Gregory Kordas
Last update: May 30, 2023

LECTURE 6
Convergence à la Mode

Abstract: We discuss convergence in distribution and in probability, and review some Central Limit Theorems and Weak Laws of Large Numbers.

1. Convergence in Distribution

The first mode of convergence to be introduced is based on the distribution function $F_X(x) = \Pr(X \leq x)$ of a random variable, rather than the random variable itself. Consider two random variables, X and Y. If $|\Pr(X \leq x) - \Pr(Y \leq x)| < \varepsilon$ for all values of x, it might be reasonable to conclude that X and Y are "near" to one another. This is in the sense that probabilities calculated from F_X and F_Y would be almost identical, thought it may be too restrictive to insist that this condition holds for all x. Indeed, as far as the calculation of probabilities is our goal we only need to require that the above approximation holds for all *continuity points* of F_X and F_Y.

Let X be a random variable with distribution function $F(x)$, and let $\{X_n\}$ be a sequence of random variables. If it is true that

$$F_n(x) = \Pr(X_n \leq x) \to \Pr(X \leq x) = F(x) \quad \text{as } n \to \infty$$

for a sufficiently large set of values of x, then F_n *converges weakly* to F, written $F_n \Rightarrow F$. By 'sufficiently large' we mean at all points x save those where F, the limiting d.f., is discontinuous. As a d.f. has at most a countable number of points of discontinuity, a major reason for excluding these points of discontinuity is so that the limiting distribution function may be a step function. This would be the case with a discrete limiting variable, for example.

In the current context, where the X's are specifically random variables, it is also often said that the associated sequence of random variables $\{X_n\}$ *converges in distribution* or *law* to the random variable X and we write $X_n \xrightarrow{d} X$. However, distribution functions need not be defined with reference to a random variable. Thus a sequence of distribution functions can be said to converge weakly without any further qualification, motivating the need for the more general $\Rightarrow$ notation.

In Figure 1 a sequence of distributions and a limit distribution that are *continuous* are illustrated. The idea is that one picks a value x, say x^*, and proceeds to check if the sequence of numbers $F_n(x^*)$ converges to the number $F(x^*)$. If this is true for all values x^*, then $F_n \Rightarrow F$.

The same information may be expressed on a density function graph, if we are willing to make the assumption that the densities of the F_n's and that of F exist. It can be shown that $F_n \Rightarrow F$ if the corresponding sequence of densities $\{f_n(x)\}$ converges to a valid density $f(x)$ except on a set of Lebesgue measure zero, i.e.

$$\{f_n(x)\} \to f(x) \text{ for almost all } x, \quad \text{implies} \quad F_n \Rightarrow F,$$

but the reverse may not be true, since convergence may occur without the existence of the relevant densities.

It is possible that the limit random variable may take on only a single value and, therefore, not be a random variable at all. In such a case $X(\omega) = c$, a constant, for all ω and the limiting random variable is said to be *degenerate*. The corresponding limiting distribution F is a step function with a unit step at c. As an example, let $\{X_n\}$ be a sequence of normal random variables with parameters $\mu_n = \mu$ and $\sigma_n^2 = \sigma^2/n$. Then $X_n \xrightarrow{d} \mu$, a degenerate distribution with a spike at μ.

The following example demonstrates what we mean by the requirement that convergence is only required at continuity points of F.

Example 1. Define a sequence of distribution functions and a limiting distribution function as follows:

$$F_n(x) = \begin{cases} 0 & \text{if } x < 1/n \\ 1 & \text{if } x \geq 1/n \end{cases} \qquad F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

Then $F_n \Rightarrow F$ even thought $F_n(0)$ does not converge to $F(0)$. ■

It will be become clear later just how important it is that random variables be *standardized* correctly when seeking a valid limiting distribution. The following example demonstrates this important point.

Example 2. Let $X_1, ..., X_n$ be independently distributed exponential random variables with distribution function $G(x) = 1 - \exp(-\alpha x)$, for $x \geq 0$. By independence, the distribution $F_n(x)$

of the *maximum* of these random variables, $M_n = \max_i\{X_i\}$, is given by

$$\begin{aligned} F_n(x) &= \Pr(M_n \leq x) = \Pr(\max_i\{X_i\} \leq x) \\ &= \Pr(X_1 < x, X_2 < x, ..., X_n < x) \\ &= \prod_{i=1}^{n} \Pr(X_i < x) = [G(x)]^n. \end{aligned}$$

Consider approximating this distribution as $n \to \infty$. As it stands, $F_n(x) \to 0$ for all x, the reason being that the maximum M_n of these exponential variables gets unboundedly large as n increases, so that the probability of it being less than any fixed value x tends to zero. However, althought $F_n(x) \to 0$ for all x, it cannot be said that F_n converges weakly to the function $F(x) = 0$, for this is not a valid distribution function.

