# ECONOMETRICS PRIMER FOR THE MATH-LITERATE

With Visualizations

Audience: Strong math background (linear algebra, probability, calculus), limited economics exposure. Each section explains the economic problem briefly, then dives into the math and code.

All estimators implemented from scratch via linear algebra / scipy.optimize. Includes Monte Carlo demonstrations, event study specifications, and a complete method selection decision tree.

Core Sections: OLS, FWL, Heteroskedasticity, IV/2SLS, Panel FE, DiD, RDD, Probit/Logit, MLE from scratch, Bootstrap inference.

Supplementary Sections: Monte Carlo OVB demonstration, Event Study diagnostics, Method Selection Flowchart, Identification Strategy Guide.

Each section includes: economic motivation, formal mathematical setup, from-scratch Python implementation, simulation results with standard errors, and diagnostic visualizations. Intuition boxes provide geometric or probabilistic intuition; policy connection boxes link methods to real-world evaluation questions.

# Section 1: OLS — Ordinary Least Squares

### Economic story

Suppose we observe wages (y) and years of schooling (x) for a sample of workers. A central question in labor economics is: what is the "return to schooling" -- how much does an additional year of education raise wages on average? OLS is the starting point for nearly every empirical analysis.

### Mathematical setup

We posit the linear model y = X*beta + epsilon, where X is the n-by-k design matrix (including a column of ones for the intercept), beta is the k-vector of unknown parameters, and epsilon is the n-vector of errors.

The classical OLS assumptions are:

```
    (A1) Linearity: the conditional expectation E[y|X] is linear in X.
    (A2) Random sampling: observations are i.i.d.
    (A3) No perfect multicollinearity: X has full column rank.
    (A4) Exogeneity (zero conditional mean): E[epsilon|X] = 0.
    (A5) Homoskedasticity: Var(epsilon|X) = sigma^2 * I.
```

Under (A1)-(A4), OLS is unbiased: E[beta_hat] = beta.
Adding (A5), OLS is BLUE (Best Linear Unbiased Estimator) -- the Gauss-Markov theorem guarantees the smallest variance among all linear unbiased estimators.

The closed-form solution:

```
    beta_hat = (X'X)^{-1} X'y
    Residuals: e_hat = y - X*beta_hat
    Estimated variance: sigma_hat^2 = e_hat'e_hat / (n - k)
    Variance of beta_hat: Var(beta_hat) = sigma_hat^2 * (X'X)^{-1}
    Standard errors: SE(beta_hat_j) = sqrt(Var(beta_hat)_{jj})
```

Geometric interpretation: OLS projects y onto the column space of X. The fitted values y_hat = X*beta_hat are the orthogonal projection, and the residuals e_hat are perpendicular to col(X), which is why X'e_hat = 0 (the "normal equations").

> ***INTUITION:*** *Why projection?*
>
> *Think of y as a vector in n-dimensional space. The columns of X span a k-dimensional subspace. OLS finds the point in that subspace closest to y (minimizing ||y - X*beta||^2). The residual vector e is the perpendicular drop from y to that subspace -- hence X'e = 0.*
> *For a mathematician, this is simply the orthogonal projection theorem*

*applied to the inner product space R^n.*

Why OLS fails here -- omitted variable bias (OVB)
In our simulation, ability affects both schooling (smarter people get more education) and wages (smarter people earn more), but we omit ability from the regression. The OVB formula is:

```
    bias = beta_ability * [Cov(schooling, ability) / Var(schooling)]
```

Since both terms are positive, the OLS estimate of the return to schooling is biased upward -- it captures part of the ability effect. This is the fundamental motivation for the IV and panel methods that follow in later sections.

Derivation of the OVB formula:

```
    Let the true model be y = X1*beta1 + X2*beta2 + u, where X2 is
    omitted. The short regression estimates:
    beta_hat_short = (X1'X1)^{-1} X1'y
    = beta1 + (X1'X1)^{-1} X1'X2 * beta2 + (X1'X1)^{-1} X1'u
    Taking expectations: E[beta_hat_short] = beta1 + delta * beta2,
    where delta = (X1'X1)^{-1} X1'X2 is the coefficient from regressing
    the omitted variable on the included variable. The sign of the bias
    is determined by the signs of delta and beta2 -- both positive here.
```

***POLICY CONNECTION:*** *Understanding OVB is essential for evaluating social programs. When we observe that people who receive mental health treatment have better outcomes, we cannot simply conclude the treatment worked: those who seek treatment may differ systematically from those who do not (in severity, motivation, access, support networks). OVB tells us exactly how such confounders bias naive estimates, and motivates the identification strategies in Sections 4-7.*

## Results (short vs long regression)

Short regression (ability omitted): wage ~ 1 + schooling

```
    beta_hat (intercept) = -3.826 SE = 0.876
    beta_hat (schooling) = 2.827 SE = 0.073
```

Long regression (ability included): wage ~ 1 + schooling + ability

```
    beta_hat (intercept) = 0.308 SE = 0.923
    beta_hat (schooling) = 2.481 SE = 0.077
    beta_hat (ability) = 1.383 SE = 0.148
```

True DGP coefficients

```
    schooling = 2.5
    ability = 1.5
```

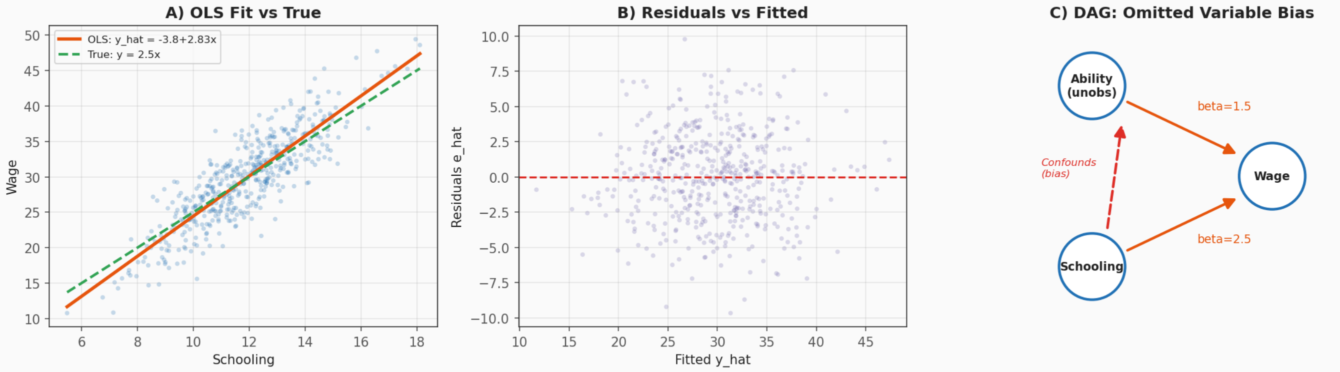
## Interpretation

Because schooling is generated as schooling = 12 + rho·ability + noise with rho = 1.0, ability is positively correlated with schooling and also directly raises wages. In the short regression, ability is absorbed into the error term, and since Cov(schooling, ability) > 0, the exogeneity condition $E[\varepsilon|\text{schooling}] = 0$ fails. The omitted-variable-bias formula implies:

```
bias( beta_hat_schooling ) = beta_ability · Cov(schooling, ability) / Var(schooling) > 0.
```

So the short-regression estimate of the “return to schooling” is biased upward on average: beta_hat(schooling) tends to exceed 2.5. The long regression restores exogeneity by controlling for ability, so beta_hat(schooling) is centered near 2.5 and beta_hat(ability) near 1.5 (up to sampling noise).

Section 1: OLS -- The Baseline Workhorse


A) OLS Fit vs True

B) Residuals vs Fitted

C) DAG: Omitted Variable Bias

## Section 1B: Monte Carlo Demonstration of Omitted Variable Bias

### Why Monte Carlo?

A single regression gives one estimate -- it might be close to or far from the truth due to sampling noise. To see whether an estimator is biased, we need to ask: "If I could repeat this experiment thousands of times, where would the estimates cluster?" Monte Carlo simulation does exactly this: generate data from a known DGP, estimate, repeat, and examine the distribution of estimates.

### Monte Carlo design

```
- 1000 simulations, n = 500 each
- DGP: schooling = 12 + 1.0*ability + noise; wage = 2.5*schooling + 1.5*ability + noise
```

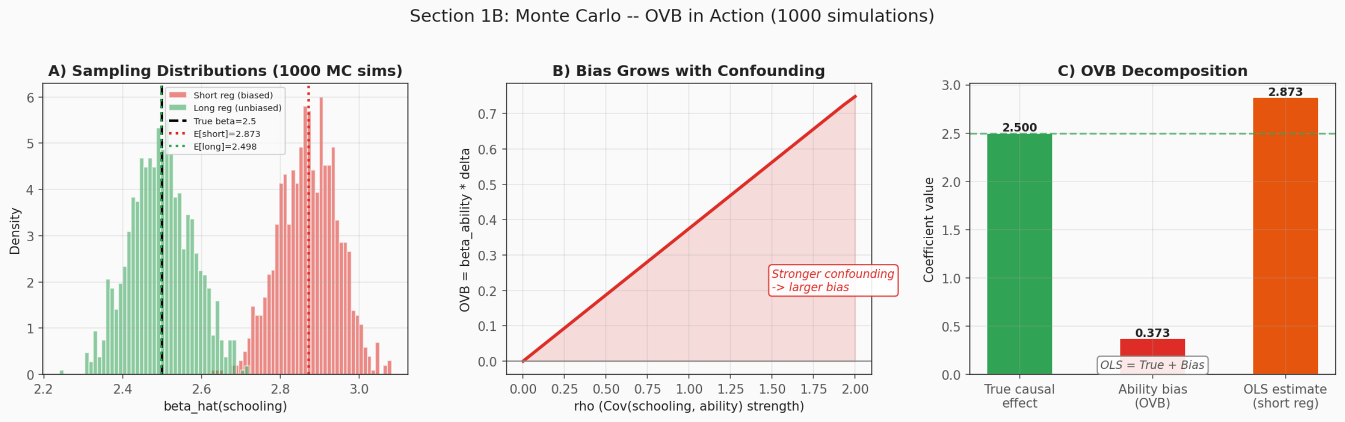
- "Short regression": regress wage on schooling only (ability omitted)
- "Long regression": regress wage on schooling + ability (correct spec)

### Results

```
Short regression: E[beta_hat] = 2.873 (biased above 2.5)
Long regression: E[beta_hat] = 2.498 (centered on 2.5)
Empirical bias = 0.373
```

Panel A shows the two sampling distributions overlaid. The short-regression distribution (red) is shifted to the right of the true value 2.5 -- this is bias. The long-regression distribution (green) is centered on 2.5 -- unbiased. Panel B shows that the bias grows with the strength of the confounding relationship (rho). Panel C decomposes the OLS estimate into its two components: the true causal effect plus the omitted variable bias.

