Introduction to Econometrics II*

Dong Woo Hahm†

Spring 2019

Contents

1 Review of Asymptotic Theory 2
- 1.1 Basic Setup . . . 2
- 1.2 Modes of Convergence . . . 2
  - 1.2.1 Covergence in Probability . . . 2
  - 1.2.2 Convergence in Distribution . . . 3
  - 1.2.3 Sufficient Conditions for Joint Weak Convergence . . . 5
- 1.3 Continuous Mapping Theorem and Delta Method . . . 6
- 1.4 Stochastic Orders . . . 7
- 1.5 Inequalities . . . 10

2 Instrumental Variable 14
- 2.1 Endogeneity . . . 14
- 2.2 Sources of Endogeneity . . . 15
  - 2.2.1 Omitted Variable Bias . . . 15
  - 2.2.2 Measurement Error . . . 16
  - 2.2.3 Simultaneous Equations . . . 16
- 2.3 Instrumental Variables . . . 17
- 2.4 Reduced Form . . . 19

References 20

*Columbia University Economics Ph.D. First Year 2018-2019. Professor Jushan Bai and Simon Lee. Mostly from Professors' lectures, Hansen (2018), Cameron and Trivedi (2005), Angrist and Pischke (2008), Wooldridge (2010), and sometimes Paul Koh's note from last year and Professor Christoph Rothe's lecture notes from 2017 Spring.

†Department of Economics, Columbia University. dongwoo.hahm@columbia.edu.

Please email me for any error. This note will be continuously updated throughout the semester. Please do not circulate outside of the class.

1 Review of Asymptotic Theory

Usually econometrics is about analyzing data from an economic context and has steps including:

- Formulating an appropriate model
- Computing estimates of unknown parameters in such a model
- Quantifying the uncertainty about those estimates
- Use these measures of uncertainty to draw empirical conclusions

Asymptotic theory is usually related to the third step, "Quantifying the uncertainty".

1.1 Basic Setup

- Data : $\{z_i\}_{i=1}^n$ with joint distribution P_n
- Model : A set $\mathcal{P}_n$ of potential candidates for P_n. Restrictions may be imposed as necessary.
- Independent and identically distributed (i.i.d) data

$$P_n = P \times P \cdots \times P$$

- Random variable (vector) of interest:

$$\hat{\theta}_n = f_n(z_1, \cdots, z_n)$$

  e.g) estimator, test statistics ...

We are interested in features of the distribution of $\hat{\theta}_n$ with a finite sample size n, which is usually impossible or impractical. Instead, we typically use asymptotics to derive approximations to the distribution of $\hat{\theta}_n$. The general idea is to think of $\hat{\theta}_n$ as the nth element of an infinite sequence and to calculate the limit of the sequence (if exists).

1.2 Modes of Convergence

1.2.1 Covergence in Probability

Definition 1.2.1. A sequence of random variables z_n is said to **converge in probability** to a random variable z if for any $\delta > 0$, $\lim_{n\to\infty} P(|z_n - z| \leq \delta) = 1$ or equivalently, $\lim_{n\to\infty} P(|z_n - z| > \delta) = 0$ and we denote as $z_n \xrightarrow{p} z$ or $p\lim_{n\to\infty} z_n = z$.

Theorem 1.2.1. (Khintchine's Weak Law of Large Numbers)
If $z_1, \cdots z_n$ are i.i.d with $E(z_i) = \mu$, then $\bar{z}_n \xrightarrow{p} \mu$ where $|\mu| < +\infty$.

Proof. Assume $Var(z_i) = \sigma^2 < +\infty$. For any $\delta > 0$,

$$\begin{aligned} P(|\bar{z}_n - \mu| > \delta) &\leq \frac{E(|\bar{z}_n - \mu|^2)}{\delta^2} && \text{: Chebyshev's Inequality} \\ &= \frac{\sigma^2}{\delta^2 \cdot n} \\ &\to 0 && \text{as } n \to \infty \end{aligned}$$

□

1.2.2 Convergence in Distribution

Definition 1.2.2. Let z_n be a sequence of random variables and define $F_n(x) = P(z_n \leq x)$. Let z a random variable of which distribution function is $F(x) = P(z \leq x)$. z_n is said to **converge in distribution** to z if $F_n(x) \to F(x)$ as $n \to \infty$ at all x where F is continuous. We denote this by $z_n \xrightarrow{d} z$.

- z is usually called the asymptotic distribution of limit distribution of z_n.
- $z_n \xrightarrow{d} z$ does not necessarily mean that z_n and z are close (only the cdfs are close) where as $z_n \xrightarrow{p} z$ implies z_n and z are close.
- $z_n \xrightarrow{p} z$ implies $z_n \xrightarrow{d} z$
- However, $z_n \xrightarrow{d} z$ don't imply $z_n \xrightarrow{p} z$
  For example, consider Y_n and Y that are mutually independent with distributions given by

$$Y_n = \begin{cases} 1 \text{ with probability } \frac{1}{2} + \frac{1}{n+1} \\ 0 \text{ with probability } \frac{1}{2} - \frac{1}{n+1} \end{cases}$$

$$Y = \begin{cases} 1 \text{ with probability } \frac{1}{2} \\ 0 \text{ with probability } \frac{1}{2} \end{cases}$$

- If $z = c \in \mathbb{R} (\Leftrightarrow P(z = c) = 1)$, that is the limit distribution z is degenerate, then $z_n \xrightarrow{p} c \Leftrightarrow z_n \xrightarrow{d} c$

Theorem 1.2.2. (Lindeberg-Levy Central Limit Thoerem)
Let $z_1, \cdots z_n$ be i.i.d with $E(z_i) = \mu, Var(z_i) = \sigma^2$ where $|\mu| < +\infty, \sigma^2 < +\infty$. Then

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

That is,

$$\lim_{n \to \infty} P(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (z_i - \mu) \leq a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2} x^2) dx.$$

for all $a \in \mathbb{R}$.

