Econometrics PhD revision Notes

Kishan Narayan

PhD Student, Public Policy with Economics, Northeastern University

narayan.ki@northeastern.edu

About the Notes

These notes are based on my Advanced Econometrics I and Advanced Econometrics II coursework taught by Jianfae Cao at Northeastern University.

The material presented here does not reproduce class notes, slides, problem sets, or teaching materials, nor does it reflect the specific structure or pedagogical style used in the classroom. The notes were independently written by me while taking the courses, drawing on standard textbooks, research papers, and extensive self-study.

The primary goal of these notes is conceptual clarity for my own understanding, rather than adherence to any particular instructional format or textbook style. As such, explanations prioritize intuition, internal consistency, and connections across topics, sometimes at the expense of formal presentation.

In addition to the assigned course readings, the following references were heavily used:

- Bruce E. Hansen, *Econometrics*
- Jeffrey M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*
- William H. Greene, *Econometric Analysis*

I also refined and stress-tested many explanations by actively questioning and iterating on ideas using AI-assisted tools (e.g., Claude.ai), alongside traditional sources.

These notes are shared publicly for learning and reference purposes only. Any errors, omissions, or interpretations are entirely my own.

Contents

1. Start with the fundamental question: What is causality?
2. Build up the potential outcomes framework (Rubin Causal Model)
3. Explain the fundamental problem of causal inference
4. Move through identification strategies
5. Cover each method in depth

Some Notations (not exhaustive):

I use Wooldridge's notation where possible. Wooldridge typically uses:

- y for outcome
- w or d for treatment indicator (he often uses w in his cross-sectional book, d in panel)
- x for covariates
- u or e for error terms
- Subscript i for individuals
- Subscript t for time

For potential outcomes, the standard notation is:

- Y(1) or Y_1 for potential outcome under treatment
- Y(0) or Y_0 for potential outcome under control
- Or y_{1_i} and y_{0_i} in Wooldridge's style

I explain the intuition first, then formalize everything mathematically, and provide examples throughout.

PART 1: THE FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

1.1 What Do We Mean by "Causality"?

Before we write any equations, let's think about what we're trying to do.

The Everyday Notion

When we say "X causes Y," we mean something like: *if we were to change X, then Y would change as a result*. This is fundamentally different from saying "X and Y move together" (correlation).

Example:

- "Education causes higher earnings" means: if we *gave* someone more education, their earnings would increase.
- This is different from observing that educated people earn more (they might earn more because of ability, family background, etc.).

The Counterfactual Definition

Modern causal inference is built on **counterfactual thinking**:

What would have happened to this individual if they had received a different treatment?

This is the question at the heart of everything we'll study.

Example:

- Maria attended a job training program and now earns $45,000/year.
- The causal effect of the program on Maria is: *what Maria earns with training* minus *what Maria would have earned without training*.
- The problem: we can never observe what Maria would have earned without training. That's the **counterfactual**.

1.2 The Potential Outcomes Framework (Rubin Causal Model)

Now let's formalize this. The framework we'll use was developed by Jerzy Neyman (1923) for randomized experiments and extended by Donald Rubin (1974) to observational studies. Wooldridge adopts this framework throughout his work.

Setup and Notation

Consider a population of individuals indexed by $i = 1,2, \dots, N$.

Treatment Indicator:

Let $w_i$denote the **treatment status** of individual i:

$$w_i = \begin{cases} 1 & \text{if individual } i \text{ receives treatment} \\ 0 & \text{if individual } i \text{ does not receive treatment (control)} \end{cases}$$

Potential Outcomes:

For each individual i, we define **two potential outcomes**:

- y_{1i}= the outcome individual $i$would experience **if treated** ($w_i = 1$)
- y_{0i}= the outcome individual $i$would experience **if not treated** ($w_i = 0$)

These are also written as $y_i(1)$and $y_i(0)$, but Wooldridge typically uses the subscript notation.

Critical Point: Both y_{1i} and y_{0i} are defined for every individual, regardless of their actual treatment status. They represent what would happen under each scenario.

The Individual Treatment Effect

For individual i, the **causal effect of treatment** is:

$$\tau_i = y_{1i} - y_{0i}$$

This is the difference between what happens to individual $i$with treatment versus without treatment.

Example:

- Maria ($i = 1$): If $y_{1,1} = 45{,}000$and $y_{0,1} = 38{,}000$, then $\tau_1 = 7{,}000$.
- The job training program caused Maria's earnings to increase by $7,000.

1.3 The Fundamental Problem of Causal Inference

Here is the central challenge, precisely:

The Fundamental Problem of Causal Inference (Holland, 1986): For any individual i, we can observe at most one of the two potential outcomes y_{1i}or y_{0i}, never both.

The Observed Outcome

What we actually observe is:

$$y_i = w_i \cdot y_{1i} + (1 - w_i) \cdot y_{0i}$$

Or equivalently:

$$y_i = \begin{cases} y_{1i} & \text{if } w_i = 1 \\ y_{0i} & \text{if } w_i = 0 \end{cases}$$

This is called the **switching equation** or **observation rule**.

What's Missing?

For treated individuals ($w_i = 1$): We observe y_{1i}but not y_{0i}. For control individuals ($w_i = 0$): We observe y_{0i}but not y_{1i}.

The unobserved potential outcome is called the **counterfactual**. We can never directly observe $\tau_i = y_{1i} - y_{0i}$for any individual.

This is not a data problem—it's a logical impossibility. We cannot observe the same individual in two mutually exclusive states at the same time.

1.4 From Individual Effects to Average Effects

Since we cannot identify individual treatment effects, we shift our focus to **average effects** across populations.

Key Causal Parameters

1. Average Treatment Effect (ATE):

$$\text{ATE} = E[y_{1i} - y_{0i}] = E[y_{1i}] - E[y_{0i}]$$

This is the average causal effect across the entire population.

2. Average Treatment Effect on the Treated (ATT):

$$\text{ATT} = E[y_{1i} - y_{0i} \mid w_i = 1] = E[y_{1i} \mid w_i = 1] - E[y_{0i} \mid w_i = 1]$$

This is the average causal effect for those who actually received treatment.

3. Average Treatment Effect on the Untreated (ATU):

$$\text{ATU} = E[y_{1i} - y_{0i} \mid w_i = 0] = E[y_{1i} \mid w_i = 0] - E[y_{0i} \mid w_i = 0]$$

Why These Different Parameters?

These can differ when treatment effects vary across individuals (**heterogeneous treatment effects**) and when selection into treatment is non-random.

Example:

- Suppose people who benefit most from job training are more likely to enroll.
- Then ATT > ATE: the effect on participants is larger than the effect would be on a random person.
- This is called **selection on gains** or **essential heterogeneity**.

1.5 The Selection Problem: Why Correlation ≠ Causation

Now let's see precisely why we can't just compare treated and untreated groups.

The Naive Comparison

Suppose we compute the difference in average outcomes between treated and untreated individuals:

$$E[y_i \mid w_i = 1] - E[y_i \mid w_i = 0]$$

This is what we can calculate from data. Let's decompose it.

Using the observation rule:

- For treated: $E[y_i \mid w_i = 1] = E[y_{1i} \mid w_i = 1]$
- For untreated: $E[y_i \mid w_i = 0] = E[y_{0i} \mid w_i = 0]$

So the naive comparison gives us:

$$E[y_{1i} \mid w_i = 1] - E[y_{0i} \mid w_i = 0]$$

The Decomposition

Now, let's add and subtract $E[y_{0i} \mid w_i = 1]$:

$$\underbrace{E[y_{1i} \mid w_i = 1] - E[y_{0i} \mid w_i = 0]}_{\text{Naive Comparison}}$$

$$= \underbrace{E[y_{1i} \mid w_i = 1] - E[y_{0i} \mid w_i = 1]}_{\text{ATT}} + \underbrace{E[y_{0i} \mid w_i = 1] - E[y_{0i} \mid w_i = 0]}_{\text{Selection Bias}}$$

The Selection Bias Term

$$\text{Selection Bias} = E[y_{0i} \mid w_i = 1] - E[y_{0i} \mid w_i = 0]$$

This measures how the treated and untreated groups would differ **even in the absence of treatment**.

