ECON 671 — Metrics

Expanded Notes

Week 5

Motivation: Why Asymptotic Statistics?

In many problems we do not know the exact finite-sample distribution of the statistic we want to use. A classical example is the one-sample t-statistic

$$T_n \;=\; \frac{\sqrt{n}\,(\bar{X}_n - \mu_0)}{S_n}, \qquad S_n^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

which is exactly t_{n-1} only under Gaussian sampling. When the data are merely i.i.d. with mean μ and variance $\sigma^2 < \infty$, exact distributions are unavailable, yet large-sample approximations make valid inference possible.

Recall. Central Limit Theorem (CLT). If $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$, then

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

Slutsky's theorem. If $Y_n \Rightarrow Y$ and $Z_n \xrightarrow{p} c \neq 0$, then $Y_n / Z_n \Rightarrow Y/c$.

Applying CLT and Slutsky, we obtain

$$T_n \;\Rightarrow\; \mathcal{N}(0, 1),$$

because $S_n \xrightarrow{p} \sigma$. Hence the usual t-test and confidence intervals for μ remain approximately valid without Gaussianity when n is large:

$$\text{Reject } H_0 : \mu = \mu_0 \;\text{ if }\; |T_n| > z_{1-\alpha/2}, \qquad \mu \in \Big[\bar{X}_n \pm z_{1-\alpha/2}\,\frac{S_n}{\sqrt{n}}\Big].$$

More broadly, *asymptotic statistics* provides a toolkit (LLN/CLT, Continuous Mapping, Slutsky, Delta method, and likelihood-based approximations) to:

- justify procedures when exact finite-sample laws are intractable;
- obtain large-sample distributions of estimators and test statistics;

- derive standard errors and confidence sets under weak regularity (e.g., non-Gaussian data).

Convergence concepts (vectors in $\mathbb{R}^k$)

We model a random vector $X = (X_1, \ldots, X_k)$ as a measurable map from a probability space to $(\mathbb{R}^k, \mathcal{B}^k)$. In simpler words, it is a vector of *real random variables*. Its (joint) distribution function is

$$F_X(x) \ := \ \mathbb{P}(X_1 \le x_1, \ldots, X_k \le x_k), \qquad x = (x_1, \ldots, x_k) \in \mathbb{R}^k.$$

Definition (Three modes of convergence)**.** Let X_n, X be $\mathbb{R}^k$-valued random vectors.

(i) **Convergence in distribution (weak convergence).** We write $X_n \rightsquigarrow X$ or $X_n \xrightarrow{d} X$ if any (hence all) of the following hold:

(a) (*Portmanteau; v.d.V. Thm. 2.1*) $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for every bounded continuous $f : \mathbb{R}^k \to \mathbb{R}$.

(b) For all continuity sets A of the law of X, $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$. Alternatively, the sequence of random vectors X_n converges in distribution to a random vector X if:

$$\mathbb{P}(X_n \le x) \to \mathbb{P}(X \le x)$$

for every x at which the limit distribution function $x \to \mathbb{P}(X \le x)$ is continuous.

(c) If $k = 1$, then $F_{X_n}(x) \to F_X(x)$ for every continuity point x of F_X.

If X has distribution L, we also write $X_n \rightsquigarrow L$.

(ii) **Convergence in probability.** We write $X_n \xrightarrow{p} X$ if

$$\forall \varepsilon > 0 : \quad \mathbb{P}(\|X_n - X\| > \varepsilon) \ \to \ 0.$$

(iii) **Almost sure convergence.** We write $X_n \xrightarrow{a.s.} X$ if

$$\mathbb{P}\Big(\lim_{n\to\infty} X_n = X\Big) = 1.$$

Proposition (Hierarchy)**.** *Almost sure convergence implies convergence in probability, and convergence in probability implies convergence in distribution:*

$$X_n \xrightarrow{a.s.} X \ \Rightarrow \ X_n \xrightarrow{p} X \ \Rightarrow \ X_n \Rightarrow X.$$

If X is a constant $c \in \mathbb{R}^k$, then $X_n \Rightarrow c$ is equivalent *to $X_n \xrightarrow{p} c$.*

Remark (Why we care)**.** Weak convergence is the language of large-sample approximations: CLTs give $X_n \Rightarrow X$ for suitable X (typically Gaussian), and the Portmanteau/continuous-mapping tools propagate limits through statistics. Almost sure and in-probability convergence control *consistency* of estimators.

Convergence in distribution

Example (Convergence in distribution)**.** Let $X_n \sim \mathrm{U}(0, 1 + \frac{1}{n})$. Its cdf is

$$F_n(x) = \begin{cases} 0, & x < 0, \\ \dfrac{x}{1 + 1/n}, & 0 \le x \le 1 + \frac{1}{n}, \\ 1, & x > 1 + \frac{1}{n}, \end{cases} \qquad F(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \le x \le 1, \\ 1, & x > 1, \end{cases}$$

so $F_n(x) \to F(x)$ for every continuity point of F, hence $X_n \rightsquigarrow \mathrm{U}(0, 1)$.

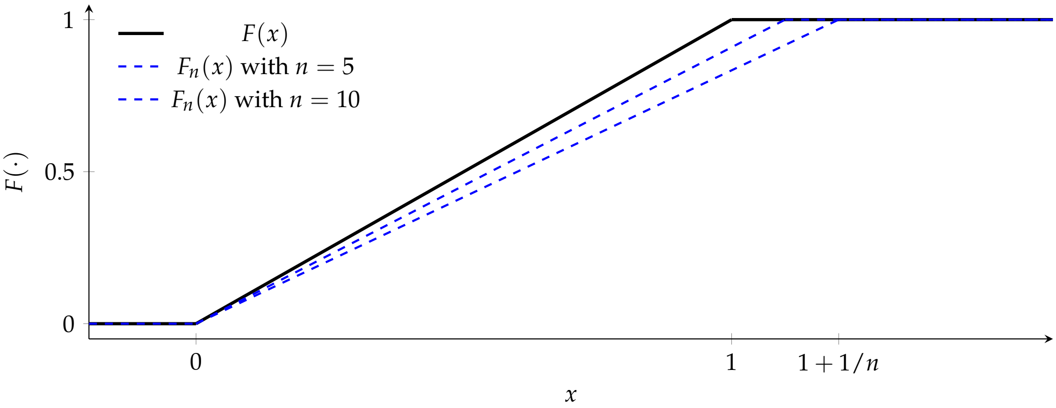

Figure 1: Convergence of F_n (dashed) to F (solid) for $X_n \sim \mathrm{U}(0, 1 + \frac{1}{n})$.

Maxima of uniforms: weak convergence and exponential limit

Example (Convergence in distribution with maxima of uniforms and exponential limit). Why is this example useful?

Let $U_1, \dots, U_n \overset{i.i.d.}{\sim} \mathrm{U}(0,1)$ and define the sample maximum

$$M_n := \max\{U_1, \dots, U_n\}.$$

Proposition (Distribution of M_n and degeneracy). *For $0 \le x \le 1$,*

$$F_{M_n}(x) = \mathbb{P}(M_n \le x) = \mathbb{P}(U_1 \le x, \dots, U_n \le x) = \prod_{i=1}^{n} \mathbb{P}(U_i \le x) = x^n.$$

Hence $f_{M_n}(x) = nx^{n-1}\mathbf{1}_{(0,1)}(x)$ (i.e., $M_n \sim Beta(n,1)$), and $M_n \xrightarrow{a.s.} 1$; without rescaling the limit is degenerate at 1.

Definition (Rescaled gap). To obtain a non-degenerate limit, zoom in at the upper endpoint by the natural $1/n$ scale:

$$Y_n := n\,(1 - M_n) \in [0, n].$$

Theorem 1 (Y_n converges to an exponential law). *For every $y \ge 0$,*

$$F_{Y_n}(y) = \mathbb{P}(Y_n \le y) = \mathbb{P}\Big(M_n \ge 1 - \frac{y}{n}\Big) = 1 - \mathbb{P}\Big(M_n \le 1 - \frac{y}{n}\Big) = 1 - \Big(1 - \frac{y}{n}\Big)^n \longrightarrow 1 - e^{-y}.$$

Therefore $Y_n \Rightarrow \mathrm{Exp}(1)$.

Proof. The display above shows pointwise convergence of the cdf to $y \mapsto 1 - e^{-y}$ on $[0, \infty)$ and 0 on $(-\infty, 0)$. This is the cdf of $\mathrm{Exp}(1)$, so $Y_n \Rightarrow \mathrm{Exp}(1)$ by the cdf characterization of weak convergence. □

Remark (Two complementary intuitions).

1. *Rare-events/Poisson heuristic.* For a fixed threshold $1 - y/n$, $\mathbb{P}(U_i > 1 - y/n) = y/n$. The number of exceedances among n trials is $\mathrm{Bin}(n, y/n) \Rightarrow \mathrm{Poisson}(y)$. The event $\{M_n \le 1 - y/n\}$ means "no exceedance", whose probability tends to e^{-y}; by complement, $F_{Y_n}(y) \to 1 - e^{-y}$.
2. *Density transformation.* With $x = 1 - y/n$,

   $$f_{Y_n}(y) = f_{M_n}(1 - y/n) \cdot \frac{1}{n} = n\Big(1 - \frac{y}{n}\Big)^{n-1} \cdot \frac{1}{n} \;\to\; e^{-y}\,\mathbf{1}_{(0,\infty)}(y).$$

   Moreover $\mathbb{E}[Y_n] = n\Big(1 - \frac{n}{n+1}\Big) = \frac{n}{n+1} \to 1$, consistent with $\mathrm{Exp}(1)$.

Remark (What this example teaches).

- For distributions with a finite upper endpoint, maxima stick to the boundary at rate $1/n$.
- After the correct rescaling, the gap to the endpoint has a *non-degenerate* limit; here it is exponential with mean 1 (a Weibull-type extreme-value limit).
- Re-scaling is the core idea of asymptotic approximations: without it, limits are often trivial.

Convergence in probability (vectors in $\mathbb{R}^k$)

Definition (Metric formulation). Let d be any metric on $\mathbb{R}^k$ that induces the usual topology (e.g., the Euclidean norm). A sequence of random vectors X_n *converges in probability* to X if

$$\forall \varepsilon > 0 : \qquad \mathbb{P}\big(d(X_n, X) > \varepsilon\big) \longrightarrow 0.$$

We write $X_n \xrightarrow{p} X$, or equivalently $d(X_n, X) \xrightarrow{p} 0$.

Remark (Independence of the metric and componentwise equivalence)**.** On $\mathbb{R}^k$, the choice of norm is immaterial. In particular,

$$X_n \xrightarrow{p} X \quad \Longleftrightarrow \quad \|X_n - X\|_2 \xrightarrow{p} 0 \quad \Longleftrightarrow \quad X_{n,j} \xrightarrow{p} X_j \text{ for all } j = 1, \dots, k,$$

where the last equivalence follows by the union bound.

Remark (Stability properties)**.** If $X_n \xrightarrow{p} X$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ is continuous, then $g(X_n) \xrightarrow{p} g(X)$ (continuous mapping). If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} c \neq 0$, then $X_n/Y_n \xrightarrow{p} X/c$ (Slutsky). If X is a constant c, then $X_n \xrightarrow{p} c$ is equivalent to $X_n \Rightarrow c$.

Theorem 2 (Weak Law of Large Numbers (WLLN))**.** *Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}X_i = \mu$ and* $\mathbf{Var}(X_i) = \sigma^2 < \infty$*, and set $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for every $\varepsilon > 0$,*

$$\mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) \longrightarrow 0, \qquad \textit{i.e.,} \quad \bar{X}_n \xrightarrow{p} \mu.$$

Proof. *TBD.*

Example (Sample proportion)**.** If $X_i \sim \text{Bernoulli}(p)$ i.i.d., then $\hat{p}_n = \bar{X}_n$ and the WLLN yields $\hat{p}_n \xrightarrow{p} p$.