Now consider the standardized random variable $M_n - \alpha^{-1}\log n$. This standardization leads to the modified maximum having a distribution function given by

$$\begin{aligned} \Pr[M_n - \alpha^{-1}\log n \leq x] &= \Pr[M_n \leq x + \alpha^{-1}\log n] = F_n[x + \alpha^{-1}\log n] \\ &= [1 - \exp(-\alpha x - \log n)]^n = [1 - n^{-1}\exp(-\alpha x)]^n \\ &\to \exp(-e^{-\alpha x}) \equiv F(x). \end{aligned}$$

Hence $F_n(x) \Rightarrow F(x)$, the *Gumbel* distribution, with the result $e^{-x} = \lim(1 - x/n)^n$ as $n \to \infty$ having been used in establishing the result. The appropriately standardized maximum, therefore, has a valid limiting distribution, whilst M_n itself had no meaningful limiting properties. ■

Statistics like the *maximum* or the *minimum* are called *order statistics*, and as we saw in the previous example, are asymptotically Gumbel distributed. In the next section we will investigate the asymptotic behavior of *sums* and *averages* of random variables, and show that they behave asymptotically like Normal random variables.

2. The Central Limit Theorem

One of the main applications of convergence in distribution is in situations in which the F_n are nit specified, but nevertheless the limiting distribution can be obtained. As one may easily imagine, this cannot be done in complete generality, but one special case of great significance arises when the sequence X_n has been formed by *summation*, or, which is the same, it is some kind of *average*. Let $X_1, X_2, ...$ be a sequence of random variables and define the sequence of

partial sums by

$$S_n = \sum_{i=1}^{n} X_i.$$

and the sequence of *averages* by

$$\bar{X}_n = \frac{1}{n} S_n.$$

The application of the concept of convergence in distribution to the sequences S_n and/or $\bar{X}_n$ leads to the famous result known as the *Central Limit Theorem* (CLT). We state the result based on $\bar{X}_n$, but the restatement of it in terms of S_n should also be obvious.

THEOREM 1. (The Lindeberg-Levy CLT).
Let $\{X_i, i = 1, ..., n\}$ be a sequence of i.i.d. random variables with finite mean μ and finite variance σ^2. Then

$$\Pr\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) dt = \Phi(x),$$

or equivalently,

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Proof. Since the first two moments μ and σ^2 exist, a Taylor series expansion of the characteristic function of X_1, $\phi_{X_1}(t)$, about $t = 0$ yields

$$\phi_{X_1}(t) = 1 + i\mu t - \frac{1}{2}\sigma^2 t^2 + o(t^2).$$

It follows that the c.f. of $(X_1 - \mu)$ has the expansion

$$\phi_{X_1 - \mu}(t) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2).$$

Then, by independence, $S_n = X_1 + \cdots + X_n$ has characteristic function $\phi_{S_n}(t) = [\phi_{X_1}(t)]^n$, and $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has characteristic function

$$\begin{aligned}
\phi_{\sqrt{n}(\bar{X}_n - \mu)/\sigma}(t) &= \left[\phi_{X_1 - \mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \\
&= \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \\
&\to \exp\left\{-\frac{1}{2}t^2\right\}, \text{ as } n \to \infty.
\end{aligned}$$

This is the characteristic function of a standard normal variable, and the result now follows from the uniqueness of the characteristic function. □

Example 3. (The De Moivre CLT). Abraham De Moivre (1667–1754) was the first mathematician to prove a CLT. He proved that for large n the Binomial may be approximated by a Normal with mean np and variance $np(1-p)$, that is, as $n \to \infty$

$$\binom{n}{x} p^x (1-p)^{n-x} \to \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(x-np)^2}{2np(1-p)}}.$$

That this is so follows immediately from Theorem 1. Given an i.i.d. sequence $\{Y_i, i = 1, ..., n\}$ of Bernoulli random variables with parameter p, the Binomial $X_n = \sum_{i=1}^n Y_i$ will satisfy Theorem 1 with

$$\Pr\left(\frac{X_n - np}{np(1-p)} \leq x\right) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dt = \Phi(x).$$