Interpretation:

- If people who choose treatment would have had better outcomes anyway → Selection Bias > 0
- If people who choose treatment would have had worse outcomes anyway → Selection Bias < 0

Example (Job Training):

- Suppose motivated, capable workers are more likely to enroll in job training.
- Even without training, these workers would earn more than those who don't enroll.
- Then $E[y_{0i} \mid w_i = 1] > E[y_{0i} \mid w_i = 0]$, so Selection Bias > 0.
- The naive comparison *overstates* the true effect of training.

Example (Medical Treatment):

- Suppose sicker patients are more likely to receive a new drug.
- Even without the drug, these patients would have worse outcomes.
- Then $E[y_{0i} \mid w_i = 1] < E[y_{0i} \mid w_i = 0]$, so Selection Bias < 0.
- The naive comparison *understates* (or even reverses) the true effect.

1.6 The Identification Problem

We now have a precise statement of the problem:

The Identification Problem:> The causal parameter (ATT or ATE) is not identified from observed data alone because we cannot observe the counterfactual mean $E[y_{0i} \mid w_i = 1]$.

"Identification" means: can we express the causal parameter in terms of observable quantities (population distributions of observed variables)?

Without additional assumptions, the answer is **no**.

The Road Ahead

The entire field of causal inference consists of strategies to solve this identification problem. Each strategy makes different **assumptions** that allow us to estimate causal effects. The key strategies we'll cover are:

Strategy	Key Assumption	Selection Bias Addressed
Randomized Experiments	Random assignment	Eliminates by design
Selection on Observables	Conditional independence	Controls for observable confounders
Instrumental Variables	Exclusion restriction	Uses exogenous variation
Difference-in-Differences	Parallel trends	Uses time variation
Regression Discontinuity	Local continuity	Exploits threshold rules
Synthetic Control	Weighted parallel trends	Constructs counterfactual

1.7 Summary of Part 1

Let's consolidate what we've learned:

Key Concepts:

1. Causal effects are defined as comparisons of potential outcomes: $\tau_i = y_{1i} - y_{0i}$
2. The fundamental problem: we only observe one potential outcome per individual
3. We focus on average effects: ATE, ATT, ATU
4. Naive comparisons conflate causal effects with selection bias
5. Identification requires additional assumptions

Key Equations:

Observed outcome (switching equation):

$$y_i = w_i \cdot y_{1i} + (1 - w_i) \cdot y_{0i}$$

Decomposition of naive comparison:

$$E[y_i \mid w_i = 1] - E[y_i \mid w_i = 0] = \text{ATT} + \text{Selection Bias}$$

Where:

$$\text{Selection Bias} = E[y_{0i} \mid w_i = 1] - E[y_{0i} \mid w_i = 0]$$

Exercises for Part 1

Before we continue, try these to solidify your understanding:

Exercise 1.1: Suppose there are only 4 individuals in the population with the following potential outcomes:

Individual	y_{0i}	y_{1i}	w_i
1	4	6	1
2	2	5	1
3	6	6	0
4	4	5	0

(a) Calculate each individual's treatment effect τ_i. (b) Calculate the ATE. (c) Calculate the ATT. (d) Calculate the naive difference in means $E[y_i \mid w_i = 1] - E[y_i \mid w_i = 0]$. (e) Calculate the selection bias and verify the decomposition.

Exercise 1.2: Give a real-world example where you would expect: (a) Positive selection bias (b) Negative selection bias

PART 2: RANDOMIZED EXPERIMENTS — THE GOLD STANDARD

2.1 Why Randomization Solves the Selection Problem

In Part 1, we established that the naive comparison of treated and untreated groups conflates the causal effect with selection bias. Now we'll see why **randomized experiments** eliminate this problem entirely.

The Setup

Consider an experiment where we:

1. Take a sample of Nindividuals from a population
2. **Randomly assign** each individual to treatment ($w_i = 1$) or control ($w_i = 0$)
3. Observe outcomes $y_i$for everyone

What Does Random Assignment Mean Formally?

Random assignment means that treatment status $w_i$is **statistically independent** of potential outcomes:

$$w_i \perp \square, \square, \square, \perp (y_{0i}, y_{1i})$$

This is read as: "w_i is independent of the pair (y_{0i}, y_{1i})."

Crucial insight: This independence is created **by design** through the physical act of randomization (coin flip, random number generator, etc.). It is not an assumption we hope is true—it is something we *make* true.

Implications of Independence

When $w_i \perp\perp (y_{0i}, y_{1i})$, several powerful results follow:

Result 1: Treatment and control groups have the same distribution of potential outcomes.

$$E[y_{0i} \mid w_i = 1] = E[y_{0i} \mid w_i = 0] = E[y_{0i}]$$
$$E[y_{1i} \mid w_i = 1] = E[y_{1i} \mid w_i = 0] = E[y_{1i}]$$

Result 2: Selection bias equals zero.

$$\text{Selection Bias} = E[y_{0i} \mid w_i = 1] - E[y_{0i} \mid w_i = 0] = 0$$

Result 3: The naive comparison identifies the ATE.

$$E[y_i \mid w_i = 1] - E[y_i \mid w_i = 0] = E[y_{1i}] - E[y_{0i}] = \text{ATE}$$

2.2 Formal Derivation

Let me prove this carefully so you see exactly how randomization works.

Step 1: Start with the naive comparison

$$E[y_i \mid w_i = 1] - E[y_i \mid w_i = 0]$$

Step 2: Apply the observation rule

For treated individuals: $y_i = y_{1i}$when $w_i = 1$For control individuals: $y_i = y_{0i}$when $w_i = 0$

So:

$$= E[y_{1i} \mid w_i = 1] - E[y_{0i} \mid w_i = 0]$$

Step 3: Apply independence

Since $w_i \perp$ ⬚⬚⬚ $\perp (y_{0i}, y_{1i})$:

$$E[y_{1i} \mid w_i = 1] = E[y_{1i}]$$
$$E[y_{0i} \mid w_i = 0] = E[y_{0i}]$$

Step 4: Conclude

$$E[y_i \mid w_i = 1] - E[y_i \mid w_i = 0] = E[y_{1i}] - E[y_{0i}] = \text{ATE}$$

This is the fundamental result of experimental economics.

2.3 Estimation: The Difference-in-Means Estimator

Now let's move from population parameters to sample estimation.