Example (Consistency of the sample variance via Chebyshev)**.** Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}(X_i) = \mu$ and $\mathbf{Var}(X_i) = \sigma^2 < \infty$, and define the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

By Chebyshev's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\big(|S_n^2 - \sigma^2| \geq \varepsilon\big) \leq \frac{\mathbb{E}\big[(S_n^2 - \sigma^2)^2\big]}{\varepsilon^2} = \frac{\mathbf{Var}(S_n^2)}{\varepsilon^2}.$$

Hence, a sufficient condition for $S_n^2 \xrightarrow{p} \sigma^2$ is that $\mathbf{Var}(S_n^2) \to 0$ as $n \to \infty$.

Remark. (Verification, optional.) Under additional moment conditions (e.g., a finite fourth moment), one can check that $\mathbf{Var}(S_n^2) = O(1/n) \to 0$.

Remark (What to remember)**.** Convergence in probability is the mode used to formalize *consistency*. It is robust under continuous transformations (continuous mapping) and algebraic combinations with deterministic limits (Slutsky). The WLLN gives a first, central example; the variance example shows how Chebyshev and a variance calculation often suffice to establish consistency.

Almost sure convergence

Definition (Almost sure convergence)**.** We say that the sequence X_n converges almost surely to X if $d(X_n, X) \to 0$ with probability one; that is,

$$\mathbb{P}\Big(\lim_{n\to\infty} d(X_n, X) = 0\Big) = 1.$$

We denote this by $X_n \xrightarrow{\text{a.s.}} X$.

Remark. Both almost sure convergence and convergence in probability require X_n and X to be defined on the same probability space. This is not required for convergence in distribution.

Example (A basic a.s. limit)**.** Let $U \sim \mathrm{U}(0, 1)$ and define $X_n = U^n$.

- If $U < 1$, then $d(X_n, 0) = d(U^n, 0) \to 0$.
- If $U = 1$, then $X_n \equiv 1$. Moreover, $\mathbb{P}(U = 1) = 0$.

Hence,

$$\mathbb{P}\Big(\lim_{n\to\infty} d(X_n, 0) = 0\Big) = 1,$$

because the only event where the limit can fail is $\{U = 1\}$, which has probability zero. Therefore $X_n \xrightarrow{\text{a.s.}} 0$.

Intuition: how to tell them apart

Almost sure (pathwise): for almost every outcome ω, the tail of the sequence $X_n(\omega)$ sticks to $X(\omega)$.
In probability (masswise): for any fixed $\varepsilon > 0$, the mass outside the ε-ball around X vanishes; occasional large deviations may still occur along many n.
In distribution (lawwise): only the laws matter; CDFs converge, regardless of the joint construction of (X_n, X).
Hierarchy: $X_n \xrightarrow{\text{a.s.}} X \Rightarrow X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$.

Example (In probability but not almost surely)**.** Let $U_n \sim \mathrm{U}(0, 1)$ i.i.d. and $X_n = \mathbf{1}\{U_n \le 1/n\}$. Then $\mathbb{P}(|X_n - 0| > \varepsilon) = 1/n \to 0$, so $X_n \xrightarrow{p} 0$. However, $\sum_n \mathbb{P}(X_n = 1) = \infty$ and $(X_n = 1)$ occurs infinitely often with probability one (Borel–Cantelli), hence $X_n \not\to 0$ a.s.

Example (In distribution but not in probability)**.** Expand. Let $X_n \sim \mathcal{N}(0, 1)$ for all n, independent of $X \sim \mathcal{N}(0, 1)$. Then $X_n \xrightarrow{d} X$ (the laws are identical), but $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|Z| > \varepsilon/\sqrt{2}) > 0$ for $Z \sim \mathcal{N}(0, 2)$, so $X_n \overset{p}{\not\to} X$.

Portmanteau Lemma

Lemma. *For any random vectors X_n and X, the following statements are equivalent:*

(i) $\mathbb{P}(X_n \le x) \to \mathbb{P}(X \le x)$ *for all continuity points x of $F_X(x) := \mathbb{P}(X \le x)$.*

Intuition. *Weak convergence is CDF convergence at the points where the limit CDF is continuous (jumps are the only obstruction).*

(ii) $\mathbb{E}\big[f(X_n)\big] \to \mathbb{E}\big[f(X)\big]$ *for all bounded, continuous functions* f.
Intuition. *Expectations of bounded continuous "test functions" probe the law; if all such probes agree in the limit, the distributions converge.*

(iii) $\mathbb{E}\big[f(X_n)\big] \to \mathbb{E}\big[f(X)\big]$ *for all bounded, Lipschitz functions* f.
Intuition. *It suffices to test with a smaller class controlling oscillations (Lipschitz). This class is convergence-determining.*

(iv) $\liminf_{n\to\infty} \mathbb{E}\big[f(X_n)\big] \geq \mathbb{E}\big[f(X)\big]$ *for all nonnegative, continuous* f.
Intuition. *A Fatou-type lower semicontinuity: mass cannot "disappear" under limits when tested with nonnegative continuous functions.*

(v) $\liminf_{n\to\infty} \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G)$ *for every open set* G.
Intuition. *Probabilities of open sets are lower semicontinuous: the limit cannot undercount the mass that stays inside opens.*

(vi) $\limsup_{n\to\infty} \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ *for every closed set* F.
Intuition. *Dually, probabilities of closed sets are upper semicontinuous: extra mass cannot suddenly appear in closed sets.*

(vii) $\mathbb{P}(X_n \in B) \to \mathbb{P}(X \in B)$ *for all Borel sets* B *with* $\mathbb{P}(X \in \delta B) = 0$, *where* $\delta B = \overline{B} - \mathring{B}$ *is the boundary of* B.
Intuition. *Convergence holds on all "continuity sets" of the limit law—sets whose boundary carries no mass.*

Remark. Any one of (i)–(vii) may be taken as the definition of weak convergence $X_n \Rightarrow X$.

Remark (Why Portmanteau matters)**. Weak convergence** means *convergence in distribution*: $X_n \Rightarrow X$ (a.k.a. $X_n \xrightarrow{d} X$). The lemma provides equivalent, easier-to-check criteria:

- *Function view* ((ii)–(iii)): convergence of $\mathbb{E}[f(X_n)]$ for bounded continuous (even bounded Lipschitz) f.
- *Set view* ((v)–(vii)): lower/upper semicontinuity for open/closed sets, and convergence on continuity sets.
- *Practical goal*: prove $X_n \Rightarrow X$ using whichever side is tractable (functions or sets), especially in $\mathbb{R}^k$ where CDFs are unwieldy.

Proof. TBW. □

Continuous Mapping Theorem

A first useful result we can prove using the Portmanteau Lemma is the Continuous Mapping Theorem.

Theorem 3. *Let X_n, X be random vectors and let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$. Then:*

(i) If $X_n \Rightarrow X$, then $g(X_n) \Rightarrow g(X)$.

(ii) If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.

(iii) If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$.

Proof. TBW. □

Example (Consistency of S via CMT)**.** In the Example *Consistency of the sample variance via Chebyshev* (p. 5) we established conditions under which $S_n^2 \xrightarrow{p} \sigma^2$. By the Continuous Mapping Theorem, picking $g(x) = \sqrt{x}$ (continuous on $[0, \infty)$),

$$S_n = g(S_n^2) \xrightarrow{p} g(\sigma^2) = \sigma.$$

Theorem 4 (Some relationships between modes of convergence)**.** *Let X_n, X and Y_n be random vectors, and let c be a constant. Then:*

(i) If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{p} X$.

(ii) If $X_n \xrightarrow{p} X$, then $X_n \Rightarrow X$.

(iii) $X_n \xrightarrow{p} c$ iff $X_n \Rightarrow c$.

(iv) If $X_n \Rightarrow X$ and $d(X_n, Y_n) \xrightarrow{p} 0$, then $Y_n \Rightarrow X$.

(v) If $X_n \Rightarrow X$ and $Y_n \xrightarrow{p} c$, then $(X_n, Y_n) \Rightarrow (X, c)$.

(vi) If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $(X_n, Y_n) \xrightarrow{p} (X, Y)$.

Proof. TBW. □

Lemma (Slutsky's Lemma)**.** *Let X_n, X and Y_n be random vectors or variables. If $X_n \Rightarrow X$ and $Y_n \Rightarrow c$ for a constant c, then:*

(i) $X_n + Y_n \Rightarrow X + c$.

(ii) $Y_n X_n \Rightarrow cX$.

(iii) $Y_n^{-1} X_n \Rightarrow c^{-1} X$ (provided $c \neq 0$).

Remark. The constant c may also be a constant vector or matrix, assuming the operations above are well-defined and of correct size.

Proof. TBW. □

Some applications

Example (t-statistic)**.** Let $Y_1, Y_2, \dots$ be i.i.d. with $\mathbb{E}(Y_1) = 0$ and $\mathbb{E}(Y_1^2) < \infty$. The t-statistic, defined as $t_n = \sqrt{n}\,\bar{Y}_n / S_n$, is asymptotically standard normal using the results above. In particular, we have already shown that

$$S_n^2 \xrightarrow{p} \sigma^2, \qquad \text{and by the continuous mapping theorem} \qquad S_n \xrightarrow{p} \sigma.$$

Next, by the central limit theorem (proved later),

$$\sqrt{n}\,\bar{Y}_n \Rightarrow \mathcal{N}(0, \sigma^2).$$

Applying Slutsky's lemma,

$$t_n = \frac{\sqrt{n}\,\bar{Y}_n}{S_n} \;\Rightarrow\; \mathcal{N}(0,1).$$

Example (Normalized statistic and asymptotic CI)**.** Suppose T_n estimates θ, and S_n estimates σ. Assume

$$\sqrt{n}\,(T_n - \theta) \Rightarrow \mathcal{N}(0, \sigma^2), \qquad S_n^2 \xrightarrow{p} \sigma^2.$$

Define the normalized statistic

$$Z_n \;=\; \frac{\sqrt{n}\,(T_n - \theta)}{S_n}.$$

By Slutsky's lemma, $Z_n \Rightarrow \mathcal{N}(0,1)$. Let z_α be the upper α-quantile of the standard normal, i.e., $\mathbb{P}(Z > z_\alpha) = \alpha$ for $Z \sim \mathcal{N}(0,1)$. Then

$$\mathbb{P}(-z_\alpha \le Z_n \le z_\alpha) \;\to\; 1 - 2\alpha.$$

Equivalently, an asymptotic $(1-2\alpha)$ confidence interval for θ is

$$\left[\, T_n - \tfrac{S_n}{\sqrt{n}}\, z_\alpha \,,\; T_n + \tfrac{S_n}{\sqrt{n}}\, z_\alpha \,\right].$$

Notation: o_p and O_p

Definition (Stochastic o and O)**.** Let (X_n) be random variables (or vectors) and (R_n) a sequence of positive scalars ("rates").

- $X_n = o_p(R_n)$ means $\dfrac{X_n}{R_n} \xrightarrow{p} 0$.
- $X_n = O_p(R_n)$ means that $\dfrac{X_n}{R_n}$ is bounded in probability, i.e., for every $\varepsilon > 0$ there exists $M < \infty$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\left| \tfrac{X_n}{R_n} \right| > M \right) < \varepsilon.$$

Equivalently, $\{\frac{X_n}{R_n}\}_n$ is tight.

Remark (Deterministic analogy)**.** For deterministic sequences (a_n) and (b_n), $a_n = o(b_n)$ means $a_n/b_n \to 0$, and $a_n = O(b_n)$ means (a_n/b_n) is bounded. The symbols $o_p(1)$ and $O_p(1)$ are the stochastic counterparts of $o(1)$ and $O(1)$.

Calculus rules for o_p and O_p

The following implications are standard and will be used without comment:

$$o_p(1) + o_p(1) = o_p(1), \qquad o_p(1) + O_p(1) = O_p(1), \qquad O_p(1)\, o_p(1) = o_p(1),$$
$$(1 + o_p(1))^{-1} = O_p(1),$$
$$o_p(R_n) = R_n\, o_p(1), \qquad O_p(R_n) = R_n\, O_p(1), \qquad o_p(O_p(1)) = o_p(1).$$

Reading rule. Each display abbreviates a statement about explicitly named sequences, e.g. $o_p(1) + o_p(1) = o_p(1)$ means: if $X_n \xrightarrow{p} 0$ and $Y_n \xrightarrow{p} 0$, then $X_n + Y_n \xrightarrow{p} 0$ (an instance of the continuous mapping theorem).

