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.

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

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").

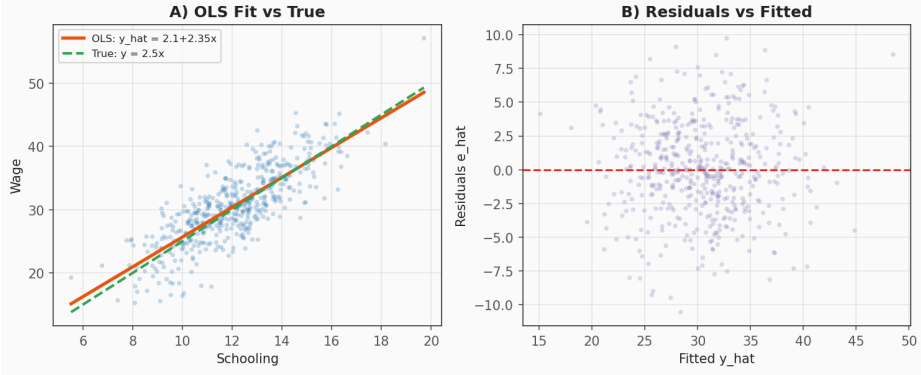
Why OLS fails here -- omitted variable bias (OVB)
In our simulation, ability affects both schooling (smarter people get more education) and wages (smarter people earn more), but we omit ability from the regression. The OVB formula is:
bias = beta_ability * [Cov(schooling, ability) / Var(schooling)]
Since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

Results

```
beta_hat (intercept) = 2.125 SE = 0.960
beta_hat (schooling) = 2.354 SE = 0.079
True coeff on schooling = 2.5 (ability bias inflates estimate)
```

Notice that beta_hat(schooling) > 2.5. The upward bias is exactly what the OVB formula predicts. The DAG in panel C of the figure makes the confounding structure visually explicit: ability is a common cause ("fork") of both schooling and wages, creating a spurious association that OLS cannot separate from the true causal effect without additional identifying assumptions or methods.

Section 1: OLS -- The Baseline Workhorse



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) : 1.997407

FWL partialled beta_hat(schooling) : 1.997407

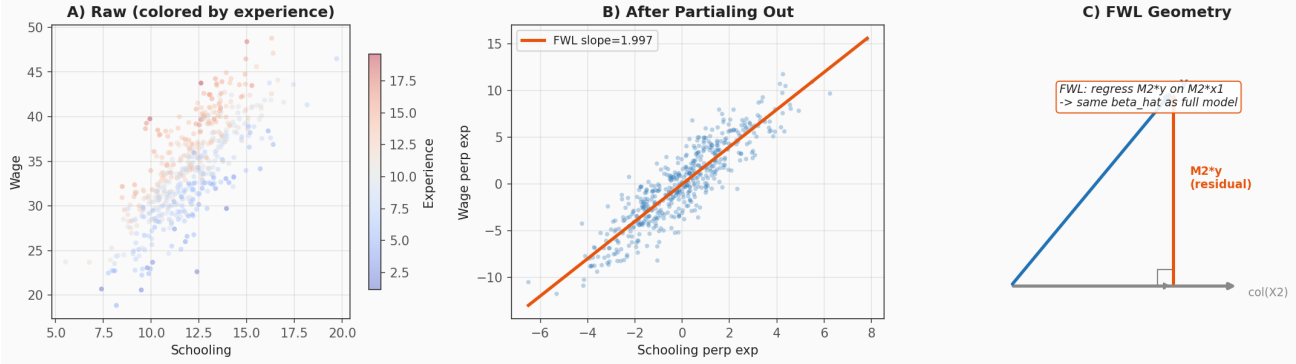
-> Identical (up to numerical precision). FWL theorem confirmed.

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



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."

