Asymptotic Theory

Large-Sample Foundations for OLS and Modern Robust Inference

by

Luis Chancí[1]

1 Introduction

In the OLS notes, we focused on finite-sample theory under the classical assumptions, especially normality. That approach delivered exact results for unbiasedness, the Gauss–Markov theorem, and the classical t and F tests. Those results are foundational, but they are also restrictive. Once we impose a full distributional assumption, we are implicitly specifying a great deal about the data-generating process, including all of its moments. That is often a strong requirement, especially in applied work, where the error distribution may be non-normal, the variance may be non-constant, or the sampling environment may be more complicated than the textbook model. In such settings, exact finite-sample theory becomes harder to use as a practical guide.

Asymptotic theory addresses this problem by asking what happens as the sample size grows. Rather than seeking an exact distribution for a fixed n, we study whether an estimator gets close to the true parameter as more data are observed, and whether a properly normalized version of that estimator approaches a tractable limiting distribution. This is the logic behind *consistency* and *asymptotic normality*. In that sense, asymptotic theory is not a replacement for finite-sample analysis, but an extension of it: when exact results are unavailable or rely on implausibly strong assumptions, large-sample approximations provide the necessary theoretical machinery to guide empirical practice.

[1]Contact me at: luis.chanci@usach.cl, luischanci@santotomas.cl, or lchanci1@binghamton.edu. These notes mirror my natural thought process for teaching this material, blending the accessible style of Wooldridge or Angrist with the formal rigor of Hansen. In the spirit of transparency, I used AI help (like Gemini) to check spelling and grammar. This version: 2026.

These notes are built around two core ideas. First, sample averages stabilize around their population counterparts. This is the content of the Law of Large Numbers (LLN). Second, properly centered and scaled averages become approximately normal. This is the content of the Central Limit Theorem (CLT). Most large-sample arguments in econometrics are built by combining these two ingredients with a small set of additional tools, especially the Continuous Mapping Theorem, Slutsky's Theorem, and the Delta Method.

Our main practical goal is to establish that

$$\sqrt{n}(\hat{\beta} - \beta)$$

is asymptotically normal. Once that result is available, we can construct confidence intervals and test statistics under conditions that are substantially weaker than those required for exact finite-sample inference. This is also the point at which modern robust inference enters naturally: heteroskedasticity-robust, HAC, and cluster-robust procedures are all fundamentally large-sample objects, and their logic is easiest to understand through the asymptotic theory developed here.

The exposition proceeds in two steps. We begin with the probabilistic tools, emphasizing both formal definitions and intuition, and then apply that machinery to the OLS model. The structure of these notes is as follows. Section 2 introduces the main modes of stochastic convergence together with the o_p and O_p notation. Sections 3 and 4 present the LLN and CLT, first in scalar form and then in the multivariate versions used most often in econometrics. Section 5 develops three auxiliary tools that connect these probabilistic limits to econometric estimators. Section 6 applies the general machinery to OLS, establishing consistency, asymptotic normality, and the large-sample role of homoskedasticity. Section 7 develops feasible variance estimation and robust standard errors.

2 Modes of Stochastic Convergence

Before proving any asymptotic result, we need a precise language for what it means for a random sequence to "converge." Unlike a deterministic sequence, an estimator is random for every sample size. As a result, convergence must be defined in probabilistic terms. Econometrics relies mainly on four modes of stochastic convergence. They are related, but they do not say exactly the same thing.

2.1 Convergence in Probability

Definition 2.1 (Convergence in Probability and Probability Limit)**.** A sequence of random variables $\{Z_n\}$ converges in probability to a constant c if

$$\lim_{n\to\infty} P\big(|Z_n - c| > \varepsilon\big) = 0 \qquad \text{for every } \varepsilon > 0.$$

In that case, we also say that c is the probability limit of Z_n, written

$$Z_n \xrightarrow{p} c \qquad \text{or} \qquad \operatorname{plim}_{n\to\infty} Z_n = c.$$

This is the central notion behind consistency. Intuitively, it says that as the sample grows, the probability that Z_n lies far from c becomes negligible. For example, when we say an estimator is consistent, we mean exactly that it converges in probability to the true parameter.

Remark 1. Convergence in probability does not say that Z_n eventually equals c or even that every sample path settles down. It says only that large deviations become increasingly unlikely.