Lemma (Rates under smooth transformations)**.** *Let $R : \mathbb{R}^k \to \mathbb{R}$ satisfy $R(0) = 0$, and let $X_n \xrightarrow{p} 0$ in $\mathbb{R}^k$. For every $p > 0$:*

(i) If $R(h) = o(\|h\|^p)$ as $h \to 0$, then $R(X_n) = o_p(\|X_n\|^p)$.

(ii) If $R(h) = O(\|h\|^p)$ as $h \to 0$, then $R(X_n) = O_p(\|X_n\|^p)$.

Characteristic functions

From the Portmanteau lemma, to show $X_n \Rightarrow X$ it suffices to verify $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, continuous f. A particularly convenient choice of such test functions is the *characteristic function*

$$\varphi_X(t) \;=\; \mathbb{E}(e^{it^\top X}), \qquad t \in \mathbb{R}^k.$$

Theorem 5 (Lévy's continuity theorem)**.** *Let X_n and X be random vectors in $\mathbb{R}^k$.*

(i) If $X_n \Rightarrow X$, then $\mathbb{E}\left(e^{it^\top X_n}\right) \to \mathbb{E}\left(e^{it^\top X}\right)$ for every $t \in \mathbb{R}^k$.

(ii) If $\mathbb{E}\left(e^{it^\top X_n}\right)$ converges pointwise to a function $\phi(t)$ that is continuous at 0, then ϕ is the characteristic function of some random vector X and $X_n \Rightarrow X$.

Remark. Part (i) is immediate from Portmanteau since $x \mapsto e^{it^\top x}$ is bounded and continuous. Part (ii) states the converse direction; we record it without proof.

Example (Binomial$(n, 1/n)$ converges to Poisson(1))**.** Let X_n be a sequence of random variables such that $X_n \sim \text{Binomial}(n, 1/n)$. Its characteristic function is

$$\varphi_{X_n}(t) = \left(1 + \tfrac{e^{it}-1}{n}\right)^n.$$

As $n \to \infty$,

$$\varphi_{X_n}(t) \longrightarrow \exp\big(e^{it} - 1\big),$$

using $\lim_{n\to\infty}\big(1 + \frac{z}{n}\big)^n = e^z$. Since $\exp(e^{it} - 1)$ is the characteristic function of Poisson(1), Lévy's theorem yields

$$X_n \Rightarrow \text{Poisson}(1).$$

Central Limit Theorem

Theorem 6 (Central Limit Theorem). *Let $Y_1, \dots, Y_n$ be i.i.d. r.v.'s with $\mathbb{E}(Y_i) = 0$ and $\mathbb{E}(Y_i^2) = 1$. Then*

$$\sqrt{n}\,\bar{Y}_n \Rightarrow \mathcal{N}(0, 1).$$

Proof. Let $\varphi_Y(t) = \mathbb{E}[e^{itY_1}]$ denote the characteristic function of Y_1. A second–order Taylor expansion of φ_Y at $t = 0$ yields

$$\varphi_Y(t) = 1 + it\,\mathbb{E}(Y_1) - \frac{t^2}{2}\mathbb{E}(Y_1^2) + o(t^2) = 1 - \frac{t^2}{2} + o(t^2),$$

since $\mathbb{E}(Y_1) = 0$ and $\mathbb{E}(Y_1^2) = 1$. Consider now the characteristic function of $\sqrt{n}\,\bar{Y}_n$. Because the Y_i are i.i.d.,

$$\varphi_{\sqrt{n}\,\bar{Y}_n}(t) = \Big(\varphi_Y\big(\tfrac{t}{\sqrt{n}}\big)\Big)^n.$$

Using the expansion above with $t/\sqrt{n}$ in place of t,

$$\varphi_Y\Big(\tfrac{t}{\sqrt{n}}\Big) = 1 - \frac{t^2}{2n} + o\big(\tfrac{1}{n}\big).$$

Hence

$$\varphi_{\sqrt{n}\,\bar{Y}_n}(t) = \Big(1 - \frac{t^2}{2n} + o\big(\tfrac{1}{n}\big)\Big)^n \longrightarrow e^{-t^2/2},$$

because $\big(1 + \frac{z}{n} + o(\frac{1}{n})\big)^n \to e^z$. The limit $e^{-t^2/2}$ is the characteristic function of $\mathcal{N}(0, 1)$. By Lévy's Continuity Theorem,

$$\sqrt{n}\,\bar{Y}_n \Rightarrow \mathcal{N}(0, 1).$$

□

Delta Method

We estimate a primitive parameter θ with T_n, yet the target is a smooth function $\phi(\theta)$. By the Continuous Mapping Theorem, if $T_n \xrightarrow{p} \theta$ and ϕ is continuous at θ, then $\phi(T_n) \xrightarrow{p} \phi(\theta)$. For inference we still need the rate and a limit law for $r_n\{\phi(T_n) - \phi(\theta)\}$ to obtain standard errors, confidence intervals, and tests. When ϕ is differentiable at θ and $r_n(T_n - \theta) \Rightarrow T$ with $r_n \to \infty$, a first-order Taylor expansion gives

$$\phi(T_n) = \phi(\theta) + \phi'_\theta(T_n - \theta) + o_p(r_n^{-1}) \quad \Rightarrow \quad r_n\{\phi(T_n) - \phi(\theta)\} \Rightarrow \phi'_\theta(T).$$

In the classical CLT case $r_n = \sqrt{n}$ and $\sqrt{n}(T_n - \theta) \Rightarrow \mathcal{N}(0, \Sigma)$, it follows that

$$\sqrt{n}\{\phi(T_n) - \phi(\theta)\} \Rightarrow \mathcal{N}(0,\ \nabla\phi(\theta)^\top \Sigma\, \nabla\phi(\theta)),$$

informally summarized as $\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big) \approx \phi'(\theta)\, \sqrt{n}\,(T_n - \theta)$.

In words, the quantity $\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big)$ converges in distribution to a mean–zero normal random variable whose variance (or covariance matrix) equals the asymptotic variance of $\sqrt{n}(T_n - \theta)$, propagated through the derivative (or gradient) of ϕ evaluated at the true parameter value θ.

Theorem 7 (Delta Method). *Let $\phi : \mathbb{R}^k \to \mathbb{R}^m$ be differentiable at θ. Suppose $r_n(T_n - \theta) \Rightarrow T$ with $r_n \to \infty$. Then*

$$r_n\big(\phi(T_n) - \phi(\theta)\big) \Rightarrow \phi'_\theta(T),$$

where ϕ'_θ denotes the derivative (linear map) of ϕ at θ, applied to T.

Proof. TBW. □

Example (Delta Method: a simple transformation). Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Consider $T_n = \bar{X}_n^2$. By the CLT,

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

Take $\phi(x) = x^2$, which is differentiable at μ with derivative $\phi'(\mu) = 2\mu$. By the Delta Method,

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

Remark (Delta Method at a zero mean). If $\mu = 0$, then $\phi'(0) = 0$ and the first–order Delta Method does not apply. Since $\sqrt{n}\,\bar{X}_n \Rightarrow Z \sim \mathcal{N}(0, \sigma^2)$, we have

$$n\,\bar{X}_n^2 = (\sqrt{n}\,\bar{X}_n)^2 \ \Rightarrow\ Z^2,$$

which is a scaled chi–square distribution:

$$Z^2 \sim \sigma^2\chi_1^2.$$

Thus, when $\mu = 0$,

$$n\,\bar{X}_n^2 \Rightarrow \sigma^2\chi_1^2,$$

instead of a normal limit.

Week 6

Point Estimation

Definition (Parameter)**.** A *parameter* θ is any (possibly vector–valued) function of the population distribution F. For example, the population mean $\mu = \mathbb{E}[X]$ is a parameter and, more generally, a functional of F (the first moment).

Definition (Estimator and estimate)**.** An *estimator* $\hat{\theta}$ of a parameter θ is a statistic—i.e., a measurable function of the sample—intended as a guess of θ. Estimators are random variables because they depend on the random sample. The *estimate* is the numerical realization of the estimator for a given dataset.

Remark (Intuition)**.** Think "target vs. recipe vs. outcome": θ is the fixed target feature of the population, $\hat{\theta}(\mathbf{X})$ is the recipe that maps the sample to a guess, and the estimate is the concrete number obtained once the sample is observed.

Method of Moments

There are many estimation methods; an important and popular one is the *Method of Moments (MoM)*.

- MoM is attractive for semi–parametric settings: it can estimate finite–dimensional parameters without fully specifying the distribution of X.
- It is often intuitive because parameters of interest frequently correspond to moments of the distribution; then closed–form estimators arise by simple matching.
- In other cases, parameters are not explicit functions of the moments and numerical solution of the moment equations is required.

Basic principle. Let m_j denote the jth sample moment and let μ_j denote the corresponding population moment. The MoM estimator sets

$$m_j = \mu_j, \qquad j = 1, 2, \ldots,$$

and solves the resulting system for the unknown parameters.

Example (Matching the first two moments)**.** Let $X_1, \ldots, X_n$ be a sample. The first two sample moments are

$$m_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

The corresponding population moments are $\mu_1 = \mathbb{E}[X]$ and $\mu_2 = \mathbb{E}[X^2]$. The MoM estimator solves $m_1 = \mu_1$ and $m_2 = \mu_2$ for the unknown parameters.

Remark (Intuition)**.** MoM "matches features of the data to features of the model." If the model says the parameter controls, say, the mean and the second moment, then we pick parameter values that make the model's mean and second moment coincide with their empirical counterparts.

Example (Normal Model (Method of Moments))**.** Suppose $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$. Take the first two data moments

$$m_1 = \bar{X}, \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

and note that the population moments are

$$\mu_1 = \mathbb{E}[X] = \theta, \qquad \mu_2 = \mathbb{E}[X^2] = \theta^2 + \sigma^2.$$

Setting $m_1 = \mu_1$ and $m_2 = \mu_2$ yields the system

$$\bar{X} = \theta, \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \theta^2 + \sigma^2,$$

whose solution is

$$\hat{\theta} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Remark (Interpretation)**.** The estimator $\hat{\theta} = \bar{X}$ matches the model's mean to the sample mean. The variance estimator matches the model's second moment to the sample second moment, leading to the *uncorrected* sample variance (with $1/n$). This reflects the moment–matching principle rather than a bias correction aim.

Example (Poisson)**.** Suppose $X_1, \dots, X_n \overset{i.i.d.}{\sim} \text{Poisson}(\lambda)$, so $\mathbb{E}[X_i] = \lambda$.

- Match the first moment: set $\mathbb{E}[X_i] = \bar{X}_n$.

Hence the MoM estimator is

$$\hat{\lambda}_{\text{MM}} = \bar{X}_n.$$

Remark (Intuition)**.** For the Poisson family, the mean equals the intensity. Matching the empirical mean therefore pins down λ directly.

Example (Gamma)**.** Suppose $X_1, \dots, X_n \overset{i.i.d.}{\sim} \text{Gamma}(\alpha, \beta)$ under the *shape–scale* parameterization, so

$$\mathbb{E}[X_i] = \alpha\beta, \qquad \textbf{Var}(X_i) = \alpha\beta^2.$$

Match model moments to the data:

$$\alpha\beta = \bar{X}_n, \qquad \alpha\beta^2 = S_n^2,$$

where $S_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is the (uncorrected) sample variance, which aligns with MoM. Solving,

$$\hat{\beta}_{\text{MM}} = \frac{S_n^2}{\bar{X}_n}, \qquad \hat{\alpha}_{\text{MM}} = \frac{\bar{X}_n^2}{S_n^2}.$$

Remark (Intuition)**.** Mean identifies the product $\alpha\beta$ and variance identifies $\alpha\beta^2$; dividing the two equations isolates β, then plug back to recover α.