Setting

We have a random sample of Nindividuals:

- $N_1 = \sum_{i=1}^{N} w_i$individuals in the treatment group
- $N_0 = \sum_{i=1}^{N} (1 - w_i) = N - N_1$individuals in the control group

The Estimator

The natural estimator for the ATE is the **difference in sample means**:

$$\hat{\tau} = \bar{y}_1 - \bar{y}_0$$

Where:

$$\bar{y}_1 = \frac{1}{N_1} \sum_{i:w_i=1} y_i = \frac{\sum_{i=1}^{N} w_i \, y_i}{\sum_{i=1}^{N} w_i}$$

$$\bar{y}_0 = \frac{1}{N_0} \sum_{i:w_i=0} y_i = \frac{\sum_{i=1}^{N} (1 - w_i) y_i}{\sum_{i=1}^{N} (1 - w_i)}$$

Properties of the Estimator

Theorem 2.1 (Unbiasedness): Under random assignment, $\hat{\tau}$is an unbiased estimator of the ATE:

$$E[\hat{\tau}] = \text{ATE}$$

Proof:

$$E[\hat{\tau}] = E[\bar{y}_1 - \bar{y}_0] = E[\bar{y}_1] - E[\bar{y}_0]$$

By the law of iterated expectations and independence:

$$E[\bar{y}_1] = E[E[\bar{y}_1 \mid w_1, \ldots, w_N]] = E[y_{1i}]$$

Similarly: $E[\bar{y}_0] = E[y_{0i}]$

Therefore: $E[\hat{\tau}] = E[y_{1i}] - E[y_{0i}] = \text{ATE}$□

2.4 Variance and Inference

Variance of the Difference-in-Means Estimator

Under random sampling and random assignment:

$$\text{Var}(\hat{\tau}) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}$$

Where:

- $\sigma_1^2 = \text{Var}(y_{1i})$is the variance of potential outcomes under treatment

- $\sigma_0^2 = \text{Var}(y_{0i})$is the variance of potential outcomes under control

Estimation of the Variance

We estimate this using sample variances:

$$\widehat{\text{Var}}(\hat{\tau}) = \frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}$$

Where:

$$s_1^2 = \frac{1}{N_1 - 1} \sum_{i:w_i=1} (y_i - \bar{y}_1)^2$$
$$s_0^2 = \frac{1}{N_0 - 1} \sum_{i:w_i=0} (y_i - \bar{y}_0)^2$$

Standard Error

$$\text{SE}(\hat{\tau}) = \sqrt{\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}}$$

Confidence Intervals and Hypothesis Testing

By the Central Limit Theorem, for large samples:

$$\frac{\hat{\tau} - \text{ATE}}{\text{SE}(\hat{\tau})} \xrightarrow{d} N(0,1)$$

95% Confidence Interval:

$$\hat{\tau} \pm 1.96 \times \text{SE}(\hat{\tau})$$

Test of H_0: ATE $= 0$:

$$t = \frac{\hat{\tau}}{\text{SE}(\hat{\tau})}$$

Reject H_0at 5% level if $\mid t \mid > 1.96$.

2.5 Regression Formulation

A key insight from Wooldridge: the difference-in-means estimator is numerically identical to OLS regression.

The Regression Model

Consider the simple linear regression:

$$y_i = \alpha + \tau w_i + u_i$$

Where:

- $y_i$is the observed outcome
- $w_i$is the treatment indicator
- $\tau$is the coefficient of interest
- $u_i$is the error term

OLS Estimation

The OLS estimator of $\tau$is:

$$\hat{\tau}^{OLS} = \frac{\sum_{i=1}^{N}(w_i - \bar{w})(y_i - \bar{y})}{\sum_{i=1}^{N}(w_i - \bar{w})^2}$$

Theorem 2.2: $\hat{\tau}^{OLS} = \bar{y}_1 - \bar{y}_0$

Proof:

Since $w_i \in \{0,1\}$:

$$\bar{w} = \frac{N_1}{N}$$
$$\sum_{i=1}^{N}(w_i - \bar{w})^2 = N_1(1 - \bar{w})^2 + N_0(0 - \bar{w})^2 = N_1 \cdot \frac{N_0^2}{N^2} + N_0 \cdot \frac{N_1^2}{N^2} = \frac{N_1 N_0}{N}$$

For the numerator:

$$\sum_{i=1}^{N}(w_i - \bar{w})(y_i - \bar{y}) = \sum_{i:w_i=1}(1 - \bar{w})(y_i - \bar{y}) + \sum_{i:w_i=0}(0 - \bar{w})(y_i - \bar{y})$$
$$= (1 - \bar{w})\sum_{i:w_i=1}(y_i - \bar{y}) - \bar{w}\sum_{i:w_i=0}(y_i - \bar{y})$$

Using $\bar{y} = \bar{w}\bar{y}_1 + (1 - \bar{w})\bar{y}_0$and simplifying:

$$= \frac{N_0}{N} \cdot N_1(\bar{y}_1 - \bar{y}) - \frac{N_1}{N} \cdot N_0(\bar{y}_0 - \bar{y}) = \frac{N_1 N_0}{N}(\bar{y}_1 - \bar{y}_0)$$

Therefore:

$$\hat{\tau}^{OLS} = \frac{N_1 N_0 / N \cdot (\bar{y}_1 - \bar{y}_0)}{N_1 N_0 / N} = \bar{y}_1 - \bar{y}_0$$ $\square$

Why Use Regression?

If OLS gives the same answer, why bother with regression? Several reasons:

1. **Standard errors** are automatically computed (though we should use heteroskedasticity-robust SEs)
2. **Covariates** can be easily added (we'll discuss this next)
3. **Multiple treatments** can be handled naturally
4. **Interactions** and heterogeneous effects can be explored

2.6 Covariate Adjustment in Experiments

Even in randomized experiments, we often include covariates. Let's understand why and how.

The Model with Covariates

$$y_i = \alpha + \tau w_i + \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

Where $\mathbf{x}_i$is a vector of pre-treatment covariates (age, gender, baseline characteristics, etc.).

Why Include Covariates?

Reason 1: Precision (Efficiency)

Including covariates that predict the outcome reduces residual variance, leading to smaller standard errors for $\hat{\tau}$.

If $\mathbf{x}_i$explains variation in y_i, then:

- $\text{Var}(u_i)$decreases
- $\text{SE}(\hat{\tau})$decreases

- Confidence intervals become tighter
- Statistical power increases

Reason 2: Balance Checking

By randomization, $\mathbf{x}_i$should be balanced across treatment and control. Including covariates allows us to verify this and adjust for any chance imbalances.

Reason 3: Subgroup Analysis

Interactions like $w_i \times x_i$allow us to examine heterogeneous treatment effects.

Does Covariate Adjustment Affect Consistency?

Important result: Under random assignment, both the simple regression (without covariates) and the regression with covariates consistently estimate the ATE.

However, the regression with covariates may be **more efficient** (smaller variance).

Formal Result: Omitted Variable Bias in Experiments

In observational studies, omitting relevant variables causes bias. In experiments, this is not the case.

Recall the omitted variable bias formula. If the true model is:

$$y_i = \alpha + \tau w_i + \mathbf{x}_i' \boldsymbol{\beta} + u_i$$

But we estimate:

$$y_i = \tilde{\alpha} + \tilde{\tau} w_i + \tilde{u}_i$$

Then:

$$\text{plim } \tilde{\tau} = \tau + \boldsymbol{\beta}' \boldsymbol{\delta}$$

Where $\boldsymbol{\delta}$is the coefficient from regressing $\mathbf{x}_i$on w_i.

Under randomization: $w_i \perp\!\!\!\perp \mathbf{x}_i$(for pre-treatment variables), so $\boldsymbol{\delta} = \mathbf{0}$

Therefore: plim $\tilde{\tau} = \tau$. No bias!

2.7 A Complete Example: The RAND Health Insurance Experiment

Here is a classic experimental study.

Background

The RAND Health Insurance Experiment (1974-1982) randomly assigned families to health insurance plans with different levels of cost-sharing to study how insurance affects healthcare utilization and health outcomes.

