Lecture Notes on Econometrics II

Zhentao Shi

This PDF document is compiled as an overview of the contents to be covered in Econ5150. Don't print this document! All chapters will be revised as the course progresses.

Contents

Chapter 1

Asymptotic Theory I

1.1 Modes of Convergence

We first review what is *convergence* for a non-random sequence, which you learned in high school. Let $z_1, z_2, \ldots$ be an infinite sequence of non-random variables.

Definition 1.1. Convergence of this non-random sequence means that for any $\varepsilon > 0$, there exists an $N(\varepsilon)$ such that for all $n > N(\varepsilon)$, we have $|z_n - z| < \varepsilon$. We say z is the limit of z_n, and write $z_n \to z$ or $\lim_{n\to\infty} z_n = z$.

Instead of a deterministic sequence, we are interested in the convergence of a sequence of random variables. Since a random variable is "random" thanks to the induced probability measure by the measurable function, we must be clear what *convergence* means. Several modes of convergence are widely used.

Definition 1.2 (Convergence in probability)**.** We say a sequence of random variables (z_n) converges in probability to z, where z can be either a random variable or a non-random constant, if for any $\varepsilon > 0$, the probability $P\{\omega : |z_n(\omega) - z| < \varepsilon\} \to 1$ (or equivalently $P\{\omega : |z_n(\omega) - z| \geq \varepsilon\} \to 0$) as $n \to \infty$. We can write $z_n \xrightarrow{p} z$ or $\text{plim}_{n\to\infty} z_n = z$.

Definition 1.3 (r-th moment convergence)**.** A sequence of random variables (z_n) converges in squared-mean to z, where z can be either a random variable or a non-random constant, if $E\left[|z_n - z|^r\right] \to 0$ for some $r \geq 1$. It is denoted as $z_n \xrightarrow{rth.m.} z$. In particular, when $r = 2$ it is called square-mean convergence, written as $z_n \xrightarrow{m.s.} z$.

In these definitions either $P\{\omega : |z_n(\omega) - z| > \varepsilon\}$ or $E\left[|z_n - z|^r\right]$ is a non-random quantity, and it converges to 0 as a non-random sequence.

Squared-mean convergence is stronger than convergence in probability. That is, $z_n \xrightarrow{rth.m.} z$ implies $z_n \xrightarrow{p} z$ but the converse is untrue. Here is an example.

Example 1.1. (z_n) is a sequence of binary random variables: $z_n = n$ with probability $1/n$, and $z_n = 0$ with probability $1 - 1/n$. Then $z_n \xrightarrow{p} 0$ but $z_n \overset{1st.m.}{\nrightarrow} 0$. To verify these claims, notice that for any $\varepsilon > 0$, we have $P(\omega : |z_n(\omega) - 0| > \varepsilon) = P(\omega : z_n(\omega) = n) = 1/n \to 0$ and thereby $z_n \xrightarrow{p} 0$. On the other hand, $E[|z_n - 0|] = n \cdot 1/n + 0 \cdot (1 - 1/n) = 1 \nrightarrow 0$, so $z_n \overset{m.s.}{\nrightarrow} 0$.

Remark 1.1. Example 1.1 highlights the difference between the two modes of convergence. Convergence in probability does not count what happens on a subset in the sample space of small

probability. Squared-mean convergence deals with the average over the entire probability space. If a random variable can take a wild value, with small probability though, it may blow away the squared-mean convergence. On the contrary, such irregularity does not undermine convergence in probability.

Both convergence in probability and squared-mean convergence are about convergence of random variables to a target random variable or constant. That is, the distribution of $z_n - z$ is concentrated around 0 as $n \to \infty$. Instead, *convergence in distribution* is about the convergence of CDF, but not the random variable. Let $F_{z_n}(\cdot)$ be the CDF of z_n and $F_z(\cdot)$ be the CDF of z.

Definition 1.4 (Convergence in distribution)**.** We say a sequence of random variables (z_n) converges in distribution to a random variable z if $F_{z_n}(a) \to F_z(a)$ as $n \to \infty$ at each point $a \in \mathbb{R}$ such that where $F_z(\cdot)$ is continuous. We write $z_n \xrightarrow{d} z$.

Convergence in distribution is the weakest mode. If $z_n \xrightarrow{p} z$, then $z_n \xrightarrow{d} z$. The converse is not true in general, unless z is a non-random constant (A constant z can be viewed as a degenerate random variables, with a corresponding "CDF" $F_z(\cdot) = 1\{\cdot \geq z\}$.

Example 1.2. Let $x \sim N(0,1)$. If $z_n = x + 1/n$, then $z_n \xrightarrow{p} x$ and of course $z_n \xrightarrow{d} x$. However, if $z_n = -x + 1/n$, or $z_n = y + 1/n$ where $y \sim N(0,1)$ is independent of x, then $z_n \xrightarrow{d} x$ but $z_n \xrightarrow{p} x$.

Example 1.3. (z_n) is a sequence of binary random variables: $z_n = n$ with probability $1/\sqrt{n}$, and $z_n = 0$ with probability $1 - 1/\sqrt{n}$. Then $z_n \xrightarrow{d} z = 0$. Because

$$F_{z_n}(a) = \begin{cases} 0 & a < 0 \\ 1 - 1/\sqrt{n} & 0 \leq a \leq n \\ 1 & a \geq n \end{cases}.$$

$F_z(a) = \begin{cases} 0, & a < 0 \\ 1 & a \geq 0 \end{cases}$. It is easy to verify that $F_{z_n}(a)$ converges to $F_z(a)$ *pointwisely* on each point in $(-\infty, 0) \cup (0, +\infty)$, where $F_z(a)$ is continuous.

So far we have talked about convergence of scalar variables. These three modes of converges can be easily generalized to random vectors. In particular, the *Cramer-Wold device* collapses a random vector into a random vector via arbitrary linear combination. We say a sequence of K-dimensional random vectors (z_n) converge in distribution to z if $\lambda' z_n \xrightarrow{d} \lambda' z$ for any $\lambda \in \mathbb{R}^K$ and $\|\lambda\|_2 = 1$.

1.2 Law of Large Numbers

(Weak) law of large numbers (LLN) is a collection of statements about convergence in probability of the sample average to its population counterpart. The basic form of LLN is:

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - E[z_i]) \xrightarrow{p} 0$$

as $n \to \infty$. Various versions of LLN work under different assumptions about some features and/or dependence of the underlying random variables.

1.2.1 Cherbyshev LLN

We illustrate LLN by the simple example of Chebyshev LLN, which can be proved by elementary calculation. It utilizes the *Chebyshev inequality*.

- *Chebyshev inequality*: If a random variable x has a finite second moment $E\left[x^2\right] < \infty$, then we have $P\{|x| > \varepsilon\} \leq E\left[x^2\right]/\varepsilon^2$ for any constant $\varepsilon > 0$.

Exercise 1.1. Show that if $r_2 \geq r_1 \geq 1$, then $E\left[|x|^{r_2}\right] < \infty$ implies $E\left[|x|^{r_1}\right] < \infty$. (Hint: use Holder's inequality.)

The Chebyshev inequality is a special case of the *Markov inequality*.

- *Markov inequality*: If a random variable x has a finite r-th absolute moment $E\left[|x|^r\right] < \infty$ for some $r \geq 1$, then we have $P\{|x| > \varepsilon\} \leq E\left[|x|^r\right]/\varepsilon^r$ any constant $\varepsilon > 0$.

Proof. It is easy to verify the Markov inequality.

$$E\left[|x|^r\right] = \int_{|x|>\varepsilon} |x|^r \, dF_X + \int_{|x|\leq\varepsilon} |x|^r \, dF_X \geq \int_{|x|>\varepsilon} |x|^r \, dF_X \geq \varepsilon^r \int_{|x|>\varepsilon} dF_X = \varepsilon^r P\{|x| > \varepsilon\}.$$

Rearrange the above inequality and we obtain the Markov inequality. □

Let the *partial sum* $S_n = \sum_{i=1}^n x_i$, where $\mu_i = E[x_i]$ and $\sigma_i^2 = \text{var}[x_i]$. We apply the Chebyshev inequality to the sample mean $z_n = \bar{x} - \bar{\mu} = n^{-1}(S_n - E[S_n])$.

$$\begin{aligned}
P\{|z_n| \geq \varepsilon\} &= P\left\{n^{-1}|S_n - E[S_n]| \geq \varepsilon\right\} \\
&\leq E\left[\left(n^{-1}\sum_{i=1}^n (x_i - \mu_i)\right)^2\right]/\varepsilon^2 \\
&= (n\varepsilon)^{-2}\left\{E\left[\sum_{i=1}^n (x_i - \mu_i)^2\right] + \sum_{i=1}^n \sum_{j\neq i} E\left[(x_i - \mu_i)(x_j - \mu_j)\right]\right\} \\
&= (n\varepsilon)^{-2}\left\{\sum_{i=1}^n \text{var}(x_i) + \sum_{i=1}^n \sum_{j\neq i} \text{cov}(x_i, x_j)\right\}. \qquad (1.1)
\end{aligned}$$

Convergence in probability holds if the right-hand side shrinks to 0 as $n \to \infty$. For example, If $x_1, \ldots, x_n$ are iid with $\text{var}(x_1) = \sigma^2$, then the RHS of (1.1) is $(n\varepsilon)^{-2}(n\sigma^2) = o(n^{-1}) \to 0$. This result gives the Chebyshev LLN:

- Chebyshev LLN: If $(z_1, \ldots, z_n)$ is a sample of iid observations, $E[z_1] = \mu$, and $\sigma^2 = \text{var}[z_1] < \infty$ exists, then $\frac{1}{n}\sum_{i=1}^n z_i \xrightarrow{p} \mu$.

The convergence in probability can be indeed maintained under much more general conditions than under iid case. The random variables in the sample do not have to be identically distributed, and they do not have to be independent either.

Exercise 1.2. Consider an inid (independent but non-identically distributed) sample $(x_1, \ldots, x_n)$ with $E[x_i] = 0$ and $\text{var}[x_i] = \sqrt{n}c$ for some constant $c > 0$. Use the Chebyshev inequality to show that $n^{-1}\sum_{i=1}^n x_i \xrightarrow{p} 0$.

Another useful LLN is the *Kolmogorov LLN*. Since its derivation requires more advanced knowledge of probability theory, we state the result without proof.

- Kolmogorov LLN: If $(z_1, \ldots, z_n)$ is a sample of iid observations and $E[z_1] = \mu$ exists, then $\frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow{a.s.} \mu$.

Compared with the Chebyshev LLN, the Kolmogorov LLN only requires the existence of the population mean, but not any higher moments. On the other hand, iid is essential for the Kolmogorov LLN.

1.3 Central Limit Theorem

The central limit theorem (CLT) is a collection of probability results about the convergence in distribution to a stable distribution. The limiting distribution is usually the Gaussian distribution. The basic form of the CLT is:

- *Under some conditions to be spelled out*, the sample average of *zero-mean* random variables $(z_1, \ldots, z_n)$ multiplied by $\sqrt{n}$ satisfies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i \xrightarrow{d} N\left(0, \sigma^2\right)$$

  as $n \to \infty$.

Various versions of CLT work under different assumptions about the random variables. *Lindeberg-Levy CLT* is the simplest CLT.

- If the sample $(x_1, \ldots, x_n)$ is iid, $E[x_1] = 0$ and $\text{var}[x_1] = \sigma^2 < \infty$, then $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \xrightarrow{d} N\left(0, \sigma^2\right)$.

Lindeberg-Levy CLT can be proved by the *moment generating function*. For any random variable x, the function $M_x(t) = E[\exp(xt)]$ is called its the *moment generating function* (MGF) if it exists. MGF fully describes a distribution, just like PDF or CDF. For example, the MGF of $N\left(\mu, \sigma^2\right)$ is $\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$.

Heuristic proof of Lindeberg-Levy CLT. If $E\left[|x|^k\right] < \infty$ for a positive integer k, then

$$M_X(t) = 1 + tE[X] + \frac{t^2}{2}E\left[X^2\right] + \ldots \frac{t}{k!}E\left[X^k\right] + O\left(t^{k+1}\right).$$

Under the assumption of Lindeberg-Levy CLT,

$$M_{\frac{X_i}{\sqrt{n}}}(t) = 1 + \frac{t^2}{2n}\sigma^2 + O\left(\frac{t^3}{n^{3/2}}\right)$$

for all i, and by independence we have

$$M_{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i}(t) = \prod_{i=1}^{n} M_{\frac{X_i}{\sqrt{n}}}(t) = \left(1 + \frac{t^2}{2n}\sigma^2 + O\left(\frac{t^3}{n^{3/2}}\right)\right)^n \to \exp\left(\frac{\sigma^2}{2}t^2\right),$$

where the limit is exactly the characteristic function of $N\left(0, \sigma^2\right)$. □

1.4 Tools for Transformations

In their original forms, LLN deals with the sample mean, and CLT handles the scaled (by $\sqrt{n}$) and/or standardized (by standard deviation) sample mean. However, most of the econometric estimators of interest are functions of sample means. For example, in the OLS estimator

$$\widehat{\beta} = \left(\frac{1}{n} \sum_i x_i x_i' \right)^{-1} \frac{1}{n} \sum_i x_i y_i$$

involves matrix inverse and the matrix-vector multiplication. We need tools to handle transformations.

- Continuous mapping theorem 1: If $x_n \xrightarrow{p} a$ and $f(\cdot)$ is continuous at a, then $f(x_n) \xrightarrow{p} f(a)$.
- Continuous mapping theorem 2: If $x_n \xrightarrow{d} x$ and $f(\cdot)$ is continuous almost surely on the support of x, then $f(x_n) \xrightarrow{d} f(x)$.
- Slutsky's theorem: If $x_n \xrightarrow{d} x$ and $y_n \xrightarrow{p} a$, then
  - $x_n + y_n \xrightarrow{d} x + a$
  - $x_n y_n \xrightarrow{d} ax$
  - $x_n / y_n \xrightarrow{d} x/a$ if $a \neq 0$.

Slutsky's theorem consists of special cases of the continuous mapping theorem 2. Only because the addition, multiplication and division are encountered so frequently in practice, we list it as a separate theorem.

- Delta method: if $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Omega)$, and $f(\cdot)$ is continuously differentiable at θ_0 (meaning $\frac{\partial}{\partial \theta} f(\cdot)$ is continuous at θ_0), then we have

$$\sqrt{n}\left(f\left(\widehat{\theta}\right) - f(\theta_0)\right) \xrightarrow{d} N\left(0, \frac{\partial f}{\partial \theta'}(\theta_0)\, \Omega \left(\frac{\partial f}{\partial \theta}(\theta_0)\right)'\right).$$

Proof. Take a Taylor expansion of $f\left(\widehat{\theta}\right)$ around $f(\theta_0)$:

$$f\left(\widehat{\theta}\right) - f(\theta_0) = \frac{\partial f(\dot{\theta})}{\partial \theta'}\left(\widehat{\theta} - \theta_0\right),$$

where $\dot{\theta}$ lies on the line segment between $\widehat{\theta}$ and θ_0. Multiply $\sqrt{n}$ on both sides,

$$\sqrt{n}\left(f\left(\widehat{\theta}\right) - f(\theta_0)\right) = \frac{\partial f(\dot{\theta})}{\partial \theta'} \sqrt{n}\left(\widehat{\theta} - \theta_0\right).$$

Because $\widehat{\theta} \xrightarrow{p} \theta_0$ implies $\dot{\theta} \xrightarrow{p} \theta_0$ and $\frac{\partial}{\partial \theta'} f(\cdot)$ is continuous at θ_0, we have $\frac{\partial}{\partial \theta'} f\left(\dot{\theta}\right) \xrightarrow{p} \frac{\partial f(\theta_0)}{\partial \theta'}$ by the continuous mapping theorem 1. In view of $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Omega)$, Slutsky's Theorem implies

$$\sqrt{n}\left(f\left(\widehat{\theta}\right) - f(\theta_0)\right) \xrightarrow{d} \frac{\partial f(\theta_0)}{\partial \theta'} N(0, \Omega)$$

and the conclusion follows.