When closed forms are unavailable. In many econometric models we can express moments as functions of parameters, but cannot invert those relations analytically. In such cases, we solve the MoM equations numerically to obtain the estimator.

Method of Moments: General Formulation

Take $\theta \in \mathbb{R}^k$ and let $m(X, \theta)$ be a $k \times 1$ vector of moment functions implied by the model. Assume the population moment condition

$$\mathbb{E}\big[m(X, \theta)\big] = 0.$$

Given data $\{x_i\}_{i=1}^n$, define the sample counterpart

$$\bar{m}_n(\theta) := \frac{1}{n}\sum_{i=1}^n m(x_i, \theta).$$

The method-of-moments estimator $\hat{\theta}$ solves the k nonlinear equations

$$\bar{m}_n(\hat{\theta}) = 0.$$

Remark (Identification)**.** When the number of moment conditions equals the dimension of θ, we have *exact identification* ("just identified"). Then, under regularity, a unique solution exists in large samples.

Example. Let $F(x) = \mathbb{P}(X \le x) = \mathbb{E}\big[1\{X \le x\}\big]$ be the distribution function of a univariate r.v. X. Given a sample $x_1, \ldots, x_n$, the MoM estimator for $F(x)$ is the fraction of observations not exceeding x:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n 1\{x_i \le x\}.$$

Remark (MoM view)**.** Here the moment condition is $\mathbb{E}[1\{X \le x\} - F(x)] = 0$ for each fixed x; replacing the expectation by the sample average yields $F_n(x)$.

Theorem 8 (Asymptotic distribution at a fixed x)**.** *If X_i are i.i.d., then for any fixed continuity point x of F,*

$$\sqrt{n}\left(F_n(x) - F(x)\right) \ \Rightarrow\ \mathcal{N}\big(0,\ F(x)\big(1 - F(x)\big)\big) .$$

Proof. Consistency follows from the WLLN since $1\{X_i \le x\}$ are i.i.d. bounded with mean $F(x)$, hence $F_n(x) \xrightarrow{p} F(x)$. For the limit distribution, apply the CLT to the Bernoulli variables $Y_i := 1\{X_i \le x\}$ with variance $F(x)\{1 - F(x)\}$:

$$\sqrt{n}\left(\bar{Y}_n - \mathbb{E}[Y_i]\right) \Rightarrow \mathcal{N}(0, \mathbf{Var}(Y_i)) ,$$

which is exactly the stated result. □

Example (Euler Equation (from Hansen, 11.11))**.** A consumer chooses consumption C_t in period t and C_{t+1} in period $t+1$. Preferences are

$$U(C_t, C_{t+1}) = u(C_t) + \frac{1}{\beta} u(C_{t+1}),$$

and the budget constraint is

$$C_t + \frac{C_{t+1}}{R_{t+1}} \le W_t,$$

where W_t is the endowment and R_{t+1} is the (uncertain) gross return on investment from t to $t+1$. Using the budget constraint to substitute $C_{t+1} = (W_t - C_t)R_{t+1}$, expected utility in period t is

$$\mathbb{E}\left[u(C_t) + \frac{1}{\beta} u\big((W_t - C_t)R_{t+1}\big)\right] .$$

The consumer chooses C_t to maximize expected utility. Assume CRRA utility $u(c) = c^{1-\alpha}/(1-\alpha)$. The first–order condition is the Euler equation

$$\mathbb{E}\left(R_{t+1}\left(\tfrac{C_{t+1}}{C_t}\right)^{-\alpha} - \beta\right) = 0.$$

Suppose we want to estimate α, the coefficient of relative risk aversion, treating β as known. Define the moment function

$$m(R_{t+1}, C_{t+1}, C_t, \alpha) := R_{t+1}\left(\tfrac{C_{t+1}}{C_t}\right)^{-\alpha} - \beta,$$

so that the population moment condition is

$$\mathbb{E}\big[m(R_{t+1}, C_{t+1}, C_t, \alpha)\big] = 0.$$

Given data $\{C_t, C_{t+1}, R_{t+1}\}_{t=1}^{n}$, the sample moment is

$$\bar{m}_n(\alpha) := \frac{1}{n}\sum_{t=1}^{n}\left[R_{t+1}\left(\tfrac{C_{t+1}}{C_t}\right)^{-\alpha} - \beta\right],$$

and the Method–of–Moments estimator $\hat{\alpha}$ is defined as the solution to

$$\bar{m}_n(\hat{\alpha}) = 0,$$

which is typically obtained numerically.

Properties of Moment Estimators

- Moment estimators are not necessarily "best" estimators, although in special cases they coincide with optimal ones.
- Under mild regularity conditions, they converge at rate $\sqrt{n}$ and are asymptotically normal.
- This asymptotic normality follows from the CLT and mean–value (Taylor) expansions of the moment equations.
- The Method of Moments is attractive because of its semi–parametric nature and robustness features.

Identification

Definition (Identification (Hansen, Def. 11.1))**.** The parameter θ is identified in Θ if there exists a unique $\theta_0 \in \Theta$ that solves the moment condition

$$\mathbb{E}\big[m(X,\theta)\big] = 0.$$

Consistency of the Method of Moments

To show that the MoM estimator is consistent, we impose the following assumptions:

1. X_i are i.i.d.
2. $\|m(x,\theta)\| \leq M(x)$ and $\mathbb{E}[M(X)] < \infty$.
3. $m(X,\theta)$ is continuous in θ with probability 1.
4. Θ is compact.
5. The population moment equation $\mathbb{E}[m(X,\theta)] = 0$ has a unique solution θ_0.

Assumptions (A1)–(A4) ensure that the sample moment function

$$\bar{m}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(X_i,\theta)$$

is a uniformly consistent estimator of its expectation. Indeed, since X_i are i.i.d., the transformed variables $m(X_i, \theta)$ are also i.i.d.; by the weak law of large numbers,

$$\bar{m}_n(\theta) \xrightarrow{p} \mathbb{E}[m(X,\theta)].$$

Adding (A5) yields consistency: the sample solution $\hat{\theta}$ to $\bar{m}_n(\hat{\theta}) = 0$ converges in probability to the population solution θ_0.

Asymptotic Distribution

Define

$$\Omega := \mathbf{Var}\big(m(X,\theta_0)\big), \qquad Q := \mathbb{E}\left[\frac{\partial}{\partial\theta} m(X,\theta_0)\right],$$

and let $\mathcal{N}$ be a neighborhood of θ_0. On top of (A1)–(A5), assume:

6. $\mathbb{E}[\|m(X,\theta_0)\|^2] < \infty$.
7. $\frac{\partial}{\partial\theta}\mathbb{E}[m(X,\theta)]$ is continuous in $\theta \in \mathcal{N}$.
8. $m(X,\theta)$ is Lipschitz–continuous in θ on $\mathcal{N}$.
9. Q is full rank.
10. θ_0 lies in the interior of Θ.

Under assumptions (A1)–(A10),

$$\sqrt{n}\,(\hat{\theta} - \theta_0) \;\Rightarrow\; \mathcal{N}(0,V), \qquad V = Q^{-1}\Omega\, Q^{-1}.$$

Why do these assumptions matter?

- (A6) ensures finite second moments, needed for the CLT.
- (A7) guarantees local invertibility of $m(X,\theta)$ near θ_0, via the inverse function theorem; this is required to apply the Delta Method.
- (A8) controls the Taylor remainder so that the linear approximation is accurate asymptotically.
- (A9) is an identification condition: it ensures that movements in θ produce unique changes in the moments.

  In exact identification (k equations, k unknowns), each moment must contribute unique information. If two moment conditions convey the same information, we cannot uniquely pin down θ.

Suppose we have a point estimate $\hat{\theta}$ for our parameter of interest. The asymptotic variance of a Method–of–Moments estimator is

$$V = Q^{-1}\Omega Q^{-1},$$

and we estimate it using plug–in quantities:

$$\hat{V} = \hat{Q}^{-1}\hat{\Omega}\hat{Q}^{-1}, \qquad \hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} m(x_i,\hat{\theta})\, m(x_i,\hat{\theta})', \qquad \hat{Q} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta} m(x_i,\hat{\theta})'.$$

Example. Euler Equation (continued) Recall the sample moment

$$\bar{m}_n(\alpha) = \frac{1}{n}\sum_{t=1}^{n}\left[R_{t+1}\left(\tfrac{C_{t+1}}{C_t}\right)^{-\alpha} - \beta\right].$$

The asymptotic variance is

$$V = Q^{-1}\Omega Q^{-1},$$

where

$$\Omega = \mathbf{Var}\left(R_{t+1}\left(\tfrac{C_{t+1}}{C_t}\right)^{-\alpha} - \beta\right), \qquad Q = \mathbb{E}\left(R_{t+1}\left(\tfrac{C_{t+1}}{C_t}\right)^{-\alpha}\log\left(\tfrac{C_{t+1}}{C_t}\right)\right).$$

The plug–in estimator is therefore

$$\hat{V} = \frac{\frac{1}{n}\sum_{t=1}^{n}\left(R_{t+1}\left(\tfrac{C_{t+1}}{C_t}\right)^{-\hat{\alpha}} - \beta\right)^2}{\left(\frac{1}{n}\sum_{t=1}^{n} R_{t+1}\left(\tfrac{C_{t+1}}{C_t}\right)^{-\hat{\alpha}}\log\left(\tfrac{C_{t+1}}{C_t}\right)\right)^2}.$$

Remarks on Method of Moments

Method of Moments is widely used in both macro and micro applications. In macro, it often appears under the label "calibration," in which case the asymptotic variance is typically not the central object of interest.

There exist many extensions of Method of Moments that appear throughout econometrics; some of them are introduced next.

Generalized Method of Moments

So far we assumed *exact identification*: as many parameters as moment conditions,

$$\theta \in \mathbb{R}^k, \qquad \mathbb{E}[m(X,\theta)] = 0,$$

with m being an $l \times 1$ function and $l = k$.

If $l > k$ (over–identification), the system cannot be solved exactly. Instead, GMM finds the θ that makes the sample moments "as close to zero as possible."

Formally, choose an $l \times l$ weighting matrix $W > 0$ and define the criterion

$$J(\theta) = n\, \bar{m}_n(\theta)' W\, \bar{m}_n(\theta).$$

The GMM estimator is

$$\hat{\theta} = \arg\min_{\theta} J(\theta).$$

An important ingredient is the choice of W. A simple option is $W = I_l$, which yields a least–squares match of the moment conditions. However, this is generally not the optimal choice of weight matrix.

Simulated Method of Moments

In some models, the moments of interest are not available in closed form. For example, a macroeconomic model may not permit a closed–form expression for $\mathbb{E}[g(X, \theta)]$, such as the correlation between output and inflation.

However, the researcher may be able to simulate the model for a given parameter θ. In that case, one can use the *Simulated Method of Moments* (SMM).

Instead of matching the empirical moment m_1 to the theoretical moment $\mu_1(\theta)$, we simulate the model S times and compute the simulated moment

$$\tilde{\mu}_1(\theta)$$

based on those S draws from the model. We then choose $\hat{\theta}$ that minimizes the discrepancy between the sample moment m_1 and the simulated moment $\tilde{\mu}_1(\theta)$.

SMM thus replaces analytically intractable expectations with simulation–based approximations.

Week 7

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a fundamental method for estimating parameters in parametric statistical models. A parametric model specifies a full probability structure for the data, indexed by an unknown parameter vector $\theta \in \Theta$. For discrete variables, the model is described by a pmf $\pi(x \mid \theta)$, and for continuous variables by a density $f(x \mid \theta)$. In both cases, the goal is to use the observed sample to learn about the underlying parameter governing the distribution.

We assume that the data $\{X_i\}_{i=1}^n$ are i.i.d. draws from the postulated model. Independence implies that the joint density factorizes as

$$f(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta),$$

and similarly for mass functions. This leads directly to the *likelihood function*,

$$L_n(\theta) = \prod_{i=1}^n f(X_i \mid \theta), \qquad \text{or} \qquad L_n(\theta) = \prod_{i=1}^n \pi(X_i \mid \theta),$$

which treats the data as fixed and evaluates how plausible each parameter value is as a generator of the observed sample.