Remark.

- WLLN and CLT :
  By WLLN, $\bar{X} - \mu \xrightarrow{p} 0$. In some sense with WLLN, $\bar{X} - \mu$ goes to 0 too fast as $n \to \infty$. Thus by multiplying $\sqrt{n}$, the CLT slows down the speed of convergence to make it converge to some non degenerate distribution, i.e, $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} Z \sim N(0, \sigma^2)$.
- It is often useful to know if $Z \sim N(0, \sigma^2)$,

$$E(X^p) = \begin{cases} 0 & \text{if } p \text{ is odd} \\ \sigma^p (p-1)!! & \text{if } p \text{ is even} \end{cases}$$

  where $n!! = \prod_{k=0}^{\lceil \frac{n}{2} \rceil - 1} (n - 2k) = n(n-2) \cdots 1$ is double factorial.

Similar things hold for random vectors, a vector of random variables. First, let's define random vectors.

Definition 1.2.3. $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m$ is called a **random vector**. We define $E(\boldsymbol{y}) = \begin{pmatrix} E(y_1) \\ \vdots \\ E(y_m) \end{pmatrix} \in \mathbb{R}^m$ and $||\boldsymbol{y}|| = \sqrt{\boldsymbol{y}^T \boldsymbol{y}} = (y_1^2 + \cdots y_m^2)^{\frac{1}{2}}$.

Theorem 1.2.3. $E||\boldsymbol{y}|| < +\infty \Leftrightarrow E|y_j| < +\infty, \forall j = 1, 2, \cdots, m$ where m is a finite natural number.

Proof. ($\Longleftarrow$) $y_1^2 + \cdots + y_m^2 \leq (|y_1| + \cdots + |y_m|)^2 \Rightarrow ||y|| \leq |y_1| + \cdots + |y_m|$. So from $E|y_i| < +\infty, E||y|| < +\infty$.
($\Longrightarrow$) $|y_j| \leq (|y_1|^2 + \cdots + |y_m|^2)^{\frac{1}{2}}, \forall j = 1, 2, \cdots, m \Rightarrow |y_j| \leq ||\boldsymbol{y}||, \forall j = 1, 2, \cdots, m$. So from $E||\boldsymbol{y}|| < +\infty, E|y_j| < +\infty, \forall j = 1, 2, \cdots, m$. □

Convergence in probability of a random vector is defined as convergence in probability of all elements in the vector. Hence, multivariate version of WLLN (Theorem 1.2.1) holds.

Theorem 1.2.4. (WLLN for Random Vectors)

Let $\boldsymbol{y}_i$ be i.i.d where $\boldsymbol{y}_i \in \mathbb{R}^m$ s.t. $E||\boldsymbol{y}_i|| < +\infty, \forall i$. Let $\bar{\boldsymbol{y}}_n = \frac{1}{n} \sum_{i=1}^n \boldsymbol{y}_i = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_m \end{pmatrix}$, then

$$\bar{\boldsymbol{y}}_n \xrightarrow{p} E(\boldsymbol{y}_i) = \begin{pmatrix} E(y_{i1}) \\ \vdots \\ E(y_{im}) \end{pmatrix}$$

Proof. By definition, $\bar{\boldsymbol{y}}_n \xrightarrow{p} E(\boldsymbol{y}_i)$ if and only if $\bar{y}_j \xrightarrow{p} \mu_j, \forall j = 1, 2, \cdots, m$. The latter holds if

$E|y_j| < +\infty, \forall j = 1, 2, \cdots, m$ which is equivalent to $E|\boldsymbol{y}_i| < +\infty$. □

Though Lindeberg-Levy CLT (Theorem 1.2.2) only provides the case for scalar random variables, it can be extended to multivariate data via Cramer-Wold device.

Theorem 1.2.5. (Cramer-Wold)
Let $\hat{\gamma}_n, \gamma_\infty$ be vector-valued random vectors. Then $\hat{\gamma}_n \xrightarrow{d} \gamma_\infty$ if and only if $\lambda'\hat{\gamma}_n \xrightarrow{d} \lambda'\gamma_\infty$ for all fixed vectors λ with $\lambda'\lambda = 1$.[a]

[a] The last condition that $\lambda'\lambda = 1$ is often omitted in some versions, and it's not necessary.

Theorem 1.2.6. (Multivariate Lindeberg-Levy Central Limit Theorem)
If $\boldsymbol{y}_i \in \mathbb{R}^k$ are independent and identically distributed and $E||\boldsymbol{y}_i||^2 < +\infty$, then as $n \to \infty$

$$\sqrt{n}(\bar{\boldsymbol{y}}_n - \mu) \xrightarrow{d} N(0, V)$$

where $\mu = E(\boldsymbol{y})$ and $V = E((\boldsymbol{y} - \mu)(\boldsymbol{y} - \mu)')$.