Setup

- y_i= annual healthcare expenditures
- $w_i = 1$if assigned to free care (no cost-sharing), $w_i = 0$if assigned to cost-sharing plan
- $\mathbf{x}_i$= demographic variables (age, income, health status at baseline)

Hypothetical Data

Suppose we have $N = 2000$families:

- $N_1 = 1000$assigned to free care
- $N_0 = 1000$assigned to cost-sharing

Results:

- $\bar{y}_1 = \$750$(average expenditure, free care)
- $\bar{y}_0 = \$550$(average expenditure, cost-sharing)
- $s_1 = \$400$, $s_0 = \$350$

Estimation

Point Estimate:

$$\hat{\tau} = \bar{y}_1 - \bar{y}_0 = 750 - 550 = \$200$$

Standard Error:

$$\text{SE}(\hat{\tau}) = \sqrt{\frac{400^2}{1000} + \frac{350^2}{1000}} = \sqrt{160 + 122.5} = \sqrt{282.5} \approx \$16.81$$

95% Confidence Interval:

$$200 \pm 1.96 \times 16.81 = [166.95,233.05]$$

t-statistic:

$$t = \frac{200}{16.81} \approx 11.9$$

Interpretation: Free healthcare causes families to spend approximately $200 more per year on healthcare, and this effect is highly statistically significant.

Regression Output

Running OLS: $y_i = \alpha + \tau w_i + u_i$

Variable	**Coefficient**	**Std. Error**	**t-stat**
Constant (α)	550.00	11.07	49.7
Free Care (τ)	200.00	16.81	11.9

$R^2 = 0.066$, $N = 2000$

The coefficient on w_i is exactly $\bar{y}_1 - \bar{y}_0 = 200$.

2.8 Threats to Validity in Experiments

Even randomized experiments can fail. Understanding these threats is essential.

Internal Validity Threats

1. Noncompliance

Not everyone follows their assignment:

- Some assigned to treatment don't take it
- Some assigned to control obtain treatment

This breaks the link between assignment and actual treatment.

Solution: Intent-to-treat (ITT) analysis or instrumental variables (we'll cover this later).

2. Attrition

People drop out of the study, and dropout may be related to treatment.

Example: In a job training study, if discouraged trainees drop out, we only observe successful trainees, biasing results upward.

Solution: Bounds analysis, careful tracking, intention-to-treat.

3. Interference (SUTVA Violation)

The Stable Unit Treatment Value Assumption (SUTVA) requires that one person's treatment doesn't affect another's outcome.

Example: In a vaccine trial, if vaccinated people reduce disease transmission, unvaccinated people benefit too, diluting the measured effect.

4. Hawthorne Effects

Subjects behave differently because they know they're being studied.

External Validity Threats

1. Sample Selection

Experimental subjects may differ from the population of interest.

2. Experimental Setting

Laboratory or artificial conditions may not reflect real-world implementation.

3. Equilibrium Effects

Small-scale experiments miss general equilibrium effects that would occur at scale.

2.9 The SUTVA Assumption

This assumption is fundamental but often unstated. Let's be explicit.

Formal Statement

Stable Unit Treatment Value Assumption (SUTVA):

1. **No interference:** The potential outcomes for individual $i$depend only on i's own treatment status, not on the treatment status of other individuals.

$$y_i(w_1, w_2, \ldots, w_N) = y_i(w_i)$$

2. **No hidden variations of treatment:** There is only one version of each treatment level.

When SUTVA Fails

Interference examples:

- Vaccination (herd immunity)
- Education (peer effects)

- Social programs (displacement effects in labor markets)

Multiple versions examples:

- Different doctors administering the same treatment
- Different quality of program implementation across sites

Notation Under SUTVA

SUTVA justifies writing potential outcomes as $y_{0i}$and y_{1i}rather than $y_i(w_1, \dots, w_N)$.

2.10 Statistical Power and Sample Size

Before running an experiment, we need to determine how many subjects we need.

The Power Calculation

Power = Probability of rejecting $H_0$when the alternative is true.

For a two-sided test at significance level α:

$$\text{Power} = \Phi\left(\frac{|\tau|}{\text{SE}(\hat{\tau})} - z_{1-\alpha/2}\right) + \Phi\left(\quad -\frac{|\tau|}{\text{SE}(\hat{\tau})} - z_{1-\alpha/2}\right)$$

For large effect sizes, the second term is negligible:

$$\text{Power} \approx \Phi\left(\frac{|\tau|}{\text{SE}(\hat{\tau})} - z_{1-\alpha/2}\right)$$

Sample Size Formula

For equal-sized treatment and control groups ($N_1 = N_0 = N/2$) and common variance σ^2:

$$\text{SE}(\hat{\tau}) = \sigma\sqrt{\frac{4}{N}}$$

To achieve power $(1-\beta)$for detecting effect τat significance α:

$$N = \frac{4\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\tau^2}$$

Example Calculation

Suppose:

- Expected effect: $\tau = 0.2\sigma$(small effect)
- Desired power: 80% ($\beta = 0.2$, so $z_{0.8} = 0.84$)
- Significance level: 5% ($\alpha = 0.05$, so $z_{0.975} = 1.96$)

$$N = \frac{4 \times 1 \times (1.96 + 0.84)^2}{0.2^2} = \frac{4 \times 7.84}{0.04} = 784$$

We need about 784 subjects total (392 per group) to detect a 0.2 standard deviation effect with 80% power.

2.11 Summary of Part 2

Key Results:

1. **Random assignment creates independence:** $w_i \perp \square\square\square \perp (y_{0i}, y_{1i})$
2. **Selection bias vanishes:**

$$E[y_{0i} \mid w_i = 1] = E[y_{0i} \mid w_i = 0]$$

3. **Difference-in-means identifies the ATE:**

$$\hat{\tau} = \bar{y}_1 - \bar{y}_0 \overset{p}{\rightarrow} \text{ATE}$$

3.

4. **OLS is equivalent:** Regressing $y_i$on $w_i$gives the same estimate.
5. **Covariates improve precision** but don't affect consistency in experiments.
6. **SUTVA is required** for potential outcomes to be well-defined.

Key Equations:

Variance of estimator:

$$\text{Var}(\hat{\tau}) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}$$

Required sample size:

$$N = \frac{4\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\tau^2}$$

Exercises for Part 2

Exercise 2.1: In a randomized experiment with 500 treated and 500 control observations, you find $\bar{y}_1 = 82, \bar{y}_0 = 78, s_1 = 15, s_0 = 12.$

(a) Estimate the ATE. (b) Calculate the standard error. (c) Construct a 95% confidence interval. (d) Test H_0: ATE $= 0$at the 5% level.

Exercise 2.2: Explain intuitively why including pre-treatment covariates in a randomized experiment improves precision but does not affect consistency.

Exercise 2.3: A researcher runs a randomized experiment but 20% of the treatment group doesn't actually receive the treatment (noncompliance). The researcher analyzes the data by comparing those who actually received treatment to those who didn't. What's wrong with this approach?

Exercise 2.4: Calculate the sample size needed to detect an effect of 0.3 standard deviations with 90% power at the 5% significance level.

Transition to Observational Methods

Randomized experiments are the gold standard, but they're often:

- **Infeasible:** We can't randomly assign education, smoking, or macroeconomic policies.
- **Unethical:** We can't randomly assign harmful treatments.
- **Expensive:** Large-scale experiments are costly.
- **Not externally valid:** Lab results may not generalize.

In the next part, we'll begin by **observational methods** that attempt to approximate experimental conditions using non-experimental data.

PART 3: SELECTION ON OBSERVABLES

structure:

1. The conditional independence assumption (CIA) / Selection on Observables
2. Regression as a tool for causal inference under CIA
3. The propensity score theorem
4. Matching estimators
5. Inverse probability weighting
6. Doubly robust estimation
7. Practical considerations and diagnostics

Regression, Matching, and Propensity Scores

3.1 The Core Idea: Controlling for Confounders

In observational studies, we cannot randomly assign treatment. Instead, individuals **select** into treatment based on their characteristics. The key insight of "selection on observables" methods is:

If we can observe and control for all the variables that jointly affect treatment selection and outcomes, we can recover causal effects.