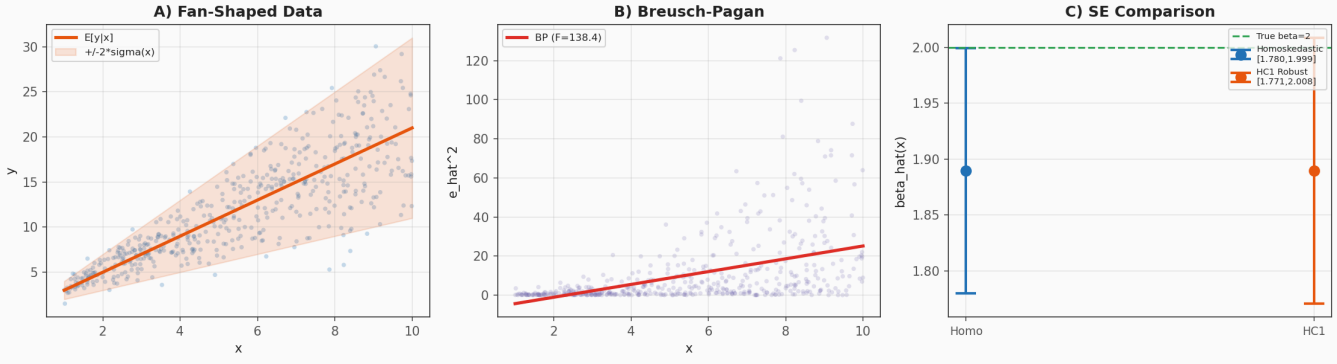
Results
Breusch-Pagan test F = 138.40, p = 0.0000
p < 0.05 -> reject homoskedasticity (the variance is not constant)

SE(x) -- homoskedastic : 0.0558
SE(x) -- HC1 robust : 0.0606

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



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.

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.

Results
First-stage F-stat (instrument strength): 245.3
Rule of thumb: F > 10 for 'strong' instrument. pass

OLS beta_hat(schooling) : 2.771 <- biased upward by ability
2SLS beta_hat(schooling) : 2.441 <- closer to truth (2.5)

Wald Estimator:

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

= -0.1224 / -0.0501

= 2.441 (True = 2.5)

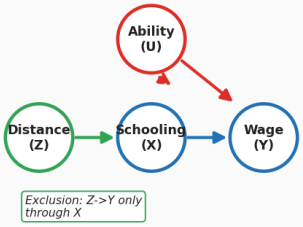
The IV estimate removes the ability bias by isolating only the variation in schooling driven by distance. The DAG in panel A shows the exclusion restriction: distance (Z) affects wages (Y) only through schooling (X), with no direct arrow from Z to Y.

Key diagnostic checklist for applied IV:

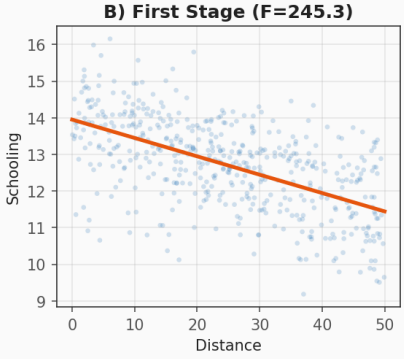
1. Report the first-stage F-stat (weak instruments bias IV toward OLS)
2. Argue the exclusion restriction on economic/institutional grounds
3. If over-identified (more instruments than endogenous vars), run the Sargan/Hansen J-test for overidentifying restrictions
4. Report both OLS and IV: if they agree,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

Section 4: Instrumental Variables / 2SLS

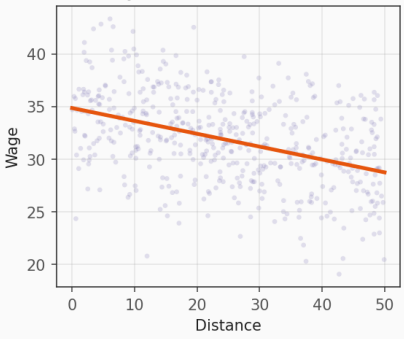
A) IV DAG



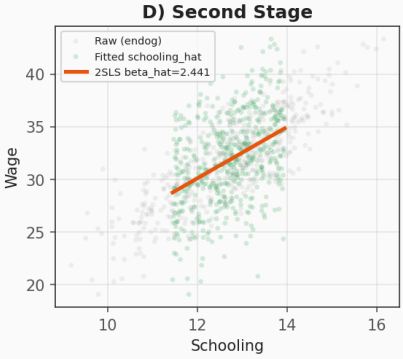
B) First Stage (F=245.3)



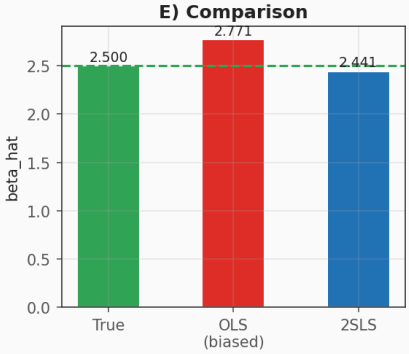
C) Reduced Form Z->Y



D) Second Stage



E) Comparison



F) Wald Estimator

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

      = -0.1224 / -0.0501
      = 2.441

      (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 α_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:
$\ddot{y}_{it} = y_{it} - \bar{y}_i$
$\ddot{X}_{it} = X_{it} - \bar{X}_i$

Then run OLS on the demeaned data:
$\hat{\beta}_{FE} = (\ddot{X}'\ddot{X})^{-1} \ddot{X}'\ddot{y}$

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."