 □

Chapter 2

CLT for inid Sequences

2.1 Notations and Definitions

A random variable z is rth *integrable* if $E[|z|^r] = \int_{-\infty}^{\infty} |z|\, dF(z) < \infty$. Equivalently,

$$\lim_{M\to\infty} E[|z|^r \mathbb{I}\left\{|z|^r > M\right\}] = 0,$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. Without referring explicitly to the rth moment, we say z is *integrable* if $r = 1$, and *square integrable* if $r = 2$.

A *triangular array* $\{(x_{1n}, x_{2n}, x_{3n}, ..., x_{r_n n})\}_{n\in\mathbb{N}}$ stacks like a triangle:

$$\begin{pmatrix} x_{11} & x_{22} & \cdots & x_{r_1 1} & & & & \\ x_{12} & x_{22} & \cdots & \cdots & x_{r_2 2} & & & \\ x_{13} & x_{23} & \cdots & \cdots & \cdots & x_{r_3 3} & & \\ \vdots & & & & & & \ddots & \\ x_{1n} & x_{2n} & \cdots & \cdots & \cdots & \cdots & \cdots & x_{r_n n} \end{pmatrix}.$$

Here r_n is an increasing number in n, and $r_n \to \infty$ as $n \to \infty$. (Think about the special case $r_n = n$, which makes an exact triangle.) Suppose for each n, the elements in $(x_{in})_{i=1}^{r_n}$ are independently non-identically distributed (inid). (Please keep a liberal mind and take "identically distributed" as a special case of "non-identically distributed".)

Without loss of generality, assume $E[x_{in}] = 0$ for all i and n and denote $\sigma_{in}^2 = E[x_{in}^2]$. Define the *partial sum* (up to n) as $S_n = \sum_{i=1}^{r_n} x_{in}$ and and (the nth) *aggregate variance* as $\tilde{\sigma}_n^2 = \sum_{i=1}^{r_n} \sigma_{in}^2$.

2.2 Lindeberg Condition

Lindeberg-Lévy Central Limit Theorem is for independently and identically distributed (iid) sequences. In this lecture we consider independent, heterogeneous sequences.

Definition 2.1. Lindeberg Condition:

$$\lim_{n\to\infty} \frac{1}{\tilde{\sigma}_n^2} \sum_{i=1}^{r_n} E[x_{in}^2 \mathbb{I}\left\{x_{in}^2 \geq \varepsilon \tilde{\sigma}_n^2\right\}] = 0$$

for all $\varepsilon > 0$.

Theorem 2.1 (Lindeberg-Feller CLT)**.** *If the triangular array* $\left\{(x_{in})_{i=1}^{r_n}\right\}_{n\in\mathbb{N}}$ *satisfies the Lindeberg condition, then*

$$\frac{S_n}{\tilde{\sigma}_n} \xrightarrow{d} N(0,1)$$

Lindeber-Feller CLT allows heterogeneity across $i = 1, ..., r_n$. It includes *Lindeberg-Lévy CLT* as a special case. To see this fact, under iid let us use z to represent the homogeneous distribution. Denote $\text{var}(z) = \sigma_z^2 \in (0, \infty)$, and equivalently $\lim_{M\to\infty} E\left[z^2 \mathbb{I}\left\{z^2 \geq M\right\}\right] = 0$ (square integrability). Set $r_n = n$, and thus $\tilde{\sigma}_n^2 = n\sigma_z^2$. As a result,

$$\begin{aligned}\frac{1}{\tilde{\sigma}_n^2}\sum_{i=1}^{n} E\left[x_{in}^2 \mathbb{I}\left\{x_{in}^2 \geq \varepsilon\tilde{\sigma}_n^2\right\}\right] &= \frac{1}{n\sigma_1^2} \times nE\left[z^2 \mathbb{I}\left\{z^2 \geq n\sigma_z^2\varepsilon\right\}\right] \\ &= const \times E\left[z^2 \mathbb{I}\left\{z^2 \geq n\sigma_z^2\varepsilon\right\}\right] \to \infty\end{aligned}$$

since $n\sigma_1^2\varepsilon \to \infty$ as $n \to \infty$.

With iid and $r_n = n$, we can drop the subscript n and write $z_i = x_{in}$. The ratio

$$\frac{S_n}{\tilde{\sigma}_n} = \frac{\sum_{i=1}^n z_i}{\sqrt{n\sigma_z^2}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{z_i}{\sigma_z}$$

retains its familiar form in CLT.

2.3 Lyapunov Condition

Lindeberg condition is mathematical artifact that is difficult to interpret. Lyapunov condition is a more interpretable sufficient condition.

Definition 2.2. Lyapunov Condition: There exists some $\delta > 0$ such that

$$\lim_{n\to\infty}\frac{1}{\tilde{\sigma}_n^{2+\delta}}\sum_{i=1}^{r_n} E[|x_{in}|^{2+\delta}] = 0$$

Lyapunov condition implies Lindeberg condition. To see this fact, we use the quantity in the Lindeberg condition as a starting point:

$$\begin{aligned}\frac{1}{\tilde{\sigma}_n^2}\sum_{i=1}^{r_n} E\left[x_{in}^2 \mathbb{I}\left\{x_{in}^2 \geq \varepsilon\tilde{\sigma}_n^2\right\}\right] &= \frac{1}{\tilde{\sigma}_n^2}\sum_{i=1}^{r_n} E\left[|x_{in}|^2 \mathbb{I}\left\{|x_{in}|^\delta \geq \varepsilon^{\frac{\delta}{2}}\tilde{\sigma}_n^\delta\right\}\right] \\ &= \frac{1}{\tilde{\sigma}_n^2}\sum_{i=1}^{r_n} E\left[\frac{|x_{in}|^{2+\delta}}{|x_{in}|^\delta}\mathbb{I}\left\{|x_{in}|^\delta \geq \varepsilon^{\frac{\delta}{2}}\tilde{\sigma}_n^\delta\right\}\right] \\ &\leq \frac{1}{\tilde{\sigma}_n^2} \times \frac{1}{\varepsilon^{\delta/2}\tilde{\sigma}_n^\delta}\sum_{i=1}^{r_n} E\left[|x_{in}|^{2+\delta}\mathbb{I}\left\{|x_{in}|^\delta \geq \varepsilon^{\frac{\delta}{2}}\tilde{\sigma}_n^\delta\right\}\right] \\ &\leq \frac{1}{\varepsilon^{\delta/2}} \times \frac{1}{\tilde{\sigma}_n^{2+\delta}}\sum_{i=1}^{r_n} E\left[|x_{in}|^{2+\delta}\right] \to 0\end{aligned}$$

as $n \to \infty$, where the limit follows from Lyapunov condition.

2.3.1 Sufficient Condition for Lyapunov Condition

- Condition1: $\sup_{i\leq r_n} E[|x_{in}|^{2+\delta}] \leq B < \infty$ for all sufficiently large n.
- Condition2: $\liminf_{n\to\infty} \bar{\sigma}_n^2 > b > 0$, where $\bar{\sigma}_n^2 = \tilde{\sigma}_n^2/r_n$ is the *average variance*.

Under Condition 1 and Condition 2 we have

$$\begin{aligned}\frac{1}{\tilde{\sigma}_n^{2+\delta}}\sum_{i=1}^{r_n} E[|x_{in}|^{2+\delta}] &\leq \frac{1}{(\sqrt{r_n b})^{2+\delta}} \times r_n \max_{i\leq r_n} E[|x_{in}|^{2+\delta}] \\ &\leq \frac{r_n B}{(\sqrt{r_n b})^{2+\delta}} = const \times r_n^{-\delta/2} \to 0\end{aligned}$$

since $r_n \to \infty$ as $n \to \infty$.

If we further assume $\bar{\sigma}_n^2 \to \sigma_*^2$ as $n \to \infty$, then under Condition 1 we have $\frac{S_n}{\sqrt{n}} \xrightarrow{d} N\left(0, \sigma_*^2\right)$.

2.3.2 Uniform CLT

If $E\left[|z|^{2+\delta}\right] \leq B < \infty$ and $\text{var}(z) \geq b > 0$ for all $f \in \mathscr{F}$, then

$$\sup_{f\in\mathscr{F}} \left| P_f\left(\frac{\sqrt{n}\left(\bar{z}_n - E(z)\right)}{\sqrt{\text{var}(z)}} \leq a\right) - \Phi(a)\right| \to 0.$$

This is a uniform CLT over a class of distributions in $\mathscr{F}$, instead of a single distribution f. Here P_f means that the probability is computed under a specific distribution f.

In a direct proof, the approximation error is controlled by B and b. The textbook uses a counter-positive argument: If the statement is false, then there is a sequence $f_1, f_2, ... \in \mathscr{F}$ that violates the convergence. That contradicts with Lyapunov CLT.

2.4 Uniform Integrability

Definition 2.3. The sequence of random variables z_n is *uniformly integrable* if

$$\lim_{M\to\infty} \sup_{n\geq 1} E\left[|z_n| \mathbb{I}\left\{|z_n| > M\right\}\right] = 0.$$

The textbook uses $\limsup_{n\to\infty}$ instead of $\sup_{n\geq 1}$ in the definition. These two notations are equivalent in our context here, as $E\left[|z_n| \mathbb{I}\left\{|z_n| > M\right\}\right] \searrow 0$ for every n as $M \to \infty$. "$\sup_{n\geq 1}$" appears more often in probability theory textbooks, and literally adheres to the notation of "uniformity".

Example 2.1. Consider a counterexample

$$z_n = \begin{cases} -\sqrt{n} & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 2/n \\ \sqrt{n} & \text{with probability } 1/n. \end{cases}$$

Notice that

$$E\left[z_n^2 \mathbb{I}\left\{z_n^2 > M\right\}\right] = 2 \times (n\mathbb{I}\{n > M\}) \times \frac{1}{n} = 2 \cdot \mathbb{I}\{n > M\}.$$

For each fixed n, this z_n is square integrable in that $2 \cdot \mathbb{I}\{n > M\} = 0$ for all $M \geq n$. However, as $\sup_{n\geq 1} E\left[z_n^2 \mathbb{I}\left\{z_n^2 > M\right\}\right] = 2\sup_{n\geq 1} \mathbb{I}\{n > M\} = 2$ for any finite M, and thus

$$\lim_{M\to\infty} \sup_{n\geq 1} E\left[z_n^2 \mathbb{I}\left\{z_n^2 > M\right\}\right] = 2 \nrightarrow 0.$$

As a result, this sequence z_n is NOT uniformly square integrable.

Definition 2.4. A triangular array of random variables is *uniformly integrable* if

$$\lim_{M\to\infty} \sup_{n\geq 1} \max_{i\leq r_n} E\left[|x_{in}| \mathbb{I}\{|x_{in}| > M\}\right] = 0.$$

Compared with Definition 2.3, we replace $E\left[|z_n| \mathbb{I}\{|z_n| > M\}\right]$ with $\max_{i\leq r_n} E\left[|x_{in}| \mathbb{I}\{|x_{in}| > M\}\right]$ in Definition 2.4 to control the worst case among the heterogeneous $(x_{in})_{i=1}^{r_n}$.

Proposition 2.1. *If Condition2 holds and the triangular array $\left\{(x_{in})_{i=1}^{r_n}\right\}_{n\in\mathbb{N}}$ is uniform square integrable, then Lindeberg condition holds.*

Proof. For any $\varepsilon > 0$, we have

$$\begin{aligned}
\frac{1}{\tilde{\sigma}_n^2} \sum_{i=1}^{r_n} E\left[x_{in}^2 \mathbb{I}\left\{x_{in}^2 \geq \varepsilon \tilde{\sigma}_n^2\right\}\right] &\leq \frac{1}{r_n b} \sum_{i=1}^{r_n} E\left[x_{in}^2 \mathbb{I}\left\{x_{in}^2 \geq \varepsilon r_n b\right\}\right] \\
&\leq \frac{1}{r_n b} \times r_n \max_{i\leq r_n} E\left[x_{in}^2 \mathbb{I}\left\{x_{in}^2 \geq r_n \varepsilon b\right\}\right] \\
&= const \times \max_{i\leq r_n} E\left[x_{in}^2 \mathbb{I}\left\{x_{in}^2 \geq r_n \varepsilon b\right\}\right] \to 0
\end{aligned}$$

by the definition of uniform integrability since $r_n \varepsilon b \to \infty$ as $n \to \infty$. □

2.4.1 Uniform Stochastic Bound

Theorem 2.2. *If*

$$\lim_{M\to\infty} \sup_{n\geq 1} \max_{i\leq r_n} E\left[|x_{in}|^r \mathbb{I}\left\{|x_{in}|^r > M\right\}\right] = 0$$

holds, then

$$r_n^{-1/r} \max_{i\leq r_n} |x_{in}| \xrightarrow{p} 0.$$

Proof. We start with the definition of convergence in probability:

$$\begin{aligned}
P\left(r_n^{-1/r} \max_{i\leq r_n} |x_{in}| > \varepsilon\right) &= P\left(\max_{i\leq n} |x_{in}|^r > r_n \varepsilon^r\right) \\
&\leq \sum_{i\leq r_n} P\left(|x_{in}|^r > r_n \varepsilon^r\right) \\
&= \sum_{i\leq r_n} E\left[\mathbb{I}\left\{|x_{in}|^r > r_n \varepsilon^r\right\}\right] \\
&\leq r_n \max_{i\leq r_n} E\left[\mathbb{I}\left\{|x_{in}|^r > r_n \varepsilon^r\right\}\right] \\
&\leq r_n \times \frac{1}{r_n \varepsilon^r} \max_{i\leq r_n} E\left[|x_{in}|^r \mathbb{I}\left\{|x_{in}|^r > r_n \varepsilon^r\right\}\right] \\
&= const \times \max_{i\leq r_n} E\left[|x_{in}|^r \mathbb{I}\left\{|x_{in}|^r \geq r_n \varepsilon^r\right\}\right] \\
&\to 0
\end{aligned}$$

under the uniform rth integrability, since $r_n \varepsilon^r \to \infty$ as $n \to \infty$. □

As a special case, if we set $r_n = n$, then $\max_{i \leq n} |x_{in}| = o_p\left(n^{1/r}\right)$ if x_{in} is rth uniformly integrable.

Chapter 3

Empirical Process Theory

3.1 Motivation

OLS and 2SLS have closed-form solutions. In general, however, GMM, NLS and MLE estimators cannot be expressed in closed-form. New asymptotic apparatus are needed for them.

Example 3.1. The maximum likelihood estimator is

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log f\left(X_i, \theta\right) = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} -\log f\left(X_i, \theta\right).$$

We will need a *uniform* law of large numbers to establish the consistency of $\widehat{\theta}$. Pointwise convergence at every $\theta \in \Theta$ is not sufficient. [to add a diagram to show pointwise convergence does not guarantee uniform convergence.]

Let $g\left(x, \theta\right) : \mathcal{X} \times \Theta \mapsto \mathbb{R}^K$. Here x is the realized value of a random variable, and θ is the parameter. The statistical model is indexed by $\theta \in \Theta$. To keep it simple, we consider Θ a subset of a finite-dimensional Euclidean space $\mathbb{R}^D$. In Example 3.1 $g\left(X_i\left(\omega\right), \theta\right) = -\log f\left(X_i\left(\omega\right), \theta\right)$.

We are interested in large sample behaviors of two objects: the *sample average*

$$\bar{g}_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(X_i, \theta\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(X_i\left(\omega\right), \theta\right)$$

and the *normalized average*

$$\nu_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \left(g\left(X_i, \theta\right) - E\left[g\left(X_i, \theta\right)\right]\right).$$

Some books call this form of ν_n *empirical process.*

Throughout this chapter, we discuss iid $\left(X_i\right)_{i=1}^n$ for simplicity. Without loss of generality, the marginal distribution of X_i can be represented by X.