■

Example 4. (Failure of the CLT). If $\{X_i\}$ is a sequence of i.i.d. standard Cauchy random variables then it can be shown that $\bar{X}_n$ also has a standard Cauchy distribution for *all* n. Since the Cauchy distribution has no moments, Theorem 1 does not apply here. Note that, not only is $\sqrt{n}\bar{X}_n$ not asymptotically normal, but $\sqrt{n}\bar{X}_n$ actually explodes as $n \to \infty$ (it is $\bar{X}_n$ itself, without normalization, that is standard Cauchy). In this extreme case, the sample mean has the exact same distribution as the individual observations, so averaging does not produce any benefit here! ■

Central Limit Theorems are not limited to random variables that are identically distributed. Extending the basic i.i.d. result of Theorem 1 to more general cases is the subject of a vast literature. The following is the most general result for independent sequences of random variables.

THEOREM 2. (The Lindeberg-Feller CLT)
Let $\{X_i, i = 1, ..., n\}$ be a sequence of independent random variables with finite means μ_i and finite variances σ_i^2, and let $\bar{\mu}_n = n^{-1} \sum_i \mu_i$, and $\bar{\sigma}_n^2 = n^{-1} \sum_i \sigma_i^2$. Then

$$\sqrt{n}(\bar{X}_n - \bar{\mu}_n) \to^d N(0, \bar{\sigma}_n^2).$$

<u>if and only if</u> for every $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{n\bar{\sigma}_n^2} \sum_{i=1}^{n} \int_{|x_i - \mu_i| \geq \epsilon \bar{\sigma}_n \sqrt{n}} (x_i - \mu_i)^2 \, dF_{X_i}(x_i) = 0. \tag{2.1}$$

Proof. See Chung (1974). □

Assumption (2.1) is known as the *Lindeberg condition* and the great thing about it is that it is *both sufficient and necessary* for asymptotic normality! It is essentially needed to rule out the possibility that the variability of one of the summands dominates that of the others. In fact the Lindeberg condition implies that $\sigma_i^2/(\sum_{i=1}^n \sigma_i^2) \to 0$ for all $i = 1, ..., n$, so one way of ensuring that the condition hold is to require that

$$\lim_{n\to\infty} \frac{\max_{1\le i\le n} \sigma_i^2}{\sum_{i=1}^n \sigma_i^2} = 0.$$

This condition is often referred to as an *asymptotic negligibility condition*, and is implied by, but does *not* imply, the Lindeberg condition. It says that no single component of $\sum_i X_i$ can contribute more than an infinitesimal amount to its total variation as n increases.

Theorem 2 contains Theorem 1 as a special case: Suppose the sequence $\{X_i\}$ is i.i.d. with mean μ and variance σ^2. Then the Lindeberg condition reduces to

$$\lim_{n\to\infty} \frac{1}{\sigma^2} \int_{|x_1-\mu|\ge\epsilon\sigma\sqrt{n}} (x_1 - \mu)^2 \, dF_{X_1}(x_1) = 0,$$

which always holds since $\{|X_1 - \mu| \ge \epsilon\sigma\sqrt{n}\} \downarrow \emptyset$, the empty set, as $n \to \infty$. Thus, Lindeberg's central limit theorem implies the central limit theorem for i.i.d. variables with finite variances.

Example 5. (Failure of the Asymptotic Negligibility Condition). Let $\sigma_i^2 = \rho^i, 0 < \rho < 1$, so that

$$\sum_i \sigma_i^2 = \rho(1 + \rho + \rho^2 + \cdots + \rho^{n-1}) = \frac{\rho(1-\rho^n)}{1-\rho}.$$

Now

$$\frac{\sigma_i^2}{n\bar{\sigma}_n^2} = \frac{\rho^i(1-\rho)}{\rho(1-\rho^n)}$$

and thus

$$\max_{1\le i\le n} \frac{\rho^i(1-\rho)}{\rho(1-\rho^n)} = \frac{1-\rho}{1-\rho^n}.$$

Hence

$$\lim_{n\to\infty} \max_{i\le n} \frac{\sigma_i^2}{n\bar{\sigma}_n^2} = \lim_{n\to\infty} \frac{1-\rho}{1-\rho^n} = 1 - \rho \neq 0,$$

so the asymptotic negligibility condition is not satisfied. The Lindeberg condition however may or may not be satisfied. ∎

We will also state one more CLT that is often used in applications.

