ECON 121 – Cohesive Notes (Draft 6)

Applied Econometrics and Data Analysis

1 Why This Course Looks Different

ECON 121 moves from "run a regression and report one coefficient" to a more disciplined empirical workflow:

1. define the target object (prediction, association, or causal effect),
2. choose an estimator that targets it under explicit assumptions,
3. quantify uncertainty with standard errors that match the data environment.

The course then builds from this foundation into binary choice models, panel methods, causal estimands, instrumental variables, and regression discontinuity designs.

Frequentist Lens

Most of these notes use repeated-sampling logic. Parameters are fixed unknown constants. Estimators are random because samples vary. Properties such as unbiasedness and consistency are statements about what happens across hypothetical repetitions of the data collection process.

2 Statistical Foundations for Econometrics

2.1 Population vs Sample Objects

Let $X_1, \ldots, X_N$ be an i.i.d. sample from a population with mean μ and variance σ^2.

- **Parameters:** μ, σ^2, β, etc. (fixed, unknown).
- **Estimators:** $\bar{X}, s^2, \hat{\beta}$, etc. (random variables).

This distinction is not cosmetic. Once we treat estimators as random variables, we can ask meaningful questions: Is the estimator centered correctly? Does it concentrate with larger N? How much uncertainty remains?

2.2 Three Benchmark Properties

- **Unbiasedness:** $E[\hat{\theta}] = \theta$.
- **Consistency:** $\hat{\theta} \xrightarrow{p} \theta$ as $N \to \infty$.
- **Precision:** lower $V(\hat{\theta})$ means tighter intervals and stronger power.

Remark 1. These need not align in finite samples. It is common to accept small bias to reduce variance enough to improve mean-squared error.

2.3 Proof 1: Mean Estimation

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Proposition 1. *If X_i are i.i.d. with mean μ and variance σ^2, then*

$$E[\bar{X}] = \mu, \qquad V(\bar{X}) = \frac{\sigma^2}{N}.$$

Proof. Linearity of expectation gives

$$E[\bar{X}] = \frac{1}{N}\sum_{i=1}^{N} E[X_i] = \mu.$$

For variance,

$$V(\bar{X}) = V\left(\frac{1}{N}\sum_{i} X_i\right) = \frac{1}{N^2}\sum_{i} V(X_i) + \frac{2}{N^2}\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

Under independence, covariance terms are zero, so

$$V(\bar{X}) = \frac{1}{N^2} \cdot N\sigma^2 = \frac{\sigma^2}{N}.$$

□

2.4 Proof 2: Why $N-1$ Appears in Sample Variance

Define

$$\tilde{s}^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2, \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2.$$

Proposition 2.

$$E[\tilde{s}^2] = \frac{N-1}{N}\sigma^2, \qquad E[s^2] = \sigma^2.$$

Proof. Use

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = \sum_{i=1}^{N}(X_i - \mu)^2 - N(\bar{X} - \mu)^2.$$

Taking expectations:

$$E\left[\sum_{i=1}^{N}(X_i - \bar{X})^2\right] = N\sigma^2 - NV(\bar{X}) = N\sigma^2 - \sigma^2 = (N-1)\sigma^2.$$

Divide by N or $N-1$ accordingly. □

2.5 Bias-Variance Decomposition

For any estimator $\hat{\theta}$,

$$E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + \left(E[\hat{\theta}] - \theta\right)^2.$$

This identity explains why a slightly biased estimator can still be preferable when variance reduction is large.

2.6 Inference Template

Most empirical inference in these notes follows

$$t = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}.$$

Using large-sample normal approximation:

- two-sided 5% test rejects when $|t| > 1.96$,
- 95% CI is $\hat{\theta} \pm 1.96\, SE(\hat{\theta})$.

Remark 2. Hypothesis tests and confidence intervals are dual. A value θ_0 is rejected at level α if and only if it is outside the $(1-\alpha)$ interval.

2.7 CEF Bridge

Textbook perspective: the conditional expectation function $m(x) = E[Y \mid X = x]$ is the minimum-MSE predictor among all measurable functions of X. Linear regression is the minimum-MSE predictor within the restricted class of linear functions.

This is why linear regression can be useful even when the true CEF is not linear: it remains the best linear summary.

3 The Bivariate Linear Model

3.1 Model and Objective

$$Y_i = \beta_0 + \beta_1 X_i + U_i, \qquad (\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{b_0, b_1} \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2.$$

3.2 Normal Equations and Closed Form

Proposition 3. *The first-order conditions are*

$$\sum_{i=1}^{N} \hat{u}_i = 0, \qquad \sum_{i=1}^{N} X_i \hat{u}_i = 0,$$

with $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$.

Proof. Differentiate the objective:

$$\frac{\partial}{\partial b_0} \sum_i (Y_i - b_0 - b_1 X_i)^2 = -2 \sum_i (Y_i - b_0 - b_1 X_i),$$

$$\frac{\partial}{\partial b_1} \sum_i (Y_i - b_0 - b_1 X_i)^2 = -2 \sum_i X_i (Y_i - b_0 - b_1 X_i).$$

Set both equal to zero at the optimum. □

Proposition 4.

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Proof. From $\sum_i \hat{u}_i = 0$, obtain $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. Substitute in the second normal equation and simplify. □

3.3 Interpretation

$\hat{\beta}_1$ is the covariance of X and Y scaled by the variance of X. This makes sign intuition immediate:

- if high X tends to come with high Y, slope is positive;
- if high X tends to come with low Y, slope is negative;
- if X barely varies, slope is unstable (denominator is small).

3.4 Unbiasedness and Consistency Logic

Assume $E[U_i \mid X_i] = 0$. Then

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_i (X_i - \bar{X}) U_i}{\sum_i (X_i - \bar{X})^2}.$$

Conditioning on X, the numerator has mean zero; this gives unbiasedness in the classical model. For consistency, the numerator is $O_p(\sqrt{N})$, denominator is $O_p(N)$, so ratio is $o_p(1)$.

3.5 Variance and Standard Errors

Under homoskedasticity and independence:

$$V(\hat{\beta}_1 \mid X) = \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}.$$

Replacing σ^2 by

$$s^2 = \frac{1}{N-2} \sum_i \hat{u}_i^2$$

gives the usual estimated standard error.

3.6 Robust vs Conventional SE

If heteroskedasticity is present, conventional SE can be inconsistent. The robust (Eicker-White) covariance estimator in matrix form is

$$\widehat{V}_{\text{rob}}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{i=1}^{N} \hat{u}_i^2 x_i x_i' \right) (X'X)^{-1}.$$

For bivariate OLS this reduces to a weighted residual-squared expression where high-leverage observations can matter a lot.