It is helpful to emphasize the opposite roles played by density functions and likelihood functions. A density $f(x \mid \theta)$ treats θ as fixed and quantifies how likely the data would be under that parameter. In contrast, the likelihood $L_n(\theta)$ treats the observed sample as fixed and asks which values of θ make the sample most plausible. Estimation via MLE therefore amounts to selecting the parameter that maximizes this likelihood, i.e., the parameter value that best explains the observed data within the assumed model.

Throughout, we assume the model is *correctly specified*: there exists a unique $\theta_0 \in \Theta$ such that $f(x \mid \theta_0)$ coincides with the true data-generating distribution $f(x)$. Uniqueness ensures that there is only one parameter value consistent with the population distribution. In contrast, under misspecification no $\theta \in \Theta$ perfectly reproduces $f(x)$, and the interpretation of MLE changes accordingly.

Under correct specification, the principle of maximum likelihood provides a natural and powerful criterion: the estimator is the value of θ that maximizes $L_n(\theta)$, making the observed sample most compatible with the assumed model.

Definition (Maximum Likelihood Estimator (MLE))**.** Given a parametric model with likelihood function $L_n(\theta)$, the *maximum likelihood estimator* of θ is any value $\hat{\theta} \in \Theta$ that maximizes the likelihood:

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

In practice, we often attempt to find $\hat{\theta}$ by differentiating the likelihood (or log–likelihood) with respect to θ, setting the derivative equal to zero, and solving the resulting first–order condition. This yields a *necessary* condition for an interior maximum, but it is not sufficient: one must still check that the candidate solution corresponds to a maximum (rather than a minimum or saddle point). Moreover, for many models the first–order condition does not admit a closed–form solution, and numerical optimization routines are required.

Typically it is analytically and numerically more convenient to work with the logarithm of the likelihood function.

Definition (Log–likelihood function)**.** The *log–likelihood* is defined as

$$\ell_n(\theta) \equiv \log L_n(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta),$$

where $f(\cdot \mid \theta)$ denotes the density (or pmf) of X_i under parameter θ.

Working with $\ell_n(\theta)$ is numerically more stable than working with $L_n(\theta)$, because the product of many densities can become extremely small, while their logarithms add to a quantity of reasonable magnitude. Importantly, the maximizer of $\ell_n(\theta)$ coincides with the maximizer of $L_n(\theta)$, since the logarithm is a strictly increasing transformation:

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

Example (Normal distribution)**.** Assume $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma_0^2)$, where $\sigma_0 > 0$ is known and we wish to estimate the mean μ. The log–density for a single observation is

$$\log f(x \mid \mu) = -\frac{1}{2}\log(2\pi\sigma_0^2) - \frac{(x-\mu)^2}{2\sigma_0^2},$$

so the sample log–likelihood is

$$\ell_n(\mu) = -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Differentiating with respect to μ and setting the derivative to zero gives the first–order condition

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\,\ell_n(\mu) = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu) = 0,$$

whose solution is

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

A second derivative check,

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\,\ell_n(\mu) = -\frac{n}{\sigma_0^2} < 0,$$

confirms that $\hat{\mu} = \bar{X}_n$ indeed maximizes the log–likelihood and is therefore the MLE for μ.

Example (Poisson distribution)**.** Assume $X_i \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ with pmf

$$f(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

The log–density for a single observation is

$$\log f(x \mid \lambda) = -\lambda + x \log \lambda - \log(x!),$$

so the sample log–likelihood is

$$\ell_n(\lambda) = \sum_{i=1}^{n} \log f(X_i \mid \lambda) = -n\lambda + \Big(\sum_{i=1}^{n} X_i\Big) \log \lambda - \sum_{i=1}^{n} \log(X_i!).$$

The first–order condition is

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\,\ell_n(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i = 0,$$

which yields the solution

$$\hat{\lambda} = \bar{X}_n.$$

The second derivative,

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\,\ell_n(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} X_i < 0 \quad \text{at } \lambda = \hat{\lambda},$$

shows that $\hat{\lambda} = \bar{X}_n$ is indeed a maximizer, and thus the MLE for the Poisson parameter λ.

Example (Linear model with Gaussian errors)**.** Assume a simple linear regression with a single regressor X_i and no intercept,

$$Y_i = \beta X_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_0^2),$$

where σ_0^2 is known. Conditional on $X_i = x_i$, the density of Y_i is

$$f(y_i \mid x_i, \beta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(y_i - \beta x_i)^2}{2\sigma_0^2}\right),$$

so that the conditional log–density is

$$\log f(y_i \mid x_i, \beta) = -\frac{1}{2}\log(2\pi\sigma_0^2) - \frac{(y_i - \beta x_i)^2}{2\sigma_0^2}.$$

The sample log–likelihood is therefore

$$\ell_n(\beta) = -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(Y_i - \beta X_i)^2.$$

Differentiating with respect to β and setting the derivative equal to zero gives the first–order condition

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\,\ell_n(\beta) = \frac{1}{\sigma_0^2}\sum_{i=1}^{n} X_i\,(Y_i - \beta X_i) = 0.$$

Solving for β yields the MLE

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2},$$

which coincides with the usual least–squares estimator in this simple regression setup.

Likelihood Analog Principle

To understand why MLE behaves well in large samples, it is useful to introduce the *expected log–density* (or population log–likelihood).

Definition (Expected log–density)**.** For a given parameter value $\theta \in \Theta$, define

$$\ell(\theta) := \mathbb{E}\big[\log f(X \mid \theta)\big],$$

where the expectation is taken under the true data–generating distribution.

Theorem 9 (Likelihood Analog Principle; Hansen, Thm. 10.2)**.** *If the model is correctly specified, there exists a unique $\theta_0 \in \Theta$ such that $f(x \mid \theta_0)$ equals the true density $f(x)$, and this true parameter maximizes the expected log–density:*

$$\theta_0 = \arg\max_{\theta\in\Theta} \ell(\theta).$$

Why is this insightful?

- The sample analog of $\ell(\theta)$ is the average log–likelihood

  $$\bar{\ell}_n(\theta) := \frac{1}{n}\,\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log f(X_i \mid \theta),$$

  which has the same maximizer as the full log–likelihood $\ell_n(\theta)$. Thus the MLE $\hat{\theta}$ maximizes $\bar{\ell}_n(\theta)$.

- In parallel, at the population level θ_0 maximizes $\ell(\theta)$. Hence we can view $\hat{\theta}$ as the *sample analog* of θ_0: the estimator solves in the sample the same optimization problem that defines the true parameter in the population, with expectations replaced by averages.

Invariance Property

A particularly convenient feature of maximum likelihood is its invariance under smooth transformations of the parameter.

Theorem 10 (Invariance of the MLE; Hansen, Thm. 10.3)**.** *Let $\hat{\theta}$ be the MLE of $\theta \in \mathbb{R}^m$. For any transformation $h : \mathbb{R}^m \to \mathbb{R}^\ell$ and $\beta = h(\theta)$, the MLE of β is*

$$\hat{\beta} = h(\hat{\theta}).$$

Proof. TBW. □

Score, Hessian, and Information

To analyze the large–sample behavior of maximum likelihood estimators, it is crucial to study how the log–likelihood reacts to local perturbations in the parameter. This sensitivity is captured by the *score* and the curvature of the log–likelihood, encoded in the *Hessian*. Throughout this section we assume that the density $f(x \mid \theta)$ is differentiable with respect to θ.

Likelihood Score. The (sample) score is the gradient of the log–likelihood,

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

When θ is a vector, $S_n(\theta)$ is a vector of partial derivatives. The score measures the direction and magnitude in which the log–likelihood increases most steeply. For any interior maximum of the log–likelihood, the score must vanish: $S_n(\hat{\theta}) = 0$.

Likelihood Hessian. The curvature of the log–likelihood is summarized by the (negative) Hessian,

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

This matrix quantifies how quickly the log–likelihood bends away from its maximum. A sharply curved log–likelihood corresponds to a large Hessian and therefore more precise estimation.

To move from sample objects to population analogs, we introduce the *efficient score*, evaluated at the true parameter value.

Efficient Score. For a single observation X, the efficient score at θ_0 is

$$S := \frac{\partial}{\partial \theta} \log f(X \mid \theta_0).$$

Under correct specification and standard regularity conditions, the efficient score plays a central role in efficiency bounds.

Theorem 11 (Hansen, Thm. 10.40)**.** *Assume the model is correctly specified, the support of X does not depend on θ, and θ_0 lies in the interior of Θ. Then the efficient score satisfies*

$$\mathbb{E}(S) = 0.$$

Proof. TBW. □

Fisher Information. The *Fisher information* is the variance of the efficient score,

$$\mathcal{I}_\theta = \mathbb{E}(SS').$$

It measures the amount of information about θ contained in a single observation and provides the benchmark for efficiency.

Expected Hessian. The population counterpart of the Hessian is defined as

$$\mathcal{H}_\theta := -\frac{\partial^2}{\partial\theta\,\partial\theta'}\,\ell(\theta_0),$$

where $\ell(\theta)$ denotes the expected log–density.

Theorem 12 (Hansen, Thm. 10.5; Information Matrix Equality)**.** *Assume the model is correctly specified and the support of X does not depend on θ. Then*

$$\mathcal{I}_\theta = \mathcal{H}_\theta,$$

where

$$\mathcal{I}_\theta := \mathbb{E}_\theta\big[S(X,\theta)S(X,\theta)'\big], \qquad \mathcal{H}_\theta := -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta\,\partial\theta'}\log f(X \mid \theta)\right].$$

This equality shows that, under correct specification, the curvature of the log–likelihood and the variability of the score encode the same information.

Proof. TBW. □

Theorem 13 (Hansen, Thm. 10.6; Cramér–Rao Lower Bound)**.** *Assume the model is correctly specified, the support of X does not depend on θ, and θ_0 lies in the interior of Θ. If $\tilde{\theta}$ is an unbiased estimator of θ, then*

$$\operatorname{var}(\tilde{\theta}) \;\geq\; (n\mathcal{I}_\theta)^{-1}.$$

Proof. TBW. □

Definition (Hansen, Def. 10.8)**.** The Cramér–Rao Lower Bound (CRLB) is $(n\mathcal{I}_\theta)^{-1}$.

Definition (Hansen, Def. 10.9). An unbiased estimator $\tilde{\theta}$ is *Cramér–Rao efficient* if it attains the lower bound:

$$\operatorname{var}(\tilde{\theta}) = (n\mathcal{I}_\theta)^{-1}.$$

This yields a fundamental conclusion: within the class of unbiased estimators, the minimal achievable variance is determined by the inverse Fisher information scaled by n. Fisher information therefore provides the natural limit for the precision of unbiased estimation.

Consistency

To study the consistency of the MLE, it is convenient to normalize the log–likelihood by the sample size and work with the *average* log–likelihood

$$\bar{\ell}_n(\theta) := \frac{1}{n}\,\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i \mid \theta).$$

If X_i are i.i.d. and $\log f(X_i \mid \theta)$ is a measurable transformation of X_i, then $\log f(X_i \mid \theta)$ are also i.i.d. By the Weak Law of Large Numbers,

$$\bar{\ell}_n(\theta) \xrightarrow{p} \ell(\theta) := \mathbb{E}[\log f(X \mid \theta)] \quad \text{for each fixed } \theta.$$

Recall from the likelihood analog principle that, under correct specification, the true parameter θ_0 maximizes the population objective $\ell(\theta)$. The MLE $\hat{\theta}$ maximizes the sample objective $\bar{\ell}_n(\theta)$. A natural question is therefore whether the maximizer of $\bar{\ell}_n$ converges to the maximizer of ℓ, i.e. whether $\hat{\theta} \to_p \theta_0$. The next theorem gives sufficient conditions.

