This version: October 19, 2016

1 Asymptotics

Asymptotic theory is concerned about the behavior of statistics when the sample size is arbitrarily large. It is a useful approximation technique to simplify complicated finite-sample analysis.

1.1 Modes of Convergence

Convergence of a deterministic 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 as $z_n \to z$.

In contrast to the convergence of a deterministic sequence, we are interested in the convergence of random variables. Since a random variable is "random", we must define clearly what "convergence" means. Several modes of convergence are often encountered.

- Convergence almost surely*
- Convergence in probability: $\lim_{n\to\infty} P(\omega : |Z_n(\omega) - z| < \varepsilon) = 1$ for any $\varepsilon > 0$.
- Squared-mean convergence: $\lim_{n\to\infty} E\left[(z_n - z)^2\right] = 0$.

Example 1. z_n is a binary random variable: $z_n = \sqrt{n}$ with probability $1/n$, and $z_n = 0$ with probability $1 - 1/n$. Then $z_n \xrightarrow{p} 0$ but $z_n \overset{m.s.}{\nrightarrow} 0$.

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, even with small probability, it may blow away the squared-mean convergence. On the contrary, such irregularity does not undermine convergence in probability.

- Convergence in distribution: $x_n \xrightarrow{d} x$ if $F(x_n) \to F(x)$ for each x on which $F(x)$ is continuous.

Convergence in distribution is about *pointwise* convergence of CDF, not the random variables themselves.

Example 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 $z_n \not\xrightarrow{p} x$.

Cramér-Wold device handles convergence in distribution for random vectors? We say a sequence of K-dimensional random vectors (X_n) converge in distribution to X if we have $\lambda' X_n \xrightarrow{d} \lambda' X$ for any $\lambda \in \mathbb{R}^K$ with $\lambda'\lambda = 1$.

1.2 Law of Large Numbers[1]

(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:

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

as $n \to \infty$. Various versions of LLN work under different assumptions about the distributions and dependence of the random variables.

- Chebyshev LLN: if $(z_1, \ldots, z_n)$ is a sample of i.i.d. observations, $E[z_1] = \mu$, and $\sigma^2 = \operatorname{var}[x_1] < \infty$ exists, then $\frac{1}{n}\sum_{i=1}^{n} z_i - \mu \xrightarrow{p} 0$.

Chebyshev LLN utilizes

- *Chebyshev inequality*: for any random variable x , we have $P(|x| > \varepsilon) \leq E\left[x^2\right]/\varepsilon^2$ for any $\varepsilon > 0$, if $E\left[x^2\right]$ exists.

Chebyshev inequality is a special case of

- *Markov inequality*: $P(|x| > \varepsilon) \leq E[|x|^r]/\varepsilon^r$ for $r \geq 1$ and any $\varepsilon > 0$, if $E[|x|^r]$ exists.

[1] Though the results in this section hold for convergence almost surely, for simplicity we state them in terms of convergence in probability.

It is easy to verify Markov inequality.

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

Consider a partial sum $S_n = \sum_{i=1}^n x_i$, where $\mu_i = E\left[x_i\right]$ and $\sigma_i^2 = \operatorname{var}\left[x_i\right]$. We apply the Chebyshev inequality to the sample mean $\overline{x} - \bar{\mu} = n^{-1}\left(S_n - E\left[S_n\right]\right)$.

$$\begin{aligned} P\left(|\bar{x}-\bar{\mu}|\geq\varepsilon\right) &= P\left(|S_n - E\left[S_n\right]|\geq n\varepsilon\right) \\ &\leq \left(n\varepsilon\right)^{-2} E\left[\sum_{i=1}^n \left(x_i-\mu_i\right)^2\right] \\ &= \left(n\varepsilon\right)^{-2} \operatorname{var}\left(\sum_{i=1}^n x_i\right) \\ &= \left(n\varepsilon\right)^{-2} \left[\sum_{i=1}^n \operatorname{var}\left(x_i\right) + \sum_{i=1}^n \sum_{j\neq i} \operatorname{cov}\left(x_i, x_j\right)\right]. \end{aligned}$$

From the above derivation, convergence in probability holds as long as the right-hand side shrinks to 0 as $n \to \infty$. Actually, the convergence can be maintained under much more general conditions than just under the i.i.d. assumption. The random variables in the sample do not have to be identically distributed, and they do not have to be independent either.