Remark 3. Practical default in applied work: report robust SE unless there is a strong reason not to.

3.7 Worked Example: Test and CI

Suppose $\hat{\beta}_1 = 0.08$, $SE(\hat{\beta}_1) = 0.03$.

$$t = \frac{0.08}{0.03} = 2.67.$$

At the 5% level, this rejects $H_0 : \beta_1 = 0$ under normal approximation. The 95% CI is

$$0.08 \pm 1.96(0.03) = [0.0212,\ 0.1388].$$

The interval excludes zero and gives effect-size uncertainty, not just significance.

4 Multiple Regression and Omitted Variable Bias

4.1 From One Regressor to Many

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i.$$

Interpretation becomes conditional: β_k is the change in Y from changing X_k by one unit while holding all other included regressors fixed.

4.2 No Perfect Multicollinearity

If one regressor is an exact linear combination of others, $X'X$ is singular and $(X'X)^{-1}$ does not exist. Econometrically, the model cannot disentangle separate effects from perfectly overlapping variation.

4.3 Core OVB Derivation

Long model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U.$$

Estimated short model:

$$Y = \delta_0 + \delta_1 X_1 + V.$$

Auxiliary relationship:

$$X_2 = \pi_0 + \pi_1 X_1 + A.$$

Proposition 5.

$$\delta_1 = \beta_1 + \beta_2 \pi_1.$$

Proof. Substitute the auxiliary equation into the long equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(\pi_0 + \pi_1 X_1 + A) + U.$$

Collect terms:

$$Y = (\beta_0 + \beta_2 \pi_0) + (\beta_1 + \beta_2 \pi_1) X_1 + (U + \beta_2 A).$$

Compare with $Y = \delta_0 + \delta_1 X_1 + V$. □

4.4 Sign Logic for OVB

Bias sign is $\text{sign}(\beta_2 \pi_1)$.

- $\beta_2 > 0, \pi_1 > 0$: upward bias.
- $\beta_2 > 0, \pi_1 < 0$: downward bias.
- $\beta_2 = 0$ or $\pi_1 = 0$: no OVB.

Example 1. True effect $\beta_1 = 0.05$, omitted ability effect $\beta_2 = 0.20$, and $\pi_1 = 0.30$ (ability correlated with schooling). Then short-regression estimand is

$$\delta_1 = 0.05 + 0.20(0.30) = 0.11.$$

Estimated return to schooling is more than double the true structural effect.

4.5 Confounders vs Mediators

- **Confounder:** causes both treatment and outcome. Omission tends to bias treatment effect estimates.
- **Mediator:** lies on the causal pathway from treatment to outcome. Including it changes the target parameter from total effect toward direct effect.

Hence "include all controls" is not a universal rule; control choice depends on the causal estimand.

4.6 Matrix Form of OLS

Stack data:

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

This compact form clarifies:

- estimation is projection of Y onto the span of X,
- variance formulas depend on $X'X$ and the error covariance structure.

4.7 Linear Combinations and Delta-Style Inference

Applied questions often involve $l'\beta$: wage gap at a profile, difference of two treatment coefficients, predicted effect at a covariate vector. Given covariance estimate $\widehat{V}(\hat{\beta})$,

$$\widehat{V}(l'\hat{\beta}) = l'\widehat{V}(\hat{\beta})l, \qquad SE(l'\hat{\beta}) = \sqrt{l'\widehat{V}(\hat{\beta})l}.$$

4.8 FWL Intuition (Proof Sketch)

Frisch-Waugh-Lovell says the coefficient on X_1 in a full regression equals the slope from:

1. residualizing Y on other controls,
2. residualizing X_1 on other controls,
3. regressing residualized Y on residualized X_1.

This explains "holding controls fixed" geometrically: we remove variation explained by controls before estimating the slope.

5 Nonparametric Density and Regression Estimation

5.1 Motivation

Linear models are low-variance but can impose too much structure. Nonparametric methods relax shape restrictions and let data speak more flexibly at the cost of greater variance and sensitivity to tuning choices.

5.2 Histogram and Kernel Density

Kernel density estimator:

$$\hat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right).$$

Interpretation:

- K gives weights as a function of distance from x_0,
- h scales "local neighborhood" width.

5.3 Bias-Variance Tradeoff in Smoothing

- Smaller h: captures local features, but noisy.
- Larger h: smoother curve, but can wash out real structure.

In finite samples, bandwidth dominates kernel choice in importance.

5.4 Boundary Issues

Near support boundaries (for example, nonnegative variables near zero), kernel neighborhoods are asymmetric, increasing bias. Local linear corrections often improve boundary behavior relative to simpler smoothers.

5.5 Nonparametric Regression

Let

$$E[Y \mid X] = g(X), \qquad Y = g(X) + U.$$

Two workhorse approaches:

- **Polynomial regression:**
$$Y = \beta_0 + \beta_1 X + \cdots + \beta_p X^p + U.$$
- **Local linear regression:** for each target x_0,
$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_{i=1}^{N} K\left(\frac{X_i - x_0}{h}\right) (Y_i - a - b(X_i - x_0))^2,$$
and $\hat{g}(x_0) = \hat{a}$.

5.6 Worked Example: Bandwidth Sensitivity

Suppose log-income density is estimated with $h = 0.15, 0.30, 0.60$:

- $h = 0.15$: many local peaks, hard to distinguish noise from structure.
- $h = 0.30$: stable shape with interpretable modality.
- $h = 0.60$: almost unimodal, tails overly smoothed.

Recommended workflow: report a baseline h, then show sensitivity across a reasonable grid.

5.7 When to Use Nonparametrics

Use when functional form is uncertain and sample size is large enough to support flexible estimation. Avoid overpromising in sparse regions; nonparametric methods are data-hungry, especially in multiple dimensions.

6 Maximum Likelihood Foundations

6.1 Likelihood Principle

Given model density $f(x;\theta)$ and i.i.d. data $X_1,\dots,X_N$:

$$L(\theta)=\prod_{i=1}^{N} f(X_i;\theta), \qquad \ell(\theta)=\sum_{i=1}^{N}\log f(X_i;\theta).$$

MLE:

$$\hat{\theta}=\arg\max_{\theta}\ell(\theta).$$

Log-likelihood is used for convenience: products become sums, derivatives are tractable, and numerical optimization is stable.