Example 3.2. A classical statistical example that motivates the Empirical Cumulative Distribution Function (ECDF)

$$F_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(X_i \leq \theta\right)$$

for $\theta \in \mathbb{R}$. The pointwise LLN and CLT yield

$$F_n(\theta) \xrightarrow{p} F(\theta) = P(X_1 \leq \theta)$$
$$\sqrt{n}(F_n(\theta) - F(\theta)) \xrightarrow{d} N(0, F(\theta)(1 - F(\theta))).$$

Can we strengthen the convergence to hold uniformly over $\theta \in \mathbb{R}$? What is the asymptotic distribution of $\sqrt{n}(F_n(\theta) - F(\theta))$ as a random function indexed by θ?

3.2 Complexity of a Family of Functions

We first work with uniform convergence. To generalize pointwise convergence to uniform convergence, it is essential to study the complexity of the class of functions

$$\mathcal{G} := \{g(\cdot, \theta) : \theta \in \Theta\}$$

under the probability distribution Q that generates the random variables X.

For a generic function $h(x)$, define the L_r-norm as

$$\|h\|_{Q,r} = \left(E_Q\left[\|h(X)\|^r\right]\right)^{1/r} = \left(\int \|h(X)\|^r \, dQ\right)^{1/r}.$$

A function $G(x)$ is called an **envelope function** of $\mathcal{G}$ if $\sup_{\theta\in\Theta} \|g(x,\theta)\| \leq G(x)$ for every $x \in \mathcal{X}$. The envelope function works as the dominating function in the celebrated dominated convergence theorem.

The **distance** between two functions $g(\cdot, \theta_1)$ and $g(\cdot, \theta_2)$ is defined as $\|g(\cdot, \theta_1) - g(\cdot, \theta_2)\|_{Q,r}$. Given the distance, we define an ε-neighborhood centered around $g(\cdot, \theta)$:

$$\mathcal{N}_{\varepsilon,r,Q}(g(\cdot,\theta)) := \left\{h \in \mathcal{G} : \|h(\cdot) - g(\cdot,\theta)\|_{Q,r} \leq \varepsilon\right\}.$$

The **covering number**, denoted as $N_r(\varepsilon, Q)$, is the smallest number of points m such that $\mathcal{G} \subseteq \bigcup_{\ell=1}^{m} \mathcal{N}_{\varepsilon,r,Q}(g(\cdot,\theta_\ell))$. In other words, those m neighborhoods cover the set of functions $\mathcal{G}$.

Remark 3.1. It is straightforward to generalize from pointwise convergence to uniform convergence if Θ consists of a finite number of singletons $\{\theta_1, \theta_2, \ldots, \theta_m\}$. If the covering number is finite, we can focus on a finite set of functions $\{g(\cdot,\theta_\ell)\}_{\ell=1}^{m}$. Those $g(\cdot,\theta)$ in the ε-neighborhood of $g(\cdot,\theta_\ell)$ will behave similarly to the center $g(\cdot,\theta_\ell)$.

If a pair of functions $l(x)$ and $u(x)$ satisfies $\|u - l\|_{Q,r} \leq \varepsilon$ and $l(x) \leq u(x)$, we call (l, u) an ε-$L_r(Q)$ **bracket**. The **bracketing number**, denoted as $N_{[]}(\varepsilon, L_r(Q))$, is the smallest number of brackets such that for each $g(\cdot,\theta) \in \mathcal{G}$ there exists a bracket (l_ℓ, u_ℓ) satisfying $l_\ell(x) \leq g(x,\theta) \leq u_\ell(x)$. The upper bracket u_ℓ or the lower bracket l_ℓ does not have to in $\mathcal{G}$.

Example 3.3. The covering number and the bracketing number on the functional space are somewhat abstract. Let us use a trivial example of the unit real interval $[0,1]$ to demonstrate the spirit behind it. If we set $\varepsilon = 0.1$, we can pick up 5 points $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ so that the 0.1-neighborhoods of these points covers the entire unit interval. The covering number in this case is 5. Regarding bracketing, $[0, 0.1], [0.1, 0.2], \ldots, , [0.9, 1]$ are 0.1-brakets and they capture all points in the unit interval. The 0.1-bracketing number is 10.

Bracketing number is more restrictive than covering number. For any fixed r, we consider a 2ε-bracket $[l, u]$. The middle line $(l + u)/2$ decomposes the 2ε-bracket into two ε-brackets $[l, \frac{l+u}{2}]$ and $[\frac{l+u}{2}, u]$ and thus $l, u \in \mathcal{N}_{\varepsilon,r,Q}\left(\frac{l+u}{2}\right)$. For any g in the bracket $[l, u]$, it must fall into $\mathcal{N}_{\varepsilon,r,Q}\left(\frac{l+u}{2}\right)$. As a result, $N_r(\varepsilon, Q) \leq N_{[]}(2\varepsilon, L_r(Q))$.

3.3 Uniform Law of Large Numbers

Definition 3.1. The sample average $\bar{g}_n(\theta)$ satisfies the **uniform law of large numbers (ULLN)** over Θ if

$$\sup_{\theta\in\Theta} |\bar{g}_n(\theta) - E[g(X_i,\theta)]| \xrightarrow{p} 0.$$

Below is a set of sufficient conditions for ULLN.

Theorem 3.1 (ULLN)**.** *If*
1. X_i are i.i.d.
2. $E[G(X)] < \infty$.
3. One of the following holds:
(i) $N_{[]}(\varepsilon, L_1(Q)) < \infty$ for all $\varepsilon > 0$.
(ii) $N_1(\varepsilon, Q) < \infty$ for all $\varepsilon > 0$.
(iii) Θ is compact and

$$P\{\omega \in \Omega : g(X(\omega),\theta) \text{ is continuous in } \theta\} = 1$$

at every $\theta \in \Theta$.

Condition 1 is for simplicity. Condition 2 allows us to invoke dominated convergence theorem. Condition 3(i) and (ii) are about finite covering number and bracketing number, respectively. See Newey and MacFadden (1994, Handbook of Econometrics, Chapter 36)'s Lemma 2.4 for Condition 3(iii).

Example 3.4. We verify that of ECDF satisfies Theorem 3.1. For $g(x,\theta) = \mathbb{I}(x \le \theta)$, an envelop function is $G(x) = 1$, which is integrable as $E[G(x)] = 1$ for all Q. Next, we construct an ε-L_1 bracket system. To be specific, consider $\varepsilon = 1/8$. We can pick up the $0/8$-th quantile $q_0 = -\infty$, $1/8$-quantile q_1, $2/8$-th quantile q_2 and so on up to the $7/8$-th quantile q_7, and $8/8$-th quantile $q_8 = \infty$, which form 8 brackets $[h_{\ell-1}, h_\ell] = [\mathbb{I}(x \le q_{\ell-1}), \mathbb{I}(x \le q_\ell)]$, for $\ell = 1, \ldots 8$. ($h_0 = 0$ and $h_8 = 1$ are outside of $\{\mathbb{I}(x \le \theta) : \theta \in \mathbb{R}\}$). It is easy to check that each $[h_{\ell-1}, h_\ell]$ is a $1/8$-L_1 bracket, and any $\mathbb{I}(x \le \theta)$ must fall into one of the brackets. In general, for any $\varepsilon > 0$, we can construct such a bracket system based on the quantiles. The bracketing number is $\lfloor 1/\varepsilon \rfloor$.

Since the bracketing system in Example 3.4 is independent of the underlying probability distribution Q, we have the well-known Glivenko-Cantelli Theorem.

Corollary 3.1 (Glivenko-Cantelli Theorem)**.** *If X_i is iid, then* $\sup_{\theta\in\mathbb{R}} |F_n(\theta) - F(\theta)| \xrightarrow{p} 0$.

3.4 Functional Central Limit Theorem

In earlier lectures, CLT is pointwise about $\nu_n(\theta)$ for a fixed θ. In this lecture, we consider $\nu_n(\theta)$ as a process indexed by θ as an entity.

For two non-random processes $\nu_1(\theta)$ and $\nu_2(\theta)$, we define a **uniform metric** $\rho(\nu_1,\nu_2) = \sup_{\theta\in\Theta} |\nu_1(\theta) - \nu_2(\theta)|$ as a measure of the distance. Let V be a class of processes $\nu : \Theta \to \mathbb{R}$.

Definition 3.2. Convergence in distribution: $\nu_n \xrightarrow{d} \nu$ if for every bounded, continuous $f : V \to \mathbb{R}$, we have $E[f(\nu_n)] \to E[f(\nu)]$, where continuity is defined with respect to the uniform metric.

Example 3.5. Consider the ECDF example

$$\nu_n = \sqrt{n}(F_n(\theta) - F(\theta))$$

[diagram] one realization. A candidate f is

$$f(\nu_n(\theta)) = \int_{[0,1]} (|\nu_n(\theta)| \wedge M)\, \mathrm{d}\theta$$

where the constant M and the integration domain $[0,1]$ set the upper bound $f(\nu_n(\theta)) \leq M$, and the integration is with respect to the Lebesgue measure. Another function can be

$$f(\nu_n(\theta)) = \int (\nu_n(\theta) \wedge M)^2 \, \mathrm{d}\Phi(\theta)$$

where $\Phi(\theta)$ is the CDF of $N(0,1)$.

Definition 3.3. A random function $S_n(\theta)$ is **stochastic equicontinuous** if for each pair of $\varepsilon, \eta > 0$, we have $\delta > 0$ such that

$$\limsup_{n\to\infty} P_r\left(\sup_{\theta\in\Theta} \sup_{\theta'\in N_\delta(\theta)} \left\|S_n(\theta) - S_n(\theta')\right\| > \eta\right) \leq \varepsilon$$

where $N_\delta(\theta) = \{\theta' \in \Theta : \|\theta - \theta'\| \leq \delta\}$is a neighbor of $\theta \in \Theta$.

[diagram]
[diagram]
If allows paths with jumps, as long as the probability associated with the jumps are small enough. It's an asymptotic and probabilistic generalization of uniform continuity.

Theorem 3.2 (FCLT). *$\nu_n \xrightarrow{d} \nu$ over $\Theta = U_{j=1}^J \Theta_j$ if and only if*

1. *$(\nu_n(\theta_1), \nu_n(\theta_2), ..., \nu_n(\theta_m)) \xrightarrow{d} (\nu(\theta_1), \nu(\theta_2), ..., \nu(\theta_m))$ for every finite set $\theta_1, \theta_2, ..., \theta_m \in \Theta$*
2. *ν_n is stochastic equicontinuous over each $\Theta_1, \Theta_2, ..., \Theta_J$.*

Stochastic equicontinuity if difficult to verify, a sufficient condition is related to bracket integral

$$J_{[]}(\delta, L_2(Q)) = \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, L_2(Q))}\mathrm{d}\varepsilon$$

Example 3.6. If $N_{[]}(\varepsilon, L_2(Q)) = \varepsilon^{-\rho}$ for some $\rho > 0$,

then $\sqrt{\log N_{[]}(\varepsilon, L_2(Q))} = \rho^{1/2}\left(\log \frac{1}{\varepsilon}\right)^{1/2}$. As $\varepsilon \to 0$, obviously $\log \frac{1}{\varepsilon} \to \infty$. The increase is slow in that $\int_0^1 \sqrt{\log \frac{1}{\varepsilon}}\mathrm{d}\varepsilon = \frac{\sqrt{\pi}}{2}$.

Theorem 3.3. *$\nu_n(\theta)$ is stochastic equicontinuous if $J_{[]}(1, L_2(Q)) < \infty$, or $J_2(1) = \int_0^1 \sqrt{\log N_2(\varepsilon, Q)}\mathrm{d}\varepsilon < \infty$ and $E\left[G^2(x)\right] < \infty$.*

3.5 Donsker's Theorem

We characterize the distribution of the empirical CDF. Let

$$g(X,\theta) = \mathbb{I}(X \leq \theta) = \begin{cases} 1, & with\ \ probability\ F(\theta) \\ 0, & with\ \ probability\ 1 - F(\theta) \end{cases}$$

and the demeaned version is $U_i(\theta) = \mathbb{I}(X \leq \theta) - F(\theta)$. The variance of $U_i(\theta)$ is

$$var(U_i(\theta)) = (1 - F(\theta)) F(\theta)$$

and the covariance is

$$Cov(U_i(\theta_1), U_i(\theta_2)) = E[\mathbb{I}(X \leq \theta_1)\mathbb{I}(X \leq \theta_2)] - F(\theta_1)F(\theta_2) = F(\theta_1 \wedge \theta_2) - F(\theta_1)F(\theta_2).$$

Since $U_i(\theta)$ is i.i.d,

$$\begin{pmatrix} \nu_n(\theta_1) \\ \nu_n(\theta_2) \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} U_i(\theta_1) \\ U_i(\theta_2) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} F(\theta_1)(1 - F(\theta_1)) & F(\theta_1 \wedge \theta_2) - F(\theta_1)F(\theta_2) \\ F(\theta_1 \wedge \theta_2) - F(\theta_1)F(\theta_2) & F(\theta_2)(1 - F(\theta_2)) \end{pmatrix} \right)$$

The joint normal distribution can be extended to for any finite $\theta_1, \theta_2, ..., \theta_m$.

In the special case of $X \sim Uniform(0,1)$, the above limit distribution becomes

$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \theta_1(1 - \theta_1) & \theta_1 \wedge \theta_2 - \theta_1\theta_2 \\ \theta_1 \wedge \theta_2 - \theta_1\theta_2 & \theta_2(1 - \theta_2) \end{pmatrix} \right)$$

This special process is called **Brownian bridge**, denoted as

$$\widetilde{B}(r) = B(r) - rB(1)$$

where $B(r)$ for $r \in [0,1]$ is a Brownian motion on the unit interval.

Zhentao Shi. Jan 31, 2023. Transcribed by Shu Shen.

Chapter 4

M-Estimators

4.1 Motivation

Let the loss function be $\rho_i(\theta) = \rho(z_i, \theta)$, where z_i is a data vector. The sample criterion is an average of $\rho_i(\theta)$:

$$S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\rho_i(\theta)$$

The m-estimator minimizes the sample criterion function:

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} S_n(\theta).$$

The m-estimator includes many examples as special cases. For example, OLS, MLE, NLS, and quantile regressions are all m-estimators.

For simplicity, in this lecture we work with iid data. Let

$$S(\theta) = E[S_n(\theta)] = E[\rho_i(\theta)]$$

be the population criterion function.

Definition 4.1. We say θ is identified if $\theta_0 = \arg\min_{\theta\in\Theta} S(\theta)$ is unique. In other words, for any $\varepsilon > 0$ there exists a $\delta = \delta(\varepsilon)$ such that $\inf_{\theta\in\Theta\setminus N_\delta(\theta_0)} S(\theta) - S(\theta_0) > \varepsilon$.