Proof. Fix some $\lambda \in \mathbb{R}^k$ such that $\lambda'\lambda = 1$. Define $u_i = \lambda'(\boldsymbol{y}_i - \mu)$. Then u_i are i.i.d with $E(u_i^2) = \lambda' V \lambda < \infty$. We have

$$\lambda'\sqrt{n}(\bar{\boldsymbol{y}}_n - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i \xrightarrow{d} N(0, \lambda' V \lambda)$$

If some random vector $\boldsymbol{z} \sim N(0, V)$ then $\lambda'\boldsymbol{z} \sim N(0, \lambda' V \lambda)$. Thus, we have

$$\lambda'\sqrt{n}(\bar{\boldsymbol{y}}_n - \mu) \xrightarrow{d} \lambda'\boldsymbol{z}$$

Since the choice of λ was arbitrary, by Cramer-Wold device (Theorem 1.2.5), we have

$$\sqrt{n}(\bar{\boldsymbol{y}}_n - \mu) \xrightarrow{d} \boldsymbol{z}$$

□

Remark. Note that having convergence in distribution to a normal distribution in each component does not imply the random vector jointly converge to joint normal. Hence for asymptotic "joint" normality results, you should make use of multivariate CLT.[1]

1.2.3 Sufficient Conditions for Joint Weak Convergence

For probability limits, it is straightforward.

Theorem 1.2.7. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ then $(X_n, Y_n) \xrightarrow{p} (X, Y)$

However for weak convergence, things are a bit more complicated. The following two theorems give some examples of sufficient conditions for joint weak convergence.

[1] Recall Pepe's PS2 Q3, PS3 Q4?

Theorem 1.2.8. If X is independent of each Y_i in $\{Y_i\}$ and $Y_n \xrightarrow{d} Y$, then $(X, Y_n) \xrightarrow{d} (X, Y)$.

Theorem 1.2.9. If X_n, Y_n and X, Y are mutually independent, then the marginal convergence in distribution implies joint convergence in distribution.

$$X_n \xrightarrow{d} X, Y_n \xrightarrow{d} Y \Rightarrow (X_n, Y_n) \xrightarrow{d} (X, Y)$$

Proof.

$$\begin{aligned}\lim_{n\to\infty} P(X_n \le x, Y_n \le y) &= \lim_{n\to\infty} P(X_n \le x)P(Y_n \le y) \\ &= \lim_{n\to\infty} P(X_n \le x) \lim_{n\to\infty} P(Y_n \le y) \\ &= P(X \le x)P(Y \le y) \\ &= P(X \le x, Y \le y)\end{aligned}$$

□

1.3 Continuous Mapping Theorem and Delta Method

Continuous Mapping theorem and Slutsky's theorem tell us how to manipulate limits in probability and distribution.

Theorem 1.3.1. (Continuous Mapping Theorem)
If $z_n \xrightarrow{p,d} z$, and $g(\cdot)$ is has the set of discontinuity points D_g such that $Pr(z \in D_g) = 0$, then $g(z_n) \xrightarrow{p,d} g(z)$.[a]

Proof. I only prove when $z = c$ is degenerate.

Since g is continuous at c, we can find a $\delta > 0$ such that if $||z_n - c|| < \delta$ then $||g(z_n) - g(c)|| \le \epsilon$ for all $\epsilon > 0$. Recall that $A \subset B$ implies $P(A) \le P(B)$. Thus $P(||g(z_n) - g(c)|| \le \epsilon) \ge P(||z_n - c|| < \delta) \to 1$ as $n \to \infty$ by the assumption that $z_n \xrightarrow{p} c$. Hence $g(z_n) \xrightarrow{p} g(c)$ as $n \to \infty$. □

[a] Note that if z is a degenerate point, the condition reduces to $g(\cdot)$ is continuous at z.

Theorem 1.3.2. (Slutsky's Theorem)
If $X_n \xrightarrow{d} Z, Y_n \xrightarrow{p} c$ as $n \to \infty$, then

1. $X_n + Y_n \xrightarrow{d} Z + c$
2. $X_n - Y_n \xrightarrow{d} Z - c$
3. $X_n Y_n \xrightarrow{d} Zc$
4. $\frac{X_n}{Y_n} \xrightarrow{d} \frac{Z}{c}$ if $c \neq 0$

Delta method is another way of approximating the distribution of "smooth" transformations of simpler objects.

Theorem 1.3.3. (Delta Method)
Suppose that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \xi \in \mathbb{R}^m$ where $g : \mathbb{R}^m \to \mathbb{R}^k, k \leq m$ is continuously differentiable at $x = \theta_0$. Then $\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} G^T\xi$ where $G = (g'(\theta_0))^T$, the transpose of Jacobian matrix of g evaluated at θ_0.

Proof. Prove only for the case $m = k = 1$.
$g(x) = g(\theta_0) + g'(\bar{x})(x - \theta_0)$ where $\bar{x}$ is in between x and θ_0.
Replace $x = \hat{\theta}_n$ then $g(\hat{\theta}_n) = g(\theta_0) + g'(\bar{\theta}_n)(\hat{\theta}_n - \theta_0)$ where $\bar{\theta}_n$ is in between $\hat{\theta}_n$ and θ_0. Thus,

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

Since $\hat{\theta}_n \xrightarrow{p} \theta_0$ and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \xi$ and g' is continuous at θ_0 and $||\bar{\theta}_n - \theta_0|| \leq ||\hat{\theta}_n - \theta_0||, g'(\bar{\theta}_n) \xrightarrow{p} g'(\theta_0)$.
Thus, $\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} g'(\theta_0)\xi$ by Slutsky's theorem. □

Note that we implicitly require that G is full-rank.