6.2 Score, Hessian, and Information

$$s(\theta)=\frac{\partial\ell(\theta)}{\partial\theta}, \qquad H(\theta)=\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta'}.$$

At an interior optimum, score is zero and Hessian is negative definite locally. Fisher information links curvature to precision:

$$\mathcal{I}(\theta)=-E[H(\theta)].$$

Sharper curvature means tighter estimator variance.

6.3 Asymptotic Properties (Regular Case)

Under standard conditions and correct specification:

- $\hat{\theta}\xrightarrow{p}\theta_0$ (consistency),
- $\sqrt{N}(\hat{\theta}-\theta_0)\xrightarrow{d}N(0,\mathcal{I}(\theta_0)^{-1})$,
- MLE is asymptotically efficient among regular estimators.

6.4 Proof Example 1: Bernoulli MLE

For $X_i\in\{0,1\}$, $P(X_i=1)=p$:

$$f(X_i;p)=p^{X_i}(1-p)^{1-X_i},$$

$$\ell(p)=\sum_i\left[X_i\log p+(1-X_i)\log(1-p)\right].$$

Proposition 6. $\hat{p}=\bar{X}$.

Proof. Differentiate:

$$\frac{d\ell}{dp} = \sum_i \left(\frac{X_i}{p} - \frac{1 - X_i}{1 - p} \right).$$

Set to zero:

$$\frac{\sum_i X_i}{p} = \frac{N - \sum_i X_i}{1 - p} \Longrightarrow \sum_i X_i = Np \Longrightarrow \hat{p} = \bar{X}.$$

Second derivative is negative at the solution:

$$\frac{d^2\ell}{dp^2} = -\sum_i \left(\frac{X_i}{p^2} + \frac{1 - X_i}{(1 - p)^2} \right) < 0.$$

□

6.5 Proof Example 2: Normal Mean with Known Variance

Suppose $X_i \sim N(\mu, \sigma^2)$ with σ^2 known.

$$\ell(\mu) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2.$$

Maximizing $\ell(\mu)$ is equivalent to minimizing $\sum_i (X_i - \mu)^2$, giving $\hat{\mu} = \bar{X}$. So least squares and MLE coincide in this canonical case.

6.6 MLE in Practice

Three common workflows:

1. analytic solution (rare beyond simple models),
2. grid search (low-dimensional teaching cases),
3. numerical optimization (Newton-Raphson, BFGS, etc.).

Always check convergence diagnostics and whether estimates sit at boundaries.

6.7 Why This Matters for the Next Topics

Logit and probit are naturally estimated by MLE. The current framework gives the mechanics and inference logic needed for those models.

7 Binary Dependent Variable Models

7.1 Linear Probability Model (LPM)

With binary outcome $D_i \in \{0, 1\}$, the linear probability model is

$$D_i = X_i'\beta + \epsilon_i.$$

Because $E[D_i \mid X_i] = P(D_i = 1 \mid X_i)$, the model implies a linear approximation to conditional probability:

$$P(D_i = 1 \mid X_i) \approx X_i'\beta.$$

Interpretation is easy:

- continuous regressor: marginal change in probability per one-unit change in X_{ki},
- binary regressor: probability difference between group $X_{ki} = 1$ and $X_{ki} = 0$, holding others fixed.

7.2 Heteroskedasticity in LPM (Derivation)

Let $p_i = P(D_i = 1 \mid X_i)$. Since D_i is Bernoulli,

$$E[D_i \mid X_i] = p_i, \qquad V(D_i \mid X_i) = p_i(1 - p_i).$$

But $\epsilon_i = D_i - E[D_i \mid X_i]$, so

$$V(\epsilon_i \mid X_i) = p_i(1 - p_i),$$

which depends on X_i. Therefore LPM errors are heteroskedastic by construction (except knife-edge cases).

Implications:

- OLS coefficients remain useful under exogeneity,
- conventional SE are generally wrong,
- robust SE are the default fix.

7.3 Why Move Beyond LPM

Two practical LPM limitations:

1. predicted probabilities can be outside $[0, 1]$,
2. linear functional form can be implausible when probability response is nonlinear.

These motivate probit and logit.

7.4 Probit and Logit

Specify

$$P(D_i = 1 \mid X_i) = G(X_i'\beta),$$

where G is a CDF mapping $\mathbb{R} \to (0, 1)$:

- probit: $G = \Phi$ (standard normal CDF),
- logit: $G(z) = \frac{e^z}{1+e^z}$ (logistic CDF).

Likelihood with independent observations:

$$L(\beta) = \prod_i G(X_i'\beta)^{D_i} \left[1 - G(X_i'\beta)\right]^{1-D_i}.$$

Log-likelihood:

$$\ell(\beta) = \sum_i \left[D_i \log G(X_i'\beta) + (1 - D_i) \log(1 - G(X_i'\beta))\right].$$

Numerical maximization yields $\hat{\beta}$.

7.5 Latent Variable Representation

An equivalent conceptual model:

$$Y_i^* = X_i'\beta + \nu_i, \qquad D_i = \mathbf{1}\{Y_i^* > 0\}.$$

Choosing ν_i normal gives probit; choosing logistic gives logit. This helps intuition: observed binary outcome is a thresholded version of latent propensity.

7.6 Interpreting Probit/Logit Estimates

Raw coefficients are not direct probability changes. Useful summaries:

- **Predicted probabilities:** evaluate $G(X'\hat{\beta})$ at profiles of interest.
- **Marginal effects:**
$$\frac{\partial P(D=1 \mid X)}{\partial X_k} = g(X'\beta)\beta_k,$$
where $g = G'$ is the PDF.
- **Odds ratios (logit):** exponentiated coefficient e^{β_k} is multiplicative change in odds for a one-unit increase in X_k, holding others fixed.

Example 2. If logit coefficient on treatment is 0.40, then odds ratio is $e^{0.40} \approx 1.49$: odds are about 49% higher, not probability necessarily 49 points higher.

8 Panel Data, Fixed Effects, and Difference-in-Differences

8.1 Error Components Setup

For unit $i = 1, \ldots, N$ and period $t = 1, \ldots, T$, write

$$Y_{it} = X_{it}'\beta + Z_i'\gamma + u_{it}, \qquad u_{it} = a_i + \epsilon_{it}.$$

Here a_i is an unobserved unit effect (time-invariant), while ϵ_{it} is idiosyncratic.

Panel data creates two econometric challenges:

- within-unit dependence in u_{it}, which affects inference,
- possible correlation $E[a_i \mid X_{i1}, \ldots, X_{iT}] \neq 0$, which affects identification.