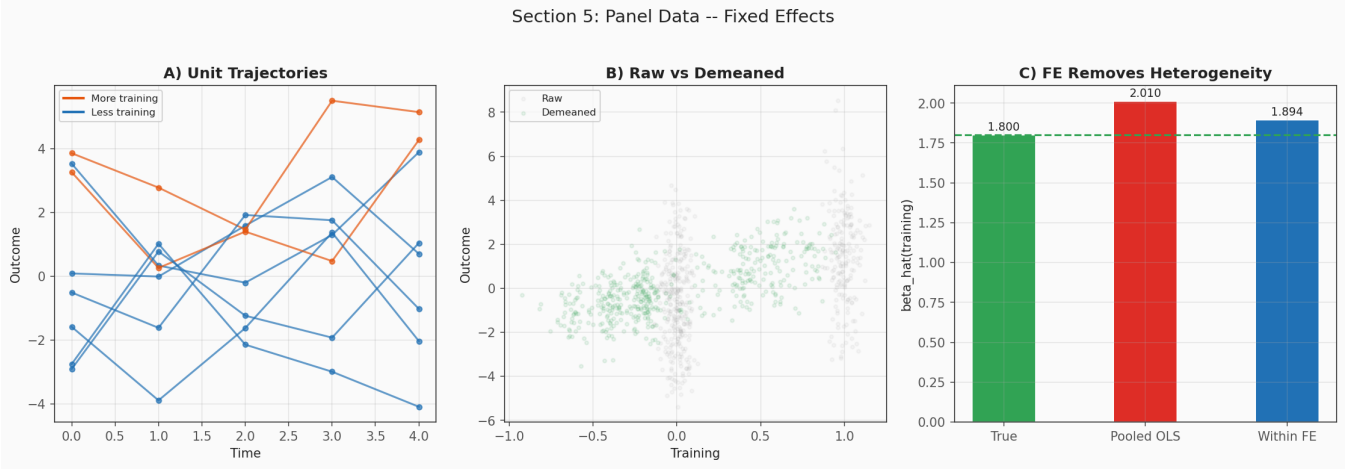
Inference note: With panel data, errors are often serially correlated within units. Standard practice is to cluster standard errors at the unit level to account for this.

Results
OLS (no FE) beta_hat(training): 2.010 <- biased by alpha_i
Within FE beta_hat(training): 1.894 <- unbiased
True effect = 1.8

The pooled OLS estimate is biased because individuals with higher

alpha_i (higher baseline outcome) may systematically differ in their training rates. The within estimator removes this confounding by comparing each individual to their own average across time -- it answers "when this person got training, how did their outcome change relative to their personal baseline?"

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
B) Raw vs Demeaned
Raw
Demeaned
Outcome
Training
C) FE Removes Heterogeneity
beta_hat(training)
1.800
2.010
1.894
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

$$\tau_{DiD} = (\bar{y}_{treat,post} - \bar{y}_{treat,pre}) - (\bar{y}_{ctrl,post} - \bar{y}_{ctrl,pre})$$

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

$$y_{it} = \beta_0 + \beta_1 \text{Treated}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Treated}_i \times \text{Post}_t) + \varepsilon_{it}$$