1.4 Stochastic Orders

Definition 1.4.1. (Stochastic Orders)
For deterministic sequences,

- $x_n = o(1) \iff x_n \to 0$
- $x_n = o(a_n) \iff \frac{x_n}{a_n} \to 0$
- $x_n = O(1) \iff \exists M < +\infty \text{ s.t. } |x_n| \leq M, \forall n$
- $x_n = O(a_n) \iff \frac{x_n}{a_n} = O(1)$

For stochastic sequences,

- $z_n = o_p(1) \iff z_n \xrightarrow{p} 0$
- $z_n = o_p(a_n) \iff \frac{z_n}{a_n} \xrightarrow{p} 0$
- $z_n = O_p(1) \iff \forall \epsilon > 0, \exists M_\epsilon > 0 \text{ s.t. } P(|z_n| > M_\epsilon) < \epsilon, \forall n$[a]
- $z_n = O_p(a_n) \iff \frac{z_n}{a_n} = O_p(1)$

[a] $z_n = O_p(1)$ is equivalent to saying that z_n is stochastically bounded which roughly means that the tail probability is small.

Example 1.4.1.

- For any consistent estimator $\hat{\beta}$ for β, $\hat{\beta} = \beta + o_p(1)$.
- For $\hat{\beta}$ such that $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(\cdot, \cdot)$, $\hat{\beta} = \beta + O_p(n^{-1/2})$.

Remark. (Some Properties of Stochastic Orders)

1. $z_n = O_p(a_n) \Rightarrow z_n = o_p(b_n)$ for any b_n such that $a_n/b_n \to 0$.

2. $z_n = o_p(1) \Rightarrow z_n = O_p(1)$
3. $z_n = o_p(1) \nLeftarrow z_n = O_p(1)$
4. $z_n \xrightarrow{p} z \Rightarrow z_n = O_p(1)$

   Proof. Fix $\epsilon > 0$.

   - $\exists M(\epsilon) > 0$ s.t. $P(|z| > \frac{M(\epsilon)}{2}) < \frac{\epsilon}{2}$ since $P(|z| > t) = F_z(-t) + (1 - F_z(t)) \to 0$.
   - Since $z_n \xrightarrow{p} z$, $\lim_n P(|z_n - z| > \frac{M(\epsilon)}{2}) < \frac{\epsilon}{2}$. Thus $\exists N \in \mathbb{N}, n > N \Rightarrow P(|z_n - z| > \frac{M(\epsilon)}{2}) < \frac{\epsilon}{2}$.
   - For $n \leq N$, $\exists M_n > 0$ s.t. $P(|z_n| > M_n) < \epsilon$
   - Let $M_\epsilon = \max\{M_1, \cdots, M_N, M(\epsilon)\}$
   - If $n \leq N$, $P(|z_N| > M_\epsilon) \leq P(|z_N| > M_n) < \epsilon$
     If $n > N$,

$$\begin{aligned} P(|z_n| > M_\epsilon) &\leq P(|z_n - z| + |z| > M_\epsilon) \\ &\leq P(|z_n - z| > \frac{M_\epsilon}{2}) + P(|z| > \frac{M_\epsilon}{2}) \\ &\leq P(|z_n - z| > \frac{M(\epsilon)}{2}) + P(|z| > \frac{M(\epsilon)}{2}) \\ &< \epsilon \end{aligned}$$

   □

5. $z_n \xrightarrow{d} z \Rightarrow z_n = O_p(1)$
6. If $E(|z_n|) \leq c < +\infty, \forall n$, then $z_n = O_p(1)$

   Proof. Use Markov inequality. Let $\epsilon > 0$. Then there exists $M(\epsilon)$ such that $\frac{c}{M(\epsilon)} < \epsilon$. Then for each $n \in \mathbb{N}, P(|z_n| > M(\epsilon)) \leq \frac{E(|z_n|)}{M(\epsilon)} \leq \frac{c}{M(\epsilon)} < \epsilon$. □

7. If $z_n = O_p(1), a_n \to \infty$ then $\frac{z_n}{a_n} = o_p(1)$.

   Proof. $\exists M_\delta > 0$ s.t. $P(|z_n| > M_\delta) < \delta, \forall n$. Fix $\epsilon > 0$ then

$$\begin{aligned} P(|z_n/a_n| > \epsilon) &= P(|z_n/a_n| > \epsilon, |z_n| > M_\delta) + P(|z_n/a_n| > \epsilon, |z_n| \leq M_\delta) \\ &\leq P(|z_n| > M_\delta) + P(|M_\delta/a_n| > \epsilon) \\ &< \delta \end{aligned}$$

   as $n \to \infty$ and since δ arbitarily chosen, we're done. □

Theorem 1.4.1. (Random Sequence with a Bounded Moment is Stochastically Bounded)
If z_n is a random vector which satisfies

$$E||z_n||^\delta = O(a_n)$$

for some sequence a_n and $\delta > 0$, then

$$z_n = O_p(a_n^{1/\delta})$$

Similarly, $E||z_n||^\delta = o(a_n)$ implies $z_n = o_p(a_n^{1/\delta})$.

Proof. The assumptions imply that there is some $M < +\infty$ such that $E||z_n||^\delta \leq Ma_n$ for all n. For any ϵ set $B = (\frac{M}{\epsilon})^{1/\delta}$. Then

$$P(a_n^{-1/\delta}||z_n|| > B) = P(||z_n||^\delta > \frac{Ma_n}{\epsilon}) \leq \frac{\epsilon}{Ma_n} E||z_n||^\delta \leq \epsilon$$

□

Theorem 1.4.2. (Simple Rules for Stochastic Orders)

1. $o_p(1) + o_p(1) = o_p(1)$
2. $o_p(1) + O_p(1) = O_p(1)$
3. $O_p(1) + O_p(1) = O_p(1)$
4. $o_p(1)o_p(1) = o_p(1)$
5. $o_p(1)O_p(1) = o_p(1)$
6. $O_p(1)O_p(1) = O_p(1)$

Proof. Below I provide proof for some of the above. Rest of them can be proved using Continuous Mapping theorem and Slutsky's theorem.