8.2 Fixed Effects: Estimator and Identification

Assumption 1 (Strict Exogeneity for FE)**.**

$$E[\epsilon_{it} \mid X_{i1}, \ldots, X_{iT}, a_i] = 0 \quad \forall t.$$

Proposition 7 (Within Transformation)**.** *Under the model above, demeaning by unit gives*

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)'\beta + (\epsilon_{it} - \bar{\epsilon}_i),$$

which eliminates both a_i and any observed time-invariant regressors Z_i.

Proof. Take unit means:

$$\bar{Y}_i = \bar{X}_i'\beta + Z_i'\gamma + a_i + \bar{\epsilon}_i.$$

Subtract from the original equation:

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)'\beta + (Z_i - \bar{Z}_i)'\gamma + (a_i - a_i) + (\epsilon_{it} - \bar{\epsilon}_i).$$

Because Z_i is time-invariant, $Z_i - \bar{Z}_i = 0$. This yields the claim. □

Define $\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$, $\tilde{X}_{it} = X_{it} - \bar{X}_i$, and stack over i, t:

$$\tilde{Y} = \tilde{X}\beta + \tilde{\epsilon}, \qquad \hat{\beta}_{FE} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}.$$

Theorem 1 (FE Consistency, Fixed T, $N \to \infty$)**.** *If strict exogeneity holds, fourth moments are finite, units are independent (or weakly dependent) across i, and*

$$\frac{1}{NT}\tilde{X}'\tilde{X} \xrightarrow{p} Q \succ 0,$$

then

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

Proof. Use

$$\hat{\beta}_{FE} - \beta = \left(\frac{\tilde{X}'\tilde{X}}{NT}\right)^{-1}\left(\frac{\tilde{X}'\tilde{\epsilon}}{NT}\right).$$

By assumption, the first factor converges to Q^{-1} in probability. For the second factor, write

$$\frac{\tilde{X}'\tilde{\epsilon}}{NT} = \frac{1}{N}\sum_{i=1}^{N} m_i, \qquad m_i := \frac{1}{T}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{\epsilon}_{it}.$$

It is enough to show $E[m_i] = 0$ and finite variance. Because $\tilde{\epsilon}_{it} = \epsilon_{it} - \bar{\epsilon}_i$,

$$E[\tilde{X}_{it}\epsilon_{it} \mid X_i, a_i] = \tilde{X}_{it}E[\epsilon_{it} \mid X_i, a_i] = 0$$

by strict exogeneity, and

$$E[\tilde{X}_{it}\bar{\epsilon}_i \mid X_i, a_i] = \tilde{X}_{it}\frac{1}{T}\sum_{s=1}^{T}E[\epsilon_{is} \mid X_i, a_i] = 0.$$

Hence $E[\tilde{X}_{it}\tilde{\epsilon}_{it} \mid X_i, a_i] = 0$, so $E[m_i] = 0$. LLN then gives $\frac{1}{N}\sum_i m_i \xrightarrow{p} 0$, i.e.

$$\frac{\tilde{X}'\tilde{\epsilon}}{NT} \xrightarrow{p} 0.$$

Combining terms yields $\hat{\beta}_{FE} \xrightarrow{p} \beta$. □

Remark 4. FE is robust to omitted variables that are fixed within unit, but it does not control for time-varying omitted variables correlated with X_{it}.

8.3 Random Effects and the Efficiency Tradeoff

Random effects assumes

$$E[a_i \mid X_{i1}, \ldots, X_{iT}] = 0.$$

Under this stronger orthogonality, RE combines within and between variation and can be more efficient than FE. If the assumption fails, RE is inconsistent while FE remains consistent under strict exogeneity.

8.4 Inference with Dependence

Because ϵ_{it} is usually serially correlated and heteroskedastic within unit, cluster-robust standard errors at the unit level are the default in panel settings.

8.5 First Differences

First differencing gives

$$\Delta Y_{it} = (\Delta X_{it})'\beta + \Delta\epsilon_{it}.$$

With $T = 2$, FE and FD coincide exactly. With $T > 2$, they differ; relative efficiency depends on the serial correlation structure of ϵ_{it}.

8.6 Difference-in-Differences: Formal Identification

In a 2x2 setup, let $g \in \{0, 1\}$ index group (treated/control), $t \in \{0, 1\}$ index time (pre/post), and treatment occur only for $g = 1, t = 1$.

Assumption 2 (No Anticipation)**.**

$$Y_{i0} = Y_{i0}(0) \quad \textit{for all } i.$$

Assumption 3 (Parallel Trends in Untreated Potential Outcomes)**.**

$$E[Y_{i1}(0) - Y_{i0}(0) \mid g_i = 1] = E[Y_{i1}(0) - Y_{i0}(0) \mid g_i = 0].$$

Theorem 2 (DiD Identifies ATT in the Post Period)**.** *Under no anticipation and parallel trends,*

$$\delta_{DiD} = (\bar{Y}_{1,1} - \bar{Y}_{1,0}) - (\bar{Y}_{0,1} - \bar{Y}_{0,0}) = E[Y_{i1}(1) - Y_{i1}(0) \mid g_i = 1].$$

Proof. For treated group:

$$E[Y_{i1} - Y_{i0} \mid g_i = 1] = E[Y_{i1}(1) - Y_{i0}(0) \mid g_i = 1].$$

For control:

$$E[Y_{i1} - Y_{i0} \mid g_i = 0] = E[Y_{i1}(0) - Y_{i0}(0) \mid g_i = 0].$$

Subtract:

$$\delta_{DiD} = E[Y_{i1}(1) - Y_{i1}(0) \mid g_i = 1]$$
$$+ E[Y_{i1}(0) - Y_{i0}(0) \mid g_i = 1] - E[Y_{i1}(0) - Y_{i0}(0) \mid g_i = 0].$$

The parallel-trends assumption sets the second line to zero. □

Equivalent regression form:

$$Y_{it} = \alpha + \eta_i + \lambda_t + \delta(Treat_i \times Post_t) + u_{it}.$$

8.7 Case Studies for DiD

Example 3 (Card and Krueger Minimum Wage)**.** Fast-food employment in New Jersey (treated) is compared to Pennsylvania (control), pre and post policy change. The identifying content is a trend comparison, not a level comparison.

Example 4 (Massachusetts Health Reform)**.** State-by-year insurance data can be estimated with a two-way FE DiD. Credibility depends on whether other states provide a valid counterfactual trend for Massachusetts absent reform.