β_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 : 2.939

True treatment effect: 3.0

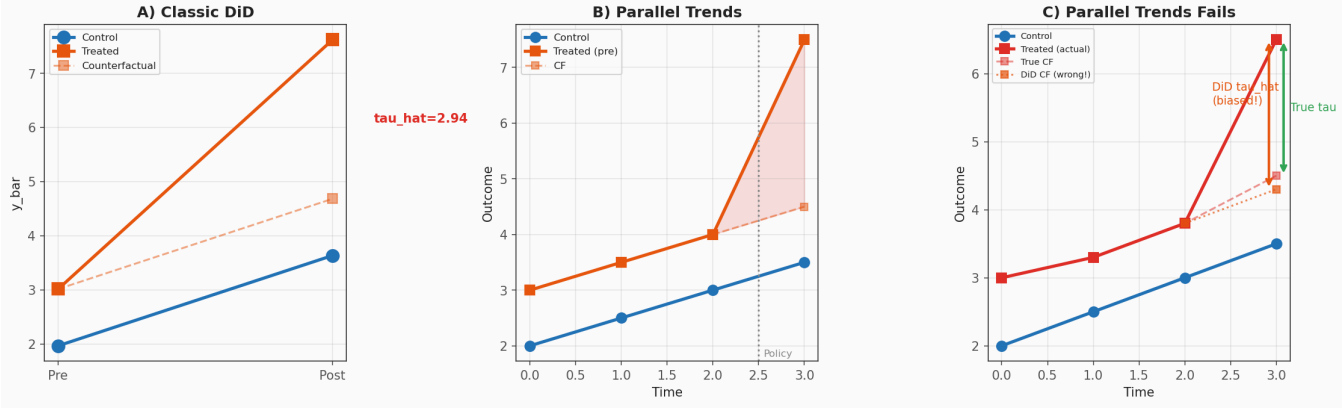
95% CI: [2.346, 3.531]

In modern practice, use two-way fixed effects DiD with clustered SEs:

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. Standard errors should be clustered at the unit level (or the level of treatment assignment) to account for serial correlation within units.

Section 6: Difference-in-Differences



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.

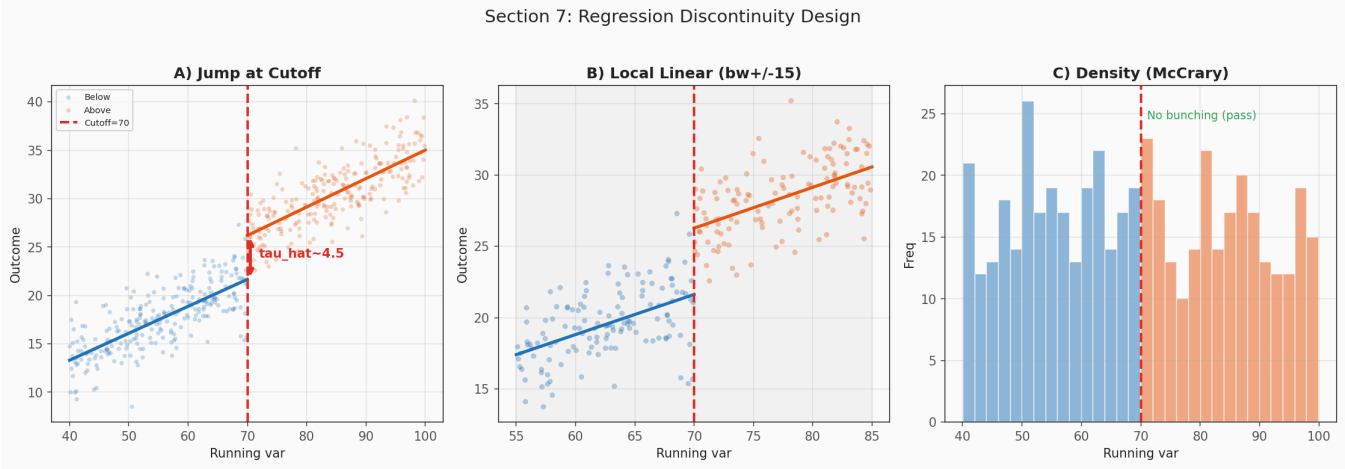
Fuzzy RDD: When the cutoff determines eligibility but not perfect compliance (some eligible don't take treatment, some ineligible do),

use a "fuzzy RDD" -- essentially an IV where crossing the cutoff
instruments for actual treatment receipt.

Results
RDD estimate of treatment effect: 4.650
True effect: 4.0
Bandwidth used: +/-15 units around cutoff

The estimate is close to the true effect of 4.0. 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,
where the treatment effect is measured as the gap between
the two intercepts at x = cutoff.
Panel C: The density histogram, which shows no suspicious bunching
at the cutoff -- consistent with no manipulation.

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.
The tradeoff is that the estimate is local to the cutoff and may
not generalize to units far from the threshold.