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

- Kolmogorov LLN: if $(z_1, \ldots, z_n)$ is a sample of i.i.d. observations and $E\left[z_1\right] = \mu$ exists, then $\frac{1}{n}\sum_{i=1}^n z_i - \mu \xrightarrow{p} 0$.

Compared to Chebyshev LLN, Kolmogorov LLN only requires the existence of the population mean, but not any higher moment. On the other hand, i.i.d. is essential for Kolmogorov LLN.

1.3 Central Limit Theorem

The central limit theorem (CLT) is a collect of probability results about the convergence in distribution to a normally distributed random variable. The basic form of the CLT is: for a sample $(z_1, \ldots, z_n)$ of *zero-mean* random variables,

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

Various versions of CLT work under different assumptions about the random variables.

Lindeberg-Levy CLT is the simplest CLT.

- If the sample is i.i.d., $E[x_1] = 0$ and $\text{var}\left[x_1^2\right] = \sigma^2 < \infty$, then (1) holds.

Lindeberg-Levy CLT is easy to verify by the characteristic function. For any random variable x, the function $\varphi_x(t) = E[\exp(ixt)]$ is called its *characteristic function*. The characteristic function fully describes a distribution, just like PDF or CDF. For example, the characteristic function of $N\left(\mu, \sigma^2\right)$ is $\exp\left(it\mu - \frac{1}{2}\sigma^2 t^2\right)$.

If $E\left[|x|^k\right] < \infty$ for a positive integer k, then

$$\varphi_X(t) = 1 + itE[X] + \frac{(it)^2}{2} E\left[X^2\right] + \ldots \frac{(it)^k}{k!} E\left[X^k\right] + o\left(t^k\right).$$

Under the assumption of Lindeberg-Levy CLT,

$$\varphi_{X_i/\sqrt{n}}(t) = 1 - \frac{t^2}{2n}\sigma^2 + o\left(\frac{t^2}{n}\right)$$

for all i, and by independence we have

$$\begin{aligned}\varphi_{\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i}(t) = \prod_{i=1}^{n} \varphi_{x_i/\sqrt{n}}(t) &= \left(1 + i \cdot 0 - \frac{t^2}{2n}\sigma^2 + o\left(\frac{t^2}{n}\right)\right)^n \\ &\to \exp\left(-\frac{\sigma^2}{2}t^2\right),\end{aligned}$$

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

- Lindeberg-Feller CLT: i.n.i.d., and *Lindeberg condition*: for any fixed $\varepsilon > 0$,
$$\frac{1}{s_n^2}\sum_{i=1}^{n}\int_{|x_i|>\varepsilon s_n} x_i^2 dPx_i \to 0$$
where $s_n = \left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}$.
- Lyapunov CLT: i.n.i.d, finite $E\left[|x|^3\right]$.

1.4 Tools for Transformations

The original forms of LLN or CLT only deal with sample means. However, most of the econometric estimators of interest are functions of sample means. Therefore, we need tools to handle transformations.

- Small op: $x_n = o_p\left(r_n\right)$ if $x_n/r_n \xrightarrow{p} 0$.
- Big Op: $x_n = O_p\left(r_n\right)$ if for any $\varepsilon > 0$, there exists a $c > 0$ such that $P\left(x_n/r_n > c\right) < \varepsilon$.
- Continuous mapping theorem 1: If $x_n \xrightarrow{p} a$ and $f\left(\cdot\right)$ is continuous at a, then $f\left(x_n\right) \xrightarrow{p} f\left(a\right)$.
- Continuous mapping theorem 2: If $x_n \xrightarrow{d} x$ and $f\left(\cdot\right)$ is continuous almost surely on the support of x, then $f\left(x_n\right) \xrightarrow{d} f\left(x\right)$.
- 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 $a \neq 0$.
- Delta method: if $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0,\Omega\right)$, and $f\left(\cdot\right)$ is continuously differentiable at θ_0, then
$$\sqrt{n}\left(f\left(\widehat{\theta}\right) - f\left(\theta_0\right)\right) \xrightarrow{d} N\left(0, \frac{\partial f}{\partial \theta'}\left(\theta\right)\Omega\left(\frac{\partial f}{\partial \theta}\left(\theta\right)\right)'\right).$$