8.8 Staggered Adoption and Modern Practice

With staggered timing and heterogeneous effects, two-way FE can produce non-convex implicit weighting across cohorts and periods. Good practice is to report cohort/event-time estimators with explicit pre-trend evidence.

9 Potential Outcomes and Causal Estimands

9.1 Core Setup

Each unit has potential outcomes $(Y_i(1), Y_i(0))$. Individual causal effect:

$$\tau_i = Y_i(1) - Y_i(0).$$

Observed outcome:

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0).$$

Only one potential outcome is observed per unit.

9.2 Population Targets

$$ATE = E[Y(1) - Y(0)], \qquad ATT = E[Y(1) - Y(0) \mid T = 1].$$

(These notes use TOT and ATT interchangeably for treated-group mean effect.)

9.3 Selection Bias Decomposition (Rigorous Form)

Proposition 8.

$$E[Y \mid T = 1] - E[Y \mid T = 0] = ATT + \Big(E[Y(0) \mid T = 1] - E[Y(0) \mid T = 0]\Big).$$

Proof. Start from observed means:

$$E[Y \mid T = 1] = E[Y(1) \mid T = 1], \quad E[Y \mid T = 0] = E[Y(0) \mid T = 0].$$

Add and subtract $E[Y(0) \mid T = 1]$:

$$E[Y(1) \mid T = 1] - E[Y(0) \mid T = 0]$$

$$= \underbrace{E[Y(1) - Y(0) \mid T = 1]}_{ATT} + \underbrace{E[Y(0) \mid T = 1] - E[Y(0) \mid T = 0]}_{\text{selection bias}}.$$

□

Assumption 4 (Conditional Unconfoundedness and Overlap)**.**

$$(Y(1), Y(0)) \perp T \mid X, \qquad 0 < P(T = 1 \mid X) < 1 \text{ a.s.}$$

Proposition 9 (Identification of ATE under Conditional Unconfoundedness)**.**

$$ATE = E[E[Y \mid T = 1, X] - E[Y \mid T = 0, X]] .$$

Proof. By conditional unconfoundedness:

$$E[Y(1) \mid X] = E[Y \mid T = 1, X], \quad E[Y(0) \mid X] = E[Y \mid T = 0, X].$$

Therefore

$$E[Y(1) - Y(0) \mid X] = E[Y \mid T = 1, X] - E[Y \mid T = 0, X].$$

Now apply the law of iterated expectations:

$$ATE = E[Y(1) - Y(0)] = E[E[Y(1) - Y(0) \mid X]]$$

$$= E[E[Y \mid T = 1, X] - E[Y \mid T = 0, X]] .$$

□

9.4 Randomized Assignment and ITT

If T is randomly assigned, then $(Y(1), Y(0)) \perp T$, and

$$E[Y \mid T = 1] - E[Y \mid T = 0] = ATE.$$

In randomized encouragement designs with eligibility Z, the primary estimand is

$$ITT = E[Y \mid Z = 1] - E[Y \mid Z = 0].$$

9.5 Compliance Types and Treatment Effects from Encouragement

Define principal strata by $(T(1), T(0))$: always-takers, never-takers, compliers, and defiers.

Assumption 5 (Encouragement Design Conditions)**.** *1. Z randomized.*

2. Exclusion: $Y(z, t) = Y(t)$.

3. Monotonicity: $T(1) \geq T(0)$ (no defiers).

Theorem 3 (Wald Ratio in Encouragement Designs)**.** *Under the encouragement design assumptions above,*

$$\frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[T \mid Z = 1] - E[T \mid Z = 0]} = E[Y(1) - Y(0) \mid T(1) > T(0)].$$

With one-sided noncompliance $T(0) = 0$, this reduces to $ATT = ITT/P(T = 1 \mid Z = 1)$.

Proof. Let $C = \{T(1) > T(0)\}$ denote compliers. By randomization and exclusion,

$$RF := E[Y \mid Z = 1] - E[Y \mid Z = 0] = E[(Y(1) - Y(0))(T(1) - T(0))].$$

Similarly,

$$FS := E[T \mid Z = 1] - E[T \mid Z = 0] = E[T(1) - T(0)].$$

Under monotonicity, $T(1) - T(0) \in \{0, 1\}$, equal to $1\{C\}$. Therefore

$$RF = E[(Y(1) - Y(0))1\{C\}] = P(C)\, E[Y(1) - Y(0) \mid C],$$

$$FS = E[1\{C\}] = P(C).$$

If $P(C) > 0$, divide:

$$\frac{RF}{FS} = E[Y(1) - Y(0) \mid C].$$

For one-sided noncompliance ($T(0) = 0$ for all units), treated units under $Z = 1$ are exactly compliers, so

$$ATT = \frac{ITT}{P(T = 1 \mid Z = 1)}.$$

□

9.6 Case Study: Encouragement-Based Policy Trials

When participation is voluntary but eligibility is randomized, ITT is policy-relevant (effect of offering access), while Wald-scaled effects recover treatment effects for induced participants.

10 Instrumental Variables and LATE

10.1 Linear IV Model

Consider

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

with endogenous X_i. A valid instrument Z_i satisfies:

$$\mathrm{Cov}(Z_i, X_i) \neq 0, \qquad \mathrm{Cov}(Z_i, u_i) = 0.$$

Proposition 10 (Population IV Identification under Homogeneous Effects)**.**

$$\beta_1 = \frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(Z, X)}.$$

Proof.

$$\mathrm{Cov}(Z, Y) = \mathrm{Cov}(Z, \beta_0 + \beta_1 X + u) = \beta_1 \,\mathrm{Cov}(Z, X) + \mathrm{Cov}(Z, u).$$

Exogeneity sets last term to zero; divide by relevance term. □

10.2 TSLS with Controls

With exogenous controls W, endogenous regressor X, and instruments Z, let $Q = [W, Z]$, $P_Q = Q(Q'Q)^{-1}Q'$. Then

$$\hat{\beta}_{2SLS} = (X'P_QX)^{-1}X'P_QY$$

for the single-endogenous-regressor case (after partialling constants as needed).

This expression can be derived from the sample moment condition

$$Q'(Y - \alpha\mathbf{1} - W\gamma - X\beta) = 0,$$

which projects residuals onto the instrument-control space. Solving normal equations after projection gives $X'P_Q(Y - X\beta) = 0$, hence the formula above.

Equivalent algorithm:

1. first stage: regress X on Q, obtain $\hat{X} = P_QX$,
2. second stage: regress Y on $\hat{X}$ (plus included exogenous controls).