Theorem 14 (Consistency of the MLE; Hansen, Thm. 10.8). *Assume the following:*

1. *X_i are i.i.d.*
2. *$\mathbb{E}(\log f(X \mid \theta)) \leq G(X)$ for some integrable function G, with $\mathbb{E}(G(X)) < \infty$.*
3. *$\log f(X \mid \theta)$ is continuous in θ almost everywhere.*
4. *The parameter space Θ is compact.*
5. *For all $\theta \neq \theta_0$, we have $\ell(\theta) < \ell(\theta_0)$.*

Then the maximum likelihood estimator is consistent:

$$\hat{\theta} \xrightarrow{p} \theta_0.$$

Remark (Role of the assumptions). Assumption (ii) guarantees that the log–density has a finite expectation, which is needed to apply the WLLN to $\log f(X_i \mid \theta)$. Assumption (iii), combined with (ii) and compactness of Θ in (iv), allows one to strengthen the pointwise LLN into a *uniform* law of large numbers for $\bar{\ell}_n(\theta)$. Finally, (v) is an identification assumption: it ensures that the population objective $\ell(\theta)$ has a unique maximizer at θ_0, so that the maximizer of the sample objective must converge to this unique population maximizer.

Asymptotic Normality

The previous result establishes consistency. To obtain distributional approximations for inference, we impose stronger smoothness conditions on the likelihood. The next theorem summarizes the classical result.

Theorem 15 (10.9 in Hansen textbook)**.** *Assume the conditions of Theorem 10.8 hold and, in addition,*

1. $\mathbb{E}\left\|\frac{\partial}{\partial\theta}\log f(X \mid \theta_0)\right\|^2 < \infty$
2. $\mathcal{H}_\theta$ *is continuous in* $\theta \in \mathcal{N}$
3. $\frac{\partial}{\partial\theta}\log f(X \mid \theta)$ *is Lipschitz-continuous in* $\mathcal{N}$
4. $\mathcal{H}_{\theta_0} > 0$
5. θ_0 *lies in the interior of* Θ
6. $I_\theta = \mathcal{H}_\theta$

Then

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

The conditions above provide the ingredients for a Taylor expansion of the score around θ_0 and ensure that both the score and Hessian behave suitably for the Central Limit Theorem to apply. The result states that the MLE is asymptotically normal with covariance equal to the inverse of the Fisher information.

Asymptotic Cramér–Rao Efficiency

Definition (10.10 in Hansen textbook)**.** An estimator $\tilde{\theta}$ is *asymptotically Cramér–Rao efficient* if

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} Z, \qquad \mathbb{E}(Z) = 0, \quad \mathrm{Var}(Z) = \mathcal{I}_{\theta_0}^{-1}.$$

Theorem 16 (10.10 in Hansen textbook)**.** *Under the conditions of Theorem 10.9, the MLE is asymptotically Cramér–Rao efficient.*

This result is important because the MLE is generally *not* unbiased in finite samples, yet asymptotically it achieves the smallest possible variance among all regular unbiased estimators. The caveat, of course, is that efficiency is derived within a parametric model; if the model is misspecified, alternative estimators may perform better.

Variance Estimation

In practice, the asymptotic variance $V = \mathcal{I}_{\theta_0}^{-1} = \mathcal{H}_{\theta_0}^{-1}$ is unknown and must be estimated. There are two common approaches:

- **Sample Hessian Estimator**:

$$\hat{V}_1 = \hat{\mathcal{H}}_\theta^{-1} = \left(\frac{1}{n} \sum_{i=1}^{n} \left(-\frac{\partial^2}{\partial\theta\partial\theta'} \log f(X_i \mid \hat{\theta}) \right) \right)^{-1} = \left(-\frac{1}{n} \frac{\partial^2}{\partial\theta\partial\theta'} \ell_n(\hat{\theta}) \right)^{-1}.$$

- **Outer Product Estimator**:

$$\hat{V}_2 = \mathcal{I}_\theta^{-1} = \left(\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial}{\partial\theta} \log f(X_i \mid \hat{\theta}) \right) \left(\frac{\partial}{\partial\theta} \log f(X_i \mid \hat{\theta}) \right)' \right)^{-1}.$$

Both estimators can be shown to be consistent for V.

Variance Estimation: Poisson Example

Consider the model $X_i \sim \text{Poisson}(\lambda)$ with pmf

$$f(x_i \mid \lambda) = \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}.$$

Then

$$\log f(X_i \mid \lambda) = -\lambda + X_i \log \lambda - \log(X_i!),$$

so the score and curvature per observation are

$$S_i(\lambda) = \frac{\partial}{\partial\lambda} \log f(X_i \mid \lambda) = -1 + \frac{X_i}{\lambda}, \qquad \frac{\partial^2}{\partial\lambda^2} \log f(X_i \mid \lambda) = -\frac{X_i}{\lambda^2}.$$

The sample Hessian estimator is

$$\hat{V}_1 = \left(\frac{1}{n} \sum_{i=1}^{n} \left(-\frac{\partial^2}{\partial\lambda^2} \log f(X_i \mid \hat{\lambda}) \right) \right)^{-1} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{\hat{\lambda}^2} \right)^{-1}.$$

Since $\hat{\lambda} = \bar{X}_n$,

$$\frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{\hat{\lambda}^2} = \frac{\bar{X}_n}{\hat{\lambda}^2} = \frac{\hat{\lambda}}{\hat{\lambda}^2} = \frac{1}{\hat{\lambda}},$$

and therefore

$$\hat{V}_1 = \hat{\lambda}.$$

Week 8

Evaluating Estimators

Different estimation procedures may produce different estimators for the same parameter. Sometimes they coincide, but often they do not. To compare them, it is useful to develop criteria that assess the statistical quality of an estimator.

In what follows, we review the standard properties used to evaluate estimators and illustrate how these criteria apply in common examples.

Bias

Definition (Bias)**.** Let W be an estimator of a parameter θ. The *bias* of W is

$$\text{Bias}_\theta(W) \equiv \mathbb{E}_\theta(W) - \theta.$$

If $\text{Bias}_\theta(W) = 0$ for all θ, the estimator is called *unbiased*. Unbiasedness is often desirable because it means the estimator hits the true parameter on average. However, an unbiased estimator is not necessarily preferable: in many situations we are willing to sacrifice unbiasedness for lower variance.

Theorem 17 (6.2 in Hansen)**.** *If $\hat{\theta}$ is an unbiased estimator of θ, then*

$$\hat{\beta} = a\hat{\theta} + b$$

is an unbiased estimator of $\beta = a\theta + b$.

Proof. TBW. □

Note. Nonlinear transformations usually do *not* preserve unbiasedness.

BLUE: Best Linear Unbiased Estimator

Theorem 18 (6.3 in Hansen; BLUE)**.** *If $\sigma^2 < \infty$, the sample mean $\bar{X}_n$ has the lowest variance among all linear unbiased estimators of μ.*

Proof. TBW. □

Mean Squared Error

Definition (Mean Squared Error)**.** The *mean squared error* (MSE) of an estimator W of θ is

$$\text{MSE}_\theta(W) = \mathbb{E}_\theta\left[(W - \theta)^2\right].$$

The MSE is the average squared deviation between W and the true parameter θ. It decomposes as

$$\mathbb{E}_\theta(W-\theta)^2 = \mathbf{Var}_\theta(W) + \big(\mathbb{E}_\theta(W)-\theta\big)^2 = \mathbf{Var}_\theta(W) + \text{Bias}_\theta(W)^2.$$

This decomposition shows that the MSE combines two aspects of estimation quality:

- *variance* (precision), and
- *bias* (accuracy).

Because these components may trade off against one another, an estimator with small bias but large variance may be worse in MSE than a biased estimator with lower variance.

For unbiased estimators, the decomposition reduces to

$$\text{MSE}_\theta(W) = \mathbf{Var}_\theta(W).$$

Example (Normal MSE)**.** Let $X_1,\ldots,X_n$ be i.i.d. $\mathcal{N}(\mu,\sigma^2)$. The statistics $\bar{X}$ and S^2 are unbiased:

$$\mathbb{E}(\bar{X}) = \mu, \qquad \mathbb{E}(S^2) = \sigma^2.$$

Hence their MSEs equal their variances:

$$\mathbb{E}(\bar{X}-\mu)^2 = \mathbf{Var}(\bar{X}) = \frac{\sigma^2}{n},$$

$$\mathbb{E}(S^2-\sigma^2)^2 = \mathbf{Var}(S^2) = \frac{2\sigma^4}{n-1}.$$

Example (Normal MSE — MLE for σ^2)**.** Consider instead the MLE for σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{n}{n-1}S^2.$$

Its expectation is

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2,$$

so the estimator is biased downward. However, the variance is

$$\mathbf{Var}(\hat{\sigma}^2) = \mathbf{Var}\left(\frac{n-1}{n}S^2\right) = \left(\frac{n-1}{n}\right)^2\mathbf{Var}(S^2) = \frac{2(n-1)\sigma^4}{n^2}.$$

Combining bias and variance gives

$$\mathbb{E}(\hat{\sigma}^2-\sigma^2)^2 = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2-\sigma^2\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4.$$

Since

$$\mathbb{E}(\hat{\sigma}^2-\sigma^2)^2 < \mathbb{E}(S^2-\sigma^2)^2,$$

the MLE for σ^2 has a *smaller* MSE than S^2, even though it is biased.

Best Unbiased Estimators

The idea of identifying a "best" estimator in terms of mean squared error (MSE) is appealing, but it is only feasible once we restrict attention to a particular class of estimators. A natural restriction is the class of *unbiased* estimators. Within this class, comparing two estimators reduces to comparing their variances: the unbiased estimator with the smallest variance is then the best.

Definition (Best Unbiased Estimator; UMVUE)**.** An estimator W^* is a *best unbiased estimator* of a function $\tau(\theta)$ if:

$$\mathbb{E}_\theta(W^*) = \tau(\theta) \quad \text{for all } \theta,$$

and for any other unbiased estimator W satisfying $\mathbb{E}_\theta(W) = \tau(\theta)$, we have

$$\mathbf{Var}_\theta(W^*) \;\leq\; \mathbf{Var}_\theta(W) \quad \text{for all } \theta.$$

Such an estimator is also called a *uniform minimum variance unbiased estimator* (UMVUE) of $\tau(\theta)$.

Note. A UMVUE may not exist, and when it exists it may not be unique.

A useful strategy for finding the best unbiased estimator is to rely on a lower bound for the variance of all unbiased estimators. Last week we studied the Cramér–Rao lower bound (CRLB). Any unbiased estimator that achieves this bound must be the best unbiased estimator.

However, the CRLB may not always be attainable. In some models, no unbiased estimator reaches the bound. A standard illustration comes from the normal distribution: the CRLB for σ^2 (equal to $2\sigma^4/n$) is attainable only when μ is known. If μ is unknown, no unbiased estimator of σ^2 can attain this lower bound.

Loss Functions

The mean squared error is one particular example of a *loss function*. Loss functions are part of the broader framework of decision theory.

After observing data $X = x$, drawn from $f(x \mid \theta)$ with $\theta \in \Theta$, the statistician chooses an action a in an action space $\mathcal{A}$. In point estimation, the action represents the proposed estimate. The loss incurred from reporting a when the true parameter is θ is denoted $L(\theta, a)$, and should be small whenever a is close to θ.

Common loss functions include:

- **Absolute error loss:**
$$L(\theta, a) = |a - \theta|.$$

- **Squared error loss:**
$$L(\theta, a) = (a - \theta)^2.$$

More general asymmetric loss functions may penalize overestimation and underestimation differently.

Risk Functions

In decision-theoretic analysis, the quality of an estimator is described by its *risk function*:

$$R(\theta, \delta) = \mathbb{E}_\theta\big[\, L(\theta, \delta(X))\, \big].$$

For a given θ, $R(\theta, \delta)$ is the expected loss incurred if estimator δ is used.