Section 7: Regression Discontinuity Design
A) Jump at Cutoff
Below
Above
Cutoff=70
tau_hat~4.5
Outcome
Running var
B) Local Linear (bw+/-15)
Outcome
Running var
C) Density (McCrary)
No bunching (pass)
Freq
Running var

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.

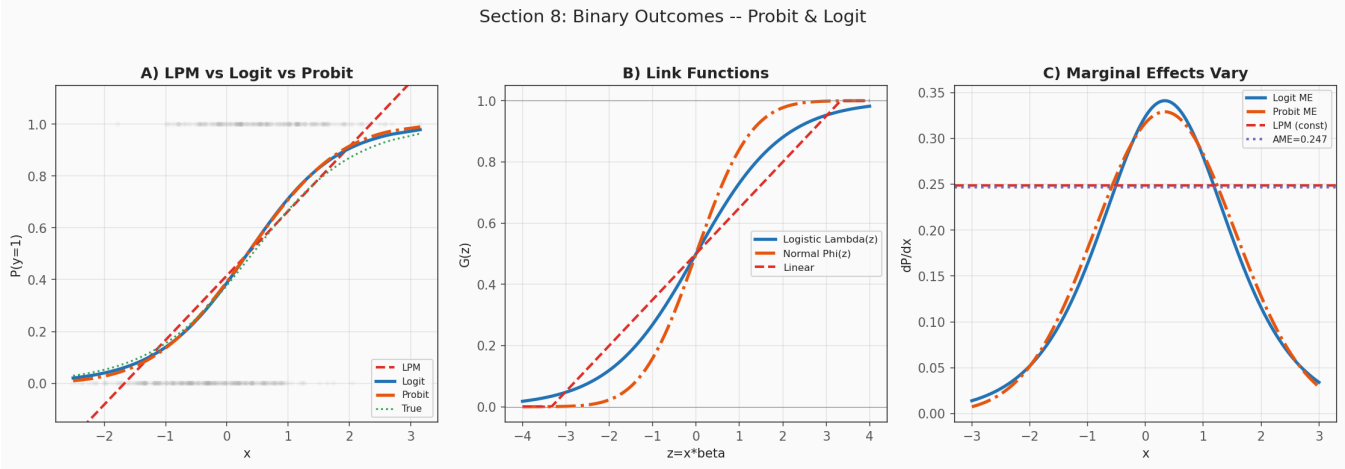
Results
Logit coefficient beta_hat(x): 1.364

Probit coefficient beta_hat(x): 0.825
(coefficients not directly comparable -- different latent scale)

Average Marginal Effect (AME):
Logit AME: 0.2467
Probit AME: 0.2491
True AME : 0.2269

AME interpretation: a 1-unit increase in x raises P(y=1) by approximately 0.247 on average across the sample.

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. The LPM, by contrast, assumes a constant marginal effect everywhere.


Section 8: Binary Outcomes -- Probit & Logit
A) LPM vs Logit vs Probit
P(y=1)
x
LPM
Logit
Probit
True
B) Link Functions
G(z)
z=x*beta
Logistic Lambda(z)
Normal Phi(z)
Linear
C) Marginal Effects Vary
dP/dx
x
Logit ME
Probit ME
LPM (const)
AME=0.247

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.