3. Let $y_n = O_p(1), z_n = O_p(1)$. Fix $\epsilon > 0$. Then $\exists M_y > 0$ s.t. $P(|y_n| > M_y) < \frac{\epsilon}{2}, \forall n$, $\exists M_z > 0$ s.t. $P(|z_n| > M_z) < \frac{\epsilon}{2}, \forall n$. Let $M_\epsilon = M_y + M_z$ then

$$\begin{aligned} P(|z_n + y_n| > M_\epsilon) &\leq P(|z_n| + |y_n| > M_\epsilon) \\ &\leq P(|z_n| > M_z) + P(|y_n| > M_y) < \epsilon \end{aligned}$$

5. Let $y_n = o_p(1), z_n = O_p(1)$. Fix $\epsilon, \delta > 0$. Let M_ϵ be s.t. $P(|z_n| > M_\epsilon) < \frac{\epsilon}{2}, \forall n$. For sufficiently large n,

$$\begin{aligned} P(|z_n y_n| > \delta) &= P(|z_n y_n| > \delta, |z_n > M_\epsilon) + P(|z_n y_n| > \delta, |z_n| \leq M_\epsilon) \\ &\leq P(|z_n| > M_\epsilon) + P(|y_n| > \frac{\delta}{M_\epsilon}) \\ &\leq \epsilon \end{aligned}$$

as $n \to \infty$ since $y_n = o_p(1)$.

6. Let $y_n = O_p(1), z_n = O_p(1)$. Fix $\epsilon > 0$. Then $\exists M_y > 0$ s.t. $P(|y_n| > M_y) < \frac{\epsilon}{2}, \forall n$, $\exists M_z > 0$ s.t. $P(|z_n| > M_z) < \frac{\epsilon}{2}, \forall n$. Let $M_\epsilon = M_y \cdot M_z$ then,

$$\begin{aligned} P(|z_n y_n| > M_\epsilon) &= P(|z_n||y_n| > M_\epsilon) \\ &\leq P(|z_n| > M_\epsilon) + P(|y_n| > M_y) \\ &< \epsilon \end{aligned}$$

□

Example 1.4.2. (Example of using Theorem 1.4.2)
$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0,1)$ implies $\hat{\beta} = \beta + o_p(1)$ so that $\hat{\beta}$ is a consistent estimator of β.

Proof.

$$\hat{\beta} - \beta = \frac{1}{\sqrt{n}} \cdot \sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \cdot O_p(1) = o_p(1) \cdot O_p(1) = o_p(1)$$

□

1.5 Inequalities[2]

Theorem 1.5.1. (Jensen's Inequality)
If $g(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is convex, then for any random vector x for which $E||x|| < +\infty$ and $E|g(x)| < +\infty$,

$$g(E(x)) \leq E(g(x)).$$

Proof. Since $g(u)$ is convex, at any point u there is a nonempty set of subderivatives (linear surfaces touching $g(u)$ at u but lying below $g(u)$ for all u). Let $a + b^t u$ be a subderivative of $g(u)$ at $u = E(x)$. Then for all u, $g(u) \geq a + b^t u$ yet $g(E(x)) = a + b^T E(x)$. Applying expectations, $E(g(x)) \geq a + b^T E(x) = g(E(x))$. □

Theorem 1.5.2. (Conditional Jensen's Inequality)
If $g(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is convex, then for any random vectors (y, x) for which $E||y|| < +\infty$ and $E||g(y)|| < +\infty$,

$$g(E(y|x)) \leq E(g(y)|x)$$

Theorem 1.5.3. (Conditional Expectation Inequality)
For any $r \geq 1$ such that $E|y|^r < +\infty$, then

$$E|E(y|x)|^r \leq E|y|^r < +\infty$$

Proof. By Conditional Jensen's inequality and the law of iterated expectations. □

[2] In this subsection I restate a number of useful equalities and their proofs from Hansen's textbook. You don't have to know the proofs unless you are interested in.

Theorem 1.5.4. (Expectation Inequality)
For any random matrix Y for which $E||Y|| < +\infty$,

$$||E(Y)|| \leq E||Y||.$$

Proof. Since matrix norm $||\cdot||$ is convex, apply Jensen's inequality. □

Theorem 1.5.5. (Hölder's Inequality)
If $p > 1$ and $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then for any random $m \times n$ matrices X and Y,

$$E||X^T Y|| \leq (E||X||^p)^{1/p}(E||Y||^q)^{1/q}.$$

Proof. Since $\frac{1}{p} + \frac{1}{q} = 1$, $exp(\cdot)$ is convex, apply Jensen's Inequality. For any real a and b,

$$\exp\left[\frac{1}{p}a + \frac{1}{q}b\right] \leq \frac{1}{p}\exp(a) + \frac{1}{q}\exp(b)$$

Now let $u = exp(a)$ and $v = exp(b)$. Then,

$$u^{1/p}v^{1/q} \leq \frac{u}{p} + \frac{v}{q}$$

Now let $u = ||X||^p/E||X||^p$ and $v = ||Y||^q/E||Y||^q$. Note that $E(u) = E(v) = 1$. By matrix Schwarz Inequality, $||X^T Y|| \leq ||X||||Y||$. Thus,

$$\begin{aligned}
\frac{E||X^T Y||}{(E||X||^p)^{1/p}(E||Y||^q)^{1/q}} &\leq \frac{E(||X||||Y||)}{(E||X||^p)^{1/p}(E||Y||^q)^{1/q}} \\
&= E(u^{1/p}v^{1/q}) \\
&\leq E(\frac{u}{p} + \frac{v}{p}) \\
&= \frac{1}{p} + \frac{1}{q} \\
&= 1
\end{aligned}$$