> ***KEY TAKEAWAY:*** *Unbiasedness is a property of the *procedure*, not any single estimate. A biased estimator systematically misses the target across repeated samples. This is why identification -- ensuring E[e|X]=0 -- is the central concern of applied econometrics.*



Section 1B: Monte Carlo -- OVB in Action (1000 simulations)

A) Sampling Distributions (1000 MC sims)

B) Bias Grows with Confounding

C) OVB Decomposition

## Section 2: Frisch-Waugh-Lovell (FWL) Theorem

### Motivation

When we run a multiple regression y ~ x1 + x2, what does it mean to say we are "controlling for x2"? The FWL theorem gives a precise, geometric answer: the coefficient on x1 is identical to what you get by first removing the linear influence of x2 from both y and x1, then regressing the residuals on each other.

### Formal statement

Partition X = [X1, X2] and beta = [beta_1, beta_2]'. Define the annihilator (residual-maker) matrix for X2:

```
M2 = I - X2 (X2'X2)^{-1} X2'
```

Then: beta_hat_1 from the full regression y = X1*beta_1 + X2*beta_2 + e equals exactly beta_hat_1 from the auxiliary regression

```
M2*y = (M2*X1) * beta_1 + residual
```

In other words, you can "partial out" X2 and get the same coefficient.

### Why it matters for econometrics

1. Understanding "controlling for": When someone says "we control for experience," FWL tells you the coefficient on schooling captures only the component of schooling orthogonal to experience -- the variation in schooling that cannot be predicted from experience.

2. Fixed effects: Entity fixed effects with hundreds of dummies are computationally expensive in a full regression. FWL says you can equivalently demean within each entity and run OLS on the demeaned data -- this is exactly the "within estimator" (Section 5).

3. Partial regression plots: Plotting M2*y against M2*x1 is the standard "added variable plot" or "partial regression plot" used in diagnostics. The slope of that scatter is beta_hat_1.

### Geometric intuition

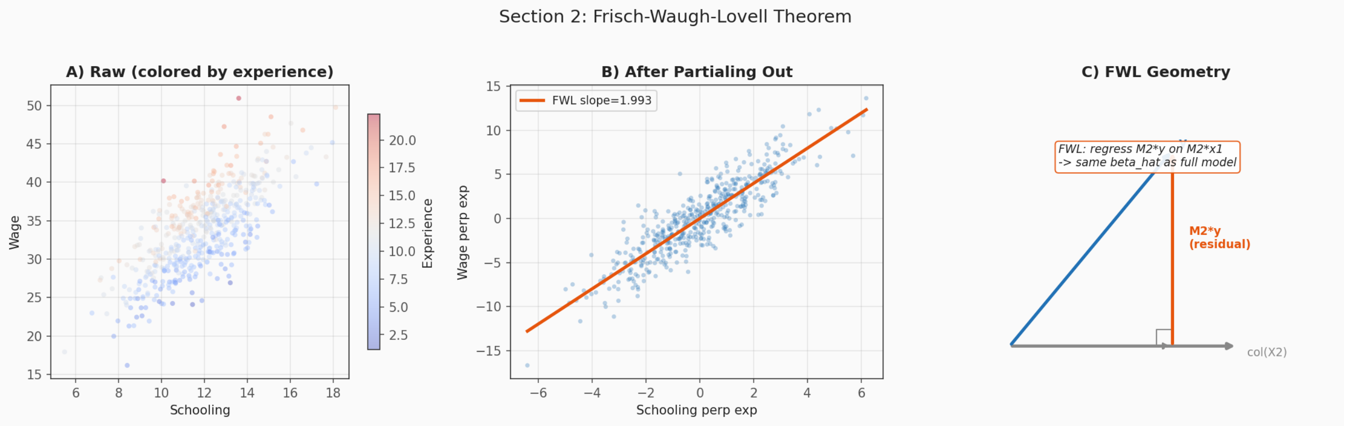
In the column space, X2 defines a subspace. M2 projects any vector onto the orthogonal complement of that subspace. FWL says: project both y and x1 out of X2's column space, then regress -- you get the same answer as fitting the full model. The figure shows this as projecting y down to its residual component M2*y, which is the part of y that X2 cannot explain.

### Results

```
    Full regression beta_hat(schooling) : 1.993459
    FWL partialled beta_hat(schooling) : 1.993459
    -> Identical (up to numerical precision). FWL theorem confirmed.
```

**Interpretation: The coefficient on schooling in the multiple regression** captures only the part of the schooling-wage association that is not explained by experience. If schooling and experience were orthogonal (uncorrelated), partialing out would make no difference. But in reality they are correlated (more experienced workers may have different education patterns), so controlling for experience changes the estimate.

This is a critical concept throughout applied econometrics: every time you add a control variable, you are implicitly partialing out. FWL makes this operation explicit and verifiable.


Section 2: Frisch-Waugh-Lovell Theorem
A) Raw (colored by experience)
Wage
50
45
40
35
30
25
20
15
6
8
10
12
14
16
18
Schooling
Experience
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
B) After Partialing Out
FWL slope=1.993
Wage perp exp
15
10
5
0
−5
−10
−15
−6
−4
−2
0
2
4
6
Schooling perp exp
C) FWL Geometry
FWL: regress M2*y on M2*x1
-> same beta_hat as full model
y
M2*y
(residual)
col(X2)

## Section 3: Heteroskedasticity — Detection and Robust Standard Errors

### Economic context

In many empirical settings, the variance of the error term is not constant across observations. A classic example: wage variance often grows with income level -- high earners have more volatile compensation (bonuses, stock options, variable pay), while minimum-wage workers cluster tightly around a fixed hourly rate. This pattern is called heteroskedasticity: Var(epsilon_i | X_i) = sigma_i^2, which varies with X.

### What breaks and what does not

Under heteroskedasticity, OLS beta_hat remains unbiased and consistent (assumptions A1-A4 still hold). However, the usual formula for standard errors, SE = sqrt(sigma_hat^2 * (X'X)^{-1}), is wrong because it assumes Var(epsilon|X) = sigma^2 * I. Using incorrect SEs means confidence intervals have wrong coverage and hypothesis tests have incorrect size -- you might reject a true null too often or too rarely.

### Detection: Breusch-Pagan test

Regress the squared OLS residuals e_hat_i^2 on X. Under H0 of homoskedasticity, the squared residuals should be unrelated to X. A significant F-stat (or chi-squared) rejects H0 and indicates heteroskedasticity.

### The fix: Heteroskedasticity-Consistent (HC) standard errors

Rather than assuming a constant sigma^2, we estimate the "sandwich" covariance matrix:

```
    V_hat_HC = (X'X)^{-1} * [Sum_i e_hat_i^2 * x_i * x_i'] * (X'X)^{-1}
```

The HC1 variant (Stata's default) applies a degrees-of-freedom correction: multiply by n/(n-k). Other variants (HC0, HC2, HC3) differ in how they adjust the residuals, but in practice with moderate n they give very similar results.

### Practical advice

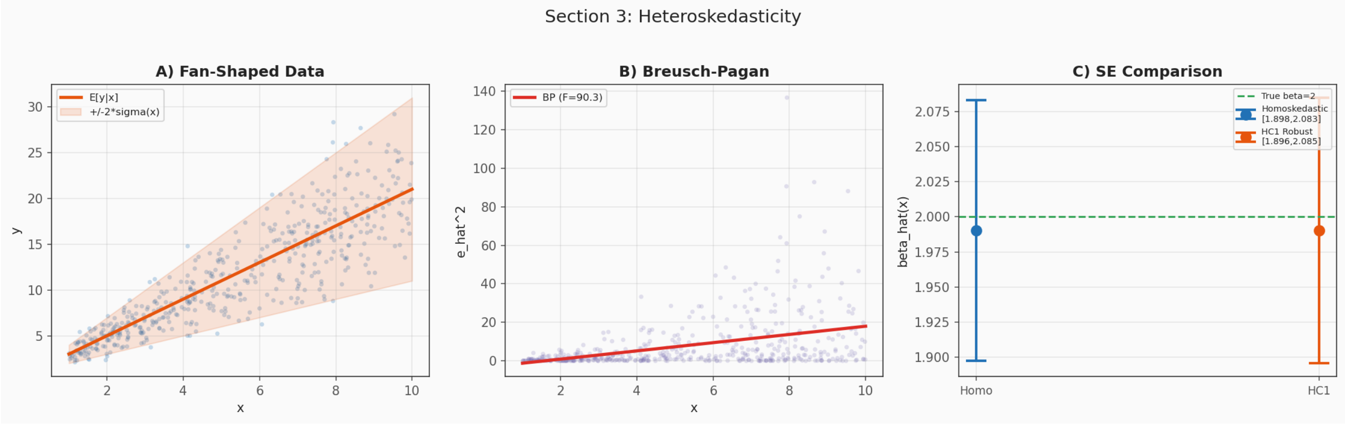
In applied econometrics, robust SEs are essentially the default. Many journals and referees expect them, and there is no real cost to using them when heteroskedasticity is absent (they are still consistent, just slightly less efficient than classical SEs).

The mantra: "Always report robust standard errors."