Both views are algebraically identical; dedicated IV commands are needed for correct standard errors.

10.3 Wald Estimator

If Z is binary:

$$\hat{\beta}_{Wald} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{X}_{Z=1} - \bar{X}_{Z=0}}.$$

Numerator is reduced form; denominator is first stage.

10.4 Weak Instruments

When first stage is weak, TSLS can have large finite-sample bias and poor normal approximations. Report first-stage strength diagnostics and use robust confidence procedures when relevance is borderline.

10.5 Heterogeneous Effects: LATE Theorem

Let treatment be D, instrument Z, and potential outcomes $Y(1), Y(0)$, $D(1), D(0)$.

Assumption 6 (LATE Conditions)**.** *1. Independence:* $(Y(1), Y(0), D(1), D(0)) \perp Z$.

2. Exclusion: $Y = Y(D)$.

3. Monotonicity: $D(1) \geq D(0)$.

Theorem 4 (Wald Equals LATE)**.** *Under the LATE conditions above and* $P(D(1) > D(0)) > 0$*,*

$$\frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]} = E[Y(1) - Y(0) \mid D(1) > D(0)].$$

Proof. By independence and exclusion:

$$E[Y \mid Z = 1] - E[Y \mid Z = 0] = E[(Y(1) - Y(0))(D(1) - D(0))].$$

Also:

$$E[D \mid Z = 1] - E[D \mid Z = 0] = E[D(1) - D(0)].$$

Under monotonicity, $D(1) - D(0) \in \{0, 1\}$, so

$$E[(Y(1) - Y(0))(D(1) - D(0))] = P(C) \cdot E[Y(1) - Y(0) \mid C],$$

where $C = \{D(1) > D(0)\}$, and

$$E[D(1) - D(0)] = P(C).$$

Since $P(C) > 0$, divide to obtain the theorem. □

10.6 IV Case Studies

Example 5 (Quarter of Birth and Schooling)**.** Compulsory-schooling exposure induced by birth timing shifts years of schooling for a subset of students. IV estimate is interpreted as return to schooling for this complier margin, not a universal return for all students.

Example 6 (Draft Lottery and Veteran Status)**.** Random lottery number shifts military service probability. Under exclusion and monotonicity, the Wald/TSLS estimate identifies service effects for those whose service status is changed by draft eligibility.

11 Regression Discontinuity Designs

11.1 Sharp RD: Identification Theorem

Let running variable R_i, cutoff c, and treatment rule $T_i = \mathbf{1}\{R_i \geq c\}$. Observed outcome:

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0).$$

Assumption 7 (Continuity at the Cutoff)**.**

$$\mu_d(r) := E[Y(d) \mid R = r] \quad \textit{is continuous at } r = c,\ d \in \{0, 1\}.$$

Theorem 5 (Sharp RD Identification)**.** *Under continuity at the cutoff,*

$$\tau_{SRD} := \lim_{r \downarrow c} E[Y \mid R = r] - \lim_{r \uparrow c} E[Y \mid R = r] = E[Y(1) - Y(0) \mid R = c].$$

Proof. For $r < c$, $T = 0$, so $E[Y \mid R = r] = E[Y(0) \mid R = r] = \mu_0(r)$. Taking left limit gives $\mu_0(c)$. For $r \geq c$, $T = 1$, so right limit gives $\mu_1(c)$. Difference is $\mu_1(c) - \mu_0(c)$. □

11.2 Fuzzy RD as Local Wald

When treatment probability jumps but not from 0 to 1, define

$$\Delta_Y := \lim_{r \downarrow c} E[Y \mid R = r] - \lim_{r \uparrow c} E[Y \mid R = r],$$

$$\Delta_T := \lim_{r \downarrow c} E[T \mid R = r] - \lim_{r \uparrow c} E[T \mid R = r].$$

Then fuzzy RD estimand is $\tau_{FRD} = \Delta_Y / \Delta_T$.

Assumption 8 (Local Monotonicity Around the Cutoff)**.** *Crossing the threshold weakly increases treatment take-up for all units in a local neighborhood of c.*

Theorem 6 (Fuzzy RD Identifies Local Complier Effect)**.** *Under continuity of potential outcomes in R, exclusion of the threshold indicator except through T, and local monotonicity:*

$$\tau_{FRD} = E[Y(1) - Y(0) \mid \textit{compliers at } R = c].$$

Proof. Define local threshold instrument $Z = \mathbf{1}\{R \geq c\}$. In a vanishing neighborhood of c, continuity assumptions imply that crossing the threshold isolates the assignment discontinuity, so

$$\Delta_Y = E[(Y(1) - Y(0))(T(1) - T(0)) \mid R = c],$$

$$\Delta_T = E[T(1) - T(0) \mid R = c].$$

Under local monotonicity, $T(1) - T(0) \in \{0, 1\}$, with value 1 on local complier set C_c. Therefore

$$\Delta_Y = P(C_c \mid R = c)\, E[Y(1) - Y(0) \mid C_c, R = c],$$

$$\Delta_T = P(C_c \mid R = c).$$

If $\Delta_T > 0$, dividing yields

$$\tau_{FRD} = \frac{\Delta_Y}{\Delta_T} = E[Y(1) - Y(0) \mid C_c, R = c].$$

□

11.3 Estimation Strategy

Preferred empirical specification is local linear on each side of c:

$$Y_i = \alpha + \tau T_i + \beta_-(R_i - c) + \beta_+ T_i(R_i - c) + u_i, \quad |R_i - c| \leq h,$$

estimated with kernel weights. Report bandwidth sensitivity and avoid high-order global polynomials unless strongly justified.

11.4 Validity Diagnostics

1. Density continuity of R at c (manipulation check).

2. Covariate continuity at c for predetermined characteristics.

3. Stability of estimated τ across reasonable bandwidths/specifications.

11.5 External Validity

RD estimates are local by construction. Interpretation should explicitly state that the identified effect pertains to units near the cutoff.

11.6 RD Case Studies

Example 7 (Close Elections)**.** Party incumbency effects can be estimated by comparing units where a party barely wins vs barely loses. Identification is local to close elections and depends on no strategic sorting/manipulation at the vote-margin cutoff.

Example 8 (Legal Drinking Age)**.** Outcomes like motor-vehicle fatalities can be studied around age 21. Credibility requires checking that no other policy discontinuities coincide with the same threshold in ways that contaminate interpretation.