2 Asymptotic Properties of OLS

We apply large sample theory to study the OLS estimator $\widehat{\beta} = (X'X)^{-1} X'Y$.

2.1 Consistency

We say $\widehat{\beta}$ is *consistent* if $\widehat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$. To verify consistency, we write

$$\widehat{\beta} - \beta = (X'X)^{-1} X'e = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i e_i. \tag{2}$$

The first term

$$\widehat{Q} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \xrightarrow{p} Q = E\left[x_i x_i' \right].$$

and the second term

$$\frac{1}{n} \sum_{i=1}^{n} x_i e_i \xrightarrow{p} 0.$$

No matter whether $(y_i, x_i)_{i=1}^n$ is an i.i.d., i.n.i.d., or dependent sample, as long as the convergence in probability holds for the above two expressions, we have $\widehat{\beta} - \beta \xrightarrow{p} Q^{-1} 0 = 0$ by the continuous mapping theorem. In other words, $\widehat{\beta}$ is a consistent estimator of β.

2.2 Asymptotic Normality

In finite sample, $\widehat{\beta}$ is a random variable. We have shown the distribution of $\widehat{\beta}$ under normality in the previous lecture. Without the restrictive normality assumption, how can we characterize the randomness of the OLS estimator? If we multiply $\sqrt{n}$ on both sides of (2), we have

$$\sqrt{n} \left(\widehat{\beta} - \beta \right) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i e_i.$$

Since $E\left[x_i e_i \right] = 0$, we apply a CLT to obtain

$$n^{-1/2} \sum_{i=1}^{n} x_i e_i \xrightarrow{d} N\left(0, \Sigma \right)$$

where $\Sigma = E\left[x_i x_i' e_i^2\right]$. By the continuous mapping theorem,

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} Q^{-1} \times N(0, \Sigma) \sim N(0, \Omega)$$

where $\Omega = Q^{-1}\Sigma Q^{-1}$ is called the *asymptotic variance*. This is the *asymptotic normality* of the OLS estimator.

Up to now we have derived the asymptotic distribution of $\widehat{\beta}$. However, to make it feasible, we still have to estimator the asymptotic variance Ω. If $\widehat{\Sigma}$ is a consistent estimator of Σ, then $\widehat{\Omega} = \widehat{Q}^{-1}\widehat{\Sigma}\widehat{Q}^{-1}$ is a consistent estimator of Ω. (Of course, there are other ways to estimate the asymptotic variance.) Then a feasible version about the distribution of $\widehat{\beta}$ is

$$\widehat{\Omega}^{-1/2}\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N(0, I_K)$$

2.3 Estimation of the Variance*

To show the finiteness of the variance, $\Sigma = E\left[x_i x_i' e_i^2\right]$. Let $z_i = x_i e_i$, so $\Sigma = E\left[z_i z_i'\right]$. Because of the Cachy-Schwarz inequality,

$$\|\Sigma\|_\infty = \max_{k=1,\dots,K} E\left[z_{ik}^2\right].$$

For each k, $E\left[z_{ik}^2\right] = E\left[z_{ik}^2 e_i^2\right] \leq \left(E\left[z_{ik}^4\right] E\left[e_i^4\right]\right)^{1/2}$.