## Results

```
Breusch-Pagan test F = 90.26, p = 0.0000
p < 0.05 -> reject homoskedasticity (the variance is not constant)

SE(x) -- homoskedastic : 0.0473
SE(x) -- HC1 robust : 0.0483
```

In this simulation, the variance grows linearly with x (the "fan" shape in panel A). The Breusch-Pagan test strongly rejects homoskedasticity. Notice that the robust SE differs from the classical SE -- using the wrong one would give misleading inference. Panel C shows that both confidence intervals cover the true value (beta=2), but in general, the homoskedastic CI can be too narrow or too wide depending on the pattern of heteroskedasticity and the distribution of X.


Section 3: Heteroskedasticity
A) Fan-Shaped Data
E[y|x]
+/-2*sigma(x)
y
x
30
25
20
15
10
5
2
4
6
8
10
B) Breusch-Pagan
BP (F=90.3)
e_hat^2
x
140
120
100
80
60
40
20
0
2
4
6
8
10
C) SE Comparison
True beta=2
Homoskedastic
[1.898,2.083]
HC1 Robust
[1.896,2.085]
beta_hat(x)
2.075
2.050
2.025
2.000
1.975
1.950
1.925
1.900
Homo
HC1

# Section 4: Instrumental Variables (IV / 2SLS)

### The endogeneity problem

In Section 1 we saw that omitting ability biases the OLS estimate of the return to schooling. More generally, whenever Cov(x, epsilon) != 0 -- due to omitted variables, measurement error, or simultaneity -- OLS is biased and inconsistent. No amount of additional data will fix it.

### The instrumental variables solution

Find a variable Z (the "instrument") that satisfies two conditions:

```
(1) Relevance: Cov(Z, X) != 0 -- Z predicts the endogenous X.
(2) Exclusion: Cov(Z, epsilon) = 0 -- Z affects Y only through X.
```

Condition (1) is testable via the first-stage F-statistic. The rule of thumb (Stock & Yogo): F > 10 for a single endogenous regressor to avoid weak-instrument bias. Condition (2) is fundamentally untestable -- it is an economic argument, not a statistical test.

Classic example: Card (1995) used geographic proximity to a four-year college as an instrument for schooling. The argument: growing up near a college lowers the cost of attending (relevance), and distance itself does not directly affect wages (exclusion). Here we simulate an analogous setup.

### The 2SLS procedure

```
Stage 1: Regress X on Z (and controls): X_hat = Z * gamma_hat
Stage 2: Regress Y on X_hat (and controls): beta_hat_IV from Y ~ X_hat
```

Equivalently, in the simple just-identified case (one instrument, one endogenous variable), the Wald estimator gives:

```
beta_hat_IV = Cov(Y, Z) / Cov(X, Z) = Reduced Form / First Stage
```

This ratio has an intuitive interpretation: the reduced form tells you how much Y changes per unit of Z, and the first stage tells you how much X changes per unit of Z. Their ratio recovers how much Y changes per unit of X, using only the Z-driven variation in X.

***INTUITION:*** *The "filtered variation" perspective*

*Think of the endogenous X as containing two components:*

*X = (part predicted by Z) + (part correlated with epsilon)*

*OLS uses ALL variation in X, including the contaminated part. 2SLS replaces X with X_hat from the first stage, which contains ONLY the*

*Z-driven variation -- by construction uncorrelated with epsilon. You lose statistical power (X_hat has less variation than X), but you gain consistency. This is the fundamental variance-bias tradeoff of IV.*

### What IV estimates: the Local Average Treatment Effect (LATE)

IV does not generally recover the population ATE. Under the Imbens-Angrist LATE framework, 2SLS estimates the causal effect for "compliers" -- units whose treatment status is shifted by the instrument. In the schooling example, LATE is the return to schooling for people whose education decision was actually affected by college proximity.

This matters: if the compliers are a non-representative subpopulation, the LATE may differ substantially from the ATE. Always discuss who the compliers are in your specific application.

### Standard error subtlety in 2SLS

The second stage uses X_hat (fitted values from stage 1) in place of X. This creates a common pitfall: running OLS on (Y, X_hat) gives the correct beta_hat but INCORRECT standard errors. The naive residuals Y - X_hat * beta overstate the error variance because they include the first-stage prediction error. The correct formula uses residuals computed with the ACTUAL X: e = Y - X * beta_hat_2SLS, then plugs into the sandwich formula with X_hat in the "bread" matrices. Most software does this automatically, but understanding why matters for custom estimators.

***POLICY CONNECTION:*** *IV is the workhorse for evaluating programs where treatment is not randomly assigned. For mental health policy, plausible instruments might include distance to the nearest provider (affecting access but not directly outcomes), random assignment to a hotline vs. in-person counselor (in a partial-compliance RCT), or policy variation across jurisdictions that shifts treatment intensity.*

### Results

```
First-stage F-stat (instrument strength): 231.3
 Rule of thumb: F > 10 for 'strong' instrument. pass

 OLS beta_hat(schooling) : 2.723 <- biased upward by ability
 2SLS beta_hat(schooling) : 2.647 <- closer to truth (2.5)

 Standard errors for 2SLS estimate:
 Naive (OLS on fitted values): 0.2369 <- WRONG
 Correct homoskedastic 2SLS: 0.1525
 Robust HC1 2SLS: 0.1643
```

```
Note: Regressing Y on X_hat gives the right beta but not the right SE.
The naive SE overstates uncertainty because it treats X_hat as data
rather than a projection. The correct formula uses sigma^2*(X_hat'X_hat)^{-1}.
Robust SEs use the sandwich formula (Angrist & Pischke 2009).
```

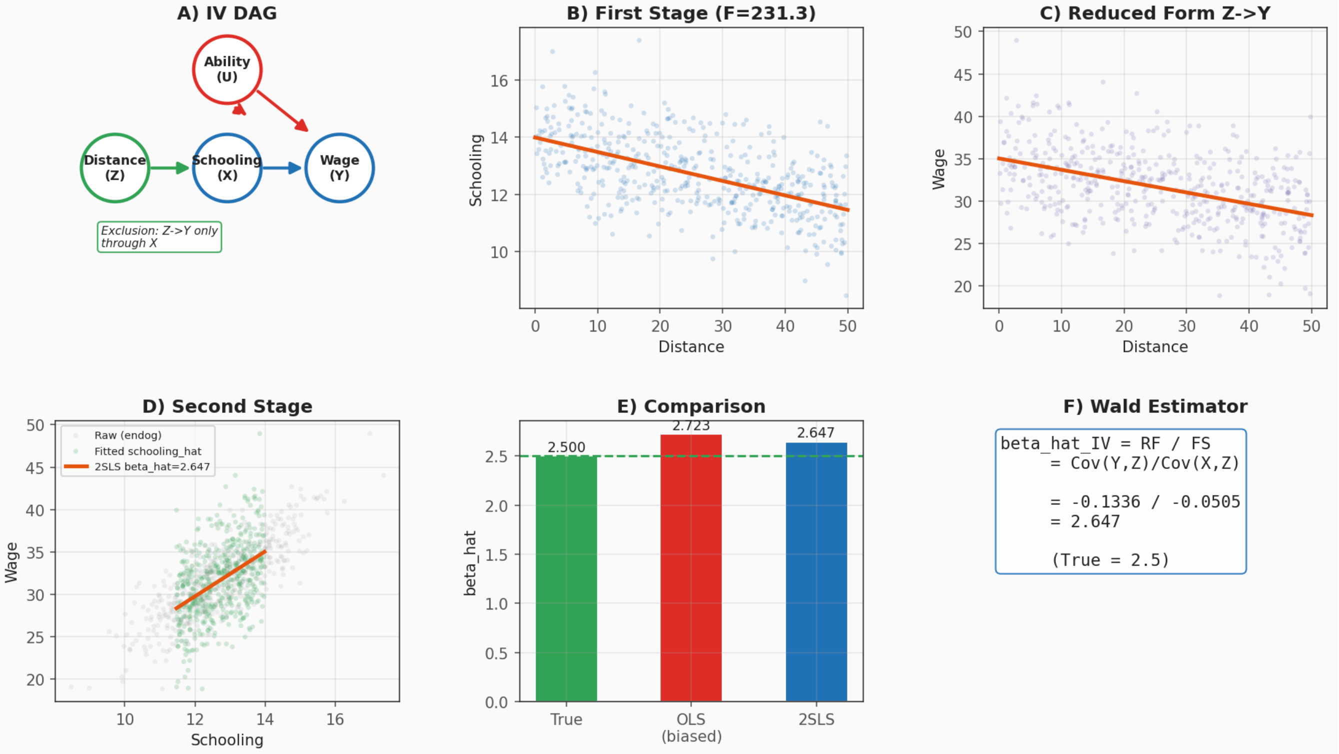
```
Wald Estimator:
beta_hat_IV = RF / FS = Cov(Y,Z)/Cov(X,Z)
= -0.1336 / -0.0505
= 2.647 (True = 2.5)
```

The IV estimate removes the ability bias by isolating only the variation in schooling driven by distance. The DAG in panel A shows the exclusion restriction: distance (Z) affects wages (Y) only through schooling (X), with no direct arrow from Z to Y.

Key diagnostic checklist for applied IV:

```
1. Report the first-stage F-stat (weak instruments bias IV toward OLS)
 2. Argue the exclusion restriction on economic/institutional grounds
 3. If over-identified (more instruments than endogenous vars), run
 the Sargan/Hansen J-test for overidentifying restrictions
 4. Report both OLS and IV: if they agree, endogeneity may be mild
```

## Section 4: Instrumental Variables / 2SLS



```
beta_hat_IV = RF / FS
     = Cov(Y,Z)/Cov(X,Z)

     = -0.1336 / -0.0505
     = 2.647

     (True = 2.5)
```

## Section 5: Panel Data — Fixed Effects (Within Estimator)

### What is panel data?

Panel (or longitudinal) data observes the same N units (people, firms, states, hospitals) across T time periods. This structure is extremely powerful because it lets us control for all time-invariant unobservable characteristics of each unit -- things like innate ability, institutional culture, or geography that we can never directly measure.