2.2 Almost Sure Convergence

Definition 2.2 (Almost Sure Convergence)**.** A sequence $\{Z_n\}$ converges almost surely to c, written $Z_n \xrightarrow{a.s.} c$, if

$$P\Big(\lim_{n\to\infty} Z_n = c\Big) = 1.$$

Almost sure convergence is stronger than convergence in probability. It means that, with probability one, the realized path of the sequence eventually stays arbitrarily close to c. For many econometric purposes this distinction is not central, but it is useful to know that some LLN results can be proved in this stronger sense.

2.3 Mean-Square Convergence

Definition 2.3 (Mean-Square Convergence)**.** A sequence $\{Z_n\}$ converges in mean square to c, written $Z_n \xrightarrow{m.s.} c$, if

$$\lim_{n\to\infty} \mathbb{E}\Big[(Z_n - c)^2\Big] = 0.$$

Mean-square convergence is often easy to verify when variances are available. For instance, if $\bar{X}_n$ is a sample mean with variance σ^2/n, then

$$\mathbb{E}\Big[(\bar{X}_n - \mu)^2\Big] = \frac{\sigma^2}{n} \to 0,$$

so $\bar{X}_n$ converges to μ in mean square. Since mean-square convergence implies convergence in probability, this is often a convenient route to consistency.

2.4 Convergence in Distribution

Definition 2.4 (Convergence in Distribution)**.** A sequence $\{Z_n\}$ converges in distribution to a random variable Z, written $Z_n \xrightarrow{d} Z$, if for every continuity point z of the cumulative distribution function F_Z,

$$\lim_{n\to\infty} F_{Z_n}(z) = F_Z(z).$$

Convergence in distribution is weaker than convergence in probability. It does not require the realizations of Z_n to get close to those of Z. Instead, it says that the distribution of Z_n approaches the distribution of Z. The CLT is the main example: standardized sample averages do not converge to a constant, but their distributions converge to the normal law.

2.5 Relationships Across Modes

The main implications among these notions are summarized in Figure 1. They are worth remembering because many asymptotic proofs use them implicitly.

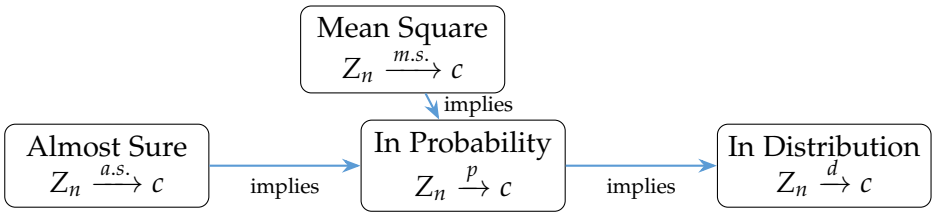

Figure 1: Main implications among common modes of stochastic convergence.

The reverse implications do not hold in general. One useful special case, however, is that if the limiting random variable is actually a constant c, then convergence in distribution to c is equivalent to convergence in probability to c.

2.6 Stochastic Order Notation: Little o_p and Big O_p

Asymptotic arguments often become cleaner when we summarize rates of convergence with stochastic order notation.

Definition 2.5 (Little o_p and Big O_p)**.**

(i) $Z_n = o_p(1)$ if $Z_n \xrightarrow{p} 0$.

(ii) $Z_n = O_p(1)$ if $\{Z_n\}$ is bounded in probability, that is, for every $\varepsilon > 0$ there exists $M < \infty$ such that

$$\sup_{n \text{ large}} P(|Z_n| > M) < \varepsilon.$$

More generally, $Z_n = o_p(r_n)$ means $Z_n/r_n = o_p(1)$, and $Z_n = O_p(r_n)$ means $Z_n/r_n = O_p(1)$.

A useful example is

$$(\hat{\beta} - \beta) = O_p(n^{-1/2}),$$

which means that the estimation error is of order $n^{-1/2}$ in probability. Equivalently,

$$\sqrt{n}(\hat{\beta} - \beta) = O_p(1).$$

This is the natural scaling, by $\sqrt{n}$, under which asymptotic normality is typically established.