4.2 Consistency

Theorem 4.1. *If (i) ULLN:* $\sup_{\theta\in\Theta}|S_n(\theta) - S(\theta)| \xrightarrow{p} 0$*; (ii)* θ_0 *is identified, then* $\hat{\theta} \xrightarrow{p} \theta_0$ *as* $n \to \infty$*.*

Proof. We start from the condition of identification.

$$\begin{aligned}
\Pr\left(|\hat{\theta} - \theta| > \delta\right) &\leq \Pr\left(S(\hat{\theta}) - S(\theta_0) > \varepsilon\right) \\
&= \Pr\left(S(\hat{\theta}) - S_n(\hat{\theta}) + S_n(\hat{\theta}) - S_n(\theta_0) + S_n(\theta_0) - S(\theta_0) > \varepsilon\right) \\
&\leq \Pr\left(S(\hat{\theta}) - S_n(\hat{\theta}) + S_n(\theta_0) - S(\theta_0) > \varepsilon\right) \\
&\leq \Pr\left(\left|S_n(\hat{\theta}) - S(\hat{\theta})\right| + |S(\theta_0) - S_n(\theta_0)| > \varepsilon\right) \\
&\leq \Pr\left(\sup_{\theta\in\Theta}|S_n(\theta) - S(\theta)| \geq \frac{\varepsilon}{2}\right) \to 0
\end{aligned}$$

where the second inequality follows from the definition of the m-estimator that $S_n(\hat{\theta}) \leq S_n(\theta_0)$ □

4.3 Asymptotic Normality

We go with a heuristic argument. Define $\bar{\psi}(\theta) = \frac{\partial}{\partial\theta} S_n(\theta)$. Taylor expansion of $\bar{\psi}(\hat{\theta})$ around θ_0 gives

$$0 = \bar{\psi}(\hat{\theta}) = \bar{\psi}(\theta_0) + \frac{\partial^2}{\partial\theta\partial\theta'} S_n(\dot{\theta})(\hat{\theta} - \theta_0)$$

where $\dot{\theta}$ lies in between $\widehat{\theta}$ and θ_0. Rearrange the above inequality,

$$\sqrt{n}(\hat{\theta} - \theta) = -\left[\frac{\partial^2}{\partial\theta\partial\theta'} S_n(\dot{\theta})\right]^{-1} \sqrt{n}\bar{\psi}(\theta_0)$$

Since $\hat{\theta} \xrightarrow{p} \theta_0$, we also have $\dot{\theta} \xrightarrow{p} \theta_0$. By the continuous mapping theorem:

$$\frac{\partial^2}{\partial\theta\partial\theta'} S(\dot{\theta}) \xrightarrow{p} \frac{\partial^2}{\partial\theta\partial\theta'} S(\theta_0) = Q$$

if $\frac{\partial^2}{\partial\theta\partial\theta'} S(\cdot)$ is continuous. In the population, $E[\bar{\psi}(\theta_0)] = E[\psi(\theta_0)] = 0$, and

$$\sqrt{n}\bar{\psi}(\theta_0) \xrightarrow{d} N(0, \Omega)$$

where $\Omega = E[\psi_i(\theta_0)\psi_i'(\theta_0)]$. As a result,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, Q^{-1}\Omega Q^{-1}\right)$$

where the asymptotic variance follows a sandwich form.

Zhentao Shi. Feb 7, 2023. Transcribed by Shu Shen.

Chapter 5

Maximum Likelihood Estimation

5.1 Parametric Model

A parametric model is a complete specification of the distribution. Once the parameter is given, the distribution function is determined. Instead, a semiparametric model only gives a few features rather than a complete description of the distribution.

Example 5.1. Semiparametric model: If we know $X \sim i.i.d.\left(\mu, \sigma^2\right)$, we can estimate μ, σ^2 by method of moments.

Parametric model: If we assume $X \sim N\left(\mu, \sigma^2\right)$, we can estimate μ, σ^2 by MLE.

Example 5.2. Conditional model: the conditioning variable can be viewed as if it is fixed and the randomness comes from the error term only.

$$y = X'\beta + \varepsilon$$

x is the conditional variable. The condition $E\left(\varepsilon|X\right) = 0$ together with a full rank $E\left[XX'\right]$can help to identify β. This is semiparametric model. However, if we assume $f\left(\varepsilon \mid X\right) \sim N\left(0, \sigma^2\right)$, then conditional parametric model as it completely describes $f\left(y \mid X\right)$ and it becomes a conditional parametric model.

Definition 5.1. Parametric model. The distribution of the data $(x_1, ..., x_n)$ is known up to a finite dimensional parameter.

Let Θ be the parameter space a researcher specifies.

Definition 5.2. A model is **correctly specified**, if the true DGP is $f\left(X \mid \theta_0\right)$ for some $\theta_0 \in \Theta$. Otherwise, the model is **misspecified**.

5.2 Likelihood

In this chapter we will mostly talk about unconditional models. The results can be carried over to conditional models. To keep the setting simple, let $(X_1, \ldots, X_n)$ be i.i.d. The **likelihood** of the sample is $\prod_{i=1}^{n} f\left(X_i \mid \theta_0\right)$. The **log-likelihood** is

$$\ell_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n} \log f\left(X_i \mid \theta\right).$$

Here, we put $1/n$ to average the log-likelihood. This scaling factor does not change the estimation at all.

In practice, we work with the log-likelihood, which is more convenient. the MLE estimator is defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell_n(\theta).$$

To justify the likelihood principle, consider the population version of the

$$\ell(\theta) = E[\log f(X \mid \theta)]$$

Theorem 5.1. *When model is correctly specified, θ_0 is the maximizer.*

Proof. Kullback-Leibler distence

$$\begin{aligned} E[\log p(\theta_0)] - E[\log p(\theta)] &= E[\log(p(\theta_0)/p(\theta))] \\ &= -E[\log(p(\theta)/p(\theta_0))] \\ &\geq -\log E[p(\theta)/p(\theta_0)] = 0 \end{aligned}$$

where the inequality holds by Jensen's inequality for the convex function $-\log(\cdot)$. □

5.3 Score, Hessian, and Information

Score:

$$\psi_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i \mid \theta)$$

Hessian:

$$\mathscr{H}_n(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta \partial \theta'} \log f(X_i \mid \theta)$$

Efficient score:

$$\psi_0 = \frac{\partial}{\partial \theta} \log f(X_i \mid \theta_0)$$

Theorem 5.2. *If the model is correctly specified, the support of X does not depend on θ, and θ_0 is in the interior of Θ, then $E(\psi_0) = 0$.*

Proof. By the Leibniz rule,

$$E(\psi_0) = E\left[\frac{\partial}{\partial \theta} \log f(X_i \mid \theta_0)\right] = \frac{\partial}{\partial \theta} E[\log f(X_i \mid \theta_0)] = 0$$

as θ_0 is the maximizer in an interior. □

Definition 5.3. Fisher information matrix:

$$\mathscr{I}_0 = E\left[\psi_0 \psi_0'\right]$$

Definition 5.4. Expected Hessian:

$$\mathscr{H}_0 = -E\left[\frac{\partial^2}{\partial \theta \partial \theta'} \log f(X \mid \theta_0)\right]$$

Theorem 5.3. *If the model is correctly specified, we have the* ***information matrix equality****:* $\mathscr{I}_0 = \mathscr{H}_0$.

Proof. Start with Hessian,

$$\begin{aligned} E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(\theta_0)\right] &= E\left[\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'}\log f(\theta_0)\right] \\ &= E\left[\frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta'}f(\theta)}{f(\theta)}\Big|_{\theta=\theta_0}\right] \\ &= E\left[\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta)}{f(\theta_0)}\right] + E\left[\frac{\frac{\partial}{\partial\theta}f(\theta)\frac{\partial}{\partial\theta'}f(\theta)}{f^2(\theta_0)}\right]. \end{aligned}$$

The first term:

$$E\left[\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta)}{f(\theta_0)}\right] = \int\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta)}{f(\theta_0)}f(\theta_0)\,dx = \int\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta)\,dx = \frac{\partial^2}{\partial\theta\partial\theta'}\int f(\theta)\,dx = \frac{\partial^2}{\partial\theta\partial\theta'}1 = 0.$$

The second term:

$$E\left[\frac{\frac{\partial}{\partial\theta}f(\theta)\frac{\partial}{\partial\theta'}f(\theta)}{f^2(\theta_0)}\right] = E\left[\frac{\partial}{\partial\theta}\log f(\theta_0)\frac{\partial}{\partial\theta'}\log f(\theta_0)\right] = E\left[\psi_0\psi_0'\right].$$

□

Notice that the information matrix equality holds only when the model is correctly specified. It fails when the model is misspecified.

5.4 Cramér-Rao Lower Bound

Theorem 5.4. *Suppose the model is correctly specified, the support of* X *does not depend on* θ*, and* θ_0 *is in the interior of* Θ*. If* $\widetilde{\theta}$ *is unbiased estimator, then var* $\left(\widetilde{\theta}\right) \geq \left(n\mathscr{I}_0\right)^{-1}$.

Proof. Because of unbiasedness,

$$\theta = E_\theta\left[\widetilde{\theta}\right] = \int\widetilde{\theta}f(\mathbf{X}\mid\theta)\,d\mathbf{x}$$

for any $\theta\in\Theta$. $\mathbf{X}$ here is for the entire sample, $f(\mathbf{X}\mid\theta) = f(X_1,...,X_n\mid\theta) = \prod_{i=1}^n f(X_i\mid\theta)$. Take derivative at the two sides. The LHS is

$$\frac{\partial\theta}{\partial\theta'} = \mathbf{I}_p$$

. The RHS:

$$\begin{aligned} \frac{\partial}{\partial\theta'}\int\widetilde{\theta}f(\mathbf{X}\mid\theta)\,d\mathbf{x} &= \int\widetilde{\theta}\frac{\partial}{\partial\theta'}f(\mathbf{X}\mid\theta)\,d\mathbf{x} \\ &= \int\widetilde{\theta}\frac{\frac{\partial}{\partial\theta'}f(\mathbf{X}\mid\theta)}{f(\mathbf{X}\mid\theta)}f(\mathbf{X}\mid\theta)\,d\mathbf{x} \\ &= \int\widetilde{\theta}\frac{\partial}{\partial\theta'}\log f(\mathbf{X}\mid\theta)f(\mathbf{X}\mid\theta)\,d\mathbf{x} \\ &= \int\widetilde{\theta}\psi_n(\theta)f(\mathbf{X}\mid\theta)\,d\mathbf{x} \end{aligned}$$

Evaluate at the true θ_0, and due to i.i.d. data

$$\mathbf{I}_p = \int \widetilde{\theta}\psi_n(\theta_0) f\left(\mathbf{X} \mid \theta_0\right) d\mathbf{x} = E\left[\widetilde{\theta}\psi_n(\theta_0)\right] = E\left[\left(\widetilde{\theta} - \theta_0\right)\psi_n(\theta_0)\right]$$

where the last equality holds by $E\left[\theta_0\psi_n(\theta_0)\right] = \theta_0 E\left[\psi_n(\theta_0)\right] = \theta_0 E\left[n\psi_0\right] = 0$. We thus have

$$var\begin{pmatrix} \widetilde{\theta} - \theta_0 \\ \psi_n(\theta_0) \end{pmatrix} = \begin{bmatrix} \boldsymbol{V} & \mathbf{I}_p \\ \mathbf{I}_p & n\mathscr{I}_0 \end{bmatrix}.$$

Pre- and post-multiply $\begin{bmatrix} \boldsymbol{I}_p & -\left(n\mathscr{I}_0\right)^{-1} \end{bmatrix}$, we have

$$\begin{bmatrix} \boldsymbol{I}_p & -\left(n\mathscr{I}_0\right)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{V} & \boldsymbol{I}_p \\ \boldsymbol{I}_p & n\mathscr{I}_0 \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_p \\ -\left(n\mathscr{I}_0\right)^{-1} \end{bmatrix} = \boldsymbol{V} - \left(n\mathscr{I}_0\right)^{-1} \geq 0.$$

□

The Cramér-Rao Lower Bound is a lower bound. It may not reachable. When it is reached, an estimator is **Cramér-Rao efficient** if it is unbiased and the variance is $\left(n\mathscr{I}_0\right)^{-1}$.

Example 5.3. Normal distribution: Let $\gamma = \sigma^2$

$$\log \ell_n\left(X \mid \mu, \sigma^2\right) = -\frac{n}{2}\log\gamma - \frac{n}{2}\log\pi - \frac{1}{2\gamma}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$$

$$\psi_n\left(\mu, \sigma^2\right) = \begin{cases} \frac{1}{\gamma}\sum_{i=1}^{n}\left(X_i - \mu\right) \\ -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 \end{cases}$$

$$\mathscr{H}_n\left(\mu, \sigma^2\right) = \begin{bmatrix} \frac{n}{\gamma} & \frac{1}{2\gamma^2}\sum_{i=1}^{n}\left(X_i - \mu\right) \\ \frac{1}{2\gamma^2}\sum_{i=1}^{n}\left(X_i - \mu\right) & -\frac{n}{2\gamma^2} + \frac{1}{\gamma^3}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 \end{bmatrix}$$

Expected Hessian:

$$E\left[\mathscr{H}_n\left(\mu, \sigma^2\right)\right] = \begin{bmatrix} \frac{n}{\gamma} & 0 \\ 0 & \frac{n}{2\gamma^2} \end{bmatrix}$$

Take inverse:

$$\begin{bmatrix} \frac{\gamma}{n} & 0 \\ 0 & 2\frac{\gamma^2}{n} \end{bmatrix}$$

This is the lower bound.
Check:
the sample mean:

$$var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n}$$

The sample mean is Cramér-Rao efficient.

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}X'\left(I - \frac{1}{n}1_n 1_n'\right)X$$

$E\left(S_n^2\right) = \sigma^2$ is unbiased

$$(n-1)\frac{s_n^2}{\sigma^2} = \left(\frac{X}{\sigma}\right)'\left(I - \frac{1}{n}1_n 1_n'\right)\left(\frac{X}{\sigma}\right) \sim \chi^2(n-1)$$

So,

$$s_n^2 = \frac{\chi^2(n-1)}{n-1}\sigma^2$$

$$var\left(s_n^2\right) = \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$$

Does not satisfy Cramér-Rao efficient.

5.5 Asymptotic Normality

MLE is a special case of m-estimator. Under regularity conditions, $\hat{\theta} \xrightarrow{p} \theta_0$, and asymptotically normal:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \mathscr{H}_0^{-1}\mathscr{I}_0\mathscr{H}_0^{-1}\right)$$

When the information equality hods, the asymptotic variance is simplified as $\mathscr{I}_0^{-1}\mathscr{I}_0\mathscr{I}_0^{-1} = \mathscr{I}_0^{-1}$, and thus

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \mathscr{I}_0^{-1}\right).$$

In other words, it achieves asymptotic efficiency.

Caveat:

1. need correct specification
2. the comparison is restricted to asymptotically unbiased estimator. There are biased estimators with better overall performance.

5.6 Kullback-Leibler Divergence

$$KLIC(f,g) = \int f(x)\log\frac{f(x)}{g(x)}dx$$

Properties:

1. $KLIC(f,f) = 0$
2. $KLIC(f,g) \geq 0$
3. $f = \arg\min_g KLIC(f,g)$

If $f(x) = f(x \mid \theta)$ is a parametric family

$$\theta_0 = \arg\min_{\theta\in\Theta} KLIC(f, f_\theta)$$

which is correctly specified model.

Pseudo-true parameter:

$$\theta_0 = \arg\min_{\theta\in\Theta} KLIC(f, f_\theta)$$

which is misspecified model.

KLIC is the distance measure of any two distributions.

$$\begin{aligned} KLIC\left(f, f_\theta\right) &= \int f\left(x\right) \log f\left(x\right) dx - \int f\left(x\right) \log f\left(x \mid \theta\right) dx \\ &= \int f\left(x\right) \log f\left(x\right) dx - E\left[\log f\left(x \mid \theta\right)\right] \\ &= \int f\left(x\right) \log f\left(x\right) dx - \ell\left(\theta\right) \end{aligned}$$

the pseudo-true value

$$\theta^* = \arg\max_{\theta \in \Theta} \ell\left(\theta\right)$$

The information equality was proved under correct specification. When the model is misspecified,

$$E\left[S\left(\theta^*\right) S\left(\theta^*\right)'\right] \neq E\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f\left(\theta^*\right)\right].$$

As a result, we will have a sandwich-form asymptotic variance in

$$\sqrt{n}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} N\left(0, \mathscr{H}_*^{-1} \mathscr{I}_* \mathscr{H}_*^{-1}\right)$$

understand that $\mathscr{I}_*$ and $\mathscr{H}_*$ are evaluated at the pseudo-true value.