The Intuition

Imagine comparing earnings of college graduates versus non-graduates. The naive comparison is biased because:

- People who attend college differ systematically from those who don't
- They may have higher ability, more motivated parents, better schools, etc.
- These factors affect both college attendance AND earnings

But: If we could compare people who are identical on all these characteristics—same ability, same family background, same schools—then among such people, college attendance might be "as good as random."

This is the logic of selection on observables.

3.2 The Conditional Independence Assumption (CIA)

Let's formalize this idea. I use Wooldridge's notation throughout.

Setup

- y_{0i}, y_{1i}: potential outcomes
- w_i: treatment indicator
- $\mathbf{x}_i$: vector of observed pre-treatment covariates (confounders)

The Assumption

Conditional Independence Assumption (CIA):

$$w_i \perp\!\!\!\perp (y_{0i}, y_{1i}) \mid \mathbf{x}_i$$

This says: conditional on $\mathbf{x}_i$, treatment assignment is independent of potential outcomes.

Alternative names for the same assumption:

- Selection on observables
- Unconfoundedness
- Ignorability (Rosenbaum & Rubin)
- Exogeneity conditional on covariates

What CIA Means

Within subpopulations defined by $\mathbf{x}_i = \mathbf{x}$:

- Treatment assignment is "as good as random"
- Treated and control units are comparable
- Selection bias is eliminated

Weaker Version: Mean Independence

Sometimes we only need the weaker assumption:

$$E[y_{0i} \mid w_i, \mathbf{x}_i] = E[y_{0i} \mid \mathbf{x}_i]$$
$$E[y_{1i} \mid w_i, \mathbf{x}_i] = E[y_{1i} \mid \mathbf{x}_i]$$

This says: conditional on $\mathbf{x}_i$, the mean potential outcomes don't depend on treatment status.

3.3 Identification Under CIA

Let's prove that CIA allows us to identify causal effects.

Theorem 3.1: Identification of ATT

Under CIA, the Average Treatment Effect on the Treated is identified:

$$\text{ATT} = E[y_{1i} - y_{0i} \mid w_i = 1] = E\{E[y_i \mid w_i = 1, \mathbf{x}_i] - E[y_i \mid w_i = 0, \mathbf{x}_i] \mid w_i = 1\}$$

Proof:

$$\text{ATT} = E[y_{1i} \mid w_i = 1] - E[y_{0i} \mid w_i = 1]$$

The first term is directly observable:

$$E[y_{1i} \mid w_i = 1] = E[y_i \mid w_i = 1]$$

For the second term, use the law of iterated expectations:

$$E[y_{0i} \mid w_i = 1] = E[E[y_{0i} \mid w_i = 1, \mathbf{x}_i] \mid w_i = 1]$$

By CIA: $E[y_{0i} \mid w_i = 1, \mathbf{x}_i] = E[y_{0i} \mid \mathbf{x}_i]$

Since $y_i = y_{0i}$for untreated: $E[y_{0i} \mid \mathbf{x}_i] = E[y_{0i} \mid w_i = 0, \mathbf{x}_i] = E[y_i \mid w_i = 0, \mathbf{x}_i]$

Therefore:

$$E[y_{0i} \mid w_i = 1] = E[E[y_i \mid w_i = 0, \mathbf{x}_i] \mid w_i = 1]$$

This is observable! It's the expected outcome of control units, at covariate values drawn from the treated population. □

Theorem 3.2: Identification of ATE

Under CIA plus an **overlap** assumption, the ATE is identified:

$$\text{ATE} = E[y_{1i} - y_{0i}] = E[E[y_i \mid w_i = 1, \mathbf{x}_i] - E[y_i \mid w_i = 0, \mathbf{x}_i]]$$

The Overlap (Common Support) Assumption:

$$0 < P(w_i = 1 \mid \mathbf{x}_i) < 1\text{for all } \mathbf{x}_i \text{ in the support}$$

This ensures that for every value of $\mathbf{x}_i$, there exist both treated and control observations.

3.4 The Selection Bias Decomposition Revisited

Let's see how CIA eliminates selection bias.

Conditional Selection Bias

Define the conditional selection bias:

$$B(\mathbf{x}) = E[y_{0i} \mid w_i = 1, \mathbf{x}_i = \mathbf{x}] - E[y_{0i} \mid w_i = 0, \mathbf{x}_i = \mathbf{x}]$$

Under CIA: $B(\mathbf{x}) = 0$for all $\mathbf{x}$.

Unconditional Selection Bias

The overall selection bias can be written as:

$$\begin{aligned}\text{Selection Bias} &= E[y_{0i} \mid w_i = 1] - E[y_{0i} \mid w_i = 0]\\ &= E[E[y_{0i} \mid w_i = 1, \mathbf{x}_i]] - E[E[y_{0i} \mid w_i = 0, \mathbf{x}_i]]\\ &= E[E[y_{0i} \mid \mathbf{x}_i] \mid w_i = 1] - E[E[y_{0i} \mid \mathbf{x}_i] \mid w_i = 0]\end{aligned}$$

Under CIA, this becomes:

$$= E[\mu_0(\mathbf{x}_i) \mid w_i = 1] - E[\mu_0(\mathbf{x}_i) \mid w_i = 0]$$

where $\mu_0(\mathbf{x}) = E[y_{0i} \mid \mathbf{x}_i = \mathbf{x}]$.

Key insight: Even under CIA, the selection bias is generally non-zero because $\mathbf{x}_i$ differs between treated and control groups! This is why we need to **adjust** for $\mathbf{x}_i$—not to make CIA true, but to account for the compositional differences it allows.

3.5 Regression as a Causal Inference Tool

Now we connect CIA to the workhorse of applied economics: linear regression.

The Linear Conditional Expectation Model

Assume the conditional expectation functions are linear:

$$\begin{aligned}E[y_{0i} \mid \mathbf{x}_i] &= \alpha + \mathbf{x}_i'\boldsymbol{\beta}\\ E[y_{1i} \mid \mathbf{x}_i] &= \alpha + \tau + \mathbf{x}_i'\boldsymbol{\beta}\end{aligned}$$

This implies a **constant treatment effect**: $E[y_{1i} - y_{0i} \mid \mathbf{x}_i] = \tau$for all $\mathbf{x}_i$.

The Regression Model

Combining these with the switching equation $y_i = w_i y_{1i} + (1 - w_i) y_{0i}$:

$$E[y_i \mid w_i, \mathbf{x}_i] = \alpha + \tau w_i + \mathbf{x}_i' \boldsymbol{\beta}$$

Adding an error term:

$$y_i = \alpha + \tau w_i + \mathbf{x}_i' \boldsymbol{\beta} + u_i$$

where $E[u_i \mid w_i, \mathbf{x}_i] = 0$by construction.

When Does OLS Identify the Causal Effect?

Theorem 3.3: Under CIA plus linearity, OLS consistently estimates the ATE:

$$\hat{\tau}^{OLS} \xrightarrow{p} \tau = \text{ATE}$$

The Crucial Requirement: $E[u_i \mid w_i, \mathbf{x}_i] = 0$

In Wooldridge's framework, this is the key exogeneity condition. It holds when:

1. CIA is satisfied (no omitted confounders)
2. The functional form is correctly specified

What If There Are Omitted Confounders?

Suppose the true model includes an unobserved variable a_i(e.g., "ability"):

$$y_i = \alpha + \tau w_i + \mathbf{x}_i' \boldsymbol{\beta} + \gamma a_i + \varepsilon_i$$

If we estimate without a_i:

$$y_i = \alpha + \tau w_i + \mathbf{x}_i' \boldsymbol{\beta} + u_i$$

where $u_i = \gamma a_i + \varepsilon_i$.