□

Theorem 1.5.6. (Cauchy-Schwarz Inequality)
For any random $m \times n$ matrices X and Y,

$$E||X^T Y|| \leq (E||X||^2)^{1/2}(E||Y||^2)^{1/2}$$

Theorem 1.5.7. (Minkowski's Inequality)
For any random $m \times n$ matrices X and Y,

$$(E||X + Y||^p)^{1/p} \leq (E||X||^p)^{1/p} + (E||Y||^p)^{1/p}$$

Proof.

$$\begin{aligned} E||X+Y||^p &= E(||X+Y||||X+Y||^{p-1}) \\ &\leq E(||X||||X+Y||^{p-1}) + E(||Y||||X+Y||^{p-1}) \\ &\leq (E||X||^p)^{1/p} E(||X+Y||^{q(p-1)})^{1/q} \\ &+ (E||Y||^p)^{1/p} E(||X+Y||^{q(p-1)})^{1/q} \\ &= ((E||X||^p)^{1/p} + (E||Y||^p)^{1/p}) E(||X+Y||^p)^{(p-1)/p} \end{aligned}$$

Divide both sides by $E(||X+Y||^p)^{(p-1)/p}$. □

Theorem 1.5.8. (Liapunov's Inequality)
For any random $m \times n$ matrix X and $1 \leq r \leq p$,

$$(E||X||^r)^{1/r} \leq (E||X||^p)^{1/p}$$

Proof. Note that function $g(u) = u^{p/r}$ is convex for $u > 0$ since $p \geq r$. Let $u = ||X||^r$ and apply Jensen's inequality. □

Theorem 1.5.9. (Markov Inequality Standard Form)
For any random vector x and non-negative function $g(x) \geq 0$,

$$P(g(x) > \alpha) \leq \frac{E(g(x))}{\alpha}$$

Theorem 1.5.10. (Markov Inequality Strong Form)
For any random vector x and non-negative function $g(x) \geq 0$,

$$P(g(x) > \alpha) \leq \frac{E(g(x)1(g(x) > \alpha))}{\alpha}.$$

Proof. Let F denote the distribution function of x. Then

$$\begin{aligned} P(g(x) \geq \alpha) &= \int_{\{g(u) \geq \alpha\}} dF(u) \\ &\leq \int_{\{g(u) \geq \alpha\}} \frac{g(u)}{\alpha} dF(u) \\ &= \alpha^{-1} \int 1(g(u) > \alpha) g(u) dF(u) \\ &= \alpha^{-1} E(g(x)1(g(x) > \alpha)) \end{aligned}$$

the inequality using the region of integration $\{g(u) > \alpha\}$. Since $1(g(x) > \alpha) \leq 1$, the final expression is less than $\frac{E(g(x))}{\alpha}$, establishing the standard form. □

Theorem 1.5.11. (Chebyshev's Inequality)
For any random variable x,

$$P(|x - E(x)| > \alpha) \leq \frac{Var(x)}{\alpha^2}$$

Proof. Define $y = (x - E(x))^2$ and note that $E(y) = Var(x)$. The events $\{|x - E(x)| > \alpha\}$ and $\{y > \alpha^2\}$ are equal, so by an application Markov's inequality we find

$$P(|x - E(x)| > \alpha) = P(y > \alpha^2) \leq \alpha^{-2} E(y) = \frac{Var(x)}{\alpha^2}$$

□

2 Instrumental Variable

2.1 Endogeneity

Consider a linear model

$$y_i = x_i'\beta + e_i \tag{1}$$

Among others, the key assumption for consistency of $\hat{\beta}$ was $E(x_i e_i) = 0$ since

$$\begin{aligned}\hat{\beta} &\xrightarrow{p} E(x_i x_i')^{-1} E(x_i y_i) \\ &= \beta + \underbrace{E(x_i x_i')^{-1} E(x_i e_i)}_{\text{endogeneity bias if } E(x_i e_i) \neq 0}\end{aligned}$$

We call there is an endogeneity in the linear model if $E(x_i e_i) = 0$ is violated.

Definition 2.1.1. (Endogeneity)
In a linear model $y_i = x_i'\beta + e_i$, we say that there is **endogeneity** in the linear model if β is the parameter of interest and

$$E(x_i e_i) \neq 0$$

It is typical to say that x_i is endogenous for β.

It turns out that for endogeneity, β_0 must describe more than just linear projection or conditional expectation. For example, suppose given i.i.d. data $\{y_i, x_i\}_{i=1}^n$, define

$$\begin{aligned}\beta^* &= E(x_i x_i')^{-1} E(x_i y_i) \\ u_i^* &= y_i - x_i'\beta^*\end{aligned}$$

So that β^* is the linear projection coefficient and u_i^* is the linear projection error. Then it is clear that

$$y_i = x_i'\beta^* + u_i^* \text{ with } E(u_i^* x_i) = 0$$

But still,

$$\begin{aligned}\beta^* &= E(x_i x_i')^{-1} E(x_i y_i) \\ &= E(x_i x_i')^{-1} E(x_i (x_i'\beta + e_i)) \\ &= \beta + E(x_i x_i)^{-1} E(x_i e_i)\end{aligned}$$

and it is not equal to β if endogeneity is present.