Zhentao Shi. Feb 14, 2023. Transcribed by Shu Shen.

Chapter 6

Quantile Regression

Quantile regression is an important topic in theoretical econometrics. It is particularly useful if we are interested in the tail of the distribution, instead of the center. For example, in financial risk management, we are concerned about some rare events, rather than everyday routines. This line of research started from Roger Koenker.

6.1 Univariate quantile estimation

Given a sample $(y_1, y_2, ..., y_n)$, we are interested in estimating its τ-th quantile, where $\tau \in (0, 1)$. To find the quantile from the sample, we can look for q such that

$$\frac{1}{n}\sum \mathbb{I}\{y_i \leq q\} \approx \tau. \tag{6.1}$$

If we ignore discreteness on the left-hand side, we solve the equation $\frac{1}{n}\sum \mathbb{I}\{y_i \leq q\} = \tau$. In this chapter, we always work with continuously distributed y. In the population model, q_τ^0 that solves $E[y \leq q] = \tau$ is the population parameter.

6.1.1 Asymptotic Result

(6.1) characterize the estimation by a method of moment. Now we cast the problem into an m-estimation. Introduce the check function

$$\rho(z) = z(\tau - \mathbb{I}(z \leq 0)).$$

Define

$$\psi(z) = \tau - \mathbb{I}(z \leq 0),$$

which be considered as a **subgradient** of $\rho(z)$. Notice $\rho(z)$ is continuous but $\psi(z)$ is discontinuous.

Let

$$S_n(q) = \frac{1}{n}\sum \rho(y_i - q) = \frac{1}{n}\sum (y_i - q)(\tau - \mathbb{I}\{y_i - q \leq 0\}).$$

The first-order condition is

$$\frac{\partial}{\partial q} S_n(q) = \frac{1}{n}\sum (\tau - \mathbb{I}\{y_i - q \leq 0\}) \times (-1) = \frac{1}{n}\sum \mathbb{I}\{y_i \leq q\} - \tau \xrightarrow{p} F(y \leq q) - \tau.$$

The "second derivative" is

$$\begin{aligned}\frac{\partial^2}{\partial q^2} S_n(q) &\xrightarrow{p} \frac{\partial}{\partial q}\{F(y \leq q) - \tau\} \\ &= \frac{\partial}{\partial q}\{F(y - q_\tau \leq q - q_\tau) - \tau\} \\ &= \frac{\partial}{\partial q}\{F(e \leq q - q_\tau) - \tau\} = f_e(q - q_\tau)\end{aligned}$$

where we define $e = y - q_\tau$, and the above heuristic calculation implicitly assumes the exchangeability between $\lim_{\delta \to 0}$ and $\xrightarrow{p}$. (See Chapter 21 of var der Vaart (1998) for a rigorous treatment.)

If true coefficient q_τ^0 is identified, then by ULLN we have

$$\hat{q} \xrightarrow{p} q_\tau^0.$$

Identification is equivalent to $f_y(q_\tau) = f_e(0) > 0$.

Evaluated at the true value $q = q_\tau$, the binary random variable $\mathbb{I}\{y_i \leq q_\tau\} - \tau$ has variance $\tau(1-\tau)$, and $f_e(q - q_\tau)|_{q=q_\tau}$. As a result,

$$\sqrt{n}(\hat{q} - q_\tau^0) \xrightarrow{d} N\left(0, \frac{\tau(1-\tau)}{f_e^2(0)}\right).$$

In the expression of the asymptotic variance, τ is known but the density $f_e^2(0)$ must be estimated based on observed "quantile residual" $\widehat{e}_i = y_i - \hat{q}$. The problem of density estimation is fundamentally a nonparametric estimation.

6.2 Quantile Regression

The above univariate quantile estimation is similar to a regression with intercept only. When other regressors X_i are present, we use $X_i'\beta$ to mimic θ in the quantile estimation:

$$S_n(\beta) = \frac{1}{n}\sum \rho_\tau(y_i - X_i'\beta)$$

The first order condition

$$\begin{aligned}\frac{\partial}{\partial \beta} S_n(\beta) &= -\frac{1}{n}\sum X_i \psi(y_i - X_i'\beta) \\ &= -\frac{1}{n}\sum X_i \psi(y_i - X_i'\beta_\tau + X_i'\beta_\tau - X_i'\beta) \\ &= -\frac{1}{n}\sum X_i \psi(e_i + X_i'\beta_\tau - X_i'\beta) \\ &\xrightarrow{p} -E\left[X\psi(e + X'\beta_\tau - X'\beta)\right] \\ &= -E\left[XE\left[\psi(e_i + X'(\beta_\tau - \beta))\,|X\right]\right] \\ &= -E\left[XE\left[\tau - \mathbb{I}\{e \leq X'(\beta - \beta_\tau)\}\,|X\right]\right] \\ &= E\left[X\left(F_{e|X}(X'(\beta - \beta_\tau)) - \tau\right)\right]\end{aligned}$$

where the fourth equality follows by the law of iterated expectations.

SOC with respect to β in the population version is $E\left[XX'f_{e|X}\left(X'\left(\beta-\beta_\tau\right)\right)\right]$. Evaluate it at $\beta=\beta_\tau$, the Hessian is $E\left[XX'f_{e|X}\left(0\right)\right]$.

Similarly, by ULLN and ID we have consistency

$$\hat{\beta}\xrightarrow{p}\beta_\tau$$

The identification condition is that $Q_\tau=E\left[XX'f_{e|X}\left(0\right)\right]$ must positive definite.

Again evaluated at $\beta=\beta_\tau$, the variance of the score function is $\Omega_\tau=E\left[XX'\psi^2\left(e\right)\right]$. We have asymptotic normality

$$\sqrt{n}\left(\hat{\beta}-\beta_\tau\right)\xrightarrow{d}N\left(0,Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\right)$$

6.2.1 Linear Conditional Quantile

Let $Q_{y|X}\left(\tau\right)$ be the τ-th conditional quantile. If the linear function is correct specified for the τ-th conditional quantile, then

$$\tau=F_{y|X}\left(X'\beta_\tau\right)=E\left[\mathbb{I}\left\{y\leq X'\beta_\tau\right\}\mid X\right]=E\left[\mathbb{I}\left\{e\leq 0\right\}\mid X\right]=F_{e|X}(0).$$

This condition simplifies the expression of the variance of the score function as

$$\Omega_\tau=E\left[XX'E\left[\left(\mathbb{I}\left\{y\leq X'\beta_\tau\right\}-\tau\right)^2\mid X\right]\right]=\tau\left(1-\tau\right)E\left[XX'\right].$$

As a result, the asymptotic variance.

$$\sqrt{n}\left(\hat{\beta}-\beta_\tau\right)\xrightarrow{d}N\left(0,\tau\left(1-\tau\right)Q_\tau^{-1}E\left[XX'\right]Q_\tau^{-1}\right)$$

If we further assume e is statistically independent of X, then the Hessian is simplified as $Q_\tau=E\left[XX'\right]f_e\left(0\right)$, and we end up with

$$\sqrt{n}\left(\hat{\beta}-\beta_\tau\right)\xrightarrow{d}N\left(0,\frac{\tau\left(1-\tau\right)}{f_e^2\left(0\right)}\left(E\left[XX'\right]\right)^{-1}\right).$$

6.3 Summary

The derivations in this chapter are heuristic, but they deliver the essence.

It is helpful to compare quantile regression with our familiar linear regression. The univariate mean model is $y=\mu+\varepsilon$, where $E\left[y\right]=u$, or equivalently $E\left[\varepsilon\right]=0$. The univariate quantile model is $y=q_\tau+e$, where $Q_y(\tau)=q_\tau$, or equivalently $Q_e\left(\tau\right)=0$.

In regression model, the conditional mean $E\left[y|X\right]$ is in general a nonlinear function of X, and we approximate it by the linear function $X'\beta$. Identification is determined by the minimum eigenvalue of $E\left[XX'\right]$. The conditional quantile $Q_{y|X}\left(\tau\right)$ is in general a nonlinear function of X too, while we approximate it with a linear function $X'\beta$ for simplicity. Identification is determined by the minimum eigenvalue of $E\left[XX'f_{e|X}\left(0\right)\right]$.

In regression models, correct specification $E\left[y|X\right]=X'\beta$ or equivalently $E\left[e|X\right]=0$ gives unbiasedness to the OLS estimator, and homoskedasticity simplifies the variance. In quantile regression, correct specification $Q_{y|X}\left(\tau\right)=X'\beta$ provides an explicit form of the variance of the score function, and independence between e and X simplifies the sandwich-form variance into one piece.

Zhentao Shi. Feb 21, 2023. Transcribed by Shu Shen.

Chapter 7

Time Series

7.1 Introduction

A random variable is a $(\Omega, \mathscr{F}) \setminus (\mathbb{R}^m, \mathscr{B})$ measure function. A time series is a sequence of random variables $(y_1(\omega), y_2(\omega), \ldots, y_n(\omega)) \in \mathbb{R}^{m\times n}$, and it can be extended to a doubly infinite sequence $(..., y_{t-1}, y_t, y_{t+1,}...) \in \mathbb{R}^{m\times\infty}$. We consider discrete time series (instead of the continuous time series). For each fixed ω, the sequence is a deterministic vector $(\omega) \in \mathbb{R}^{m\times n}$; for each fixed t, $y_t(\omega)$ is a common random vector in $\mathbb{R}^m$.

7.2 Stationarity

In reality, we have only one realized sequence, but statistics needs repeated observations. We introduce the concept *stationarity* to produce "repeated" observations.

Definition 7.1. (y_t) is **covariance stationarity** or **weakly stationarity** if the mean $\mu = E[y_t]$, covariance $\Sigma = E\left[(y_t - \mu)(y_t - \mu)'\right]$ and autocovariance $\Gamma(\ell) = E\left[(y_t - \mu)(y_{t-\ell} - \mu)'\right]$ are independent of t.

- For a vector-valued weakly stationarity time series, $\Sigma = \Gamma(0)$ is a positive-definite symmetric matrix. The autovariance $\Gamma(\ell), \ell \neq 0$ is not symmetric in general, and

$$\Gamma(-\ell) = E\left[(y_t - \mu)(y_{t+\ell} - \mu)'\right] = E\left[(y_{t-\ell} - \mu)(y_t - \mu)'\right] = \Gamma(\ell)'.$$

- When $m = 1$ (scalar time series), we use $\gamma(0), \gamma(1), ...$, for the autocovariance, and we define *autocorrelation* as $\rho(\ell) = \gamma(\ell)/\gamma(0)$. By the Cauchy-Schwarz inequality $\rho(\ell) \in [-1, 1]$.

Definition 7.2. (y_t) is *strictly stationarity*, if for every $\ell \in \mathbb{Z}^+$, joint distribution of $(y_t, y_{t+1}, \ldots, y_{t+\ell})$ is independent of t.

When mentioning "stationarity", the default is "strictly stationarity".

- If (y_t) is i.i.d, then it is strictly stationarity.
- If (y_t) is strictly stationarity, its transformation $x_t \in \phi(y_t, y_{t-1}, ...) \in \mathbb{R}^q$ is also strictly stationarity. In other words, strictly stationarity is preserved by transformation.

Series: $x_t = \sum_{j=0}^{\infty} a_j y_{t-j}$

- The infinite series x_t is convergent if the partial sum $\sum_{j=1}^{N} a_j y_{t-j}$ has a finite limit as $N \to \infty$ almost surely.
- If y_t is strictly stationarity, $E\|y\| < \infty$ and $\sum_{j=0}^{N} |a_j| < \infty$ (absolutely summable), then x_t is convergent and strictly stationarity.

7.3 Ergodicity

A time series $\{y_t\}$ is *ergodic* if all invariant events are trivial. Any event unaffected by time shift is of probability 0 or 1. "invariant" means the sequence of a random variable gets stuck somewhere. Ergodicty is preserved by transformation. If $\{y_t\}$ is stationarity and erdogic, the same is for $x_t \in \phi(y_t, y_{t-1}, ...)$ (function with infinite terms).

Example 7.1. If $x_t = \sum_{j=0}^{\infty} a_j y_{t-j}$ if convergent and (y_t) is erdogic, then x_t is also erdogic.

(Cesaro means) If $a_j \to a$ as $j \to \infty$, then $\frac{1}{n}\sum_{j=0}^{\infty} a_j \to a$ as $n \to \infty$.

Theorem 7.1. *If $y_t \in \mathbb{R}^m$ is stationarity and erdogic, and var $(y_t) < \infty$, then $\frac{1}{n}\sum_{\ell=1}^{n} cov(y_t, y_{t+\ell}) \to 0$ as $n \to \infty$*

Definition 7.3. Formal definitions

Let $\widetilde{y}_t = (..., y_{t-1}, y_t, y_{t+1,}...)$ an event $A \in \{\widetilde{y}_t \in G\}$ for some $G \subseteq \mathbb{R}^{m\times\infty}$.

The ℓ-th time shift is $\widetilde{y}_{t+\ell} = (..., y_{t-1+\ell}, y_{t+\ell}, y_{t+\ell+1,}...)$ and a time shift of the event is $A_\ell \in \{\widetilde{y}_{t+\ell} \in G\}$.

An event is **invariant** if $A_\ell = A$

An event is **trivial** if $P(A) = 0$ or $P(A) = 1$.

Theorem 7.2. *A stationarity $\{y_t\}$ is erdogic if for all events A and B,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} P(A_\ell \cap B) = P(A)P(B)$$

Let $B = A$, and then we solve $P(A) = [P(A)]^2 \Rightarrow P(A) = 0$ or 1.

A "sufficient" condition for ergodicity is $P(A_\ell \cap B) \to P(A)P(B)$ as $\ell \to \infty$, according to Cesaro means. This sufficient condition is called "mixing".

- Mixing says that separate events (any A and B) are asymptotically independent when A is shifted to A_ℓ as $\ell \to \infty$.
- Ergodicity is slightly weaker than mixing (weak dependence), in the sense that the independence is "on average" in the form of $\frac{1}{n}\sum_{\ell=1}^{n} P(A_\ell \cap B)$.

Theorem 7.3. *Ergodic Theorem:*

$y_t \in \mathbb{R}^m$ is stationarity and erdogic, and $E\|y\| < \infty$, then $E\|\bar{y} - \mu\| \to 0$ and $\bar{y} \xrightarrow{p} \mu$.

Interpretation: Convergence in the 1st mean implies $\xrightarrow{p}$.

7.4 Information Set

- for a univariate time series, definite $E_{t-1}[y_t] = E[y_t \mid y_{t-1}, y_{t-2}, ...]$ as the condition expectation of y_t given the past history $(y_{t-1}, y_{t-2}, ...)$
- More generally, we write $\mathscr{F}_t$ as the smallest σ-field generated by the information up to time t. $\mathscr{F}_t$ is called an "information set".

$$E[y_t \mid \mathscr{F}_{t-1}] = E_{t-1}[y_t]$$

- Information sets are nested $\mathscr{F}_{t-1} \subseteq \mathscr{F}_t \subseteq \mathscr{F}_{t+1}, ...$
- Depends on the definition, when multiple random variables are involved

$$\sigma(y_t, y_{t-1}, ...) \neq \sigma(y_t, x_t, y_{t-1}, x_{t-1}, ...)$$

7.5 Martingale Difference Sequence (MDS)

- Let $\{e_t\}$ be a time series, and $\mathscr{F}_t$ be an information set, $\{e_t\}$ is **adapted** to $\mathscr{F}_t$ if $E[e_t \mid \mathscr{F}_t] = e_t$ ($\mathscr{F}_t$ contain the complete information of e_t. A **natural filtration** is $\mathscr{F}_t = \sigma(e_t, e_{t-1}, ...)$.)
- MDS: a process $\{e_t, \mathscr{F}_t\}$ is MDS if

1. e_t is adapted to $\mathscr{F}_t$
2. $E|e_t| < \infty$
3. $E[e_t \mid \mathscr{F}_{t-1}] = 0$

Interpretation: unforeseeable.