3 Laws of Large Numbers (LLN)

The Law of Large Numbers (LLN) formalizes a simple but powerful idea: sample averages stabilize around their population means. In econometrics, this is what allows sample moments such as

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} X_i u_i$$

to converge to their population counterparts as n grows. Without this step, consistency arguments for OLS and related estimators would not go through. In what follows, we present several versions of the LLN.

3.1 A First Version: Chebyshev's Inequality

The quickest route to a weak LLN uses Chebyshev's inequality.

Proposition 3.1 (Chebyshev's Inequality)**.** *If a random variable Z has mean μ and finite variance σ^2, then for every $\varepsilon > 0$,*

$$P(|Z - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Theorem 3.2 (Weak LLN: Chebyshev Version)**.** *Let* $\{X_i\}$ *be i.i.d. with* $\mathbb{E}[X_i] = \mu$ *and* $\text{Var}(X_i) = \sigma^2 < \infty$*. Then*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$

Proof sketch. Since $\mathbb{E}[\bar{X}_n] = \mu$ and $\text{Var}(\bar{X}_n) = \sigma^2/n$, Chebyshev's inequality gives

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0.$$

□

This proof also shows why finite variance is enough: the dispersion of the sample mean shrinks at rate $1/n$.

3.2 A More General Version: Khinchine

Finite variance is convenient, but it is not essential for a weak LLN.

Theorem 3.3 (Weak LLN: Khinchine Version)**.** *Let* $\{X_i\}$ *be i.i.d. with* $\mathbb{E}|X_i| < \infty$ *and mean* μ*. Then*

$$\bar{X}_n \xrightarrow{p} \mu.$$

This version is more general because it allows heavy-tailed distributions with finite first moments but infinite variances. In practice, most econometric applications use finite second moments anyway, since asymptotic normality usually requires them later on.

3.3 Multivariate LLN

Since econometric estimators are built from vectors and matrices, we need LLN results in multivariate form.

Theorem 3.4 (Multivariate WLLN)**.** *Let* $\{Z_i\}$ *be i.i.d. random vectors with* $\mathbb{E}\|Z_i\| < \infty$ *and mean* μ*. Then*

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{p} \mu.$$

A direct implication is that, when $\mathbb{E}\|X_i\|^2 < \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{p} \mathbb{E}[X_i X_i'] \equiv \mathbf{Q}_{XX}.$$

This is one of the key ingredients in proving OLS consistency.

3.4 Strong LLN

Theorem 3.5 (Strong LLN). *Let* $\{X_i\}$ *be i.i.d. with* $\mathbb{E}|X_i| < \infty$. *Then*

$$\bar{X}_n \xrightarrow{a.s.} \mu.$$

The strong LLN gives almost sure convergence rather than convergence in probability. In many econometric proofs this stronger version is not essential, but it is useful to know that sample averages often converge in an even stronger sense than the one required for consistency.

4 Central Limit Theorems

While the LLN explains why estimators converge to population quantities, the CLT explains why their sampling fluctuations become approximately normal. This is the main bridge from consistency to inference. Once we understand the normal approximation for sample averages, we can transport it to estimators such as OLS.

4.1 The i.i.d. CLT

Theorem 4.1 (Lindeberg–Lévy CLT). *Let* $\{X_i\}$ *be i.i.d. with* $\mathbb{E}[X_i] = \mu$ *and* $\mathrm{Var}(X_i) = \sigma^2 \in (0, \infty)$. *Then*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1).$$

Equivalently,

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

The normalization by $\sqrt{n}$ is not arbitrary. Since variances shrink at rate $1/n$, standard deviations shrink at rate $1/\sqrt{n}$. That is why large-sample approximations almost always scale estimators by $\sqrt{n}$.

4.2 Beyond Identical Distributions

The i.i.d. CLT is often enough for introductory econometrics, but it is useful to know that more general results are available when observations are independent but not identically distributed.

Theorem 4.2 (Lyapunov CLT). *Let $\{X_i\}$ be independent with means μ_i, variances σ_i^2, and*

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2.$$

If for some $\delta > 0$,

$$\frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left[|X_i - \mu_i|^{2+\delta}\right] \to 0,$$

then

$$\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0,1).$$

Here s_n denotes the standard deviation of the *sum* $\sum_{i=1}^{n}(X_i - \mu_i)$, not the sample standard deviation of the individual observations. The Lyapunov condition essentially prevents a single observation from dominating the entire sum.