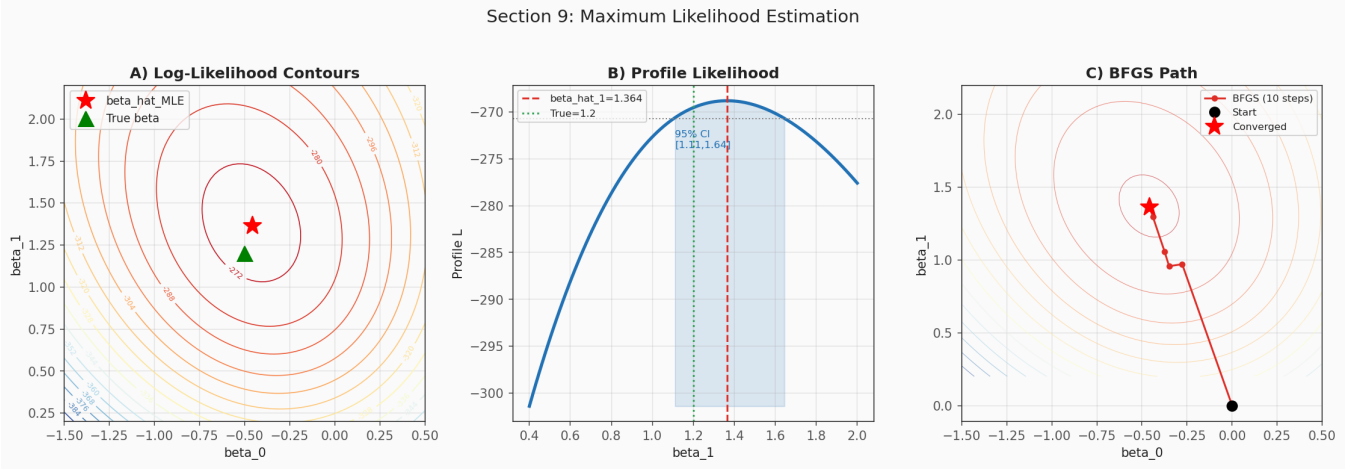
Profile likelihood and confidence intervals
The profile likelihood for a single parameter beta_j is obtained by maximizing the log-likelihood over all other parameters for each fixed value of beta_j. The 95% confidence interval can be read off as the set of beta_j values where the profile likelihood 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): -0.4584 (true: -0.5)
beta_hat (x) : 1.3643 (true: 1.2)
SE via observed Fisher info: intercept=0.1080, x=0.1386

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
beta_hat_MLE
True beta
beta_0
beta_1
B) Profile Likelihood
beta_hat_1=1.364
True=1.2
95% CI
[1.11,1.64]
Profile L
beta_1
C) BFGS Path
BFGS (10 steps)
Start
Converged
beta_0
beta_1

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.

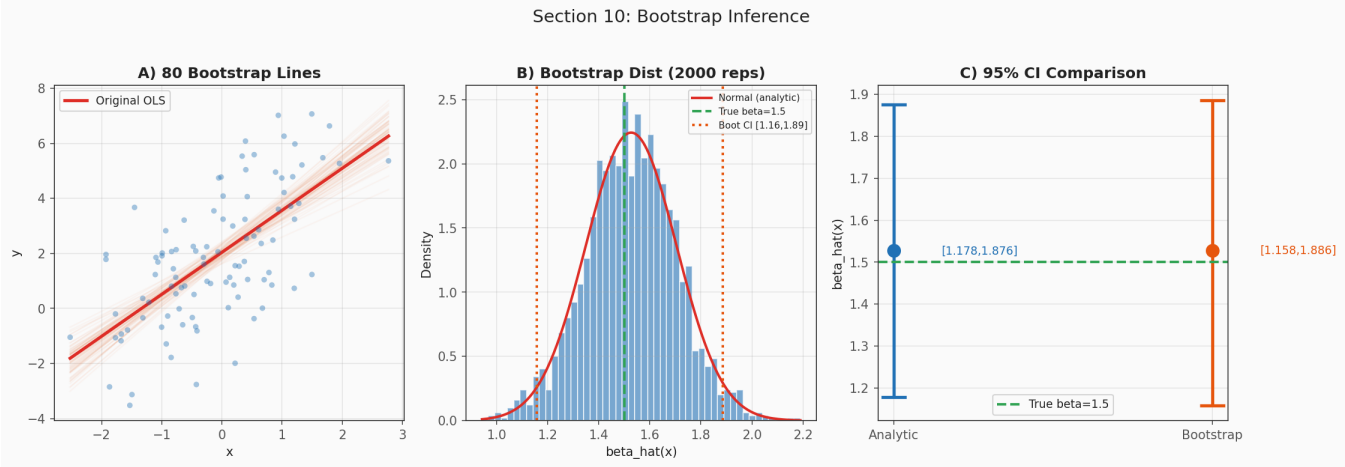
Results

Analytic OLS SE(x) : 0.1779
Bootstrap SE(x) : 0.1787 (2000 replications)
Bootstrap 95% CI : [1.1582, 1.8856]
Analytic 95% CI : [1.1784, 1.8757]

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
x
y
B) Bootstrap Dist (2000 reps)
Normal (analytic)
True beta=1.5
Boot CI [1.16,1.89]
beta_hat(x)
Density
C) 95% CI Comparison
[1.178,1.876]
[1.158,1.886]
True beta=1.5
Analytic
Bootstrap
beta_hat(x)

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