Mean independence. But it does not rule our predictability in other moments.

Example 7.2. $e_t = u_t u_{t-1}$, $u_t \sim i.i.d.\ N(0,1)$
e_t is MDS, but not i.i.d.
The covariance of e_t^2 and e_{t-1}^2 is not 0.
The filtration here is $\mathscr{F}_t = \sigma(u_t, u_{t-1}, ...)$, which subsumes $\sigma(e_t, e_{t-1}, ...)$

$$\begin{aligned} cov(e_t, e_{t-k}) &= E[e_t e_{t-k}] = E[E[e_t e_{t-k} \mid \mathscr{F}_{t-1}]] \\ &= E[e_{t-k} E[e_t \mid \mathscr{F}_{t-1}]] = 0 \end{aligned}$$

- A MDS $(e_t, \mathscr{F}_t)$ is a homoskedastic martingale difference sequence if $E\left[e_t^2 \mid \mathscr{F}_{t-1}\right] = \sigma^2$. $e_t = u_t u_{t-1}$ is MDS, but not homoskedastic.

Theorem 7.4. *CLT for MDS: If $\{u_t\}$ is strictly stationary, ergodic and MDS, then*

$$S_n = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} u_t \xrightarrow{d} N(0, \sum)$$

where $\sum = E[u_t u_t']$. There is the t.s. counterpart of the Lindeberg-Lévy CLT.

7.6 Mixing

we will loose the restriction of MDS. The price are stronger assumptions on the dependence than ergodicity.

- $\alpha(A,B) = |P(AB) - P(A)P(B)|$
- Let two σ-fields be $\mathscr{F}_{-\infty}^{t} = \sigma(...,y_{t-1},y_t)$ and $\mathscr{F}_{t}^{\infty} = \sigma(y_t,y_{t+1},...)$
- Strong mixing coefficients
$$\alpha(\ell) = \sup_{A\in\mathscr{F}_{-\infty}^{t-\ell}, B\in\mathscr{F}_{t}^{\infty}} \alpha(A,B)$$
y_t is strong mixing if $\alpha(\ell)\to 0$ as $\ell\to\infty$.
- In general, the α-coefficients should have a sup over t
$$\alpha(\ell) = \sup_{t}\sup_{A\in\mathscr{F}_{-\infty}^{t-\ell}, B\in\mathscr{F}_{t}^{\infty}} \alpha(A,B)$$
- A mixing process is ergodic.
- Absolute regularity (β-mixing)
$$\beta(\ell) = \sup_{A\in\mathscr{F}_{t}^{\infty}} \left| P\left(A \mid \mathscr{F}_{-\infty}^{t-\ell}\right) - P(A)\right|$$
β mixing is stronger than α mixing.
- Strong mixing is preserved by finite transformation.

Theorem 7.5. *y_t has mxing coefficients $\alpha_y(\ell)$. $x_t = \sigma(y_t, y_{t-1},...,y_{t-q})$*
Then $\alpha_x(\ell) < \alpha_y(\ell - q)$ for $\ell \geq q$.
The α-coefficients satisfy the same rate and summation properties.

- Rate conditions $\alpha(\ell) = O(e^{-r})$. Summation restriction $\sum_{\ell=0}^{\infty}\alpha(e)^r < \infty$ or $\sum_{\ell=0}^{\infty} e^s\alpha(e)^r < \infty$.
- Thm 14.13 bounds covariances with functions of α-coefficients.

7.7 CLT for Correlated Variables

$$\begin{aligned} var\left(S_n\right) &= var\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} y_t\right) \\ &= \frac{1}{n}\mathbf{I}_N' E\left[YY'\right]\mathbf{I}_N \\ &= \frac{1}{n}\mathbf{I}_N' \begin{bmatrix} \sigma^2 & \gamma(1) & \gamma(2) & & \\ \gamma(1) & \sigma^2 & \gamma(1) & & \\ \gamma(2) & \gamma(1) & \sigma^2 & & \\ & & & \ddots & \\ & & & & \sigma^2 \end{bmatrix} \mathbf{I}_N \\ &= \frac{1}{n}\left(n\sigma^2 + 2\left(n-1\right)\gamma(1) + 2\left(n-2\right)\gamma(2) + ... + 2\gamma\left(n-1\right) + 2\times 0\times\gamma(n)\right) \\ &= \sigma^2 + 2\sum_{\ell=1}^{n}\left(1-\frac{\ell}{n}\right)\gamma(\ell) \end{aligned}$$

As $\gamma(-\ell) = \gamma(\ell)$, $var\left(S_n\right) = \sum_{\ell=-n}^{n}\left(1-\frac{|\ell|}{n}\right)\gamma(\ell)$

In vector case, similarly we have

$$var\left(S_n\right) = \Gamma(0) + \sum_{\ell=1}^{n}\left(1-\frac{\ell}{n}\right)\left(\gamma(\ell) + \gamma(\ell)'\right) = \sum_{\ell=-n}^{n}\left(1-\frac{|\ell|}{n}\right)\gamma(\ell)$$

- For CLT to work, $var\left(S_n\right)$ must be convergent in the limit

$$\begin{aligned} \sum_{\ell=1}^{n}\left(1-\frac{\ell}{n}\right)\gamma(\ell) &= \frac{1}{n}\sum_{\ell=1}^{n}\left(n-\ell\right)\gamma(\ell) \\ &= \frac{1}{n}\sum_{\ell=1}^{n-1}\sum_{j=1}^{\ell}\gamma(j) \\ &\to \sum_{j=1}^{\infty}\gamma(j) = \sum_{\ell=1}^{\infty}\gamma(\ell) \end{aligned}$$

  by the Theorem of Cesaro means if $\sum_{\ell=1}^{\infty}\gamma(\ell)$ is convergent.

Necessary condition: $\gamma(\ell)\to 0$ as $\ell\to\infty$.

Sufficient: $\sum_{\ell=1}^{\infty}|\gamma(\ell)| < \infty$

It can be show if $E\left\|u_t\right\|^r < \infty$ and $\sum_{\ell=0}^{\infty}\alpha(\ell)^{1-2/\gamma} < \infty$ for some $\gamma > 2$, then $\sum_{\ell=1}^{\infty}|\Gamma(\gamma)| < \infty$ is absolutely convergent.

Theorem 7.6. *(CLT) If y_t is strictly stationarity with α-mixing coefficients $\sum_{\ell=0}^{\infty}\alpha(\ell)^{1-2/\gamma} < \infty$ and $E\left\|u_t\right\|^r < \infty$ for some $\gamma > 2$, $E\left[u_t\right] = 0$, then $S_n \xrightarrow{d} N(0,\Omega)$ where $\Omega = \sum_{\ell=-\infty}^{\infty}\Gamma(\gamma)$ is the long-run variance.*

7.8 Linear Projection

- In regression problems, $\mathscr{P}\left(y \mid X\right) = X\beta^* = X'\left(E\left[XX'\right]\right)^{-1}E\left[Xy\right]$

- Extend to a projection to the infinite past history $\tilde{y}_{t-1} = (y_{t-1}, y_{t-2}, ...)$

Denote $\mathscr{P}_{t-1}(y_t) = \mathscr{P}[y_t \mid \tilde{y}_{t-1}]$, and the projection error $e_t = y_t - \mathscr{P}_{t-1}(y_t)$

Theorem 7.7. *Projection Theorem:*
If $y_t \in \mathbb{R}$ is covariance stationarity, then the projection error statistics
(1) $E[e_t] = 0$
(2) $\sigma^2 = E[e_t^2] \leq E[y_t^2]$
(3) $E[e_t e_{t-j}] = 0$ for all $j \geq 1$.
*In other words, $\{e_t\}$ is a **white noise**.*

- If $\{y_t\}$ is strictly stationarity, then $\{e_t\}$ is strictly stationarity.

Definition 7.4. A time series is a white noise if it is covariance stationarity with 0 autocovariance.

It is helpful to imagine the projection as a linear combination

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + e_t$$

The nature of projection ensures e_t is uncorrelated with all regressions.
e_{t-j} is a linear combination $y_{t-j} - \alpha_1 y_{t-j-1} - \alpha_2 y_{t-j-2} - ....$
Then e_t is uncorrelated with e_{t-j}.

7.9 Wold Decomposition

- If y_t is covariance stationarity, and the linear projection error has $\sigma^2 > 0$, then $y_t = u_t + \sum_{j=0}^{\infty} b_j e_{t-j}$, $b_0 = 1$, and $u_t = \lim_{m\to\infty} \mathscr{P}_{t-m}(y_t)$

Project y_t onto the orthogonal elements $e_t, e_{t-1}, e_{t-2}, ...$ For simplicity, we can consider the case $\mu_t = \mu$.

Definition 7.5. Lag operator: $\mathrm{L}y_t = y_{t-1}$, $\mathrm{L}^2 y_t = \mathrm{L}(\mathrm{L}y_t) = \mathrm{L}y_{t-1} = y_{t-2}$, and so on.

$$\begin{aligned} y_t &= \mu + \sum_{j=0}^{\infty} b_j e_{t-j} \\ &= \mu + \left(b_0 + b_1 \mathrm{L} + b_2 \mathrm{L}^2 + ...\right) e_t \\ &= \mu + b(\mathrm{L}) e_t \end{aligned}$$

where $b(\mathrm{L})$ is an infinite-order polynomial.

- Autoregressive Wold Representation: If y_t is covariance stationarity with $y_t = u_t + b(\mathrm{L}) e_t$, then with some additional technical restrictions, $y_t = \mu + \sum_{j=1}^{\infty} a_j y_{t-j} + e_j$.

Chapter 8

ARMA Models

8.1 AR(p) Processes

We have learned Wold decomposition in the previous lecture. Let e_t be strictly stationary ergodic white noise. The ARMA are the classical approach to model a univariate time series.

- MA(1)

$$y_t = \mu + e_t + \theta e_{t-1}$$

mean: $E[y_t] = \mu$
variance: $var(y_t) = \theta^2 + 1$
autocovariance: $E[e_t e_{t-1}] = \theta$

- MA(∞)

$$y_t = \mu + \sum_{j=1}^{\infty} b_j e_{t-j}$$

where $b_0 = 1$

- AR(1)

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + e_t$$

mean: $E[y_t] = \frac{\alpha_0}{1-\alpha_1}$
variance: $var(y_t) = \frac{\sigma^2}{1-\alpha_1^2}$
MA(∞) regression: $y_t = \mu + \sum_{j=1}^{\infty} \alpha_1^j e_{t-j}$, where $\mu = \frac{\alpha_0}{1-\alpha_1}$.

To facilitate the notation, we introduce the **lag operator** L. Its effect is to push any time series observation one period to the past. That is, $Lx_t = x_{t-1}$. An AR(1) can be written as

$$(1 - \alpha L) y_t = \alpha_0 + e_t$$
$$y_t = (1 - \alpha L)^{-1} (\alpha_0 + e_t).$$

For stationarity, the AR coefficient $|\alpha| < 1$. If $\alpha = 1$, it becomes a unit root process, which is very different from stationary time series.

- AR(p)

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + +\alpha_p y_{t-p} + e_t$$

lag operator: $(1 - \alpha(L)) y_t = \alpha_0 + e_t$, where $\alpha(z)$ is a polynomial. Stationarity requires that : all roots of $1 - \alpha(z) = 0$ are strictly outside of the unit circle. That is, all the p roots (on the complex plain) must have their modulus strictly greater than 1.

8.2 ARMA and ARIMA Processes

- ARMA: $(1 - \alpha(\mathrm{L}))\, y_t = b(\mathrm{L})\, e_t$
- ARIMA(p,d,q): $(1 - \alpha(\mathrm{L}))\,(1 - \mathrm{L})^d\, y_t = b(\mathrm{L})\, e_t$

8.3 Estimation and Asymptotic Distribution

Estimate AR: take $X_t = (1, y_{t-1}, ..., y_{t-p})$, run OLS:

$$\hat{\alpha} = \left(\frac{X'X}{n}\right)^{-1} \frac{X'y}{n}$$

Theorem 8.1. *If y_t is strictly stationary, ergodic, $E[y_t^2] < \infty$, then $\hat{\alpha} \xrightarrow{p} \alpha$ and $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$*

Asymptotic normality: If e_t is MDS, with $\mathscr{F}$ including X_t, then

$$E[X_t e_t \mid \mathscr{F}_{t-1}] = X_t E[e_t \mid \mathscr{F}_{t-1}] = 0$$

then $\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{p} N(0, Q^{-1} \sum Q^{-1})$,
where $Q = E[X_t X_t']$ and $\sum = E[X_t X_t' e_t^2]$.
Under conditional homoskedasticity $E[e_t^2 \mid \mathscr{F}_{t-1}] = \sigma^2$, then the variance is simplified to

$$\begin{aligned} \sum &= E[X_t X_t' e_t^2] = E[X_t X_t' E[e_t^2 \mid \mathscr{F}_{t-1}]] \\ &= E[X_t X_t'] \sigma^2 = Q\sigma^2 \end{aligned}$$

then $\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{p} N(0, Q^{-1}\sigma^2)$

Without MDS, $z_t = X_t e_t$ can be serially correlated, we need to estimate the long-run variance $\Omega = \sum_{\ell=-\infty}^{\infty} E[X_t X_{t-\ell}' e_t e_{t-\ell}]$

8.4 Model Selection

$$\mathrm{AIC} = \log \hat{\sigma}^2 + 2\frac{p}{n}$$

$$\mathrm{BIC} = \log \hat{\sigma}^2 + \frac{p}{n} \log n$$

8.5 Regression with Time Series Data

Observe $(y_t, X_t)_{t=1}^T$, want to run regression

$$y_t = X_t'\beta + e_t$$

where X_t can include lagged dependent variables.

AR(p)

By the definition of projection, $E[X_t e_t] = 0$

The OLS estimator is $\hat{\beta} = (X'X)^{-1} X'y$

The uncorrelation is necessary for asymptotic normality.

If we impose MDS, $E[e_t \mid \mathscr{F}_{t-1}] = 0$, where $\mathscr{F}_{t-1}$ is adapted to (X_t, e_{t-1}), then we have MDS CLT, because

$$E[X_t e_t \mid \mathscr{F}_{t-1}] = X_t E[e_t \mid \mathscr{F}_{t-1}] = X_t \cdot 0 = 0$$

is also MDS.

Under MDS

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{p} N\left(0, Q_X^{-1} \sum Q_X^{-1}\right)$$

where $\Omega = E\left[X_t X_t' e_t^2\right]$

Under $E[X_t e_t] = 0$, we need conditions about the α-mixing coefficient, then

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{p} N\left(0, Q_X^{-1} \sum Q_X^{-1}\right)$$

where Ω is the long-run variance of $\{X_t e_t\}$.

8.6 Regression with Deterministic Trend

$y_t = T_t + u_t$, where T_t is a deterministic trend and u_t is a random error term.