Theorem 4.3 (Lindeberg–Feller CLT). *Under the same independent setup, if for every $\varepsilon > 0$,*

$$\frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\Big[(X_i - \mu_i)^2 \mathbf{1}\{|X_i - \mu_i| > \varepsilon s_n\}\Big] \to 0,$$

then

$$\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{d} \mathcal{N}(0,1).$$

This result is more general than Lyapunov's theorem. For our purposes, however, the main lesson is that asymptotic normality does not depend on identical distributions alone; it depends on whether the individual terms remain well behaved relative to the size of the whole sum.

4.3 Multivariate CLT

Econometrics needs a vector version because the score terms that drive OLS and GMM are multivariate.

Theorem 4.4 (Multivariate CLT). *Let $\{Z_i\}$ be i.i.d. random vectors with $\mathbb{E}[Z_i] = 0$ and finite covariance matrix $\Sigma = \mathbb{E}[Z_i Z_i']$. Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

In the OLS application, we will take $Z_i = X_i u_i$. Then the limiting covariance matrix becomes

$$S = \mathbb{E}[X_i X_i' u_i^2],$$

which will appear as the middle component of the asymptotic sandwich variance formula.

5 Auxiliary Tools

The LLN and CLT provide the raw probabilistic limits, but we still need other tools that allow us to manipulate those limits through transformations, sums, products, and smooth nonlinear maps. Three results play this role repeatedly in econometrics.

5.1 Continuous Mapping Theorem

Theorem 5.1 (Continuous Mapping Theorem)**.** *Let g be continuous at the relevant limit point.*

(i) If $Z_n \xrightarrow{p} c$, then $g(Z_n) \xrightarrow{p} g(c)$.

(ii) If $Z_n \xrightarrow{d} Z$, then $g(Z_n) \xrightarrow{d} g(Z)$.

This theorem justifies many steps that otherwise look informal. For example, once we know that

$$\frac{1}{n} X'X \xrightarrow{p} \mathbf{Q}_{XX},$$

we can conclude that

$$\left(\frac{1}{n} X'X\right)^{-1} \xrightarrow{p} \mathbf{Q}_{XX}^{-1},$$

provided $\mathbf{Q}_{XX}$ is positive definite. In asymptotic proofs, that one step is often decisive.

5.2 Slutsky's Theorem

Theorem 5.2 (Slutsky's Theorem)**.** *If $Z_n \xrightarrow{d} Z$ and $A_n \xrightarrow{p} a$, where a is a constant, then*

$$A_n Z_n \xrightarrow{d} aZ \qquad \text{and} \qquad A_n + Z_n \xrightarrow{d} a + Z.$$

Slutsky's Theorem is the reason consistent nuisance estimators can be substituted into asymptotic distributions without changing the limit. This will matter later when

we replace unknown population variance terms by feasible sample analogs inside asymptotic t and Wald statistics.

5.3 Delta Method

Theorem 5.3 (Delta Method)**.** *Suppose*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, V),$$

and let g be continuously differentiable at θ_0 with Jacobian

$$G = \left.\frac{\partial g(\theta)}{\partial \theta'}\right|_{\theta=\theta_0}.$$

Then

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta_0)\big) \xrightarrow{d} \mathcal{N}(0, GVG').$$

The logic is a first-order Taylor expansion. Near θ_0, the nonlinear function $g(\hat{\theta}_n)$ behaves approximately like a linear function of $\hat{\theta}_n$. Once that approximation is in place, the asymptotic normality of $\hat{\theta}_n$ transfers to $g(\hat{\theta}_n)$.

Example 1 (A Ratio and the Delta Method). Suppose $\hat{\mu}_x$ and $\hat{\mu}_y$ are sample means with positive limits μ_x and μ_y. If we care about the ratio $g(\mu_x, \mu_y) = \mu_y/\mu_x$, then the Delta Method lets us derive the asymptotic distribution of

$$\sqrt{n}\left(\frac{\hat{\mu}_y}{\hat{\mu}_x} - \frac{\mu_y}{\mu_x}\right)$$

by linearizing the ratio around (μ_x, μ_y). This is much easier than trying to derive its distribution directly.