Omitted Variable Bias Formula:

$$\text{plim } \hat{\tau}^{OLS} = \tau + \gamma \cdot \delta_{aw \cdot \mathbf{x}}$$

where $\delta_{aw\cdot\mathbf{x}}$is the coefficient on $w_i$from regressing $a_i$on $w_i$and $\mathbf{x}_i$.

The bias is zero if:

- $\gamma = 0$: the omitted variable doesn't affect the outcome, OR
- $\delta_{aw\cdot\mathbf{x}} = 0$: the omitted variable is uncorrelated with treatment after controlling for $\mathbf{x}_i$

3.6 A Detailed Example: Returns to Education

Let's apply these concepts to a classic problem.

The Question

What is the causal effect of an additional year of schooling on wages?

The Model

$$\ln(wage_i) = \alpha + \tau \cdot educ_i + \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

where:

- $\ln(wage_i)$= log hourly wage
- $educ_i$= years of schooling
- $\mathbf{x}_i$= observed characteristics (experience, gender, region, etc.)
- τ= causal return to education

The Selection Problem

People with more education likely differ in unobserved ways:

- Higher innate ability
- Greater motivation
- Better family connections

Let a_i= unobserved ability. If $a_i$affects both education and wages:

$$E[u_i \mid educ_i, \mathbf{x}_i] \neq 0$$

CIA is violated! OLS is biased.

Direction of Bias

If higher-ability people get more education ($\delta_{a,educ\cdot\mathbf{x}} > 0$) and ability raises wages ($\gamma > 0$), then:

$$\text{Bias} = \gamma \cdot \delta_{a,educ\cdot\mathbf{x}} > 0$$

OLS **overstates** the causal return to education.

Empirical Evidence

Studies comparing OLS to experimental/quasi-experimental estimates typically find:

- OLS estimates: ~7-10% return per year
- IV/experimental estimates: ~5-7% return per year

The OLS estimates are indeed upward biased, though perhaps less than we might expect.

3.7 The Propensity Score

When $\mathbf{x}_i$is high-dimensional, conditioning on all variables becomes impractical. The propensity score provides an elegant solution.

Definition

The Propensity Score is the probability of receiving treatment given covariates:

$$p(\mathbf{x}_i) = P(w_i = 1 \mid \mathbf{x}_i) = E[w_i \mid \mathbf{x}_i]$$

The Propensity Score Theorem (Rosenbaum & Rubin, 1983)

Theorem 3.4: If CIA holds given $\mathbf{x}_i$, then CIA also holds given $p(\mathbf{x}_i)$:

$$w_i \perp\!\!\!\perp (y_{0i}, y_{1i}) \mid \mathbf{x}_i \implies w_i \perp\!\!\!\perp (y_{0i}, y_{1i}) \mid p(\mathbf{x}_i)$$

Proof:

We need to show that $w_i \perp\!\!\!\perp y_{0i} \mid p(\mathbf{x}_i)$. (The argument for $y_{1i}$is symmetric.)

It suffices to show: $E[y_{0i} \mid w_i = 1, p(\mathbf{x}_i)] = E[y_{0i} \mid w_i = 0, p(\mathbf{x}_i)]$

Step 1: By the law of iterated expectations:

$$E[y_{0i} \mid w_i, p(\mathbf{x}_i)] = E[E[y_{0i} \mid w_i, \mathbf{x}_i] \mid w_i, p(\mathbf{x}_i)]$$

Step 2: By CIA:

$$E[y_{0i} \mid w_i, \mathbf{x}_i] = E[y_{0i} \mid \mathbf{x}_i]$$

Step 3: Since $p(\mathbf{x}_i)$is a function of $\mathbf{x}_i$:

$$E[E[y_{0i} \mid \mathbf{x}_i] \mid w_i, p(\mathbf{x}_i)] = E[E[y_{0i} \mid \mathbf{x}_i] \mid p(\mathbf{x}_i)]$$

The last equality holds because, conditional on $p(\mathbf{x}_i)$, $w_i$provides no additional information about $E[y_{0i} \mid \mathbf{x}_i]$.

Why? Consider any two individuals with the same propensity score p. Even if they have different $\mathbf{x}_i$values, the theorem shows that their expected potential outcomes must average to the same value. □

The Power of This Result

Instead of conditioning on a potentially high-dimensional $\mathbf{x}_i$, we can condition on a **single scalar** $p(\mathbf{x}_i)$.

This is a **dimension reduction** result: from $\dim(\mathbf{x}_i) = k$to $\dim(p(\mathbf{x}_i)) = 1$.

3.8 Estimating the Propensity Score

The propensity score must be estimated from data.

Logit/Probit Models

The most common approach uses binary response models:

Logit:

$$p(\mathbf{x}_i) = P(w_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\gamma})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\gamma})} = \Lambda(\mathbf{x}_i'\boldsymbol{\gamma})$$

Probit:

$$p(\mathbf{x}_i) = P(w_i = 1 \mid \mathbf{x}_i) = \Phi(\mathbf{x}_i'\boldsymbol{\gamma})$$

where $\Lambda(\cdot)$is the logistic CDF and $\Phi(\cdot)$is the standard normal CDF.

Estimation Procedure

1. Run logit/probit regression of w_i on $\mathbf{x}_i$

2. Obtain fitted values: $\hat{p}_i = \hat{p}(\mathbf{x}_i)$
3. Use $\hat{p}_i$ in subsequent analysis

Example: Job Training Program

Suppose we're evaluating a job training program. Covariates include:

- age_i: age in years
- $educ_i$: years of education
- $married_i$: marital status (1 = married)
- $black_i$: race indicator
- $earn74_i$: earnings in 1974 (pre-program)

Propensity Score Model:

$$P(w_i = 1 \mid \mathbf{x}_i) = \Lambda(\gamma_0 + \gamma_1 age_i + \gamma_2 educ_i + \gamma_3 married_i + \gamma_4 black_i + \gamma_5 earn74_i)$$

Hypothetical Logit Output:

Variable	**Coefficient**	**Std. Error**	**z-stat**
Constant	-1.82	0.45	-4.04
Age	-0.03	0.01	-3.00
Education	0.08	0.03	2.67
Married	-0.45	0.18	-2.50
Black	0.62	0.20	3.10
Earnings 1974	-0.0001	0.00003	-3.33

Interpretation:

- Younger workers more likely to participate
- More educated workers more likely to participate
- Unmarried workers more likely to participate
- Black workers more likely to participate
- Workers with lower prior earnings more likely to participate

3.9 Matching Estimators

Matching is the most intuitive implementation of selection on observables: find control units that “look like” treated units and compare their outcomes.

The Basic Idea

For each treated unit i, find one or more control units $j$with similar covariates, and impute the counterfactual:

$$\hat{y}_{0i} = y_j \text{where } j \text{ is similar to } i$$

Then estimate the treatment effect as:

$$\hat{\tau}_i = y_i - \hat{y}_{0i}$$

Exact Matching

Definition: Unit $j$is an exact match for unit $i$if $\mathbf{x}_j = \mathbf{x}_i$.

Problem: With continuous or high-dimensional $\mathbf{x}_i$, exact matches may not exist.

Nearest-Neighbor Matching

Definition: Match each treated unit to the control unit(s) with the smallest distance:

$$j(i) = \arg \min_{j:w_j=0} \parallel \mathbf{x}_i - \mathbf{x}_j \parallel$$

where $\parallel\cdot\parallel$is some distance metric.