We prefer estimator δ_1 to δ_2 if

$$R(\theta, \delta_1) < R(\theta, \delta_2) \qquad \text{for all } \theta \in \Theta.$$

For squared error loss, the risk coincides with the mean squared error:

$$R(\theta, \delta) = \mathbb{E}_\theta\big[(\delta(X) - \theta)^2\big] = \mathbf{Var}_\theta(\delta(X)) + \big(\mathrm{Bias}_\theta(\delta(X))\big)^2.$$

Example (Stein's Loss)**.** Assume $X_1, \ldots, X_n$ are i.i.d. with finite variance σ^2. Stein's loss for estimating σ^2 is

$$L(\sigma^2, a) = \frac{a}{\sigma^2} - 1 - \log\Big(\frac{a}{\sigma^2}\Big).$$

Consider estimators of the form $\delta_b = bS^2$, where S^2 is the usual unbiased estimator of variance. Then the risk under Stein's loss is

$$R(\sigma^2, \delta_b) = \mathbb{E}\bigg(\frac{bS^2}{\sigma^2} - 1 - \log\Big(\frac{bS^2}{\sigma^2}\Big)\bigg) = b\,\mathbb{E}\Big(\frac{S^2}{\sigma^2}\Big) - 1 - \mathbb{E}\Big[\log\Big(\frac{bS^2}{\sigma^2}\Big)\Big].$$

Since $\mathbb{E}(S^2/\sigma^2) = 1$,

$$R(\sigma^2, \delta_b) = b - \log b - 1 - \mathbb{E}\Big[\log\Big(\frac{S^2}{\sigma^2}\Big)\Big].$$

The expression is minimized at $b = 1$, so $\delta_1 = S^2$ is optimal among estimators of the form bS^2.

Bayes Risk

From Week 7, recall that a Bayesian framework evaluates estimators by averaging over the posterior distribution. Given a prior $\pi(\theta)$ and data X, the posterior is $\pi(\theta \mid X)$. The *Bayes risk* of an estimator T is the expected posterior loss:

$$R(T \mid X) = \int_\Theta \ell(T, \theta)\, \pi(\theta \mid X)\, d\theta.$$

For a given loss function $\ell(T, \theta)$, the optimal Bayes estimator is the one that minimizes $R(T \mid X)$.

Example (7.3.28)**.** With squared error loss,

$$\int_{\Theta} (\theta - a)^2 \, \pi(\theta \mid X) \, d\theta = \mathbb{E}\big[(\theta - a)^2 \mid X\big].$$

This is minimized at $a = \mathbb{E}(\theta \mid X)$, the posterior mean. Thus, under squared error loss, the Bayes estimator is the posterior expectation.

Week 9: Linear Regression

Introduction to Simple Linear Regression

In simple linear regression, we study the functional dependence of one variable on another. Given observed pairs (X_i, Y_i), the model postulates a linear relationship of the form

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

where:

- Y_i is the response (dependent) variable,
- X_i is the predictor (independent) variable,
- α is an unknown intercept,
- β is an unknown slope,
- ε_i is a disturbance term with $\mathbb{E}[\varepsilon_i] = 0$.

Population Regression Function

The expected value of Y_i conditional on $X_i = x_i$ is

$$\mathbb{E}(Y_i \mid X_i = x_i) = \alpha + \beta x_i.$$

This expression is known as the *population regression function*. It represents the conditional expectation of Y given $X = x$, under the assumption that the relationship between Y and X is linear in the parameters.

Summarizing Sample Data

Given sample data $(x_1, y_1), \ldots, (x_n, y_n)$, we define the sample means

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

We also define the sums of squares:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

and the cross-product sum

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Least Squares Estimation

A residual e_i measures the vertical distance between each data point (x_i, y_i) and the fitted line:

$$e_i = y_i - (\alpha + \beta x_i).$$

The *sum of squared residuals* (SSR) is

$$\text{SSR} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2.$$

The least squares estimators of α and β minimize SSR with respect to both parameters.

Derivative with Respect to the Intercept

The first-order condition with respect to α is:

$$\frac{\partial \text{SSR}}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i)) = 0.$$

Simplifying,

$$\sum_{i=1}^{n} y_i = n\alpha + \beta \sum_{i=1}^{n} x_i,$$

which yields

$$\alpha = \bar{y} - \beta \bar{x}.$$

Derivative with Respect to the Slope

The first-order condition with respect to β is:

$$\frac{\partial \text{SSR}}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (y_i - (\alpha + \beta x_i)) = 0.$$

Substituting $\alpha = \bar{y} - \beta \bar{x}$ and simplifying gives

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}.$$

Least Squares Estimates

Putting everything together:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}.$$

These coefficients define the fitted line $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, which minimizes the sum of squared residuals.

Residuals and Their Properties

Residuals are defined as

$$e_i = y_i - \hat{\alpha} - \hat{\beta} x_i.$$

Two key properties follow directly from the first-order conditions:

$$\sum_{i=1}^{n} e_i = 0, \qquad \sum_{i=1}^{n} (x_i - \bar{x}) e_i = 0.$$

These identities express that the residuals sum to zero and are orthogonal to the regressor (after centering).

Three Sums of Squares

To evaluate the performance of least squares, we compare the sum of squared residuals with the total variation in y_i.

Starting from the fitted model

$$y_i = \hat{\alpha} + \hat{\beta} x_i + e_i,$$

we subtract $\bar{y}$ and use $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ to obtain

$$y_i - \bar{y} = \hat{\beta}(x_i - \bar{x}) + e_i.$$

This decomposition implies

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} e_i^2.$$

We define:

$$\text{SST} = \text{SSE} + \text{SSR},$$

where:

- SST: Total Sum of Squares,
- SSE: Explained Sum of Squares,
- SSR: Sum of Squared Residuals.

Coefficient of Determination R^2

The coefficient of determination measures the proportion of variance explained by the model:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = \frac{\hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

Alternatively, using S_{xy} and S_{xx}:

$$R^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}.$$

Another common expression is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

This makes clear that $0 \leq R^2 \leq 1$, and that minimizing the sum of squared residuals is equivalent to maximizing R^2.

Interpretation of the Least Squares Line

If x is the predictor and y is the response, the least squares line yields predictions of y based on x.

The fitted line minimizes the total vertical distance between observed points and the regression line.

Importantly, least squares is primarily a method for fitting; without further assumptions, it does not automatically provide statistical inference such as confidence intervals, hypothesis testing, or causal interpretation.
⇒ The method provides a best-fitting line for the data.
⇒ Additional assumptions are needed for inference.

Least Squares in Matrix Form

The linear model can be written compactly as

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

and the residual vector as

$$e = y - X\hat{\beta}.$$

Matrix dimensions:

- y: $n \times 1$ vector of observations,
- X: $n \times k$ matrix of predictors,
- β: $k \times 1$ vector of parameters,
- e: $n \times 1$ vector of residuals.

Least Squares Criterion

The sum of squared residuals in matrix form is

$$S(\hat{\beta}) = e'e = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - y'X\hat{\beta} - \hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}.$$

Taking the derivative with respect to $\hat{\beta}$:

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0.$$

Solving:

$$X'X\hat{\beta} = X'y, \qquad \hat{\beta} = (X'X)^{-1}X'y.$$

This is the least squares estimator of β.

Note: For $(X'X)^{-1}$ to exist, X must have full column rank k, which requires $n \geq k$.

Least Squares as a Projection

The least squares estimator can be viewed as an *orthogonal projection* of the data vector $y \in \mathbb{R}^n$ onto the column space of X, denoted by

$$S(X) = \{Xa : a \in \mathbb{R}^k\}.$$

Intuitively, $S(X)$ is the set of all linear combinations of the regressors: every element of $S(X)$ is a "candidate fitted value" vector that one can obtain by choosing some coefficient vector a.

We know that the least squares estimator satisfies

$$\hat{\beta} = (X'X)^{-1}X'y, \qquad \hat{y} = X\hat{\beta}.$$

Substituting the expression for $\hat{\beta}$, we obtain

$$\hat{y} = X(X'X)^{-1}X'y.$$

This shows that the fitted values can be written as a linear transformation of y:

$$\hat{y} = Hy,$$

where

$$H = X(X'X)^{-1}X'$$

is called the *hat matrix* or projection matrix onto $S(X)$.

Residual Vector and Annihilator Matrix

The residual vector is

$$e = y - X\hat{\beta} = y - \hat{y}.$$

Using the expression for $\hat{y}$, we can rewrite this as

$$e = y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y.$$

Define the *annihilator matrix* M by

$$M = I - X(X'X)^{-1}X'.$$

Then

$$e = My.$$

The terminology "annihilator" reflects that M kills (annihilates) the component of any vector lying in $S(X)$: it removes the part that is explainable by X, leaving only the orthogonal residual part.

Algebraic Properties of M

The matrix M has the following properties:

- **Symmetric:** $M = M'$.
- **Idempotent:** $M^2 = M$.

Idempotence captures the idea of a projection: once a vector has been projected, projecting it again does nothing.

Moreover,

$$MX = 0,$$

so the columns of X lie in the null space of M. This means that the residuals are orthogonal to the space spanned by X.

Using $e = My$ and $MX = 0$, we have

$$X'e = X'My = 0,$$

which expresses the familiar least squares normal equations: each regressor is orthogonal to the residual vector. Geometrically, the regression plane $S(X)$ and the residual vector e meet at a right angle.

Properties of the Projection Matrix and Residuals

We have already defined the hat matrix

$$H = X(X'X)^{-1}X',$$

so that $\hat{y} = Hy$ and $e = My$ with $M = I - H$.

The matrix H satisfies:

- $H = H'$ (symmetric),
- $H^2 = H$ (idempotent),
- $H + M = I$,
- $HM = MH = 0$.

Thus H and M are complementary orthogonal projections:

- H projects onto $S(X)$,
- M projects onto $S^{\perp}(X)$, the subspace orthogonal to $S(X)$.

Using $H + M = I$, we can decompose y as

$$y = Hy + My = \hat{y} + e.$$

Because H and M project onto orthogonal subspaces, we also have

$$\hat{y}'e = 0,$$

which means that the fitted values and the residuals are orthogonal vectors in $\mathbb{R}^n$.

Geometric Interpretation of Least Squares

The least squares method admits a clear geometric interpretation in $\mathbb{R}^n$:

- Think of the data vector y as a point in $\mathbb{R}^n$.
- The subspace $S(X)$, spanned by the columns of X, is the set of all linear combinations of the regressors. Any vector Xa in $S(X)$ represents what the data would look like if it were *perfectly* explained by X with coefficients a.
- The residual vector is $e = y - \hat{y}$, where $\hat{y} \in S(X)$. The length squared of the residual is

  $$e'e = ||e||^2,$$

  the squared distance from y to the regression subspace $S(X)$.

The least squares estimator chooses $\hat{y} = X\hat{\beta}$ in such a way that $e'e$ is minimized. Geometrically, this means:

- $\hat{y}$ is the orthogonal projection of y onto $S(X)$;
- e is the component of y lying in the orthogonal complement $S^{\perp}(X)$.

Let

$$S(X) = \{Xa : a \in \mathbb{R}^k\}, \qquad S^{\perp}(X) = \{z \in \mathbb{R}^n : X'z = 0\}.$$

Then:

- H projects any y onto $S(X)$: $\hat{y} = Hy \in S(X)$;
- M projects any y onto $S^{\perp}(X)$: $e = My \in S^{\perp}(X)$;
- The decomposition

  $$y = \hat{y} + e, \qquad \hat{y} \in S(X),\ e \in S^{\perp}(X),$$

  expresses y as the sum of two orthogonal components: the explained part and the unexplained (residual) part.

Because of orthogonality, we have a Pythagorean identity:

$$\|y\|^2 = \|\hat{y}\|^2 + \|e\|^2,$$

which, after suitable centering, underlies the decomposition

$$\text{SST} = \text{SSE} + \text{SSR}.$$

Thus the usual variance decomposition in regression is just a geometric statement about the lengths of orthogonal vectors in $\mathbb{R}^n$.

The following picture gives a schematic geometric interpretation of least squares as an orthogonal projection of y onto $S(X)$:

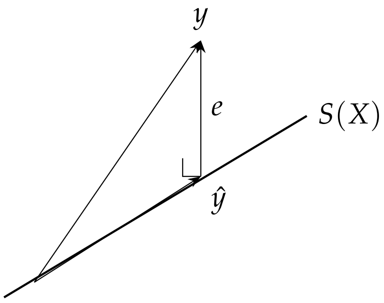

Figure 2: Geometric view: y decomposed into $\hat{y} \in S(X)$ and $e \in S^{\perp}(X)$.