12 Integrated Review: Full-Course Checklist

12.1 Conceptual Checklist

1. What parameter am I trying to learn?
2. Under which assumptions does my estimator identify it?
3. Which uncertainty formula is valid here?
4. Are my controls chosen for prediction, causal identification, or both?
5. Are key conclusions robust to standard alternatives (for example robust SE, alternate bandwidth, alternate control group)?

12.2 Technical Checklist

- Verify normal equations and residual orthogonality in OLS.
- Check collinearity before interpreting multivariate coefficients.
- Evaluate possible OVB direction using $\beta_2\pi_1$ logic.
- In binary models, distinguish coefficient scale from marginal effects.
- In panel models, cluster at the policy/group level and justify FE vs RE.
- In DiD, defend parallel trends.
- In IV, report and discuss first-stage strength and instrument validity.
- In RD, report bandwidth sensitivity and diagnostic checks around cutoff.
- In nonparametrics and MLE, inspect tuning/convergence sensitivity.

12.3 Common Failure Modes

- Treating significance as magnitude.
- Ignoring omitted confounders while over-controlling for mediators.
- Using default conventional SE under clear heteroskedasticity.
- Reading logit coefficients directly as probability-point effects.

- Treating FE and RE as interchangeable without an identifying argument.
- Presenting DiD estimates without credible trend evidence.
- Interpreting IV/LATE as universal treatment effects.
- Treating RD estimates as globally generalizable.
- Reading one nonparametric curve as "the truth" without bandwidth checks.
- Trusting numerical MLE output without diagnostics.

13 Extended Worked Examples and Derivations

13.1 Worked OVB Sign Analysis with Economic Story

Suppose we want the effect of schooling (S) on log wages (W):

$$W = \beta_0 + \beta_1 S + \beta_2 A + U,$$

where A is latent ability. If ability is omitted, the short model is

$$W = \delta_0 + \delta_1 S + V.$$

Using $\delta_1 = \beta_1 + \beta_2 \pi_1$, where π_1 is slope in $A = \pi_0 + \pi_1 S + \text{noise}$:

- If ability raises wages ($\beta_2 > 0$) and correlates positively with schooling ($\pi_1 > 0$), δ_1 overstates return to schooling.
- If ability raises wages but is negatively correlated with schooling in a particular sample, bias can be downward.

Numerical sensitivity table:

$$\beta_1 = 0.06, \quad \beta_2 = 0.25, \quad \pi_1 \in \{0.1, 0.2, 0.3, 0.4\}.$$

Then

$$\delta_1 \in \{0.085, 0.11, 0.135, 0.16\}.$$

Even moderate correlation between omitted ability and schooling can materially change the estimated slope.

13.2 Detailed Derivation: Bivariate OLS Unbiasedness

Starting from

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i (X_i - \bar{X}) U_i}{\sum_i (X_i - \bar{X})^2},$$

condition on observed X:

$$E[\hat{\beta}_1 \mid X] = \beta_1 + \frac{\sum_i (X_i - \bar{X}) E[U_i \mid X]}{\sum_i (X_i - \bar{X})^2}.$$

If $E[U_i \mid X] = 0$, then $E[\hat{\beta}_1 \mid X] = \beta_1$, hence $E[\hat{\beta}_1] = \beta_1$. This is the cleanest statement of exogeneity in linear regression: zero conditional mean of the error.

13.3 Detailed Derivation: Bivariate OLS Variance Under Homoskedasticity

Given $E[U_i \mid X] = 0$, $V(U_i \mid X) = \sigma^2$, and conditional independence:

$$V(\hat{\beta}_1 \mid X) = V\left(\frac{\sum_i (X_i - \bar{X}) U_i}{\sum_i (X_i - \bar{X})^2} \;\middle|\; X \right) = \frac{\sum_i (X_i - \bar{X})^2 V(U_i \mid X)}{\left(\sum_i (X_i - \bar{X})^2 \right)^2} = \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}.$$

So the variance shrinks when:

- sample size increases (usually increases spread sum),
- regressor variation is larger.

13.4 Conventional vs Robust Inference in Practice

Suppose estimates are:

$$\hat{\beta}_1 = 0.05, \quad SE_{\text{conv}} = 0.015, \quad SE_{\text{rob}} = 0.025.$$

Then

$$t_{\text{conv}} = 3.33, \qquad t_{\text{rob}} = 2.00.$$

Both are positive, but inferential strength changes materially. This is why reporting both can be informative early in a project, even if final tables report robust by default.

13.5 Worked Nonparametric Regression Comparison

Suppose true CEF is mildly concave and sample size is moderate ($N = 400$).

- A linear model captures average trend but misses curvature.
- A 4th-order polynomial captures curvature but oscillates in tails.
- Local linear with moderate bandwidth tracks curvature without tail explosions.

Interpretation principle: choose the simplest estimator that captures economically relevant features robustly across nearby tuning choices.

13.6 Bandwidth Selection Workflow

One practical workflow:

1. Start with software default bandwidth.
2. Plot results for $0.5h, h, 1.5h, 2h$.
3. Keep interpretations that are stable across this range.
4. If conclusions are unstable, report that instability explicitly.

This avoids false precision and prevents over-reading small visual wiggles.

13.7 MLE Derivation: Normal Mean and Variance Unknown

Suppose $X_i \sim N(\mu, \sigma^2)$ with both parameters unknown:

$$\ell(\mu, \sigma^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2.$$

FOC for μ:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_i (X_i - \mu) = 0 \Longrightarrow \hat{\mu} = \bar{X}.$$

FOC for σ^2:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (X_i - \hat{\mu})^2 = 0 \Longrightarrow \hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_i (X_i - \bar{X})^2.$$

Notice the denominator is N, not $N-1$: MLE for variance is biased in finite samples but consistent. This is a useful reminder that unbiasedness is not the only design criterion.

13.8 Likelihood Curvature and Uncertainty

Near the optimum, a quadratic approximation gives:

$$\ell(\theta) \approx \ell(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})'[-H(\hat{\theta})](\theta - \hat{\theta}).$$

Steep curvature (large $-H(\hat{\theta})$) implies the likelihood drops quickly away from $\hat{\theta}$, corresponding to smaller standard errors.

13.9 Model Interpretation Discipline

Before finalizing any empirical result, write four lines:

1. **Estimand:** what exact parameter do we claim to estimate?
2. **Assumptions:** which assumptions map estimator to estimand?
3. **Inference:** which covariance estimator is being used and why?
4. **Sensitivity:** what is the minimal robustness check?

This habit dramatically improves clarity and helps prevent over-claims.