Example 8.1. $T_t = \beta_0 + \beta_1 t$ (linear trend) or $T_t = \beta_0 + \beta_1 t + +\beta_2 t^2$ (quadratic trend)

Fact:

$$\frac{1}{n^{1+r}} \sum_{t=1}^{n} t^r = \frac{1}{n} \sum_{t=1}^{n} \left(\frac{t}{n}\right)^r \to \int_0^1 x^r dx = \frac{1}{1+r} x^{r+1} \,|_0^1 = \frac{1}{1+r}$$

Thus, $\frac{1}{n^2}\sum_{t=1}^n t = \frac{1}{2}$, $\frac{1}{n^3}\sum_{t=1}^n t^2 = \frac{1}{3}$

OLS estimator

$$\hat{\beta} - \beta = (X'X)^{-1} X'u = \begin{pmatrix} n & \sum_{t=1}^n t \\ \sum_{t=1}^n t & \sum_{t=1}^n t^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum u_t \\ \sum t u_t \end{pmatrix}$$

Let $D_n = \begin{pmatrix} n^{\frac{1}{2}} & 0 \\ 0 & n^{\frac{3}{2}} \end{pmatrix}$

$$\begin{aligned} D_n(\hat{\beta} - \beta) &= D_n \begin{pmatrix} n & \sum_{t=1}^n t \\ \sum_{t=1}^n t & \sum_{t=1}^n t^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum u_t \\ \sum t u_t \end{pmatrix} \\ &= D_n \begin{pmatrix} n & \sum_{t=1}^n t \\ \sum_{t=1}^n t & \sum_{t=1}^n t^2 \end{pmatrix}^{-1} D_n D_n^{-1} \begin{pmatrix} \sum u_t \\ \sum t u_t \end{pmatrix} \\ &= \left(D_n^{-1} \begin{pmatrix} n & \sum_{t=1}^n t \\ \sum_{t=1}^n t & \sum_{t=1}^n t^2 \end{pmatrix} D_n^{-1}\right)^{-1} \begin{pmatrix} \frac{1}{\sqrt{n}} \sum u_t \\ \frac{1}{n^{3/2}} \sum t u_t \end{pmatrix} \\ &= \begin{pmatrix} 1 & \frac{1}{n^2}\sum_{t=1}^n t \\ \frac{1}{n^2}\sum_{t=1}^n t & \frac{1}{n^3}\sum_{t=1}^n t^2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sqrt{n}} \sum u_t \\ \frac{1}{n^{3/2}} \sum t u_t \end{pmatrix} \end{aligned}$$

The denominator

$$\begin{pmatrix} 1 & \frac{1}{n^2}\sum_{t=1}^n t \\ \frac{1}{n^2}\sum_{t=1}^n t & \frac{1}{n^3}\sum_{t=1}^n t^2 \end{pmatrix} \to \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix}$$

The numerator is

$$\begin{pmatrix} \frac{1}{\sqrt{n}} \sum u_t \\ \frac{1}{n^{3/2}} \sum t u_t \end{pmatrix} = \frac{1}{\sqrt{n}} \sum \begin{pmatrix} 1 \\ \frac{t}{n} \end{pmatrix} u_t = \frac{1}{\sqrt{n}} \sum X_t u_t$$

where $X_t = \begin{pmatrix} 1 \\ \frac{t}{n} \end{pmatrix}$.

$$var\left(\frac{1}{\sqrt{n}}\sum X_t u_t\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j' E\left[u_i u_j\right]$$

In the special case when u_i is a white noise,

$$\begin{aligned} var\left(\frac{1}{\sqrt{n}}\sum X_t u_t\right) &= \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j'\right)\sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} 1 & \frac{t}{n} \\ \frac{t}{n} & \frac{t^2}{n^2} \end{pmatrix}\sigma^2 \xrightarrow{d} \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix}\sigma^2. \end{aligned}$$

Zhentao Shi. Mar 21, 2023. Transcribed by Shu Shen.

Chapter 9

Nonstationary Times Series

9.1 Partial Sum Process and Functional Convergence

Let $y_t \in \mathbb{R}^m$ follow a random walk $y_t = y_{t-1} + e_t$, where $(e_t, \mathscr{F}_t)$ is a vector mds. Iterative substitution makes $y_t = y_0 + \sum_{i=1}^t e_i = y_0 + S_t$, where

$$S_t = \sum_{i=1}^t e_i$$

is the *partial sum*. We define the *standardized partial sum* as

$$S_n(r) = \frac{1}{\sqrt{n}} S_{\lfloor nr \rfloor} = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nr \rfloor} e_t$$

for some real number $r \in [0,1]$. For a finite n, $S_n(r)$ is a step function in r.

Recall **Convergence in distribution**: we say $\nu_n(r) \xrightarrow{d} \nu(r)$ if $E[f(\nu_n(r))] \to E[f(\nu(r))]$ for any bounded, continuous function $f : \nu \to \mathbb{R}$, where continuity is defined with respect to the uniform metric $\rho(\nu_1, \nu_2) = \sup_{0 \le r \le 1} \|\nu_1(r) - \nu_2(r)\|$. The definition of convergence in distribution is abstract and difficult to verify. It is easier to verify its equivalent conditions: (i) for any finite $r_1, ..., r_m$, we have $(\nu_n(r_1), ..., \nu_n(r_m)) \xrightarrow{d} (\nu(r_1), ..., \nu(r_m))$; (ii) $\nu_n(r)$ is stochastically equicontinuous.

As $n \to \infty$, asymptotically, the maximal jump size $\frac{1}{\sqrt{n}} \max_{i \le n} \|e_t\| = O_p(1)$, so jumps vanish and S_n is stochastically equicontinuous. Now we verify its finite joint distribution. For $S_n(r)$, we have

1. $S_n(0) = 0$
2. For any r, $S_n(r) \xrightarrow{d} N(0, r\Sigma)$
3. For $r_1 < r_2$, $S_n(r_1)$ and $S_n(r_2) - S_n(r_1)$ are asymptotically independent.

The second point holds as

$$S_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nr \rfloor} e_t = \sqrt{\frac{\lfloor nr \rfloor}{n}} \frac{1}{\sqrt{\lfloor nr \rfloor}} \sum_{t=1}^{\lfloor nr \rfloor} e_t \xrightarrow{d} N(0, r\Sigma).$$

And the third point holds as

$$\begin{pmatrix} S_n(r_1) \\ S_n(r_2) - S_n(r_1) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} r_1\Sigma & 0 \\ 0 & (r_2 - r_1)\Sigma \end{pmatrix}\right).$$

The above joint distribution are written for any two points $r_1, r_2 \in [0,1]$, and it is easy to see that the asymptotic normality can be extended to any $r_1, \ldots, r_m \in [0,1]$ with a finite m.

Notice that the initial value y_0 does not affect the asymptotic behavior, since $\frac{1}{\sqrt{n}} y_{\lfloor nr \rfloor} = S_n(r) + \frac{1}{\sqrt{n}} y_0$ with the initial value $\frac{1}{\sqrt{n}} y_0 = O_p\left(\frac{1}{\sqrt{n}}\right) = o_p(1)$. (For simplicity, we can simply assume $y_0 = 0$.)

Next, we introduce the Brownian motion.

Definition 9.1. A vector **Brownian motion** satisfies (of variance $var(B(1)) = \Sigma$)

1. $B(0) = 0$
2. $B(r) \sim N(0, r\Sigma)$
3. $B(r_1)$ is independent of $B(r_2) - B(r_1)$ for $r_1 < r_2$.

We find the limiting behavior of $S_n(r)$ in any finite coordinates coincides with the Brownian motion, and thus we have the following functional CLT.

Theorem 9.1. *if $(e_t, \mathscr{F}_t)$ is strictly stationary, ergodic mds with $\Sigma < \infty$, then $S_n(r) \xrightarrow{d} B(r)$*

9.2 Beveridge-Nelson Decomposition

So far we discussed mds innovation, which is a special case. In general, we want to allow the the innovations to be serially correlated. Let the innovation be $e_t = \Theta(L) u_t$, where u_t is mds and the polynomial $\Theta(z) = \theta_0 + \theta_1 z + \theta_2 z^2 + \theta_3 z^3 + ...$ Obviously,

$$e_t = \Theta(L) u_t = \Theta(1) u_t + (\Theta(L) - \Theta(1)) u_t.$$

Notice

$$\begin{aligned} \Theta(1) - \Theta(z) &= \theta_0 + \theta_1 + \theta_2 + \theta_3 + ... - \left(\theta_0 + \theta_1 z + \theta_2 z^2 + \theta_3 z^3 + ...\right) \\ &= \theta_1(1 - z) + \theta_2\left(1 - z^2\right) + \theta_3\left(1 - z^3\right) + ... \\ &= (1 - z)\left[\theta_1 + \theta_2(1 + z) + \theta_3\left(1 + z + z^2\right) + ...\right] \\ &= (1 - z)\Theta^*(z) \end{aligned}$$

Replacing the dummy z by L, we write

$$\begin{aligned} e_t &= \Theta(1) u_t + (1 - L)\left[-\Theta^*(L) u_t\right] \\ &= \Theta(1) u_t + (1 - L)\nu_t \\ &= \Theta(1) u_t + \nu_t - \nu_{t-1} \end{aligned}$$

where $\nu_t = -\Theta^*(L) u_t$. As a result,

$$y_t = \sum_{s=1}^{t} e_s + y_0 = \Theta(1) \sum_{s=1}^{t} u_s + \nu_t + (y_0 - \nu_0)$$

where the first term is the permanent component, the second term the transitory component, and the third term in the parenthesis is the initial value.

The MA form of e_t ensures that it is stationary, with long-run variance

$$\begin{aligned} var\left(\frac{1}{\sqrt{n}}\sum_{s=1}^{n} e_s\right) &= var\left(\Theta(1)\frac{1}{\sqrt{n}}\sum_{s=1}^{t} u_s + \frac{v_t}{\sqrt{n}} - \frac{v_0}{\sqrt{n}}\right) \\ &= \Theta(1)\Sigma\Theta'(1) + o(1) \\ &\to \Theta(1)\Sigma\Theta'(1) \end{aligned}$$

where $\Sigma = var(u_s)$. In other word, the effect of the MA representation is multiply with the white noise variance by a factor $\Theta(1)$.

9.3 Functional CLT

Consider the representation

$$y_t = S_t + u_t + (y_0 - v_0).$$

Define $S_n(r) = \frac{1}{\sqrt{n}} S_{\lfloor nr \rfloor}$ and

$$z_n(r) = \frac{1}{\sqrt{n}} y_{\lfloor nr \rfloor} = S_n(r) + \frac{1}{\sqrt{n}} u_{\lfloor nr \rfloor} + \frac{1}{\sqrt{n}}(y_0 - v_0).$$

If u_t is mds, we have

$$z_n(r) = S_n(r) + o_p(1) \xrightarrow{d} B(r)$$

where $B(1) \sim N(0, \Theta(1)\Sigma\Theta'(1))$.

Linear projection ensures the innovations e_t in the Wold decomposition are white noise, but may not necessarily be mds. If u_t is not mds, we impose assumptions on the α-mixing coefficient so that we can still apply FCLT to conclude

$$z_n(r) \xrightarrow{d} B(r)$$

where $B(1) \sim \Omega$ with Ω being the long-run variance of Δy_t.

9.4 Orders of Integration

We say a time series y_t is $I(0)$ if y_t is weakly stationary with positive long-run variance. We say it is $I(d)$ if $\Delta^d y_t \sim I(0)$.

What happens if we "over differentiate" y_t? Suppose $y_t = \Theta(L) u_t$ in MA(∞) representation

$$\Delta y_t = (1 - L)\Theta(L) u_t.$$

Consider $(1 - L)\Theta(L)$ as an entity for the MA(∞) representation, and then the long-run $var(\Delta y_t) = (1 - 1)\Theta(L) var(u_t) = 0$.

9.5 Means

By the continuous mapping theorem, if $z_n(r) \xrightarrow{d} B(r)$, then $f(z_n) \xrightarrow{d} f(B)$ for continuous functional f. Notice $\frac{1}{\sqrt{n}} y_{\lfloor nr \rfloor} = z_n(r)$ is a step function.

$$\frac{1}{\sqrt{n}}\overline{y}_n = \frac{1}{n}\sum_{t=1}^{n}\frac{y_t}{\sqrt{n}} = \frac{1}{n}\sum_{r\in\left\{\frac{1}{n},\frac{2}{n},..,\frac{n-1}{n},1\right\}} z_n(r) = \int_0^1 z_n(r)\,\mathrm{d}r$$

for any finite n. We conclude

$$\frac{1}{\sqrt{n}}\overline{y}_n \xrightarrow{d} \int_0^1 B(r)\,\mathrm{d}r$$

is an average of a Brownian motions over $[0,1]$.

9.6 Regression with intercept and time trend

If we fit a unit root process y_t with a deterministic trend $y_t = \beta_0 + \beta_1 t + error_t$, we can denote the regressor as $X_t = \begin{pmatrix} 1 \\ t \end{pmatrix}$, and thus the OLS estimator is

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X'X)^{-1} X'y.$$

As we have seen before, if we set $D_n = \begin{pmatrix} \sqrt{n} & 0 \\ 0 & n^{\frac{3}{2}} \end{pmatrix}$, then

$$\frac{1}{n} D_n \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = D_n (X'X)^{-1} D_n D_n^{-1} X'y = \left(D_n^{-1}(X'X) D_n^{-1}\right)^{-1} \frac{X'y}{n}.$$

The denominator

$$\begin{pmatrix} \frac{n}{n} & n^{-2}\sum t \\ n^{-2}\sum t & n^{-3}\sum t^2 \end{pmatrix} \to \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix}.$$

The numerator

$$D_n^{-1}\frac{1}{n}\sum_{t=1}^{n} X_t y_t = \frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{n} \end{pmatrix} X_t \frac{y_t}{\sqrt{n}} = \frac{1}{n}\sum_{r\in\left\{\frac{1}{n},\frac{2}{n},..,\frac{n-1}{n},1\right\}} X_n(r) z_n(r) = \int_0^1 X(r) z(r)\,\mathrm{d}r.$$

We conclude

$$\frac{1}{n} D_n \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} n^{-\frac{1}{2}}\hat{\beta}_0 \\ n^{-\frac{1}{2}}\hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix}^{-1} \begin{pmatrix} \int_0^1 B(r)\,\mathrm{d}r \\ \int_0^1 rB(r)\,\mathrm{d}r \end{pmatrix}.$$

The behavior of the OLS estimator is very different from our familiar iid cases. The intercept $\hat{\beta}_0 = O_p\left(\sqrt{n}\right)$ is explosive, whereas $\hat{\beta}_1 = O_p\left(n^{-\frac{1}{2}}\right)$. In particular, the trend coefficient matches the order of the two sides, but the estimated right-end of the trend is $n\hat{\beta}_1 = O_p\left(n^{-\frac{1}{2}}\right)$ is also explosive.

9.7 Demeaning and Detrending

When we witness a trend in a time series, one may attempt to detrend it. Have we investigate the consequence of demean and detrending if the true $\{y_t\}$ is a unit root process.

- demean: $y_t^* = y_t - \overline{y}_n$ is irrelevant of the initial value.