### The model

```
    y_it = alpha_i + X_it * beta + epsilon_it
```

Here alpha_i is the unit-specific "fixed effect" -- a constant unique to each unit that absorbs everything about that unit that does not change over time. The key assumption is strict exogeneity conditional on the fixed effect: E[epsilon_it | alpha_i, X_i1, ..., X_iT] = 0.

### The within estimator (demeaning)

Rather than estimating N dummy variables (computationally expensive and sometimes infeasible), subtract unit means:

```
    y_ddot_it = y_it - y_bar_i
    X_ddot_it = X_it - X_bar_i
```

Then run OLS on the demeaned data:

```
    beta_hat_FE = (X_ddot'X_ddot)^{-1} X_ddot'y_ddot
```

This is algebraically identical to including N unit dummies (by the FWL theorem from Section 2!) but computationally much cheaper.

### What FE does and does not solve

FE removes bias from any time-invariant confounder. In our simulation, if some workers always get more training because of innate traits (captured by alpha_i), FE removes that selection bias. However, FE cannot address time-varying confounders -- if a worker's motivation increases in the same period they get training, that bias remains.

FE also means you cannot estimate the effect of time-invariant regressors (e.g., race, sex in a person-level panel). Those are absorbed into alpha_i and "differenced away."

### Inference note: With panel data, errors are often serially correlated

within units. Standard practice is to cluster standard errors at the unit level to account for this.

***INTUITION:** Why demeaning works*

*Consider a worker observed over 5 years. Their wage in year t is:*

*wage_it = alpha_i + beta*training_it + epsilon_it*

*Taking the worker-level mean: wage_bar_i = alpha_i + beta*tr_bar_i + eps_bar_i*

*Subtracting: (wage_it - wage_bar_i) = beta*(training_it - tr_bar_i) + (eps_it - eps_bar_i)*

*The fixed effect alpha_i cancels! We identify beta purely from within-worker variation -- comparing each worker to themselves across time. If a worker gets more training in some years, did their outcome improve in those same years?*

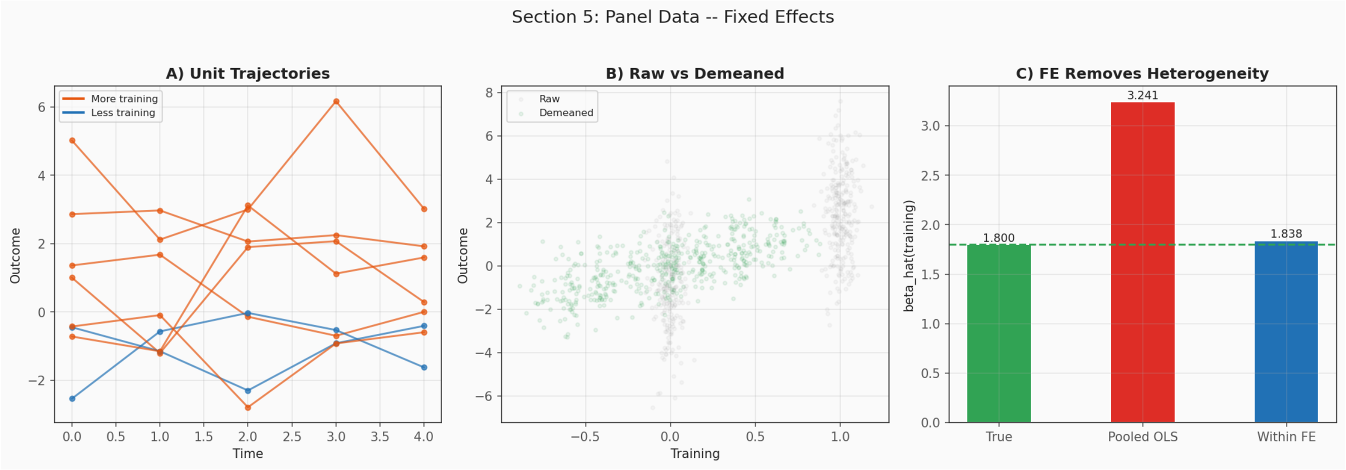
***POLICY CONNECTION:** Panel FE is invaluable for mental health policy research. A panel of states observed annually could control for all stable differences (culture, climate, demographics, institutional history) while estimating the effect of new funding policies on outcomes like hospitalization rates, employment, or suicide rates. The key limitation: FE cannot address time-varying confounders, so if states that expanded mental health funding also experienced economic booms, the FE estimate would be biased.*

### Results

```
 DGP: P(training=1) = logistic(0.5 * alpha_i), so Cov(alpha_i, tr) > 0.
 Corr(alpha_i, mean(tr_i)) = 0.695

 OLS (no FE) beta_hat(training): 3.241 <- biased by alpha_i selection
 Within FE beta_hat(training): 1.838 <- unbiased
 True effect = 1.8

 Inference on FE estimate:
Homoskedastic SE: 0.0937
 Clustered SE (unit): 0.1012
 With serial correlation within units, cluster SEs (Arellano 1987)
 are larger -- using homoskedastic SEs would overstate precision.
```

The pooled OLS estimate is biased because units with higher alpha_i are more likely treated (selection on unobservables). The within estimator removes this confounding by comparing each individual to their own average across time.

This is a powerful identification strategy for policy evaluation. For example, in mental health research, a panel of states observed over years could use state fixed effects to control for all stable differences across states (culture, demographics, historical funding levels) while estimating the effect of a new funding policy on outcomes like hospitalization rates or employment.


Section 5: Panel Data -- Fixed Effects
A) Unit Trajectories
More training
Less training
Outcome
Time
−2
0
2
4
6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
B) Raw vs Demeaned
Raw
Demeaned
Outcome
Training
−6
−4
−2
0
2
4
6
8
−0.5
0.0
0.5
1.0
C) FE Removes Heterogeneity
beta_hat(training)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1.800
3.241
1.838
True
Pooled OLS
Within FE

## Section 6: Difference-in-Differences (DiD)

### The idea

DiD is one of the most widely used quasi-experimental designs in applied economics and policy evaluation. A policy or treatment is rolled out to some units ("treated group") but not others ("control group") at a particular point in time. We cannot simply compare treated vs control after the policy (selection bias) or treated before vs after (time trends). DiD combines both comparisons to difference out confounds.

### The estimand

```
    tau_DiD = (y_bar_treat,post - y_bar_treat,pre)
```

- (y_bar_ctrl,post - y_bar_ctrl,pre)

The first difference removes the level difference between groups (selection). The second difference removes the common time trend. What remains is (under assumptions) the causal effect of the treatment.

### Regression formulation

```
    y_it = beta_0 + beta_1*Treated_i + beta_2*Post_t
    + beta_3*(Treated_i * Post_t) + epsilon_it

    beta_3 is the DiD estimator -- the coefficient on the interaction term.
```

### The parallel trends assumption

This is the crucial identifying assumption: absent treatment, the treated and control groups would have followed the same trend. It is fundamentally untestable for the post-treatment period, but we can check pre-treatment trends as a falsification test. If the groups were trending differently before the policy, DiD is biased.

Panel C of the figure illustrates this failure mode: when the treated group has a steeper pre-trend, the parallel trends counterfactual (extrapolating from the control group's trend) gives a biased estimate.

### Modern extensions

The canonical 2x2 DiD (one treated group, one time period) extends to staggered adoption designs where different units adopt treatment at different times. Recent econometric research (Callaway & Sant'Anna 2021, Sun & Abraham 2021, de Chaisemartin & d'Haultfoeuille 2020) shows that the standard two-way fixed effects estimator can be biased under treatment effect heterogeneity. Modern DiD uses group-time specific ATTs and robust aggregation.

In practice, always:

```
1. Plot pre-trends and test for parallel trends
2. Use clustered standard errors (at the group level)
3. Consider event-study specifications that show dynamic effects
4. Be transparent about the parallel trends argument
```

### Results

```
DiD coefficient beta_hat_3 : 3.308
True treatment effect: 3.0
```

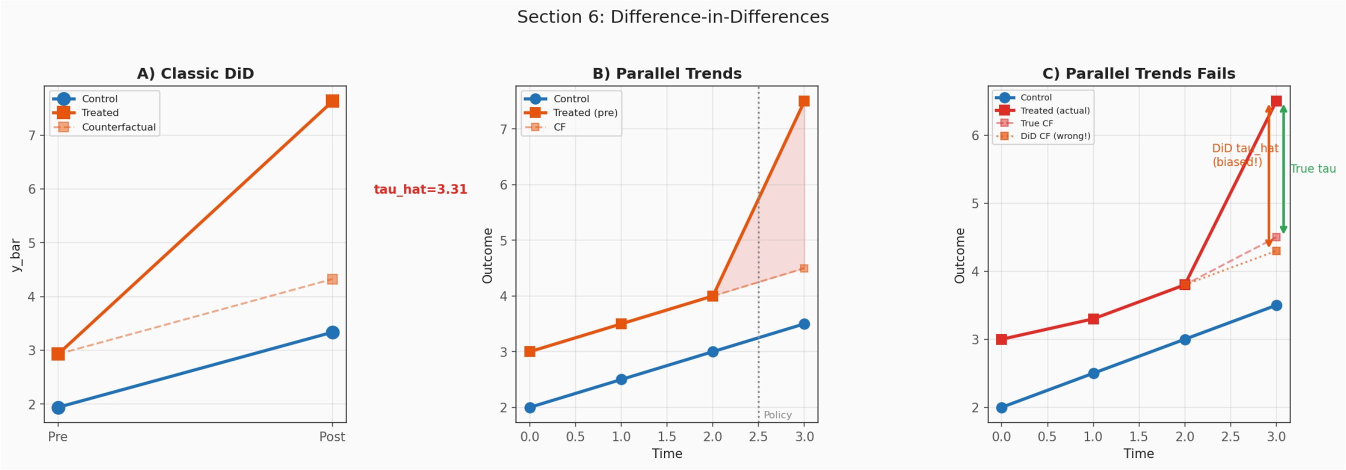
```
Homoskedastic SE: 0.2703
HC1 Robust SE: 0.2678
95% CI (homoskedastic): [2.778, 3.838]
```

In panel DiD settings, cluster standard errors at the unit level (Bertrand, Duflo & Mullainathan 2004). With serial correlation within units, naive SEs dramatically overreject -- Bertrand et al. show that clustering can increase SEs by a factor of 2-3x in typical applications.