6 Asymptotic Theory of OLS

We now bring together all the tools developed above and apply them to the OLS estimator. The strategy is similar in spirit to the finite-sample chapter, but the logic is different. There, we imposed strong assumptions, especially normality, to obtain exact results. Here, we allow weaker assumptions and derive approximate results that become accurate as n grows. This is attractive in econometrics because, as applied

researchers, we usually prefer to rely on assumptions that are less restrictive and therefore more plausible in practice.

6.1 Setup and Large-Sample Assumptions

Let us consider the linear model developed in the previous set of class notes (OLS).

$$y_i = X_i'\beta + u_i, \qquad i = 1, \dots, n,$$

where X_i is a $k \times 1$ regressor vector, $\beta \in \mathbb{R}^k$, and u_i is the disturbance.

Assumption 1 (i.i.d. Sampling and Moment Exogeneity)**.** The observations $\{y_i, X_i\}_{i=1}^n$ are i.i.d.,

$$\mathbb{E}[X_i u_i] = 0,$$

$\mathbf{Q}_{XX} = \mathbb{E}[X_i X_i']$ exists and is positive definite, and $\mathbb{E}[u_i^2] < \infty$.

This assumption deserves a short comparison with the finite-sample OLS notes. There, unbiasedness was obtained under the stronger condition

$$\mathbb{E}[u \mid X] = 0,$$

which conditions on the full regressor matrix. Here we ask for the weaker orthogonality condition $\mathbb{E}[X_i u_i] = 0$, which is enough for consistency. The distinction is important. Large-sample theory can validate OLS even when exact finite-sample unbiasedness is not available. In other words, asymptotic consistency is often obtained under weaker conditions than those needed for exact finite-sample results.

At the same time, the weaker assumption should not be romanticized. It is still an identifying condition, and if it fails OLS will generally converge to the wrong object. For instance, in the presence of endogeneity, a standard concern in causal inference, this problem affects both finite-sample and large-sample analysis. The difference is simply that consistency relies on a weaker exogeneity condition than the full strict exogeneity structure used earlier to establish finite-sample unbiasedness.

6.2 OLS as a Sample Moment Estimator

Starting from the formula we derived in OLS, $\hat{\beta} = (X'X)^{-1}(X'Y)$, we have

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

and substituting $y_i = X_i'\beta + u_i$, we obtain

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

This representation is the large-sample starting point for OLS. It shows that the estimator is built from two sample moments: a second-moment matrix of the regressors and a sample covariance between regressors and disturbances. Once we know the limits of those two objects, the behavior of OLS follows almost mechanically.

6.3 Consistency of OLS

Theorem 6.1 (Consistency of OLS)**.** *Under Assumption 1,*

$$\hat{\beta} \xrightarrow{p} \beta.$$

Proof sketch. By the multivariate LLN,

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{p} \mathbf{Q}_{XX} \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} X_i u_i \xrightarrow{p} \mathbb{E}[X_i u_i] = 0.$$

Since $\mathbf{Q}_{XX}$ is positive definite, the matrix inverse is continuous at $\mathbf{Q}_{XX}$, so the Continuous Mapping Theorem gives

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \xrightarrow{p} \mathbf{Q}_{XX}^{-1}.$$

Combining these two convergences and applying the Continuous Mapping Theorem again,

$$\hat{\beta} - \beta = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i u_i\right) \xrightarrow{p} \mathbf{Q}_{XX}^{-1} \cdot 0 = 0.$$

□

The proof is worth pausing over because it illustrates a general pattern in econometrics: consistency is often the result of an LLN step, followed by one or two applications of the Continuous Mapping Theorem.

[Homework: Suppose the true model is

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i,$$

with $\mathbb{E}[u_i \mid x_{2i}, x_{3i}] = 0$, but we estimate the short regression that omits x_{3i}. Show that the probability limit of the short-regression slope is

$$\operatorname{plim} \tilde{\beta}_2 = \beta_2 + \beta_3 \delta_{32},$$

where

$$\delta_{32} = \operatorname{plim} \frac{\sum_{i=1}^{n}(x_{3i} - \bar{x}_3)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}.$$

Interpret δ_{32} as the slope from projecting x_3 on x_2, and explain why consistency fails when the omitted regressor is correlated with the included one and $\beta_3 \neq 0$.**]**

6.4 Additional Moment Condition for Asymptotic Normality

To move from consistency to asymptotic normality, we need the CLT to apply to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i u_i.$$

That requires finite second moments of $X_i u_i$, which in turn is conveniently ensured by finite fourth moments.