14 Mini Problem Set (With Short Solutions)

Problem 1

If $\hat{\beta}_1 = 0.12$, $SE = 0.05$, test $H_0 : \beta_1 = 0$ at 5%.

Solution.

$$t = \frac{0.12}{0.05} = 2.4 > 1.96.$$

Reject H_0 under normal approximation.

Problem 2

Given $\beta_2 < 0$ and $\pi_1 > 0$, what is OVB direction when omitting X_2?

Solution. Bias sign is $\text{sign}(\beta_2\pi_1) < 0$: downward bias.

Problem 3

Why can robust SE exceed conventional SE in practice?

Solution. Conventional SE assumes constant conditional variance. If variance grows with X or fitted values, conventional formulas understate uncertainty. Robust formulas account for heteroskedasticity.

Problem 4

Why is a single nonparametric bandwidth plot usually insufficient?

Solution. Because local shape can be bandwidth-driven. Inference about curvature should rely on stability across a plausible bandwidth neighborhood.

Problem 5

For Bernoulli MLE, what is the estimator and why?

Solution. $\hat{p} = \bar{X}$. It solves the log-likelihood first-order condition and maximizes likelihood because second derivative is negative.

Problem 6

In a logit model, $\hat{\beta}_k = 0.7$ on a binary regressor. What is the odds ratio?

Solution.

$$OR = e^{0.7} \approx 2.01.$$

Odds are about doubled when the regressor moves from 0 to 1, holding others fixed.

Problem 7

In a two-group two-period setting, treated group changes by +6 and control group changes by +2. What is DiD?

Solution.

$$\hat{\delta}_{DiD} = 6 - 2 = 4.$$

Problem 8

The observed treated-control mean gap is 10. Estimated selection bias is 3. What is TOT?

Solution.

$$TOT = 10 - 3 = 7.$$

Problem 9

Reduced form estimate is 0.12, first stage is 0.30. What is Wald/IV estimate?

Solution.

$$\hat{\beta}_{IV} = \frac{0.12}{0.30} = 0.40.$$

Problem 10

In fuzzy RD, outcome jump is 2.5 and treatment-probability jump is 0.5. What is the local effect at cutoff?

Solution.

$$\hat{\tau}_{FRD} = \frac{2.5}{0.5} = 5.$$

15 Formula Sheet (Full-Course Core)

Estimation and Inference

$$\bar{X} = \frac{1}{N}\sum_i X_i, \qquad s^2 = \frac{1}{N-1}\sum_i (X_i - \bar{X})^2$$

$$t = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}, \qquad \text{CI}_{95\%} : \hat{\theta} \pm 1.96\, SE(\hat{\theta})$$

Bivariate OLS

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$s^2 = \frac{1}{N-2}\sum_i \hat{u}_i^2, \qquad \widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{s^2}{\sum_i (X_i - \bar{X})^2}}$$

Multivariate OLS

$$\hat{\beta} = (X'X)^{-1}X'Y, \qquad \widehat{V}_{\text{rob}}(\hat{\beta}) = (X'X)^{-1}\left(\sum_i \hat{u}_i^2 x_i x_i'\right)(X'X)^{-1}$$

$$\widehat{V}(l'\hat{\beta}) = l'\widehat{V}(\hat{\beta})l$$

OVB

Long model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$, omitted X_2:

$$\delta_1 = \beta_1 + \beta_2 \pi_1, \qquad \pi_1 = \text{slope from } X_2 \text{ on } X_1$$

Bias sign is $\text{sign}(\beta_2 \pi_1)$.

Nonparametrics

$$\hat{f}(x_0) = \frac{1}{Nh}\sum_i K\left(\frac{X_i - x_0}{h}\right)$$

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_i K\left(\frac{X_i - x_0}{h}\right)\left(Y_i - a - b(X_i - x_0)\right)^2, \quad \hat{g}(x_0) = \hat{a}$$

Maximum Likelihood

$$L(\theta) = \prod_i f(X_i;\theta), \qquad \ell(\theta) = \sum_i \log f(X_i;\theta), \qquad \hat{\theta} = \arg\max_\theta \ell(\theta)$$

Bernoulli case:

$$\ell(p) = \sum_i \left[X_i \log p + (1 - X_i)\log(1-p)\right], \qquad \hat{p} = \bar{X}$$

Binary Outcomes

$$P(D_i = 1 \mid X_i) = G(X_i'\beta), \qquad \frac{\partial P(D_i = 1 \mid X_i)}{\partial X_{ki}} = g(X_i'\beta)\beta_k$$

$$V(\epsilon_i \mid X_i) = p_i(1-p_i) \quad \text{in LPM}$$

Logit odds ratio:

$$OR_k = e^{\beta_k}.$$

Panel and DiD

$$Y_{it} = X_{it}'\beta + \eta_i + \lambda_t + u_{it}$$

$$\hat{\delta}_{DiD} = (\bar{Y}_{T,post} - \bar{Y}_{T,pre}) - (\bar{Y}_{C,post} - \bar{Y}_{C,pre})$$

Clustered inference: cluster at group/policy-assignment level in panel settings.

Potential Outcomes

$$Y_i = T_iY_i(1) + (1-T_i)Y_i(0), \qquad ATE = E[Y(1) - Y(0)], \qquad TOT = E[Y(1) - Y(0) \mid T = 1]$$

$$E[Y \mid T = 1] - E[Y \mid T = 0] = TOT + \text{selection bias}.$$

IV and LATE

$$\beta_{IV} = \frac{\mathrm{Cov}(Z,Y)}{\mathrm{Cov}(Z,X)}, \qquad \hat{\beta}_{Wald} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{X}_{Z=1} - \bar{X}_{Z=0}}$$

$$LATE = E[Y(1) - Y(0) \mid \text{compliers}]$$

under independence, exclusion, and monotonicity.

Regression Discontinuity

Sharp RD:

$$\tau_{SRD} = \lim_{x \downarrow c} E[Y \mid X = x] - \lim_{x \uparrow c} E[Y \mid X = x]$$

Fuzzy RD:

$$\tau_{FRD} = \frac{\lim_{x\downarrow c} E[Y \mid X = x] - \lim_{x\uparrow c} E[Y \mid X = x]}{\lim_{x\downarrow c} E[T \mid X = x] - \lim_{x\uparrow c} E[T \mid X = x]}.$$

16 Roadmap

The full course coverage is now in place. The next optional pass can focus on:

- tightening prose and reducing repetition while preserving length,
- adding one end-to-end empirical case study that threads OLS, DiD, IV, and RD,
- generating a short "exam sprint" version (5-7 pages) from this master document.