The regression formulation with two-way fixed effects:

```
y_it = alpha_i + gamma_t + tau * D_it + epsilon_it
```

This absorbs unit fixed effects (alpha_i) and time fixed effects (gamma_t), with tau capturing the treatment effect.


Section 6: Difference-in-Differences
A) Classic DiD
Control
Treated
Counterfactual
y_bar
7
6
5
4
3
2
Pre
Post
tau_hat=3.31
B) Parallel Trends
Control
Treated (pre)
CF
Outcome
7
6
5
4
3
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Time
Policy
C) Parallel Trends Fails
Control
Treated (actual)
True CF
DiD CF (wrong!)
DiD tau_hat
(biased!)
True tau
Outcome
6
5
4
3
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Time

## Section 6B: Event Study Specifications and Diagnostics

### What is an event study?

The classic 2x2 DiD gives a single treatment effect estimate. The event study generalizes this by estimating separate effects for each time period relative to the treatment date. This serves two critical purposes:

1. PRE-TREND TEST: If the estimated effects in pre-treatment periods (leads) are statistically indistinguishable from zero, this supports the parallel trends assumption. If they trend away from zero, DiD is likely biased.

2. DYNAMIC EFFECTS: Post-treatment coefficients reveal how the effect evolves over time. Does it appear immediately? Grow over time? Fade out? This matters enormously for policy evaluation.

The event study regression:

```
y_it = alpha_i + gamma_t + Sum_{k != -1} tau_k * D_it^k + epsilon_it
```

where D_it^k = 1 if unit i is k periods from treatment at time t. Period k = -1 is the reference period (normalized to zero). The coefficients {tau_k} trace out the dynamic treatment effect path.

### Reading the event study plot (Panel A):

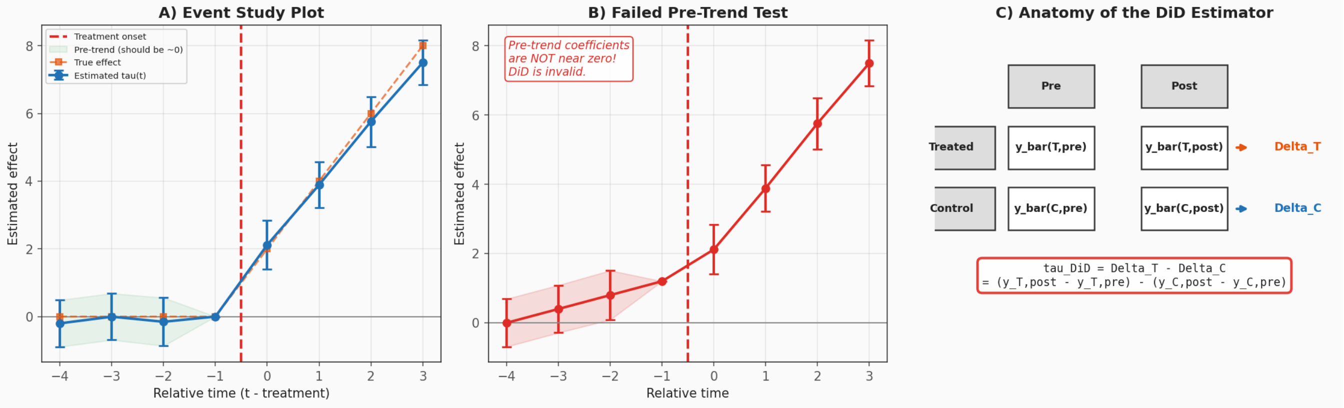
- Pre-treatment points near zero: parallel trends supported
- Post-treatment points showing an effect: treatment is working
- Confidence intervals crossing zero: not statistically significant
- Growing post-treatment effects: the treatment effect builds over time

### When pre-trends fail (Panel B):

If pre-treatment coefficients trend upward (or downward), it means the treated and control groups were already diverging before treatment. Any post-treatment difference could be a continuation of that pre-existing divergence rather than a treatment effect. In this case, DiD is invalid, and you should consider alternative identification strategies or at minimum attempt to detrend the data (though this introduces its own assumptions).

***POLICY CONNECTION:*** *Event studies are the gold standard for evaluating policy interventions. For mental health policy -- e.g., a state expanding Medicaid coverage for psychiatric services -- the event study would show whether hospitalization rates, employment, or crisis incidents changed after the policy, while verifying that treated and control states were trending similarly beforehand.*

Section 6B: Event Study & DiD Diagnostics



## Section 7: Regression Discontinuity Design (RDD)

### The setup

In many policy contexts, treatment is assigned based on whether a "running variable" (or "forcing variable") crosses a known cutoff. Examples: students receive a scholarship if their test score >= 70; districts receive federal aid if poverty rate > some threshold; patients receive treatment if a biomarker exceeds a clinical threshold.

The key insight: for units just above and just below the cutoff, assignment is "as good as random." A student scoring 70.1 is essentially identical to one scoring 69.9 in all respects except treatment status. This local randomization provides a credible causal estimate.

### Sharp RDD estimand

```
tau_RDD = lim_{x->c+} E[y|x] - lim_{x->c-} E[y|x]
```

This is a local average treatment effect (LATE) at the cutoff -- it tells us the causal effect of treatment for units right at the threshold. It does not generalize to units far from the cutoff without additional assumptions.

### Estimation: local linear regression

Fit separate linear regressions on each side of the cutoff within a bandwidth h:

```
Below: y = alpha_L + beta_L*(x - c) + epsilon for x in [c-h, c)
Above: y = alpha_R + beta_R*(x - c) + epsilon for x in [c, c+h]
```

The treatment effect estimate is tau_hat = alpha_R - alpha_L, i.e., the jump in the intercept at the cutoff.

### Bandwidth choice is critical: too narrow and you have too few

observations (high variance); too wide and the linear approximation breaks down (high bias). Optimal bandwidth selectors (Imbens & Kalyanaraman 2012, Calonico, Cattaneo & Titiunik 2014) balance this bias-variance tradeoff formally.

### Validity threats and diagnostics

```
1. Manipulation: If agents can precisely control the running variable
```

```
to sort above/below the cutoff, the design fails. The McCrary
(2008) density test checks for bunching at the cutoff.
2. Covariate smoothness: Pre-treatment covariates should be smooth
through the cutoff. A jump in covariates suggests confounding.
3. Bandwidth sensitivity: Results should be robust to reasonable
bandwidth choices. Report estimates across a range of bandwidths.
```

### Fuzzy RDD: When the cutoff determines eligibility but not perfect

compliance (some eligible don't take treatment, some ineligible do), use a "fuzzy RDD" -- essentially an IV where crossing the cutoff instruments for actual treatment receipt.

***INTUITION:*** *Why "local randomization" at the cutoff? Imagine lining up all students by test score. At score 69.9 vs 70.1, students are essentially identical in every way -- ability, motivation, background -- except that one barely missed the cutoff and one barely made it. Any difference in their outcomes must be due to the treatment (scholarship), not to pre-existing differences. This is why RDD is considered one of the most credible quasi-experimental designs: the identifying assumption (continuity) is mild and partially testable.*

***POLICY CONNECTION:*** *RDD naturally arises in many mental health policy contexts. Eligibility for subsidized treatment often depends on income thresholds, clinical severity scores, or age cutoffs. If a community mental health program serves patients with GAD-7 scores above 10, we can estimate its causal effect by comparing patients just above and below that threshold -- provided the score cannot be manipulated.*

## Results

```
RDD estimate (local linear): 3.399
Bias-corrected (quadratic) tau_hat=2.808, CI=[1.504, 4.112]
True effect: 4.0
Bandwidth used: +/-15 units around cutoff

McCrary density test at cutoff:
z-stat = 3.018, p-value = 0.003
No bunching detected (fail to reject H0)
```

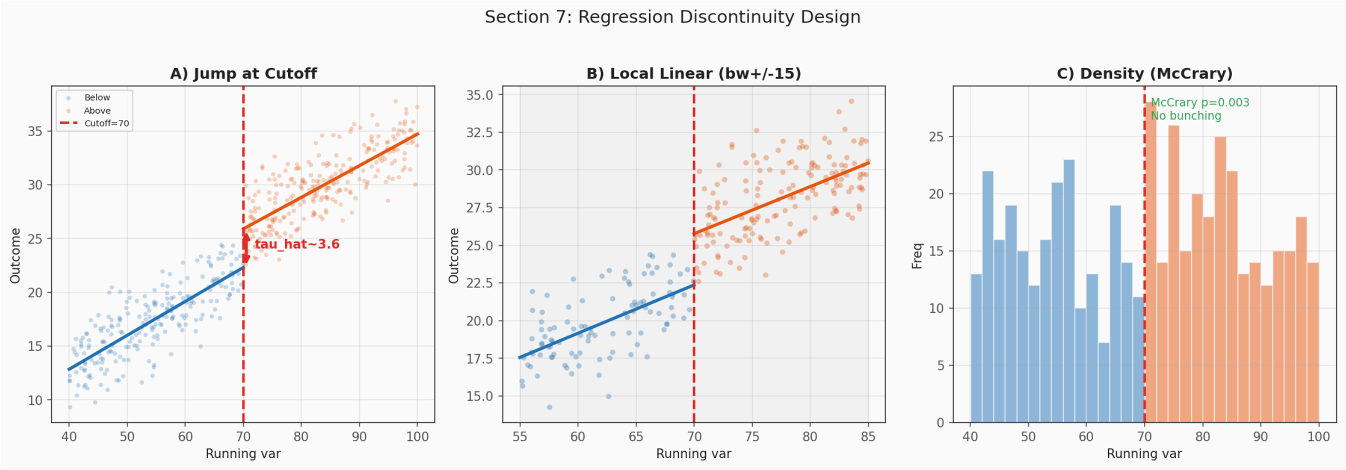
The local linear estimate is close to the true effect of 4.0. The bias-corrected estimate (using quadratic fits or rdrobust) provides a robust confidence interval following Calonico, Cattaneo & Titiunik

```
(2014). The McCrary (2008) density test formally checks for bunching
at the cutoff -- here we use a tighter significance level (alpha = 0.01) so a
p-value greater than alpha indicates no evidence of manipulation.
```

The figure shows:

```
Panel A: The clear jump in the outcome at the cutoff, with separate
linear fits on each side.
Panel B: The local linear regression zoomed into the bandwidth window.
Panel C: The density histogram with McCrary test result.
```

RDD is often considered one of the most credible quasi-experimental designs because the identifying assumption (continuity of potential outcomes at the cutoff) is relatively mild and partially testable.


