Introduction to Econometrics II*

Dong Woo Hahm†

Spring 2019

Contents

---

*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 Professor Christoph Rothe's lecture notes.

†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,

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

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.

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

$\square$

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 $n \to \infty$. $\square$

[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.[2]

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 $k = 1$.
$g(x) = g(\theta_0) + g'(\bar{x})(x - \theta_0)$ 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,

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

For stochastic sequences,

- $z_n = o_p(1) \iff z_n \xrightarrow{p} 0$
- $z_n = o_p(a_n) \iff \frac{z_n}{a_n} \xrightarrow{p} 0$
- $z_n = O_p(1) \iff \forall \epsilon > 0, \exists M_\epsilon > 0$ 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, $\hat{\beta} = \beta + o_p(1)$.

[2] The method used in proof of Delta method using Taylor expansion is extremely useful in econometrics. We will get to see this later in course for example in proving asymptotic normality of any type of extremum estimators, LM tests, or LR tests.

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,

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

□

1.5 Inequalities[3]

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. □

[3] 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. $\square$

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,

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

$\square$

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

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.

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

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