For the estimation of variance, homoskedastic,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left(e_i + x_i'\left(\widehat{\beta} - \beta\right)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n e_i^2 + \left(\frac{2}{n}\sum_{i=1}^n e_i x_i\right)'\left(\widehat{\beta} - \beta\right) + \frac{1}{n}\sum_{i=1}^n e_i^2 \left(\widehat{\beta} - \beta\right)' x_i x_i' \left(\widehat{\beta} - \beta\right).
\end{aligned}$$

The second term

$$\left(\frac{2}{n}\sum_{i=1}^n e_i x_i\right)'\left(\widehat{\beta} - \beta\right) = o_p(1)\, o_p(1) = o_p(1).$$

The third term

$$\left(\widehat{\beta}-\beta\right)\left(\frac{1}{n}\sum_{i=1}^{n}e_i^2x_ix_i'\right)\left(\widehat{\beta}-\beta\right)=o_p\left(1\right)O_p\left(1\right)o_p\left(1\right)=o_p\left(1\right).$$

As $\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2=\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2+o_p\left(1\right)$ and $\frac{1}{n}\sum_{i=1}^{n}e_i^2=\sigma_e^2+o_p\left(1\right)$, we have $\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2=\sigma_e^2+o_p\left(1\right)$. In other words, $\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2\xrightarrow{p}\sigma_e^2$.

For general heteroskedasticity,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\left(e_i+x_i'\left(\widehat{\beta}-\beta\right)\right)^2\\&=\frac{1}{n}\sum_{i=1}^{n}x_ix_i'e_i^2+\frac{1}{n}\sum_{i=1}^{n}x_ix_ie_ix_i'\left(\widehat{\beta}-\beta\right)+\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\left(\left(\widehat{\beta}-\beta\right)'x_i\right)^2.\end{aligned}$$

The third term is bounded by

$$\begin{aligned}&\operatorname{trace}\left(\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\left(\left(\widehat{\beta}-\beta\right)'x_i\right)^2\right)\\\leq\;&K\max_k\frac{1}{n}\sum_{i=1}^{n}x_{ik}^2\left[\left(\widehat{\beta}-\beta\right)'x_i\right]^2\\\leq\;&K\left\|\widehat{\beta}-\beta\right\|_2^2\max_k\frac{1}{n}\sum_{i=1}^{n}x_{ik}^2\left\|x_i\right\|_2^2\\\leq\;&K\left\|\widehat{\beta}-\beta\right\|_2^2\frac{1}{n}\sum_{i=1}^{n}\left\|x_i\right\|_2^2\left\|x_i\right\|_2^2\\=\;&K\left\|\widehat{\beta}-\beta\right\|_2^2\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K}x_{ik}^2\right)^2\\\leq\;&K\left\|\widehat{\beta}-\beta\right\|_2^2K\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}x_{ik}^4=o_p\left(1\right)O_p\left(1\right)=o_p\left(1\right).\end{aligned}$$

where the third inequality follows by $\left(a_1+\cdots+a_K\right)^2\leq K\left(a_1^2+\cdots+a_K^2\right)$.

The second term is bounded by

$$\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^{n} x_{ik} x_{ik'} e_i x_i' \left(\widehat{\beta} - \beta \right) \right| \\
\leq & \max_k \left| \widehat{\beta}_k - \beta_k \right| K \max_{k,k',k''} \left| \frac{1}{n} \sum_{i=1}^{n} e_i x_{ik} x_{ik'} x_{ik''} \right| \\
\leq & \left\| \widehat{\beta} - \beta \right\|_2 \left(\frac{1}{n} \sum_{i=1}^{n} e_i^4 \right)^{1/4} K \max_{k,k',k''} \left(\frac{1}{n} \sum_{i=1}^{n} \left(x_{ik} x_{ik'} x_{ik''} \right)^{4/3} \right)^{3/4} \\
\leq & \left\| \widehat{\beta} - \beta \right\|_2 K \max_k \left(\frac{1}{n} \sum_{i=1}^{n} x_{ik}^4 \right)^{3/4} = o_p(1) O_p(1)
\end{aligned}$$

where the second and the third inequality hold by the Holder's inequality.