As another example, suppose given i.i.d. data $\{y_i, x_i\}_{i=1}^n$, define

$$\begin{aligned}E(y_i|x_i) &= x_i'\beta^* \\ u_i^* &= y_i - x_i'\beta^*\end{aligned}$$

then it is clear that

$$y_i = x_i'\beta^* + u_i^* \text{ with } E(u_i^*|x_i) = 0$$

and $E(u_i^* x_i) = E(E(u_i^* x_i|x_i)) = E(E(u_i^*|x_i)x_i) = 0$.

Hence to distinguish Equation (1) from linear projection or conditional models, we call Equation (1) a **structural equation** and β a **structural parameter**.

In most cases, β_0 in a model with endogeneity is an interpretable parameter in an idealized model that cannot be estimated due to practical limitations.

2.2 Sources of Endogeneity

Three major examples of sources of endogeneity are omitted variable bias, measurement error and simultaneous equations.

2.2.1 Omitted Variable Bias[3]

Suppose that we want to estimate the linear model

$$y_i = x_i'\beta + w_i'\delta + e_i \text{ with } E(e_i(x_i', w_i)') = 0$$

However, we do not observe (y_i, x_i, w_i) but only (y_i, x_i). That is, by defining

$$u_i = w_i'\delta + e_i$$

we can only at best estimate the model

$$y_i = x_i'\beta + u_i$$

Then there is endogeneity of x_i for β if x_i and w_i are correlated and $\delta \neq 0$ since

$$E(x_i u_i) = E(x_i w_i')\delta \neq 0$$

If x_i and w_i are uncorrelated, basically we cannot distinguish between w_i and e_i so there is no problem.

As a consequence, the probability limit of the OLS estimator that only uses $\{y_i, x_i\}_{i=1}^n$ is

$$\begin{aligned}\beta^* &= \beta + E(x_i x_i')^{-1} E(x_i w_i')\delta \\ &\neq \beta\end{aligned}$$

so the OLS estimator is inconsistent.

The direction of the (asymptotic) bias depends on the sign of $E(x_i w_i')$ and δ, which we can guess in some applications. If $E(x_i w_i')\delta$ is positive, we call that $\hat{\beta}$ is **over-estimated** and if $E(x_i w_i')\delta$ is negative, we call that $\hat{\beta}$ is **under-estimated**. Especially in the case where parameter estimates are to be given a causal interpretation, controlling for omitted variables bias is necessary. Since too many regressors cause little harm, but too few regressors can lead to inconsistency, microeconometric models estimated from large data sets tend to include many regressors.

Example 2.2.1. Suppose wage y_i is determined by education x_i and unobserved ability w_i as $y_i = x_i\beta + w_i\delta + e_i, E(e_i(x_i, w_i)') = 0$ but at best we can only estimate the model $y_i = x_i\beta + u_i$. If wages are affected by unobserved ability ($\delta \neq 0$) and individuals with high ability self-select into higher education ($E(x_i w_i') \neq 0$) then the error term contains unobserved ability ($u_i = w_i\delta + e_i$) so education and error will be positively correlated ($E(u_i x_i) \neq 0$). Hence we have education x_i endogenous for β and th e linear projection coefficient will be upward biased relative to the structural coefficient β.

Remark. (Inclusion of Irrelevant Regressors)
Related form of misspecification is inclusion of irrelevant regressors. In this case it is straightforward to show that OLS is consistent, but there is a loss of efficiency.

[3]Though it is a convention to write as omitted variable "bias", one should be aware that the bias here is an asymptotic concept.

2.2.2 Measurement Error

Suppose that we want to estimate the linear model

$$y_i = x_i^{*\prime}\beta + e_i \text{ with } E(e_i x_i^*) = 0$$

However, we do not observe x_i^* but only $x_i = x_i^* + v_i$ with v_i a mean zero measurement error that is independent of x_i^* and e_i. This case is known as **classical measurement error**, which means x_i is a noisy but unbiased measure of z_i.

Define $u_i = e_i - v_i'\beta$, we find that

$$\begin{aligned} y_i &= x_i^{*\prime}\beta + e_i \\ &= (x_i - v_i)'\beta + e_i \\ &= x_i'\beta + u_i \end{aligned}$$

Then there is endogeneity of x_i for β because

$$E(x_i u_i) = E((x_i^* + v_i)(e_i - v_i'\beta_0)) = -E(v_i v_i')\beta \neq 0$$

in general.

As a consequence, the probability limit of the OLS estimator that uses only $\{y_i, x_i\}_{i=1}^n$ is

$$\begin{aligned} \beta^* &= \beta - E(x_i x_i')^{-1} E(v_i v_i')\beta \\ &= \beta - [E(x_i^* x_i^{*\prime}) + E(v_i v_i')]^{-1} E(v_i v_i')\beta && : x_i^*, v_i \text{ independent} \\ &= [E(x_i^* x_i^{*\prime}) + E(v_i v_i')]^{-1} E(x_i^* x_i^{*\prime})\beta \end{aligned}$$

So OLS estimator is again inconsistent. Suppose that x_i and x_i^* are both scalar random variables. Then we find the probability limit of OLS estimator as

$$\underbrace{\frac{E(x_i^* x_i^{*\prime})}{E(x_i^* x_i^{*\prime}) + E(v_i v_i')}}_{\in(0,1)} \beta$$

hence the it shrinks the structural parameter β towards zero. This is called **measurement error bias** or **attenuation bias**.

Remark.

- Note that additive measurement error in dependent variable y_i with all assumptions above is not a problem since the measurment error just adds to the error e_i.
- If we have one more variable which is a noisy measure of x_i^*, say $w_i = x_i^* + \eta_i$, we can use as an instrument for x_i.