Theorem 3. (The Lyapounov CLT)
Let $\{X_i\}$ be a sequence of independent random variables with finite means $EX_i = \mu_i$, and finite variances σ_i^2, let $\bar{\mu}_n = n^{-1}\sum_i \mu_i$ and $\bar{\sigma}_n^2 = n^{-1}\sum_i \sigma_i^2$, and assume that $E|X_i|^{2+\delta} < \infty$ for some positive δ. Then

$$\sqrt{n}(\bar{X}_n - \bar{\mu}) \to^d N(0, \bar{\sigma}_n^2),$$

if for every $\epsilon > 0$ and all $i = 1, \cdots, n$

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{\bar{\sigma}_n^{2+\delta}} E|X_i - \mu_i|^{2+\delta} = 0. \tag{2.2}$$

Assumption (2.2) is often referred to as the *Lyapounov condition*. The Lyapounov condition implies the Lindeberg one, so Theorem 3 is also a special case of Theorem 2. However, the Lyapounov condition is often easier is verify than the Lindeberg condition, making Theorem 3 a favorite in applications.

3. Convergence in Probability

Convergence in distribution says little about the values of the random variable in the sequence. However, it may be the case that these values are relevant in deciding whether a sequence of random variables gets 'near to' a limit random variable X. This will require us to look at $|X_n(\omega) - X(\omega)|$ for ω in the sample space Ω. This way of thinking takes us back to recalling that a random variable is in reality not a 'variable' at all, but a function (to be accurate a measurable function) of the basic outcomes ω in a sample space Ω. Convergence of a sequence of random variables $X_n(\omega)$ to a random variable $X(\omega)$ can then viewed as akin to the concept of convergence of a sequence of real functions $f_n(x)$ to a real function $f(x)$.

One way of doing this is to fix n and compute $\Pr(\omega : |X_n(\omega) - X(\omega)| < \varepsilon)$, that is, for fixed n, find all those ω for which $|X_n(\omega) - X(\omega)|$ is less than a positive number ε, and then compute the probability measure of these ω's. If this probability tends to zero as $n \to \infty$ we say that X_n converges in probability to X.

Definition 1. *A sequence of random variables X_n converges in probability to a random variable X if for all $\varepsilon > 0$*

$$\lim_{n\to\infty} \Pr(\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon) = 0.$$

We write $X_n \xrightarrow{p} X$.

In terms of our analogy to a sequence of real functions $f_n(x)$, this definition says that $f_n(x)$ converges in measure (we replace 'probability' with 'measure' because $f_n(x)$ are not random variables) to a function $f(x)$ if the set of points for which their absolute distance exceeds $\varepsilon > 0$ has measure zero as $n \to \infty$, i.e. if for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathscr{M}(x : |f_n(x) - f(x)| \geq \varepsilon) = 0.$$

4. The Weak Law of Large Numbers

As we show the CLT provides an important example of convergence in distribution and is one of the most important pillars on which statistical inference is based. Here we will discuss another fundamental result, the *Weak Law of Large Numbers* (WLLN). The starting point is what is known as *Markov's inequality* which states that, provided the relevant moment exists,

$$\Pr(|Z| \geq \lambda) \leq \frac{E(|Z|^r)}{\lambda^r}$$

for any positive λ and r. If we let $Z = X_n - E(X_n)$ which suggests that the random variable in question is a member of a sequence then, putting $r = 2$ and relabelling λ as ε, the above inequality becomes the well-known *Chebyshev inequality*

$$\Pr(|X_n - E(X_n)| \geq \varepsilon) \leq \frac{V(X_n)}{\varepsilon^2}.$$

Since $E(X_n)$ and $V(X_n)$ are just numbers, all that needs to be done to prove that $X_n \xrightarrow{p} c$ is to show that $E(X_n) = c$ and that $V(X_n) \to 0$. In effect a statement concerning convergence of random variables (i.e. functions) has been converted to one about numbers.