The standardized version

$$Z_n^*(r) = \frac{1}{\sqrt{n}} y_{\lfloor nr \rfloor} - \frac{1}{\sqrt{n}} \overline{y}_n = z_n(r) - \int_0^1 z(r)\,\mathrm{d}r \xrightarrow{d} B(r) - \int_0^1 B(r)\,\mathrm{d}r =: B^*(r)$$

demeaned B-motion

- detrending

$$\begin{aligned} Z_n^{**}(r) &= \frac{1}{\sqrt{n}} y_{\lfloor nr \rfloor} - \frac{1}{\sqrt{n}} X_{\lfloor nr \rfloor} \hat{\beta} \\ &= Z_n(r) - \frac{1}{\sqrt{n}} X'_{\lfloor nr \rfloor} n D_n^{-1} \frac{1}{n} D_n \hat{\beta} \\ &\xrightarrow{d} Z_n(r) - X'(r) \left(\int_0^1 XX' \right)^{-1} \left(\int_0^1 XB \right) =: B^{**}(r) \end{aligned}$$

detrended B-motion

- First difference

if $y_t = \beta_0 + \beta_1 t + z_t$, then $\Delta y_t = \beta_0 + \Delta z_t$
if β_1 is estemated by sample mean, then $\overline{\Delta y_n} = \frac{1}{n}\sum_{t=1}^n \Delta y_t = \frac{y_n - y_\infty}{n}$
And normalization $z_0 = 0$ gives $y_0 = \beta_0$

$$\widetilde{y}_t = y_t - y_0 - \frac{t}{n}(y_n - y_0)$$

this is the residual after (β_0, β_1) are estimated.
Standardization:

$$\widetilde{z}_n(r) = \frac{1}{\sqrt{n}} y_{\lfloor nr \rfloor} - \frac{y_0}{\sqrt{n}} - \frac{\lfloor nr \rfloor}{n} \frac{(y_n - y_0)}{\sqrt{n}} = \frac{1}{\sqrt{n}} y_{\lfloor nr \rfloor} - \frac{\lfloor nr \rfloor}{n} y_n + o_p(1) \xrightarrow{d} B(r) - rB(1) =: V(r)$$

Brownian bridge

9.8 Stochastic Integral

The Riemann-Stieltijes integral (deterministic) in $[0,1]$ is defined as

$$\int_0^1 g(X)\,\mathrm{d}f(X) = \lim_{N\to\infty} \sum_{i=0}^{N-1} g\left(\frac{i}{n}\right)\left(f\left(\frac{i+1}{N}\right) - f\left(\frac{i}{N}\right)\right).$$

The key difference of the **stochastic integral** is that the measure for integration is a random:

$$\int_0^1 X\mathrm{d}z' = \int_0^1 X(r)\,\mathrm{d}z(r)' = \mathrm{plim}_{N\to\infty} \sum_{i=0}^{N-1} X\left(\frac{i}{n}\right)\left(z\left(\frac{i+1}{N}\right) - z\left(\frac{i}{N}\right)\right)$$

This RHS limit is a usually random variable, not a constant.

Consider (X_t, e_t), where e_t is a mds and X_t is non-stationary. If $X_n(r) = D_n^{-1} X_{\lfloor nr \rfloor}$ for some deterministic D_n and $X_n(r) \to X(r)$ then

$$\begin{aligned}\frac{1}{\sqrt{n}} D_n^{-1} \sum_{t=0}^{n-1} X_t e_{t+1}' &= \sum_{t=0}^{n-1} \left(D_n^{-1} X_t\right) \frac{e_{t+1}'}{\sqrt{n}} \\ &= \sum_{t=0}^{n-1} \left(D_n^{-1} X_t\right) \left(S_n\left(\frac{t+1}{N}\right) - S_n\left(\frac{t}{N}\right)\right) = \int_0^1 X_n \mathrm{d}S_n'\end{aligned}$$

Theorem 9.2. *If $(e_t, \mathscr{F}_t)$ is mds, $E(e_t e_t') = \sum < \infty$, $X_t \in \mathscr{F}_t$, and $(X_n(r), S_n(r)) \xrightarrow{d} X(r), B(r)$, then*

$$\int_0^1 X_n \mathrm{d}S_n' \xrightarrow{d} \int_0^1 X_n \mathrm{d}B'$$

Example 9.1. if $X_n(r) = S_n(r)$ and $S_t = \sum_{i=0}^t e_t$, where e_t is mds, then

$$\frac{1}{n} \sum_{t=0}^{n-1} S_t e_{t+1}' = \sum_{t=0}^{n-1} \frac{S_t}{\sqrt{n}} \frac{e_{t+1}'}{\sqrt{n}} \xrightarrow{d} \int B \mathrm{d}B'$$

If e_t is serially correlated, then

$$\frac{1}{n} \sum_{t=0}^{n-1} S_t e_{t+1}' \xrightarrow{d} \int B \mathrm{d}B' + \Lambda$$

where $\Lambda = \sum_{j=1}^{\infty} \left[z_{t-j} z_t'\right]$

proof : use BN-decomposition for $e_t = \zeta_t + u_t - u_{t-1}$

9.9 AR(1) Regression

Let us start with the simplest model, an AR(1) regression with no intercept:

$$y_t = \alpha y_{t-1} + e_t$$

where e_t is a homoskedastic mds. Obviously, the OLS estimator satisfies

$$\hat{\alpha} - \alpha = \left(\sum_{t=0}^{n-1} y_t^2\right)^{-1} \sum_{t=0}^{n-1} y_t e_{t+1}$$

and proper scaling yields

$$n(\hat{\alpha} - \alpha) = \frac{1}{n} \sum_{t=0}^{n-1} y_t e_{t+1} / \frac{1}{n^2} \sum_{t=0}^{n-1} y_t^2.$$

The numerator in the last expression is

$$\begin{aligned}\sum_{t=0}^{n-1} \frac{y_t}{\sqrt{n}} \frac{y_{t+1} - y_t}{\sqrt{n}} &= \sum S_n(r) \left(S_n\left(r + \frac{1}{N}\right) - S(r)\right) \\ &= \int S_n(r) \, \mathrm{d}S_n(r) \xrightarrow{d} \int_0^1 B \mathrm{d}B = \sigma^2 \int W \mathrm{d}W\end{aligned}$$

and the denominator is

$$\frac{1}{n} \sum_{t=0}^{n-1} \left(\frac{y_t}{\sqrt{n}}\right)^2 = \sum_{t=0}^{n-1} \frac{1}{n} S_n^2(r) \xrightarrow{d} \int_0^1 B^2 = \sigma^2 \int_0^1 W^2.$$

Theorem 9.3. *if $(e_t, \mathscr{F}_t)$ is stationary, ergodic mds, then*

$$n(\hat{\alpha}-1) \xrightarrow{d} \int_0^1 W\mathrm{d}W / \int_0^1 W^2$$

This estimator is super-consistent, in the sense that its rate of convergence is n, instead of $\sqrt{n}$ as in the iid case.

The stochastic integral $\int_0^1 W\mathrm{d}W = \frac{1}{2}\left(W^2(1)-1\right)$ is an Ito integral. "-1" is present because $W_n(r)\left[W_n\left(r+\frac{1}{N}\right)-W_n(r)\right]$ is a mds.

Next, we usually use the t-statistic to infer the slope coefficient. Notice that the residual $\hat{e}_t = y_t - \hat{\alpha}y_{t-1}$ gives

$$\hat{\sigma}^2 = \frac{1}{n}\sum \hat{e}_t^2 = \frac{1}{n}\sum \hat{e}_t^2 + o_p(1) \xrightarrow{d} \sigma^2$$

Assume $(e_t, \mathscr{F}_t)$ homoskedastic mds. We have $\hat{var}(\hat{\alpha}) = \frac{\hat{\sigma}^2}{\sum y_t^2}$. The t-statistic is

$$\begin{aligned} t = \frac{\hat{\alpha}-1}{s.e.(\hat{\alpha})} &= \frac{\left(\sum_{t=0}^{n-1} y_t^2\right)^{-1}\sum_{t=0}^{n-1} y_t e_{t+1}}{\hat{\sigma}/\sqrt{\sum y_t^2}} = \frac{\sum_{t=0}^{n-1} y_t e_{t+1}/\hat{\sigma}}{\sqrt{\sum y_t^2}} \\ &\xrightarrow{d} \frac{\sigma\int_0^1 W\mathrm{d}W/\sigma}{\sqrt{\int_0^1 W^2}} = \int_0^1 W\mathrm{d}W / \sqrt{\int_0^1 W^2} \end{aligned}$$

The above calculation is demonstrated by a regression with no intercept. For the regression with an intercept, $y_t = \mu + \alpha y_{t-1} + e_t$, by the Frisch-Waugh-Lovell Theorem the slope coefficient will be numerically equivalent to running OLS with $y_t = \alpha(y_t - \overline{y}_n) + e_t$, and thus

$$n(\hat{\alpha}-1) \xrightarrow{d} \int_0^1 W^*\mathrm{d}W / \int_0^1 W^{*2}$$

where W^* is the demeaned Brownian motion. Similarly, if the regression has both an intercept and a time trend, then

$$n(\hat{\alpha}-1) \xrightarrow{d} \int_0^1 W^{**}\mathrm{d}W / \int_0^1 W^{**2}$$

where W^{**} is the demeaned-and-detrended Brownian motion.

9.10 AR(p) Models with a Unit Root

If the true DGP is $e_t = a(L)\Delta y_t = a(L)(1-L)y_t$, then

$$\begin{aligned} y_t &= a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + e_t \\ &= (a_1, a_2, ..., a_p)\left(y_{t-1}, y_{t-2}, ..., y_{t-p}\right)' + e_t \\ &= (a_1, a_2, ..., a_p) AA^{-1}\left(y_{t-1}, y_{t-2} ..., y_{t-p}\right)' + e_t \\ &= (\rho, \beta_1 ... \beta_{p-1})\left(y_{t-1}, \Delta y_{t-1}, ..., \Delta y_{t-p-1}\right)' + e_t \end{aligned}$$

where

$$A=\begin{bmatrix} 1 & & & \cdots & 0 \\ 1 & -1 & & & \vdots \\ \vdots & \vdots & -1 & & \\ & & & \ddots & \\ 1 & -1 & & \cdots & -1 \end{bmatrix}, \text{ and } A^{-1}=\begin{bmatrix} 1 & & & \cdots & 0 \\ 1 & -1 & & & \vdots \\ & 1 & -1 & & \\ \vdots & & \ddots & \ddots & \\ 0 & \cdots & & 1 & -1 \end{bmatrix}$$

transforms keeps only one level variable y_{t-1} while transforms all other further lagged level variables $(y_{t-2},...,y_{t-p})$ into differenced variables $X_{t-1}=(\Delta y_{t-1},...,\Delta y_{t-p-1})$. If y_t is unit root, we have $a(1)=a_1+...+a_p=1$.

The transformation separates the regressors into two types: one nonstationary variable and the other stationary variables. The OLS estimator of the transformed equation satisfies

$$\begin{pmatrix} n(\hat{\rho}-1) \\ \sqrt{n}(\hat{\beta}-\beta) \end{pmatrix}=\begin{pmatrix} \frac{1}{n^2}\sum_{t=p+1}^{n} y_{t-1}^2 & \frac{1}{n^{3/2}}\sum_{t=p+1}^{n} y_{t-1}X_{t-1}' \\ \frac{1}{n^{3/2}}\sum_{t=p+1}^{n} y_{t-1}X_{t-1}' & \frac{1}{n}\sum_{t=p+1}^{n} X_{t-1}X_{t-1}' \end{pmatrix}\begin{pmatrix} \frac{1}{n}\sum_{t=p+1}^{n} y_{t-1}e_t \\ \frac{1}{\sqrt{n}}\sum_{t=p+1}^{n} X_{t-1}e_t \end{pmatrix}.$$

notice

$$\frac{1}{n^{3/2}}\sum_{t=p+1}^{n} y_{t-1}X_{t-1}'=\frac{1}{n}\sum_{t=p+1}^{n}\frac{y_{t-1}}{\sqrt{n}}X_{t-1}'=\frac{1}{n}\sum_{t=p+1}^{n} S_n(r)X_{t-1}' \xrightarrow{p} 0$$

as $E[X_{t-1}]=0$

- Alternatively, we understand it as $\frac{y_{t-1}}{\sqrt{n}}=\frac{y_{t-p}+y_{t-p+1}+...+y_{t-1}}{\sqrt{n}}$.

The denominator

$$\begin{pmatrix} \frac{1}{n^2}\sum_{t=p+1}^{n} y_{t-1}^2 & \frac{1}{n^{3/2}}\sum_{t=p+1}^{n} y_{t-1}X_{t-1}' \\ \frac{1}{n^{3/2}}\sum_{t=p+1}^{n} y_{t-1}X_{t-1}' & \frac{1}{n}\sum_{t=p+1}^{n} X_{t-1}X_{t-1}' \end{pmatrix}\xrightarrow{d}\begin{pmatrix} \omega^2\int_0^1 W^2(r) & 0 \\ 0 & Q \end{pmatrix}$$

The numerator

$$\begin{pmatrix} \frac{1}{n}\sum_{t=p+1}^{n} y_{t-1}e_t \\ \frac{1}{\sqrt{n}}\sum_{t=p+1}^{n} X_{t-1}e_t \end{pmatrix}\xrightarrow{d}\begin{pmatrix} \omega\sigma\int_0^1 W\mathrm{d}W \\ N(0,\Omega) \end{pmatrix}$$

where ω is the long-run variance of Δy_t, $Q=E\left[X_{t-1}X_{t-1}'\right]$, Ω is the variance of $X_{t-1}e_t$

9.11 Test for Unit Root: ADF Test

When $(e_t,\mathscr{F}_t)$ is mds, we have $\omega=\sigma$. If we are interested in the null hypothesis that y_t is a unit root process, we have the celebrated Dicky-Fuller test.

Theorem 9.4. *Assume $a(L)\Delta y_t=e_t$, where $a(z)$ is p-1 order with $a_1+...+a_p=1$. $(e_t,\mathscr{F}_t)$ is stationary mds with finite constant variance σ^2. Then*

$$\text{ADF}=\frac{\hat{\alpha}-\alpha}{s.e.(\hat{\alpha})}\xrightarrow{d}\frac{\int_0^1 u\mathrm{d}W}{\sqrt{\int_0^1 u^2}},$$

where u depends on the specification of the deterministic part.

9.12 Test for a Unit Root: KPSS Stationarity Test

Kwiatkowski, Phillips, Schmidt, and Shin (1992) is an alternative test for nonstationarity. Its null hypothesis is that y_t is a stationary time series. Consider the model

$$y_t = \mu + S_t + e_t,$$

where $S_t = \sum_{s=1}^{t} u_t$. If $\sigma_u^2 = 0$, then S_t drops out and y_t is stationary as $y_t = \mu + e_t$.

The null hypothesis $H_0 : \sigma_u^2 = 0$ vs. $H_1 : \sigma_u^2 > 0$: we have the KPSS test statistic defined as

$$\text{KPSS} = \frac{1}{n^2\hat{\omega}^2}\sum_{i=1}^{n}\sum_{t=1}^{i}\hat{e}_t^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{t=1}^{i}\frac{\hat{e}_t}{\sqrt{n}\hat{\omega}}\right]^2$$

It is a sample average of the square of the standardized paritial sum $\sum_{t=1}^{\lfloor nr\rfloor}\frac{\hat{e}_t}{\sqrt{n}\hat{\omega}} \xrightarrow{d} W(r) - rW(1) = V(r)$ is a Brownian Bridge.

To see this point, consider the simple case when e_t is mds so $\sigma = \omega$

$$\sum_{t=1}^{\lfloor nr\rfloor}\frac{\hat{e}_t}{\sqrt{n}\sigma} = \sum_{t=1}^{\lfloor nr\rfloor}\frac{t - \frac{1}{n}\sum_{t=1}^{n} e_t}{\sqrt{n}\sigma} = \sum_{t=1}^{\lfloor nr\rfloor}\frac{e_t}{\sqrt{n}\sigma} - \frac{\lfloor nr\rfloor}{n}\sum_{t=1}^{n}\frac{e_t}{\sqrt{n}\sigma} = S_n(r) - rS(1)$$

as $\hat{e}_t = y_t - \overline{y}_n$. Thus KPSS $\xrightarrow{d} \int_0^1 V(r)\,\mathrm{d}r$.

If a trend is added in the form $y_t = \mu + \theta S_t + e_t$, then

$$\text{KPSS} \xrightarrow{d} \int_0^1 V_2(r)\,\mathrm{d}r$$

where $V_2(r)$ is a 2nd-type Brownian bridge.

$$V_2(r) = W(r) - \left(\int_0^r X(S)\,\mathrm{d}S\right)'\left(\int_0^1 XX'\right)^{-1}\int_0^1 X\mathrm{d}W$$

where $X(S) = \begin{pmatrix} 1 \\ S \end{pmatrix}$.

Zhentao Shi. Apr 11, 2023. Transcribed by Shu Shen.