Section 7: Regression Discontinuity Design
A) Jump at Cutoff
Below
Above
Cutoff=70
tau_hat~3.6
Outcome
Running var
10
15
20
25
30
35
40
50
60
70
80
90
100
B) Local Linear (bw+/-15)
Outcome
Running var
15.0
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
55
60
65
70
75
80
85
C) Density (McCrary)
McCrary p=0.003
No bunching
Freq
Running var
0
5
10
15
20
25
40
50
60
70
80
90
100

## Section 8: Binary Outcomes — Probit and Logit

### The problem with OLS for binary outcomes

When the outcome y takes values in {0, 1} (e.g., employed or not, enrolled or not, treated or not), OLS -- the "Linear Probability Model" (LPM) -- has two issues:

```
1. Predicted probabilities can fall outside [0, 1].
2. The error term is necessarily heteroskedastic since
Var(y|x) = P(y=1|x) * (1 - P(y=1|x)), which depends on x.
```

The LPM is still commonly used in applied work because its coefficients are directly interpretable as marginal effects, and with robust SEs the heteroskedasticity is handled. But for prediction and when the probability is near 0 or 1, nonlinear models are preferred.

### Logit and Probit models

Both model P(y=1|x) = G(x*beta), where G is a CDF that maps the linear index x*beta to [0, 1]:

```
Logit: G(z) = Lambda(z) = 1 / (1 + e^{-z}) (logistic CDF)
Probit: G(z) = Phi(z) (standard normal CDF)
```

The two CDFs are very similar in shape -- the logistic has slightly heavier tails. In practice they almost always give substantively identical results. The logit is more common in epidemiology and machine learning (due to the odds-ratio interpretation); probit is more common in some areas of economics.

Estimation is by Maximum Likelihood (see Section 9 for details).

### Interpreting coefficients: marginal effects

The raw coefficients beta_hat are NOT marginal effects. Because G is nonlinear, the effect of a one-unit change in x on P(y=1) depends on where you are on the curve:

```
dP/dx = G'(x*beta) * beta
```

Two common summaries:

```
(a) Marginal Effect at the Mean (MEM): evaluate at x = x_bar.
(b) Average Marginal Effect (AME): compute dP/dx for each obs,
then average. AME is generally preferred because it does not
depend on a potentially unrepresentative "average" individual.
```

### Practical guidance

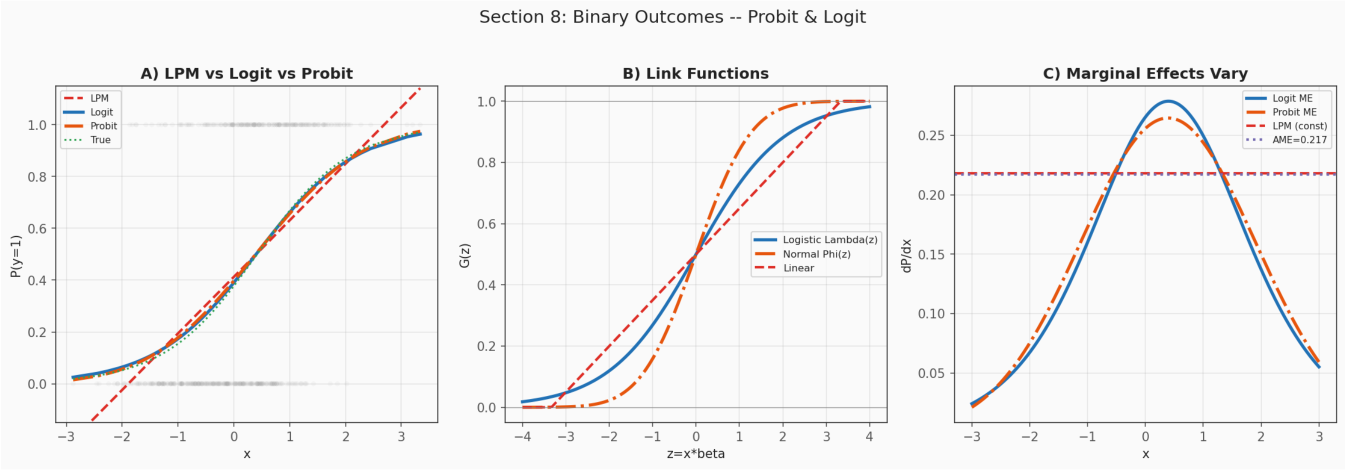
In applied work, always report AME (or MEM) alongside raw coefficients so readers can interpret the economic magnitude. The LPM coefficient is a rough approximation to AME when probabilities are in the middle of [0, 1], but diverges near the extremes.

## Results

```
  Logit coefficient beta_hat(x): 1.115
  Probit coefficient beta_hat(x): 0.664
  (coefficients not directly comparable -- different latent scale)

  Average Marginal Effect (AME):
  Logit AME: 0.2173
  Probit AME: 0.2161
  True AME : 0.2267

  AME Standard Errors (logit):
Delta method SE: 0.0161
Bootstrap SE: 0.0159 (500 reps)

  AME interpretation: a 1-unit increase in x raises P(y=1) by
  approximately 0.217 +/- 0.016 on average across the sample.
```

The delta method and bootstrap SEs are similar, which validates the asymptotic approximation. In practice, report AME with SEs (via delta method or bootstrap) so readers can assess statistical significance of the marginal effect, not just the raw logit/probit coefficient.

Note: The logit and probit AMEs are nearly identical, confirming that the choice of link function rarely matters for substantive conclusions. Panel C in the figure shows how the marginal effect varies with x -- it is largest when P(y=1) is near 0.5 (where the CDF is steepest) and shrinks toward 0 and 1.


Section 8: Binary Outcomes -- Probit & Logit
A) LPM vs Logit vs Probit
LPM
Logit
Probit
True
P(y=1)
x
−3
−2
−1
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
B) Link Functions
Logistic Lambda(z)
Normal Phi(z)
Linear
G(z)
z=x*beta
−4
−3
−2
−1
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
C) Marginal Effects Vary
Logit ME
Probit ME
LPM (const)
AME=0.217
dP/dx
x
−3
−2
−1
0
1
2
3
0.05
0.10
0.15
0.20
0.25

## Section 9: Maximum Likelihood Estimation — from scratch

### Why MLE matters

Nearly every estimator beyond OLS -- logit, probit, Poisson regression, Tobit, mixed-effects models -- is estimated via Maximum Likelihood. MLE is the bridge between specifying a probability model and obtaining parameter estimates that best explain the observed data.

### The principle

Given a parametric model with density f(y_i | x_i; beta), the likelihood function is the joint density of the observed sample viewed as a function of the parameters:

```
    L(beta) = Product_{i=1}^{n} f(y_i | x_i; beta)
```

Maximizing L is equivalent to maximizing the log-likelihood (which turns products into sums, improving numerical stability):

```
    l(beta) = Sum_{i=1}^{n} log f(y_i | x_i; beta)
```

For the logit model specifically:

```
    P(y_i=1 | x_i) = Lambda(x_i' beta) = 1 / (1 + e^{-x_i' beta})
    l(beta) = Sum [ y_i log Lambda(x_i' beta)
    + (1-y_i) log(1 - Lambda(x_i' beta)) ]
```

This is a globally concave function (for logit), so any gradient-based optimizer will find the unique maximum.

### Standard errors from the Fisher Information

The asymptotic distribution of the MLE is:

```
    beta_hat_MLE ~approx N(beta_0, I(beta_0)^{-1})
```

where I(beta) = -E[d^2 l / d beta d beta'] is the Fisher Information matrix. In practice, we use the observed information (the negative Hessian of the log-likelihood evaluated at beta_hat) and invert it to get the variance-covariance matrix.

### Numerical optimization

We use scipy.optimize.minimize with BFGS (a quasi-Newton method that approximates the Hessian from gradient evaluations). The BFGS path in panel C shows convergence from the starting point [0, 0] to the MLE in a few iterations -- the global concavity of the logit likelihood makes optimization straightforward.