Common distance metrics:

1. Euclidean distance: $\parallel \mathbf{x}_i - \mathbf{x}_j \parallel = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$

Mahalanobis distance: $\parallel \mathbf{x}_i - \mathbf{x}_j \parallel_M = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \mathbf{x}_j)}$

2. where $\hat{\boldsymbol{\Sigma}}$is the sample covariance matrix of $\mathbf{x}$.
3. **Propensity score distance:** $\mid p(\mathbf{x}_i) - p(\mathbf{x}_j) \mid$

The Matching Estimator for ATT

Single nearest-neighbor matching:

$$\hat{\tau}_{ATT}^{match} = \frac{1}{N_1} \sum_{i:w_i=1} [y_i - y_{j(i)}]$$

where $j(i)$is the matched control for treated unit i.

Matching with Replacement vs. Without

With replacement: Each control unit can be matched to multiple treated units.

- Pro: Better matches (lower bias)
- Con: Fewer distinct controls used (higher variance)

Without replacement: Each control unit can be matched at most once.

- Pro: More distinct controls (lower variance)
- Con: May have to accept poor matches (higher bias)

Multiple Matches (k-Nearest Neighbors)

Match each treated unit to $k$nearest controls and average:

$$\hat{y}_{0i} = \frac{1}{k} \sum_{j \in \mathcal{J}_k(i)} y_j$$

where $\mathcal{J}_k(i)$is the set of $k$nearest control neighbors of i.

3.10 Propensity Score Matching

Matching on the propensity score is particularly attractive due to the dimension reduction.

The Estimator

1. Estimate propensity scores $\hat{p}_i$for all units
2. For each treated unit i, find control(s) with closest $\hat{p}_j$
3. Compute the ATT

$$\hat{\tau}_{ATT}^{PSM} = \frac{1}{N_1} \sum_{i:w_i=1} \left[y_i - \sum_{j \in \mathcal{J}(i)} \omega_{ij}\, y_j \right]$$

where $\omega_{ij}$are matching weights (often equal weights for k-NN matching).

Caliper Matching

To avoid poor matches, impose a maximum distance (**caliper**):

$$| p_i - p_j |< c$$

Units without matches within the caliper are dropped.

Trade-off:

- Tighter caliper → better matches but smaller sample
- Looser caliper → larger sample but worse matches

Common Support (Overlap)

A crucial requirement: we can only match where both treated and control observations exist.

Common Support Condition:

$$\mathcal{S} = \{p: f_{p|w=1}(p) > 0 \text{ and } f_{p|w=0}(p) > 0\}$$

where $f_{p|w}$is the density of propensity scores conditional on treatment status.

In practice:

- Examine the distribution of $\hat{p}_i$for treated and controls
- Trim observations outside the region of overlap
- Report how many observations are dropped

3.11 Inverse Probability Weighting (IPW)

An alternative to matching that uses all observations.

The Idea

Reweight observations to create a “pseudo-population” where treatment is independent of covariates.

The IPW Estimator for ATE

$$\hat{\tau}_{ATE}^{IPW} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{w_i y_i}{\hat{p}_i} - \frac{(1-w_i)y_i}{1-\hat{p}_i}\right]$$

Derivation

Under CIA and overlap:

$$E[y_{1i}] = E\left[\frac{w_i y_i}{p(\mathbf{x}_i)}\right]$$

Proof:

$$E\left[\frac{w_i y_i}{p(\mathbf{x}_i)}\right] = E\left[E\left[\frac{w_i y_i}{p(\mathbf{x}_i)} \mid \mathbf{x}_i\right]\right]$$
$$= E\left[\frac{1}{p(\mathbf{x}_i)} E[w_i y_i \mid \mathbf{x}_i]\right]$$
$$= E\left[\frac{1}{p(\mathbf{x}_i)} E[w_i \mid \mathbf{x}_i] E[y_i \mid w_i = 1, \mathbf{x}_i]\right]$$

$$= E\left[\frac{p(\mathbf{x}_i)}{p(\mathbf{x}i)} E[y{1i} \mid \mathbf{x}i]\right] = E[E[y{1i} \mid \mathbf{x}i]] = E[y{1i}]$$ $\square$

Similarly: $E[y_{0i}] = E\left[\frac{(1-w_i)y_i}{1-p(\mathbf{x}_i)}\right]$

Therefore: ATE $= E[y_{1i}] - E[y_{0i}]$is identified by the IPW formula.

Normalized IPW (Hajek Estimator)

The basic IPW can be unstable. A normalized version:

$$\hat{\tau}_{ATE}^{NIPW} = \frac{\sum_{i=1}^{N} \frac{w_i y_i}{\hat{p}_i}}{\sum_{i=1}^{N} \frac{w_i}{\hat{p}_i}} - \frac{\sum_{i=1}^{N} \frac{(1-w_i)y_i}{1-\hat{p}_i}}{\sum_{i=1}^{N} \frac{1-w_i}{1-\hat{p}_i}}$$

This ensures the weights sum to one within each group.

IPW for ATT

$$\hat{\tau}_{ATT}^{IPW} = \frac{1}{N_1}\sum_{i=1}^{N} w_i\, y_i - \frac{\sum_{i=1}^{N} \frac{(1-w_i)\hat{p}_i y_i}{1-\hat{p}_i}}{\sum_{i=1}^{N} \frac{(1-w_i)\hat{p}_i}{1-\hat{p}_i}}$$

Intuition: The weights $\frac{\hat{p}_i}{1-\hat{p}_i}$upweight control units that “look like” treated units (high propensity score).

3.12 The Problem of Extreme Propensity Scores

IPW estimators can be highly variable when propensity scores are close to 0 or 1.

Why It’s a Problem

If $\hat{p}_i \approx 0$for a treated unit: weight $\frac{1}{\hat{p}_i} \approx \infty$If $\hat{p}_i \approx 1$for a control unit: weight $\frac{1}{1-\hat{p}_i} \approx \infty$

A single observation can dominate the entire estimate!

Solutions

1. Trimming: Drop observations with $\hat{p}_i < c$or $\hat{p}_i > 1 - c$(e.g., $c = 0.01$or 0.05).

2. Truncation: Replace extreme weights with a maximum value:

$$\widetilde{w}_i = \min\left(\frac{1}{\hat{p}_i}, M\right)$$

3. Overlap weights (Li, Morgan, Zaslavsky 2018):

$$\omega_i = w_i(1-\hat{p}_i) + (1-w_i)\hat{p}_i$$

These weights emphasize the region of overlap and downweight extreme propensity scores.

3.13 Doubly Robust Estimation

A powerful approach that combines regression and propensity scores.

The Idea

What if our regression model is wrong? What if our propensity score model is wrong? Doubly robust estimators are consistent if **either** model is correct.

The Augmented IPW (AIPW) Estimator

$$\hat{\tau}^{AIPW} = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\mu}_1(\mathbf{x}_i) - \hat{\mu}_0(\mathbf{x}_i) + \frac{w_i(y_i - \hat{\mu}_1(\mathbf{x}_i))}{\hat{p}_i} - \frac{(1-w_i)(y_i - \hat{\mu}_0(\mathbf{x}_i))}{1-\hat{p}_i}\right]$$

Where:

- $\hat{\mu}_1(\mathbf{x}_i) = \hat{E}[y_i \mid w_i = 1, \mathbf{x}_i]$: estimated outcome regression for treated
- $\hat{\mu}_0(\mathbf{x}_i) = \hat{E}[y_i \mid w_i = 0, \mathbf{x}_i]$: estimated outcome regression for control
- $\hat{p}_i$: estimated propensity score

Understanding the Components

The AIPW estimator has two parts:

Part 1: Regression imputation

$$\frac{1}{N}\sum_{i=1}^{N}[\hat{\mu}_1(\mathbf{x}_i) - \hat{\mu}_0(\mathbf{x}_i)]$$

This would be consistent if the outcome model is correct.