Assumption 2 (Finite Fourth Moments)**.**

$$\mathbb{E}\|X_i\|^4 < \infty \qquad \text{and} \qquad \mathbb{E}[u_i^4] < \infty.$$

Under this assumption,

$$S = \mathbb{E}[X_i X_i' u_i^2]$$

is finite. This matrix will become the middle component of the asymptotic sandwich variance.

6.5 Asymptotic Normality of OLS

Theorem 6.2 (Asymptotic Normality of OLS)**.** *Under Assumptions 1 and 2,*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, V_\beta),$$

where

$$V_\beta = \mathbf{Q}_{XX}^{-1} S \mathbf{Q}_{XX}^{-1}, \qquad S = \mathbb{E}[X_i X_i' u_i^2].$$

Proof sketch. Starting from the sample-moment representation of OLS,

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

The first factor converges in probability to $\mathbf{Q}_{XX}^{-1}$ by the consistency argument above. The second factor converges in distribution to $\mathcal{N}(0, S)$ by the multivariate CLT, because the vectors $X_i u_i$ are i.i.d. with mean zero and finite covariance matrix S. Slutsky's Theorem then implies

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbf{Q}_{XX}^{-1} S \mathbf{Q}_{XX}^{-1}).$$

□

The structure of the asymptotic variance is worth memorizing because it reappears throughout modern econometrics. The matrices

$$\mathbf{Q}_{XX}^{-1} \qquad \text{and} \qquad S \qquad \text{and} \qquad \mathbf{Q}_{XX}^{-1}$$

are often called the *bread–meat–bread* decomposition of the sandwich formula. The *bread* reflects the curvature or information in the regressors, while the *meat* reflects the variance of the score terms $X_i u_i$. This terminology is standard and will return later in GMM, MLE, and robust inference.

6.6 Homoskedastic Simplification

If we strengthen the setup by assuming homoskedasticity,

$$\mathbb{E}[u_i^2 \mid X_i] = \sigma^2,$$

then the sandwich simplifies substantially. In that case,

$$S = \mathbb{E}[X_i X_i' u_i^2] = \mathbb{E}\big[\mathbb{E}[X_i X_i' u_i^2 \mid X_i]\big] = \sigma^2 \mathbb{E}[X_i X_i'] = \sigma^2 \mathbf{Q}_{XX}.$$

Substituting into the general formula gives

$$V_\beta^0 = \sigma^2 \mathbf{Q}_{XX}^{-1}.$$

Thus, under homoskedasticity,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \mathbf{Q}_{XX}^{-1}).$$

This is the large-sample counterpart of the finite-sample variance formula derived in the OLS notes.

6.7 Consistency of the Error Variance Estimator

Theorem 6.3 (Consistency of s^2)**.** *Under Assumption 1, the residual variance estimator*

$$s^2 = \frac{1}{n-k} \sum_{i=1}^{n} e_i^2$$

satisfies

$$s^2 \xrightarrow{p} \sigma^2, \qquad \sigma^2 = \mathbb{E}[u_i^2].$$

Proof sketch. Write the residual as

$$e_i = u_i - X_i'(\hat{\beta} - \beta).$$

Then

$$e_i^2 - u_i^2 = -2u_i X_i'(\hat{\beta} - \beta) + \big(X_i'(\hat{\beta} - \beta)\big)^2.$$

Averaging across i, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}(e_i^2 - u_i^2) = -2\left(\frac{1}{n}\sum_{i=1}^{n} u_i X_i'\right)(\hat{\beta} - \beta) + (\hat{\beta} - \beta)'\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)(\hat{\beta} - \beta).$$

The first term is $o_p(1)$ because

$$\frac{1}{n}\sum_{i=1}^{n} u_i X_i \xrightarrow{p} \mathbb{E}[u_i X_i]$$

by the LLN, so it is bounded in probability, while $\hat{\beta} - \beta = o_p(1)$ by consistency. The second term is also $o_p(1)$ because