### Profile likelihood and confidence intervals

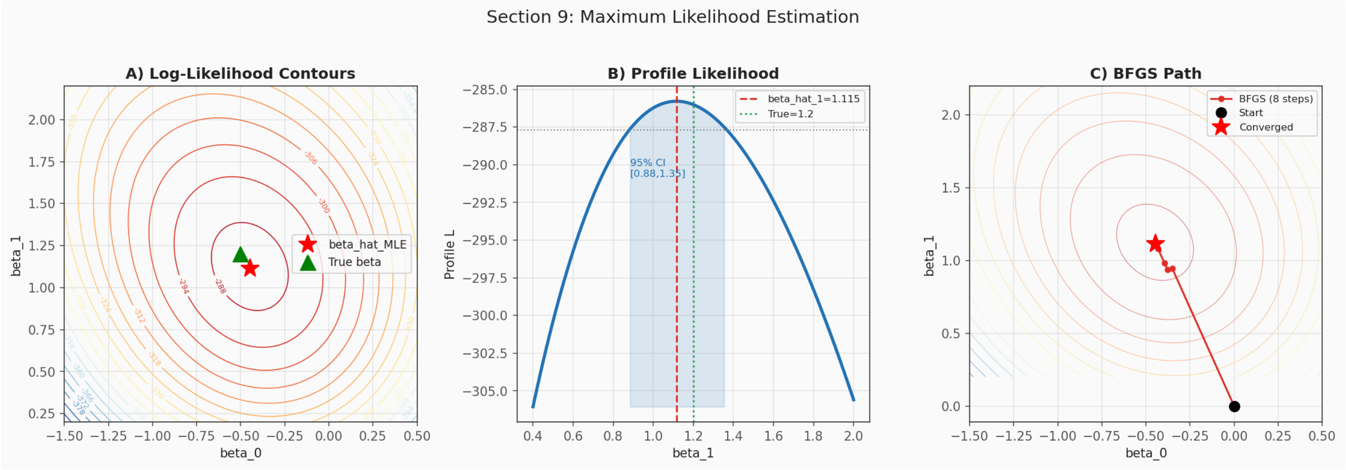
The profile likelihood for a single parameter beta_j is obtained by maximizing the log-likelihood over all other parameters for each fixed value of beta_j. The 95% confidence interval can be read off as the set of beta_j values where the profile likelihood is within 1.92 (= chi^2_1(0.95)/2) of its maximum. This is an alternative to Wald-type intervals (beta_hat +/- 1.96*SE) and can be more accurate in small samples or with nonlinear models.

### Results

```
MLE via scipy.minimize (BFGS):
beta_hat (intercept): -0.4454 (true: -0.5)
beta_hat (x) : 1.1155 (true: 1.2)
SE via observed Fisher info: intercept=0.1035, x=0.1244
```

The contour plot (panel A) shows the log-likelihood surface. The MLE (red star) sits at the peak. The profile likelihood (panel B) shows the 95% confidence interval for beta_1 as the range where the profile drops by at most 1.92 from its peak. The BFGS path (panel C) shows rapid convergence -- only a handful of iterations are needed because the logit likelihood is globally concave.

Being able to code MLE from scratch matters because many research questions require custom likelihoods that don't come in a package: mixture models, structural models with latent variables, models with non-standard censoring or selection. Understanding the mechanics -- write the likelihood, take derivatives (or let scipy approximate them), invert the Hessian for SEs -- is a transferable skill.


Section 9: Maximum Likelihood Estimation

A) Log-Likelihood Contours

B) Profile Likelihood

C) BFGS Path

## Section 10: Bootstrap Inference

### Motivation

Classical inference relies on asymptotic approximations: we derive the limiting distribution of an estimator (usually normal) and use it for CIs and tests. But these approximations can be poor when:

- The sample is small
- The estimator is nonlinear or complex (e.g., median, quantile

```
    regression, Gini coefficient, IV in small samples)
```

- The test statistic has an unknown or complicated distribution
- Standard errors involve complicated covariance structures

The bootstrap lets us approximate the sampling distribution of any statistic directly from the data, without relying on closed-form asymptotic results.

### The nonparametric bootstrap algorithm

```
    For b = 1, ..., B:
    1. Draw a sample of size n *with replacement* from the original data.
    2. Compute the estimator theta_hat*_b on this bootstrap sample.
    The collection {theta_hat*_1, ..., theta_hat*_B} approximates the
    sampling distribution of theta_hat.
```

From the bootstrap distribution we get:

```
    SE_boot = std(theta_hat*_b)
  Percentile CI: [quantile(2.5%), quantile(97.5%)]
```

For more refined intervals, the bias-corrected accelerated (BCa) bootstrap or the bootstrap-t method offer better coverage in finite samples.

### Why "with replacement" matters

Drawing with replacement means some observations appear multiple times and some not at all in each bootstrap sample. On average, each bootstrap sample contains about 63.2% unique observations (= 1 - 1/e). This mimics the randomness of repeated sampling from the population.

### When does the bootstrap fail?

- Extremely heavy-tailed distributions (the CLT is also slow here)
- Non-smooth statistics in small samples
- Dependent data (need block bootstrap or cluster bootstrap)

- When the parameter is on the boundary of the parameter space

### Practical guidance

B = 1000-2000 replications is standard for SE estimation. For percentile CIs, B >= 2000 is preferred. For BCa intervals, B >= 5000. The computational cost is usually trivial on modern hardware.

### Results

```
Bootstrap unique obs fraction: ~0.634 (theoretical 1-1/e = 0.632)
Each bootstrap sample contains ~63.2% unique observations (Efron 1979).
```

```
Analytic OLS SE(x) : 0.2061
Bootstrap SE(x) : 0.1953 (2000 replications)
```

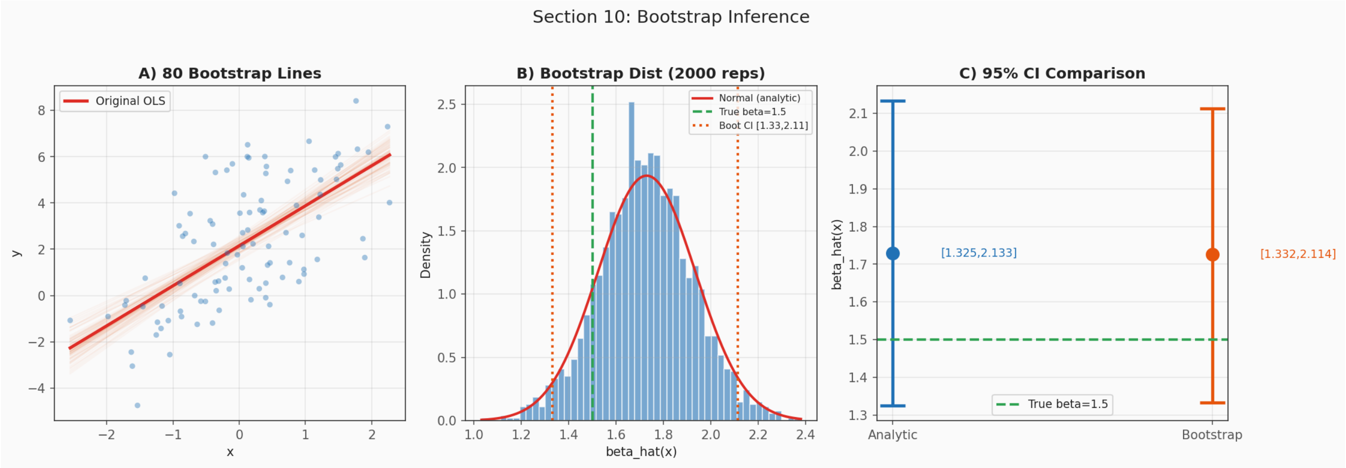
```
Bootstrap 95% CI : [1.3325, 2.1137]
Analytic 95% CI : [1.3254, 2.1334]
```

```
Note: np.random.seed(0) is set before this section for reproducibility.
Running the script twice yields identical bootstrap results.
```

In this well-behaved OLS setting, the bootstrap SE and analytic SE are very close, and both CIs cover the true value of 1.5. This is reassuring -- the bootstrap reproduces what theory predicts when theory is applicable.

The bootstrap really shines in situations where analytic SEs are unavailable or unreliable. For example:
- Ratio estimators (e.g., the Wald/IV estimator in small samples)
- Non-smooth statistics (e.g., sample median, quantile regression)
- Complex test statistics (e.g., testing equality of Gini coeffs)
- Cluster-robust inference with few clusters (wild cluster bootstrap)

Panel A shows 80 bootstrap regression lines overlaid on the original data, visualizing the uncertainty in the slope. Panel B shows the bootstrap distribution of beta_hat alongside the analytic normal approximation. Panel C compares the two 95% confidence intervals.


Section 10: Bootstrap Inference
A) 80 Bootstrap Lines
Original OLS
y
x
8
6
4
2
0
−2
−4
−2
−1
0
1
2
B) Bootstrap Dist (2000 reps)
Normal (analytic)
True beta=1.5
Boot CI [1.33,2.11]
Density
beta_hat(x)
2.5
2.0
1.5
1.0
0.5
0.0
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
C) 95% CI Comparison
beta_hat(x)
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
[1.325,2.133]
[1.332,2.114]
True beta=1.5
Analytic
Bootstrap

# Method Selection Flowchart

The flowchart above provides a practical decision tree for choosing an econometric method. Start at the top and follow the branches:

1. CAN YOU RANDOMIZE? If yes, run an RCT. OLS on treatment assignment gives the ATE directly. This is the gold standard for causal inference because randomization ensures E[epsilon|treatment] = 0 by design.

2. IS THERE A VALID INSTRUMENT? If you have a variable that predicts treatment but affects outcomes only through treatment, use IV/2SLS. Always check the first-stage F-statistic (F > 10 for strong instruments) and argue the exclusion restriction carefully.