2.2.3 Simultaneous Equations

Consider a simple linear supply-and-demand system:

$$\begin{aligned} q_i &= -\beta p_i + e_{1i} && : \text{demand} \\ q_i &= \beta p_i + e_{2i} && : \text{supply} \end{aligned}$$

Assume that $e_i = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$ is i.i.d., $E(e_i) = 0$ and $E(e_i e_i') = I_2$[4] for simplicity.

Simple algebra shows that in equilibrium,

$$\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix} \begin{pmatrix} q_i \\ p_i \end{pmatrix} = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$$

and hence

$$\begin{aligned} \begin{pmatrix} q_i \\ p_i \end{pmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\beta_2 e_{1i} + \beta_1 e_{2i}}{\beta_1 + \beta_2} \\ \frac{e_{1i} - e_{2i}}{\beta_1 + \beta_2} \end{pmatrix} \end{aligned}$$

Hence, in both demand and supply equations the regressor p_i is correlated with the errors

$$\begin{aligned} E(p_i e_{1i}) &\neq 0 \\ E(p_i e_{2i}) &\neq 0 \end{aligned}$$

and hence we have endogeneity of p_i for both β_1 and β_2. The projection of q_i on p_i yields

$$\begin{aligned} q_i &= \beta^* p_i + e_i^* \\ E(p_i e_i^*) * &= 0 \end{aligned}$$

where

$$\beta^* = \frac{E(p_i q_i)}{E(p_i^2)} = \frac{\beta_2 - \beta_1}{2}$$

and hence the OLS estimate $\hat{\beta} \xrightarrow{p} \beta^*$ and it does not equal either β_1 and β_2. This is called the **simultaneous equations bias**. In general, when both the dependent variable and independent variable are simultaneously determined then they should be treated as endogenous.

2.3 Instrumental Variables

Again consider a linear model with endogeneity:

$$y_i = x_i'\beta + e_i \text{ with } E(e_i x_i) \neq 0$$

How can we identify and estimate β? An obvious way is through a randomized experiment, but for most economics applications such experiments are too expensive or even impossible. One popular approach is through the use of so-called instruments or instrumental variables (IV). This is to generate only *exogenous* variation in x.

In most applications we only treat a subset of the regressors as endogenous; most of the regressors will be treated as exogenous, meaning that they are assumed uncorrelated with the equation error. To be specific, we make the partition

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}$$

and similarly

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}$$

[4] I_k is an identity matrix of dimension k.

and assume

$$E(x_{1i}e_i) = 0$$
$$E(x_{2i}e_i) \neq 0$$

where $k_1 + k_2 = k$ and so that x_{1i} is **exogenous** and x_{2i} is **endogenous** for the structural parameter β. The **structural equation** is

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ &= x_{1i}'\beta_1 + x_{2i}'\beta_2 + e_i \end{aligned}$$

or equivalently using matrix notation,

$$\begin{aligned} y &= X\beta + e \\ &= X_1\beta + X_2\beta_2 + e \end{aligned}$$

To consistently estimate β, we require additional information. One type of information which is commonly used in economic applications are what we call instruments.

Definition 2.3.1. (Instrumental Variables)
An l-vector z_i is a vector of IVs for x_i in a regression of the form

$$y_i = x_i'\beta + e_i$$

if it satisfies

1. Exogeneity : $E(z_i u_0) = 0$
2. Relevance : $rank(E(z_i x_i')) = k$
3. $E(z_i z_i') > 0$

Roughly speaking, instrument z_i should be correlated with x_i and uncorrelated with u_i. Exogeneity condition[5] requires the instruments are uncorrelated with the regression error, and the relevance condition is essential for the identification of the model and a necessary condition follows, $l \geq k$. Third one is a normalization which excludes linearly redundant instruments.

Note that even if $E(x_i e_i) \neq 0$, x_{1i} is still uncorrelated with e_i (for example, the intercept) and they should be included as instrumental variables, as a subset of variables z_i. Hence,

$$z_i = \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} = \begin{pmatrix} x_{1i} \\ z_{2i} \end{pmatrix} = \begin{matrix} k_1 \\ l_2 \end{matrix}$$

where $k_1 + l_2 = l$.

Many authors simply label x_{1i} as the "exogenous variables", x_{2i} as the "endogenous variables" and z_{2i} as the "instrumental variables".

Example 2.3.1. (Example 2.2.1 Continued)
An ideal instrument affects the choice of the regressor (education) but does not directly influence the dependent variable (wages) except through the indirect effect on the regressor. Card (1993) suggested if a potential student lives close to a college, this reduces the cost of attendence and thereby raises the likelihood that the student will attend college (increase education, 'relevance'). However, college proximity does not directly affect a students skills or abilities, so should not have a direct effect on his or her market wage ('exogenous'). In this case college proximity can be used as instrument for education.

[5] Exogenous variables mean they are determined outside of the model for y_i.

Definition 2.3.2. (Identification)
In a model with instruments, we call the model is **just-identified** if $l = k$ (i.e., $l_2 = k_2$) and is **over-identified** if $l > k$ (i.e., $l_2 > k_2$). If $l < k$, the model is unidentified.

2.4 Reduced Form

References

Angrist, J. D., and J.-S. Pischke (2008): *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Cameron, A. C., and P. K. Trivedi (2005): *Microeconometrics: methods and applications*. Cambridge university press.

Card, D. (1993): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," *NBER Working Paper Series*.

Hansen, B. E. (2018): *Econometrics*. University of Wisconsin.

Wooldridge, J. M. (2010): *Econometrics Analysis of Cross Sections and Panel Data*. MIT press.