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{p} \mathbf{Q}_{XX},$$

so this matrix is $O_p(1)$, and the quadratic form is therefore the product of an $O_p(1)$ matrix and two $o_p(1)$ vectors. Hence,

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 - \frac{1}{n}\sum_{i=1}^{n} u_i^2 \xrightarrow{p} 0.$$

By the LLN,

$$\frac{1}{n}\sum_{i=1}^{n} u_i^2 \xrightarrow{p} \mathbb{E}[u_i^2] = \sigma^2.$$

Finally, since $(n-k)/n \to 1$, replacing $1/n$ by $1/(n-k)$ does not change the limit. □

7 Variance Estimation and Robust Standard Errors

The asymptotic normality result is only the first half of modern inference. To make that result operational, we also need a feasible estimator of the asymptotic variance. In the previous section, we showed that the asymptotic variance of the OLS estimator takes the sandwich form

$$V_\beta = \mathbf{Q}_{XX}^{-1} S \mathbf{Q}_{XX}^{-1}$$

Therefore, the next step is to estimate this object from the data.

This issue is central because asymptotic normality by itself is not enough to conduct inference. To construct standard errors, confidence intervals, and test statistics, we need a consistent estimator of V_β. The discussion below shows how different assumptions about the disturbance term lead to different feasible estimators of this variance matrix.

7.1 The Basic Logic

Estimating the bread is straightforward:

$$\widehat{\mathbf{Q}_{XX}} = \frac{1}{n} X'X \xrightarrow{p} \mathbf{Q}_{XX}.$$

The more delicate object is the meat,

$$S = \mathbb{E}[X_i X_i' u_i^2],$$

because the disturbances u_i are not observed. The standard solution is to replace them with residuals, $\hat{u}_i$.

7.2 The Homoskedastic Estimator

Under homoskedasticity,

$$V_\beta^0 = \sigma^2 \mathbf{Q}_{XX}^{-1},$$

so a natural estimator is

$$\hat{V}_\beta^{hom} = s^2 \widehat{\mathbf{Q}_{XX}}^{-1} = s^2 (X'X/n)^{-1}.$$

Equivalently, the estimated covariance matrix of $\hat{\beta}$ is

$$\widehat{\text{Var}}(\hat{\beta}) = s^2 (X'X)^{-1}.$$

This is the familiar textbook formula. It is consistent when the disturbances are homoskedastic, but not in general.

7.3 White's Estimator

To remain valid under heteroskedasticity, White's idea is to estimate the meat directly:

$$\hat{S}_{HC0} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i' \hat{u}_i^2.$$

This leads to the heteroskedasticity-consistent estimator

$$\hat{V}_\beta^{HC0} = \widehat{\mathbf{Q}_{XX}}^{-1} \hat{S}_{HC0} \widehat{\mathbf{Q}_{XX}}^{-1}.$$

Under Assumptions 1 and 2, this estimator converges in probability to V_β.

7.4 Beyond Heteroskedasticity: HAC and Cluster-Robust Standard Errors

The heteroskedasticity-robust estimator discussed above allows the variance of the disturbances to vary across observations, but it still treats observations as conditionally uncorrelated. In many empirical settings, that is not enough. In time-series or panel applications, disturbances may display serial dependence. In grouped data, observations within the same cluster may be correlated even if observations across clusters are independent. In such cases, the asymptotic variance of OLS must be modified accordingly. This is precisely the point anticipated in the earlier OLS notes: once we move beyond the classical spherical-disturbance case, inference becomes fundamentally a large-sample problem. The key object is still the sandwich variance,

$$V_\beta = \mathbf{Q}_{XX}^{-1} S \mathbf{Q}_{XX}^{-1},$$

but now the middle matrix S must be estimated in a way that reflects the relevant dependence structure in the data.

HAC standard errors. Suppose the data are ordered in a way that makes serial dependence plausible, as in a time series or in a panel after suitable aggregation. Then the matrix S is no longer determined only by the contemporaneous terms $\mathbb{E}[X_i X_i' u_i^2]$, because autocovariances across observations also matter. A convenient representation is

$$S = \Gamma_0 + \sum_{\ell=1}^{\infty} (\Gamma_\ell + \Gamma_\ell'), \qquad \Gamma_\ell = \mathbb{E}[X_i u_i \, u_{i-\ell} X_{i-\ell}'].$$

This is often called the *long-run variance* of the score process $X_i u_i$.