Example 6. Let X_i be a sequence of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Define $S_n = \sum_{i=1}^n X_i$ and attempt to apply Chebychev's inequality to S_n. This is unproductive, for $V(S_n)$ approaches infinity. The Lindeberg CLT indicates that $S_n/\sqrt{n}$ converges in distribution to a normal random variable with mean μ and variance σ^2. To obtain convergence in probability to a constant sufficient scaling must be applied to render the limiting distribution degenerate. Indeed applying Chebychev's inequality to S_n/n gives

$$\Pr(|S_n/n - \mu| \geq \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2}$$

and the right-hand side converges to zero as $n \to \infty$. Thus $S_n/n = \bar{X} \xrightarrow{p} \mu$. ∎

We have proven the following theorem.

THEOREM 4. (Weak Law of Large Numbers) *If $\{X_i, i = 1, ..., n\}$ is a sequence of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$, then $\bar{X} \xrightarrow{p} \mu$ as $n \to \infty$.*

More refined results that do not presume the existence of variances can be produced with a more delicate use of inequalities. We state without proof the following theorem that gives a *necessary and sufficient* condition for the WLLN.

THEOREM 5. (Weak Law of Large Numbers II) *Let $\{X_i, i = 1, ..., n\}$ is a sequence of i.i.d. F random variables. In order that there exist constants μ_n such that $S_n/n - \mu_n \xrightarrow{p} 0$, it is necessary and sufficient that*

$$x[1 - F(x) + F(-x)] \to 0 \qquad \text{as } x \to \infty.$$

In this case $\mu_n = \int_{-n}^{n} x dF(x)$ works.

There is also a WLLN for independent but not-identically distributed random variables.

Example 7. Let X_i be a sequence of independently distributed random variables with means μ_i and variances σ_i^2. With $S_n = \sum_{i=1}^{n} X_i$, an application of Chebychev's inequality to S_n/n gives

$$\Pr(|S_n/n - n^{-1}\sum_{i=1}^{n} \mu_i| \geq \varepsilon) \leq \frac{n^{-1}\sum \sigma_i^2/n}{\varepsilon^2} = \frac{\sum \sigma_i^2/n^2}{\varepsilon^2}$$

where $\bar{\mu} = n^{-1}\sum \mu_i$ and $\bar{\sigma}^2 = n^{-1}\sum \sigma_i^2$. The right hand side of this expression cannot be shown to converge to zero without some assumption about the $\{\sigma_i\}$. It would not id $\sigma_i = i$, for example. Two conditions that would be sufficient to ensure that the variance of S_n/n goes to zero are that the maximum variance, $\sigma^2_{\max}$, be bounded, so that $\sum_{i=1}^{n} \sigma_i^2/n \leq n\sigma^2_{\max}/n^2 \to 0$ as $n \to \infty$, or if the average of the variances converged to some number σ_0^2, say. In either of these cases the WLLN will hold. ■

We have proven the following theorem.

THEOREM 6. (Weak Law of Large Numbers III) *If $\{X_i, i = 1, ..., n\}$ is a sequence of independently distributed random variables with means μ_i and variances σ_i^2 such that $\max_i \sigma_i^2 < \infty$, then $\bar{X} \xrightarrow{p} \bar{\mu}$ as $n \to \infty$, where $\bar{\mu} = n^{-1}\sum_{i=1}^{n} \mu_i$.*

Just as the Lindeberg and asymptotic negligibility conditions in the context of the CLT, it is seen that some mechanism to control the relative behavior of the variances is also needed for the WLLN.

An instance of a WLLN for dependent random variables is provided in the next example.

Example 8. Let X_i be i.i.d. normal random variables with mean μ and variance σ^2. Put $W_i = (X_i - \sum_{i=1}^n X_i/n)^2$ and $S_n = \sum_{i=1}^n W_i$. It is not difficult to show that $E(S_n) = (n-1)\sigma^2$ and that $V(S_n) = 2(n-1)\sigma^4$. However, the W_i are neither independent nor uncorrelated because of the common presence of $\sum_{i=1}^n X_i$. Nevertheless, applying Markov's inequality with $Z = n^{-1}S_n - \sigma^2$, $r = 2$ and $\lambda = \varepsilon$ gives

$$\begin{aligned}
\Pr(|S_n/n - \sigma^2| \geq \varepsilon) &\leq \varepsilon^{-2} E(n^{-1}S_n - \sigma^2)^2 \\
&= \varepsilon^{-2}\Big\{V(n^{-1}S_n) + [n^{-1}(n-1)\sigma^2 - \sigma^2]^2\Big\} \\
&= \varepsilon^{-2}\Big\{2n^{-2}(n-1)\sigma^4 + [n^{-1}(n-1)\sigma^2 - \sigma^2]^2\Big\}
\end{aligned}$$

The middle line uses the result that the mean squared error is the variance plus the squared bias. Both terms approach zero as $n \to \infty$ so S_n/n converges in probability to σ^2 and the statistic is a consistent estimator of the variance. ■

In the last example the statistic S_n/n provides a consistent estimator of the σ^2, even though $E(S_n/n) \neq \sigma^2$ for any finite n. It illustrates a general point that sufficient conditions for consistency are that the bias and the variance of the statistic should both go to zero with n. It should be kept in mind, however, that although sufficient these conditions are not necessary.

