$
3. $\frac{X_n}{Y_n} \xrightarrow{p} \frac{X}{Y}$, *provided that* $\Pr(Y = 0) = 0$

However, the Slutsky's Theorem does not work well when both random variables converge in distribution, but not in probability. For such cases, we need to consider whether the two different random variables JOINTLY converge to another vector of random variable.

Example 2.1 (Counter-example)**.** *Let* $Z_n \xrightarrow{d} Z$, *where* $Z \sim N(0,1)$. *Also define* $Y_n = (-1)^n Z_n$. *Then,* $Y_n \xrightarrow{d} N(0,1)$ *as well. However,* $Z_n + Y_n$ *displays following patterns*

$$Z_n + Y_n = \begin{cases} 2Z_n & n \text{ being even} \\ 0 & n \text{ being odd} \end{cases}$$

Thus, for even numbered n, the distribution converges to $N(0,4)$. *For odd numbered n, it converges to a distribution centered at 0. Therefore,* $X_n + Y_n$ *fails to converge in distribution to* $2N(0,1) = N(0,4)$

Example 2.2 (Relationship between convergence of marginal and joint distribution)**.** *Suppose* $\mathbf{X}_n = (A_n\ B_n)'$ *and* $\mathbf{X} = (A\ B)'$, *where* A_n, B_n, A, B *are random variables.*

- *Suppose that* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$. *From Cramer-Wold Device, let* $\lambda = (1\ 0)'$. *By applying Cramer-Wold Device, we can obtain that* $A_n \xrightarrow{d} A$. *We can obtain similar results for* $B_n \xrightarrow{d} B$.
- *Convergence in distribution for marginal distributions does not imply the same for joint distribution. Let* A_n *and* B_n *be independent standard normal random variables for all n. Thus,* $A_n, B_n \xrightarrow{d} Z \sim N(0,1)$ *individually. Jointly, however,* $(A_n\ B_n)'$ *is bivariate normal with correlation 0 while* $(Z\ Z)'$ *is bivariate normal with correlation 1. Therefore, we cannot say* $(A_n\ B_n)'$ *converges in distribution to* $(Z\ Z)'$
- *However, if it is the case that* $A_n \xrightarrow{d} A$ *and* $B_n \xrightarrow{d} B$, *where* A_n *is independent of* B_n *for all n and A is independent of B, it can be said that* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$. *This is because*

$$F_{A_n,B_n}(a,b) = F_{A_n}(a)F_{B_n}(b) \rightarrow F_A(a)F_B(b) = F_{A,B}(a,b)$$

for all a, b that are the continuity points of F_A and F_B

The delta method provides us with another way to approximate the distribution of a smooth transformations of simpler objects.

Theorem 2.10 (Delta Method (Univariate)). *Suppose that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim N(0, \sigma^2)$ for some $\sigma^2 > 0$. If $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable at θ_0 with a nonzero derivative, then*

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} g'(\theta_0)Z \sim N(0, (g'(\theta_0))^2\sigma^2)$$

Proof. Take first order Taylor approximation of $g(\cdot)$ around θ_0, so that $g(x) = g(\theta_0) + g'(\bar{x})(x - \theta_0)$ for some $\bar{x}$ in between θ_0 and x. Then, let $x = \hat{\theta}_n$ and $\bar{x} = \bar{\theta}_n$ that is in between θ_0 and $\hat{\theta}_n$. We can then obtain

$$g(\hat{\theta}_n) - g(\theta_0) = g'(\bar{\theta}_n)(\hat{\theta}_n - \theta_0) \iff \sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) = g'(\bar{\theta}_n)\sqrt{n}(\hat{\theta}_n - \theta_0)$$

Since $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z$, we have $\hat{\theta}_n \xrightarrow{p} \theta_0$. In addition, $|\bar{\theta}_n - \theta_0| \leq |\hat{\theta}_n - \theta_0|$. In this context, $\bar{\theta}_n \xrightarrow{p} \theta_0$. By Continuous mapping theorem (and the fact that g' is continuous at θ_0), $g'(\bar{\theta}_n) \xrightarrow{p} g'(\theta_0)$. By Slutsky theorem (part (2)),

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} g'(\theta_0)Z = N(0, (g'(\theta_0)^2\sigma^2)$$

□

Theorem 2.11 (Delta Method (Multidimensional)). *Suppose that θ_0 is an m-dimensional vector and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim N(0, \Sigma)$ as $n \to \infty$. Let $g : \mathbb{R}^m \to \mathbb{R}^n$ be a function that is differentiable at θ_0 with a continuous partial derivatives at θ_0. Define $G(\theta_0)$ as*

$$G(\theta_0) = \frac{\partial g(\theta)}{\partial \theta'} |_{\theta=\theta_0} = \left[\frac{\partial g(\theta_0)}{\partial \theta_1}, ..., \frac{\partial g(\theta_0)}{\partial \theta_m}\right] \in \mathbb{R}^{n \times m}$$

Then, we can get

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} G(\theta_0)Z \sim N(0, G(\theta_0)\Sigma G(\theta_0)') \text{ as } n \to \infty$$

2.3 Problem with Standard Asymptotics

While the asymptotic methods introduced here are the most commonly used approach, it is not without problems. There could be more problems here, but I will introduce one of them. When the functions transforming the random variable to a non-linear form, the accuracy of the approximation may fall.

Example 2.3 (Non-linear Transformation). *If $X_i \sim N(\mu, \sigma^2)$, so that $\bar{X} \sim N(\mu, \sigma^2/n)$, and $g(\cdot) : x \to x^2$, delta method suggests that*

$$g(\bar{X}) \sim N\left(g(\mu), g'(\mu)^2 \frac{\sigma^2}{n}\right)$$

However, we have

$$E[g(\bar{X})] = E[\bar{X}^2] = var(\bar{X}) + \{E[\bar{X}]\}^2 = \mu^2 + \sigma^2/n$$

whereas $g(\mu) = \mu^2$. This implies that $g(\bar{X})$ has a bias of σ^2/n.

This happens since delta method is based on first-order Taylor approximation. Therefore, the accuracy of the approximation hinges on the performance of the first-order approximation of a non-linear function. One of the possible solutions is to use high-order approximation to obtain a smaller bias.