A common estimator is the Newey–West HAC (*heteroskedasticity and autocorrelation consistent*) estimator. It extends White's estimator by allowing both changing variance and serial dependence in the disturbances:

$$\widehat{S}_{HAC} = \widehat{\Gamma}_0 + \sum_{\ell=1}^{L} w_\ell (\widehat{\Gamma}_\ell + \widehat{\Gamma}_\ell'),$$

where

$$\widehat{\Gamma}_\ell = \frac{1}{n} \sum_{i=\ell+1}^{n} X_i \hat{u}_i \hat{u}_{i-\ell} X_{i-\ell}',$$

L is a bandwidth parameter, and $\{w_\ell\}$ are kernel weights, typically chosen so that the influence of more distant autocovariances is downweighted. The corresponding HAC

variance estimator is

$$\widehat{V}_{\beta,HAC} = \widehat{\mathbf{Q}_{XX}}^{-1} \widehat{S}_{HAC} \widehat{\mathbf{Q}_{XX}}^{-1}.$$

Cluster-robust standard errors. A different form of dependence arises when observations are naturally grouped into clusters, such as students within schools, workers within firms, or repeated observations within geographic units. In these settings, it is often reasonable to allow arbitrary correlation within each cluster while maintaining independence across clusters.

Let $g = 1, \ldots, G$ index clusters, and let X_g and $\hat{u}_g$ denote the regressor matrix and residual vector for cluster g. Then the cluster-robust estimator of the asymptotic variance is

$$\widehat{V}_{\beta,CR} = (X'X)^{-1} \left(\sum_{g=1}^{G} X_g' \hat{u}_g \hat{u}_g' X_g \right) (X'X)^{-1},$$

or, in asymptotic notation,

$$\widehat{V}_{\beta,CR} = \widehat{\mathbf{Q}_{XX}}^{-1} \left(\frac{1}{n} \sum_{g=1}^{G} X_g' \hat{u}_g \hat{u}_g' X_g \right) \widehat{\mathbf{Q}_{XX}}^{-1},$$

up to scaling conventions. This estimator allows unrestricted dependence within each cluster and is therefore much more flexible than the heteroskedasticity-robust estimator when grouped dependence is present.

Remark 2. The choice between heteroskedasticity-robust, HAC, and cluster-robust standard errors should be guided by the dependence structure one is willing to allow in the data. HAC is designed for ordered dependence, especially serial correlation over time. Cluster-robust inference is designed for grouped dependence, where arbitrary correlation is allowed within clusters but not across them. In both cases, the logic is asymptotic: these estimators do not deliver exact finite-sample inference, but rather consistent large-sample approximations to the true variance of the OLS estimator. In the cluster-robust case, a key regularity condition is that the number of clusters G grows to infinity. Large cluster sizes by themselves are not enough to justify the estimator.

7.5 Feasible Asymptotic Inference and Slutsky's Theorem

Once a consistent estimator of the asymptotic variance is available, asymptotic normality can be translated into a feasible test statistic. Under a null hypothesis such as $H_0 : \beta_j =$

$\beta_{j,0}$, we have

$$\frac{\sqrt{n}(\hat{\beta}_j - \beta_{j,0})}{\sqrt{[V_\beta]_{jj}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Of course, the matrix V_β is unknown. However, if $\hat{V}_\beta \xrightarrow{p} V_\beta$, then Slutsky's Theorem implies that replacing the unknown variance by a consistent estimator does not affect the limiting distribution. Therefore,

$$\frac{\sqrt{n}(\hat{\beta}_j - \beta_{j,0})}{\sqrt{[\hat{V}_\beta]_{jj}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Equivalently, since the asymptotic variance of $\hat{\beta}_j$ is of order $1/n$, this can be written in the more familiar form

$$\frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{[n^{-1}\hat{V}_\beta]_{jj}}} \xrightarrow{d} \mathcal{N}(0,1).$$

This is the asymptotic justification for feasible t-statistics. In other words, Slutsky's Theorem is what allows us to replace unknown population quantities with consistent estimators, such as the homoskedastic variance estimator, the HAC estimator, or the cluster-robust variance estimator, without changing the limiting normal distribution.