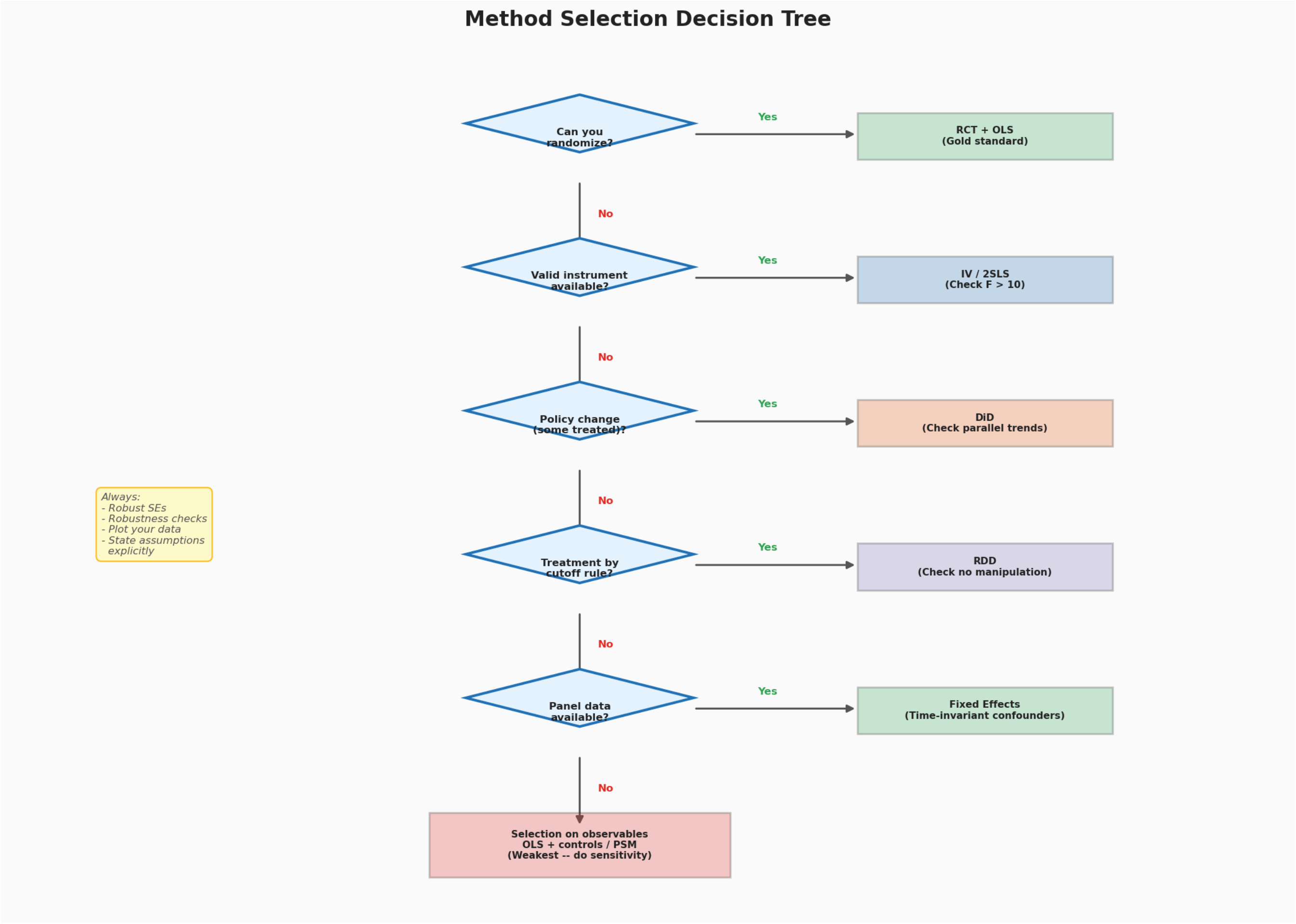
3. IS THERE A POLICY CHANGE? If a treatment rolls out to some units but not others at a point in time, DiD is your tool. Always check pre-trends via an event study specification.

4. IS TREATMENT ASSIGNED BY A CUTOFF? If there is a threshold rule (score >= c gets treatment), use RDD. Check for manipulation with a McCrary density test, and verify covariate smoothness at cutoff.

5. DO YOU HAVE PANEL DATA? Fixed effects remove time-invariant confounders. Combine with DiD if there is also a policy shock.

6. NONE OF THE ABOVE? You are in "selection on observables" territory. Use OLS with careful controls, propensity score methods, and always conduct sensitivity analysis (Oster 2019, Cinelli and Hazlett 2020) to bound how large unobserved confounding would need to be.

The Identification Credibility Ladder (right panel) summarizes the rough hierarchy of causal credibility, along with the key assumption each method requires. Note the fundamental tradeoff: methods with high internal validity (RDD, IV) often have limited external validity because they estimate local treatment effects for specific subpopulations.



Method Selection Decision Tree

Can you randomize?
Yes
RCT + OLS (Gold standard)
No
Valid instrument available?
Yes
IV / 2SLS (Check F > 10)
No
Policy change (some treated)?
Yes
DiD (Check parallel trends)
No
Treatment by cutoff rule?
Yes
RDD (Check no manipulation)
No
Panel data available?
Yes
Fixed Effects (Time-invariant confounders)
No
Selection on observables OLS + controls / PSM (Weakest -- do sensitivity)
Always:
- Robust SEs
- Robustness checks
- Plot your data
- State assumptions explicitly

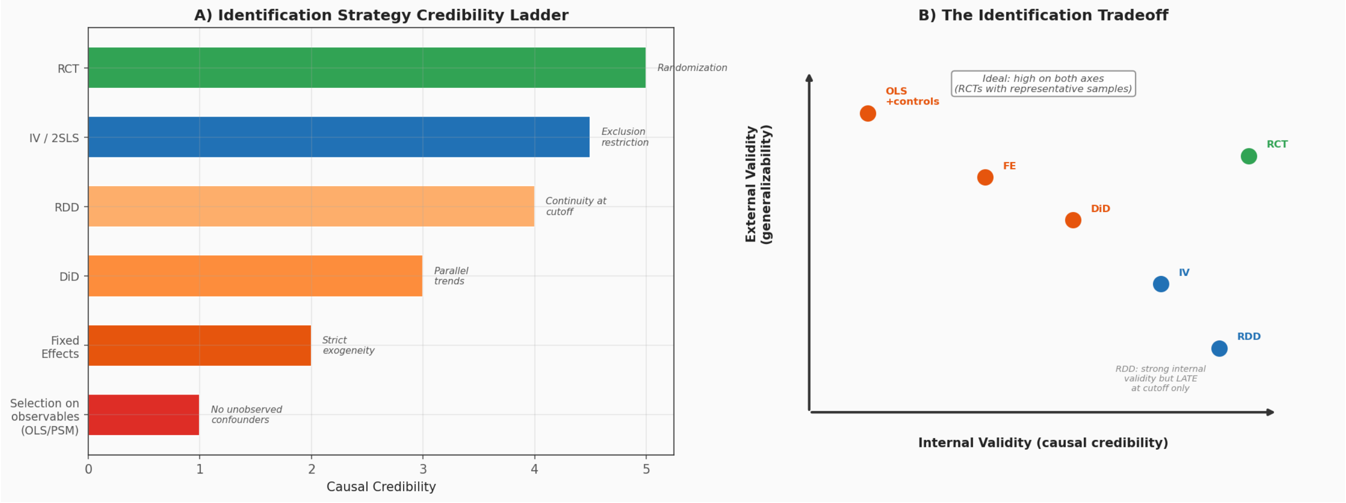
# The Identification Tradeoff: Internal vs External Validity

A critical concept for policy research is the tradeoff between internal validity (did we correctly estimate the causal effect for the population studied?) and external validity (does the finding generalize to the populations we care about?).

- RCTs can achieve both if the sample is representative, but many RCTs use convenience samples (volunteers, specific clinics).
- IV estimates the LATE for compliers, who may not represent the full population. Distance-to-college IV estimates the return to schooling for marginal students, not all students.
- RDD estimates the effect at the cutoff, which may differ from the effect at other points in the distribution.
- DiD estimates the ATT (average treatment effect on the treated), which may differ from the ATE if treatment effects are heterogeneous.
- OLS with controls has high external validity (uses the full sample) but questionable internal validity (unobserved confounders).

For policy, this matters enormously: if you want to know "what would happen if we expanded this program to everyone?" but your estimate is a LATE for a narrow subgroup, you need to think carefully about extrapolation. This is an active area of research in econometrics (Angrist & Fernandez-Val 2013, Mogstad, Santos & Torgovitsky 2018).

Summary: Causal Identification Strategy Guide



## Summary: When to Use What

SUMMARY: WHEN TO USE WHAT

========================

This primer covered ten core methods in the applied econometrician's toolkit. Here is a condensed guide for choosing among them.

| Method | When / What it solves |
|---|---|
| OLS | Baseline; assume exogeneity (no omitted variables) |
| Robust SEs | Always; protects against heterosk. without changing beta_hat |
| FWL / Partialing out | Understand what "controlling for X" actually does |
| IV / 2SLS | Endogenous regressor; need a valid instrument |
| Fixed Effects | Panel; time-invariant unobservables; within-unit variation |
| DiD | Policy rollout; treated/control groups; parallel trends |
| RDD | Sharp cutoff in assignment; LATE at threshold |
| Probit / Logit | Binary outcome; report marginal effects (not raw coefs) |
| MLE from scratch | Non-standard likelihood; fully custom models |
| Bootstrap | Small n, complex estimator, non-standard asymptotics |

### Key identification hierarchy (roughly strongest to weakest):

```
Randomized Experiment > IV > DiD > RDD > Selection-on-observables (OLS/PSM)
```

This ordering reflects how credible the causal claim typically is. An RCT directly controls treatment assignment; IV uses exogenous variation from an instrument; DiD exploits differential timing under parallel trends; RDD exploits a cutoff under continuity; and selection-on-observables (OLS with controls, or propensity score matching) assumes no unobserved confounders -- the strongest and least credible assumption.

### Choosing a method in practice -- a decision tree:

1. Can you randomize?

```
-> Yes: Run an experiment. OLS on treatment assignment estimates the ATE.
```

2. Is there an instrument?

```
-> Yes + strong first stage + credible exclusion: Use IV/2SLS.
```

3. Is there a policy change affecting some units but not others?

```
-> Yes + parallel trends plausible: Use DiD.
-> If staggered adoption: Use modern DiD estimators (Callaway-Sant'Anna).
```

4. Is treatment assigned by a cutoff in a running variable?

-> Yes + no manipulation: Use RDD (sharp or fuzzy).

5. Do you have panel data?

-> Yes: Use Fixed Effects to remove time-invariant confounders.

-> Combine with DiD if there is also a policy shock.

6. None of the above?

-> Selection-on-observables: OLS with careful controls, propensity score methods. Be transparent that unobserved confounding is a threat. Sensitivity analysis (e.g., Oster 2019, Cinelli & Hazlett 2020) can bound how large the bias from unobservables would need to be to overturn your results.

## Regardless of method, always:

- Report robust standard errors (or cluster as appropriate)
- Show the main specification AND robustness checks
- Be explicit about identifying assumptions and threats to validity
- Plot your data -- diagnostics catch problems that tests miss
- Remember: the goal is not statistical significance, but a credible estimate of a quantity that matters for the question you are asking