Derivation of R^2

Definition of R^2. The coefficient of determination R^2 measures the fraction of the total sample variation in y that is explained by the model.

In matrix form, the total sample variation can be written as

$$\text{SST} = y'Ny,$$

where

$$N = I - \frac{1}{n}11'$$

is idempotent and 1 is the $n \times 1$ vector of ones. Note that Ny has components $y_i - \bar{y}$, so $y'Ny = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

Decomposition of total variation (SST). Using the regression decomposition

$$y = X\hat{\beta} + e = \hat{y} + e,$$

we obtain

$$y'Ny = \hat{\beta}'X'NX\hat{\beta} + e'e,$$

because $\hat{\beta}'X'Ne = 0$ (since $Ne = e$ and $X'e = 0$). We then define

- $\text{SSE} = \hat{\beta}'X'NX\hat{\beta}$: Sum of Squares Explained,
- $\text{SSR} = e'e$: Sum of Squares Residual.

Hence,

$$\text{SST} = \text{SSE} + \text{SSR}.$$

Coefficient of Determination R^2 and Adjusted R^2

Definition of R^2. Using the decomposition above,

$$R^2 = \frac{\text{SSE}}{\text{SST}} = \frac{\hat{\beta}'X'NX\hat{\beta}}{y'Ny} = 1 - \frac{e'e}{y'Ny} = 1 - \frac{\text{SSR}}{\text{SST}}.$$

Interpretation.

- R^2 represents the proportion of the variance in y that is explained by the model.
- $0 \leq R^2 \leq 1$, with higher values indicating a better in-sample fit.

Adjusted R^2. To account for the number of regressors k and penalize overfitting, we define the adjusted coefficient of determination:

$$\bar{R}^2 = 1 - \frac{e'e/(n-k)}{y'Ny/(n-1)} = 1 - \frac{n-1}{n-k}(1 - R^2).$$

Adjusted R^2 is therefore more appropriate for comparing models with different numbers of predictors, since it increases only when the added variables improve the fit sufficiently after accounting for the loss of degrees of freedom.

Frisch–Waugh–Lovell (FWL) Theorem

There are several additional results that can be derived without imposing further statistical assumptions on the error term. A key one in econometrics is the *Frisch–Waugh–Lovell (FWL) Theorem*, which characterizes how OLS behaves when we include control variables.

Linear Regression Model and Partition of Regressors

Consider the linear regression model

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

where:

- y: dependent variable (an $n \times 1$ vector);
- X_1: regressors of interest;
- X_2: control variables;
- ε: error term.

We can think of $X = [X_1 \ \ X_2]$ as the full regressor matrix, partitioned into variables whose coefficients we care about directly (β_1) and variables that we include only as controls (β_2).

Key Statement of the FWL Theorem

Instead of estimating the full system at once, the OLS coefficient $\hat{\beta}_1$ can be obtained by the following three-step procedure:

1. Regress y on X_2 and obtain the residuals $\tilde{y}$.
2. Regress each column of X_1 on X_2 and obtain the residuals $\tilde{X}_1$.
3. Regress $\tilde{y}$ on $\tilde{X}_1$ to obtain $\hat{\beta}_1$.

Why does this work?

- The FWL theorem uses the orthogonality properties of OLS residuals.
- By "partialling out" X_2 from both y and X_1, we remove the influence of X_2, isolating the remaining linear relationship between y and X_1.
- The result holds because OLS minimizes the residual sum of squares and enforces orthogonality between regressors and residuals.

Algebraic Proof of the FWL Theorem

Let $X = [X_1\ X_2]$ and consider

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Define

$$P_2 = X_2(X_2'X_2)^{-1}X_2', \qquad M_2 = I - P_2,$$

so that P_2 projects onto the column space of X_2 and M_2 projects onto its orthogonal complement.

The OLS first-order conditions for the full model are

$$X'(y - X\hat{\beta}) = 0,$$

which, in block form, are equivalent to

$$\begin{cases} X_1'(y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2) = 0, \\ X_2'(y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2) = 0. \end{cases}$$

Now pre-multiply the first equation by M_2:

$$X_1'M_2(y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2) = X_1'(I - P_2)(y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2) = 0.$$

Using linearity and the fact that $M_2X_2 = 0$ (since M_2 annihilates the space spanned by X_2), we obtain

$$X_1'M_2y - X_1'M_2X_1\hat{\beta}_1 = 0 \quad \Longrightarrow \quad X_1'M_2X_1\hat{\beta}_1 = X_1'M_2y. \tag{$*$}$$

Define

$$\tilde{X}_1 := M_2X_1, \qquad \tilde{y} := M_2y.$$

Since M_2 is symmetric, $M_2' = M_2$, we can rewrite $(*)$ as

$$\tilde{X}_1'\tilde{X}_1\hat{\beta}_1 = \tilde{X}_1'\tilde{y}.$$

These are exactly the normal equations of the regression of $\tilde{y}$ on $\tilde{X}_1$. Therefore, the coefficient $\hat{\beta}_1$ obtained from the full regression on $[X_1\ X_2]$ coincides with the coefficient from the regression of residualized y on residualized X_1.

Implications and Applications

The FWL theorem is widely used in applied econometrics, particularly for understanding causal relationships and for interpreting regression coefficients with controls.

- Suppose you are interested in the relationship between y and X_1, but you know that there is a set of important control variables X_2.

- Instead of plotting y against X_1, you can:

  1. residualize y with respect to X_2, obtaining $\tilde{y}$,
  2. residualize X_1 with respect to X_2, obtaining $\tilde{X}_1$,
  3. plot $\tilde{y}$ against $\tilde{X}_1$.

  This plot visualizes the "partialled-out" relationship, i.e. the part of y and X_1 that remains after removing linear effects of X_2.

- For instance, one might plot residualized wages (after controlling for age, gender, race, etc.) against residualized years of education: the slope in this plot corresponds to the education coefficient in the full regression.

OLS: from Line Fitting to Statistics

Up to now, least squares has been introduced as a purely geometric or algebraic line-fitting procedure: choose $\hat{\beta}$ to minimize the sum of squared residuals and obtain the projection $\hat{y} = X\hat{\beta}$.

To turn this into a *statistical* procedure, we now add assumptions about the data-generating process and interpret OLS as an estimator of an underlying population parameter β. In particular, we show how OLS arises as:

- a method of moments (MoM) estimator,
- an unbiased estimator under suitable exogeneity assumptions,
- an estimator with a well-defined (finite-sample or asymptotic) variance, which can be computed under homoskedasticity or in a heteroskedasticity-robust way.

OLS as a Method of Moments Estimator

Moment Conditions and Estimator

Ordinary Least Squares (OLS) can be interpreted as a method of moments estimator.

Assume the linear model

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

and impose the moment (exogeneity) condition

$$\mathbb{E}[X'\varepsilon] = 0 \quad \Longleftrightarrow \quad \mathbb{E}[X'(y - X\beta)] = 0.$$

Intuitively, this assumption says that the regressors X are uncorrelated with the error term ε; there is no systematic relationship between the regressors and the part of y left unexplained by the model.

The method of moments idea is to replace the population expectation by its sample analog and then solve for β. The sample analog of the moment condition is

$$\frac{1}{n}\sum_{i=1}^{n} X_i(y_i - X_i'\beta) = 0 \quad \Longleftrightarrow \quad X'(y - X\beta) = 0.$$

Solving these equations yields

$$\hat{\beta} = (X'X)^{-1}X'y,$$

which is exactly the OLS estimator. Thus OLS is the method of moments estimator associated with the (vector) moment condition $\mathbb{E}[X'(y - X\beta)] = 0$.

Unbiasedness of OLS

The OLS estimator can be written as

$$\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon.$$

To study its expectation, it is convenient to condition on X and then apply the iterated law of expectations:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}\big[\,\mathbb{E}[\hat{\beta} \mid X]\,\big].$$

Substituting the expression for $\hat{\beta}$,

$$\mathbb{E}[\hat{\beta} \mid X] = \beta + (X'X)^{-1}X'\mathbb{E}[\varepsilon \mid X].$$

Assume exogeneity: $\mathbb{E}[\varepsilon \mid X] = 0$. Then

$$\mathbb{E}[\hat{\beta} \mid X] = \beta \quad \Longrightarrow \quad \mathbb{E}[\hat{\beta}] = \beta.$$

Hence OLS is unbiased under the conditional mean independence assumption $\mathbb{E}[\varepsilon \mid X] = 0$.

Variance of the OLS Estimator

We now derive the variance of $\hat{\beta}$. Using the law of total variance,

$$\mathrm{Var}(\hat{\beta}) = \mathbb{E}\big[\mathrm{Var}(\hat{\beta} \mid X)\big] + \mathrm{Var}\left(\mathbb{E}[\hat{\beta} \mid X]\right).$$

From the previous step, $\mathbb{E}[\hat{\beta} \mid X] = \beta$, so

$$\mathrm{Var}\left(\mathbb{E}[\hat{\beta} \mid X]\right) = 0$$

and

$$\mathrm{Var}(\hat{\beta}) = \mathbb{E}\big[\mathrm{Var}(\hat{\beta} \mid X)\big].$$

Using $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$, we obtain

$$\operatorname{Var}(\hat{\beta} \mid X) = (X'X)^{-1}X' \operatorname{Var}(\varepsilon \mid X)X(X'X)^{-1}.$$

Homoskedasticity. Suppose

$$\operatorname{Var}(\varepsilon \mid X) = \sigma^2 I$$

for some scalar σ^2 (errors have constant variance and are uncorrelated across observations). Then

$$\operatorname{Var}(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1},$$

and consequently

$$\operatorname{Var}(\hat{\beta}) = \sigma^2 \mathbb{E}\big[(X'X)^{-1}\big].$$

In practice, we estimate σ^2 by the residual-based estimator

$$s^2 = \frac{1}{n-k} e'e,$$

and use

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

as the estimated covariance matrix under homoskedasticity.

Heteroskedasticity. More generally, we may have

$$\operatorname{Var}(\varepsilon \mid X) = \Omega,$$

where Ω is an $n \times n$ covariance matrix that allows for heteroskedasticity (and possibly correlation) across observations. Then

$$\operatorname{Var}(\hat{\beta} \mid X) = (X'X)^{-1}X'\Omega X(X'X)^{-1}.$$

This expression motivates heteroskedasticity-robust variance estimators.

OLS as a Method of Moments Estimator (Sandwich Variance)

Finally, we connect OLS to the general method of moments / GMM variance formula.

Recall: for a just-identified MoM/GMM estimator $\hat{\theta}$ with moment function $m(Z_i, \theta) \in \mathbb{R}^k$, the asymptotic variance can be written as

$$\hat{V} = (\hat{Q}^{-1})' \hat{\Omega}\, \hat{Q}^{-1},$$

where

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\theta}) m(Z_i, \hat{\theta})', \qquad \hat{Q} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta} m(Z_i, \hat{\theta})'.$$

For OLS, take $Z_i = (X_i, y_i)$ and

$$m(X_i, y_i, \beta) = X_i(y_i - X_i'\beta),$$

so that the moment condition $\mathbb{E}[m(X_i, y_i, \beta)] = 0$ is exactly $\mathbb{E}[X_i(y_i - X_i'\beta)] = 0$.

With residuals $e_i = y_i - X_i'\hat{\beta}$, we have

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 X_i X_i'.$$

The Jacobian with respect to β is

$$\frac{\partial}{\partial\beta} m(X_i, y_i, \beta)' = -X_i X_i',$$

so

$$\hat{Q} = -\frac{1}{n}\sum_{i=1}^{n} X_i X_i' = -\frac{1}{n}X'X.$$

Substituting these into the GMM variance formula yields the heteroskedasticity-robust (sandwich) covariance estimator:

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

This expression coincides with the familiar Eicker–White heteroskedasticity-robust variance estimator for OLS: it replaces $\sigma^2 I$ with an empirical estimate of Ω constructed from the squared residuals e_i^2 and the regressors X_i.