We close this section with a result that is very useful in applications, the Continuous Mapping Theorem (CMT).

THEOREM 7. (Continuous Mapping Theorem) *If $X_n \xrightarrow{p} c$ as $n \to \infty$ and $g(\cdot)$ is continuous at c, then $g(X_n) \xrightarrow{p} g(c)$ as $n \to \infty$.*

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

5. OLS ASYMPTOTICS

In this section we apply the theory developed so far to prove the consistency and asymptotic normality of the OLS estimator. Consider the linear regression model

$$y_i = x_i\beta + \varepsilon_i, \qquad i = 1, ..., n$$

where y_i is a scalar, x_i is a k vector of exogenous variables, and ε_i is a scalar disturbance. The OLS estimator is give by

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i' \right)^{-1} \sum_{i=1}^{n} x_i y_i$$

We will prove first consistency and then asymptotic normality.

5.1. Consistency

The following conditions are sufficient for consistency.

Assumption 1. *Assume that*

(i) $E(\varepsilon_i) = 0$*;*
(ii) $E(x_i \varepsilon_i) = 0$*;*
(iii) $E(\varepsilon_i^2) = \sigma_\varepsilon^2 < \infty$*;*
(iv) $E(x_i' x_i) < \infty$*;*
(v) $Q = E(x_i x_i')$ *is positive semidefinite;*

Theorem 8. *Under Assumption 1,* $\hat{\beta} \xrightarrow{p} \beta$ *as* $n \to \infty$*.*

Proof. Write

$$\hat{\beta} = \beta + \left(\sum_{i=1}^{n} x_i x_i' \right)^{-1} \sum_{i=1}^{n} x_i \varepsilon_i.$$

Assumption 1 and the WLLN imply that

$$\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \xrightarrow{p} E(x_i x_i') = Q$$

and

$$\frac{1}{n} \sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{p} E(x_i \varepsilon_i) = 0.$$

Then by the CMT,

$$\hat{\beta} \xrightarrow{p} \beta + Q^{-1} 0 = \beta.$$

□

5.2. Asymptotic Normality

To prove asymptotic normality we need extra assumptions.

Assumption 2. *Assume that*

(i) $E(\varepsilon_i^4) < \infty$*;*
(ii) $E(|x_i|^4) < \infty$*.*

THEOREM 9. *Under Assumptions 1 and 2,*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \sigma_\varepsilon^2 E(x_i x_i')^{-1}\right).$$

Proof. Write

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

By Assumption 1 and the WLLN,

$$\frac{1}{n}\sum_{i=1}^n x_i x_i' \xrightarrow{p} E(x_i x_i')$$

while by Assumptions 1 and 2 and the CLT

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i \varepsilon_i \xrightarrow{d} N\left(0, E(x_i x_i' \varepsilon_i^2)\right).$$

Assumption 2 guarantees that $E(x_i x_i' \varepsilon_i^2)$ exists, since by the Cauchy-Schwarz inequality

$$E|x_i x_i' \varepsilon_i^2| \leq \left(E|x_i x_i'|^2\right)^{1/2} \left(E|\varepsilon_i|^4\right)^{1/2} = \left(E|x_i|^4\right)^{1/2} \left(E|\varepsilon_i|^4\right)^{1/2} < \infty.$$

By Assumption 1(ii),

$$E(x_i x_i' \varepsilon_i^2) = E(x_i x_i') E(\varepsilon_i^2) = \sigma_\varepsilon^2 E(x_i x_i'),$$

and now taking everything together and invoking the CMT, we obtain

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} E(x_i x_i')^{-1} N\left(0, \sigma_\varepsilon^2 E(x_i x_i')\right) = N\left(0, \sigma_\varepsilon^2 E(x_i x_i')^{-1}\right).$$

□