Part 2: IPW correction for regression errors

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{w_i(y_i - \hat{\mu}_1(\mathbf{x}_i))}{\hat{p}_i} - \frac{(1-w_i)(y_i - \hat{\mu}_0(\mathbf{x}_i))}{1-\hat{p}_i}\right]$$

This corrects the bias in Part 1 using IPW.

The Double Robustness Property

Theorem 3.5: $\hat{\tau}^{AIPW}$is consistent for the ATE if either:

- The propensity score model $\hat{p}(\mathbf{x}_i)$is correctly specified, OR
- The outcome model $\hat{\mu}_w(\mathbf{x}_i)$is correctly specified

Intuition:

- If outcome model is correct: $y_i - \hat{\mu}_w(\mathbf{x}_i) \approx 0$, so Part 1 dominates

- If propensity model is correct: Part 2 corrects any bias in Part 1

Practical Advantage

We get "two chances" to be right. This is especially valuable because:

- We rarely know the true functional forms
- Model misspecification is the norm, not the exception

3.14 Assessing the Plausibility of CIA

CIA cannot be tested directly—it involves unobserved potential outcomes. But we can do several checks.

1. Covariate Balance

After matching or weighting, treated and control groups should look similar on observables.

Balance statistics:

For each covariate x_k:

$$\text{Standardized Difference} = \frac{\bar{x}_{k,1} - \bar{x}_{k,0}}{\sqrt{(s_{k,1}^2 + s_{k,0}^2)/2}}$$

Rule of thumb: Differences > 0.1 (or 0.25) standard deviations are concerning.

2. Propensity Score Distribution

Examine overlap in propensity scores between treated and control groups.

- Plot histograms or densities of $\hat{p}_i$by treatment status
- Substantial non-overlap suggests extrapolation is required
- Lack of common support indicates fundamental identification problems

3. Sensitivity Analysis

How much unobserved confounding would be needed to overturn results?

Rosenbaum bounds: Assume an unobserved confounder $u_i$affects treatment odds by at most factor Γ:

$$\frac{1}{\Gamma} \leq \frac{P(w_i = 1 \mid \mathbf{x}_i, u_i)/P(w_i = 0 \mid \mathbf{x}_i, u_i)}{P(w_j = 1 \mid \mathbf{x}_j, u_j)/P(w_j = 0 \mid \mathbf{x}_j, u_j)} \leq \Gamma$$

for matched pairs $i, j$with $\mathbf{x}_i = \mathbf{x}_j$.

Report: "Results are robust to unobserved confounding up to $\Gamma = 2$" (meaning an unobserved factor would need to double the odds of treatment to explain away the effect).

4. Placebo Tests

Test for "effects" on outcomes that shouldn't be affected:

- Pre-treatment outcomes (should show no effect)
- Outcomes determined before treatment

Finding "effects" on placebo outcomes suggests confounding.

3.15 Complete Example: LaLonde's Job Training Data

Let's work through a famous example.

Background

Robert LaLonde (1986) evaluated the National Supported Work (NSW) demonstration, a job training program for disadvantaged workers.

The setup:

- Randomized experiment conducted in the 1970s
- Treatment: Job training program
- Outcome: Post-program earnings (1978)
- LaLonde's innovation: Compare experimental estimates to observational estimates using non-experimental comparison groups

The Data

Experimental sample:

- 185 treated (program participants)

- 260 control (randomly assigned to control)

Observational comparison groups:

- PSID (Panel Study of Income Dynamics): ~2,490 non-participants
- CPS (Current Population Survey): ~15,992 non-participants

Experimental Benchmark

From the randomized experiment:

$$\hat{\tau}^{exp} = \bar{y}_{1978,treat} - \bar{y}_{1978,control} = \$1,794$$

This is our “true” causal effect (subject to sampling error).

Naive Observational Estimate

Comparing treated to PSID comparison group:

$$\hat{\tau}^{naive} = \bar{y}_{treat} - \bar{y}_{PSID} = \$5,976 - \$21,554 = -\$15,578$$

The naive estimate is massively **negative**—it suggests training *reduced* earnings by $15,000!

What went wrong? The PSID sample consists of typical American workers, while NSW participants were severely disadvantaged. Enormous selection bias.

Propensity Score Matching

Step 1: Estimate propensity scores

Using covariates: age, education, Black, Hispanic, married, no degree, earnings in 1974, earnings in 1975.

Logit regression of treatment on covariates.

Step 2: Check overlap

The propensity score distributions barely overlap:

- Treated: most have $\hat{p}_i \in [0.01, 0.20]$
- PSID: most have $\hat{p}_i < 0.001$

This is a severe common support problem.

Step 3: Match and estimate

After matching treated to nearest PSID controls (with caliper):

$$\hat{\tau}^{PSM} \approx \$1{,}691$$

Much closer to the experimental benchmark!

Lessons from LaLonde

1. Selection bias can be enormous in observational data
2. Propensity score methods can help but require:
   - Good overlap (common support)
   - Relevant covariates (CIA must be plausible)
3. When treatment and comparison groups are very different, causal inference is difficult
4. Experimental benchmarks are invaluable for evaluating observational methods

3.16 Summary of Part 3

Key Assumptions

Conditional Independence (CIA):

$$w_i \perp\!\!\!\perp (y_{0i}, y_{1i}) \mid \mathbf{x}_i$$

Overlap (Common Support):

$$0 < P(w_i = 1 \mid \mathbf{x}_i) < 1$$

Key Results

Propensity Score Theorem:

$$w_i \perp\!\!\!\perp (y_{0i}, y_{1i}) \mid \mathbf{x}_i \implies w_i \perp\!\!\!\perp (y_{0i}, y_{1i}) \mid p(\mathbf{x}_i)$$

Identification under CIA:

$$\text{ATE} = E[E[y_i \mid w_i = 1, \mathbf{x}_i] - E[y_i \mid w_i = 0, \mathbf{x}_i]]$$

Key Estimators

Estimator	Formula	Key Feature
Regression	OLS with covariates	Simple, parametric
Matching	Compare matched pairs	Nonparametric, local
IPW	Reweight by $1/\hat{p}_i$	Uses all data
AIPW	Combines regression + IPW	Doubly robust

Practically:

1. **Think carefully about CIA:** What confounders exist? Can you measure them?
2. **Check overlap:** Ensure common support exists
3. **Assess balance:** After adjustment, groups should be similar
4. **Use multiple methods:** If results are robust across methods, more credible
5. **Conduct sensitivity analysis:** How much confounding could overturn results?

Exercises for Part 3

Exercise 3.1: Prove that the IPW estimator for the ATT is:

$$E[y_{1i} - y_{0i} \mid w_i = 1] = E[y_i \mid w_i = 1] - E\left[\frac{(1 - w_i)p(\mathbf{x}_i)y_i}{(1 - p(\mathbf{x}_i))P(w_i = 1)}\right] / E\left[\frac{(1 - w_i)p(\mathbf{x}_i)}{(1 - p(\mathbf{x}_i))P(w_i = 1)}\right]$$

Exercise 3.2: Suppose you estimate a propensity score model and find that 20% of treated observations have $\hat{p}_i > 0.95$. What does this imply for identification? What would you do?

Exercise 3.3: A researcher claims CIA is satisfied because she “controlled for all observable variables.” Critique this claim.

Exercise 3.4: Explain intuitively why the doubly robust estimator is consistent if either model is correct, but not necessarily efficient.

PART 4: INSTRUMENTAL VARIABLES

Selection on observables methods work when we can measure all confounders. But what if we can't? What if there are important unobserved factors that affect both treatment and outcomes?

Part 4: Instrumental Variables, how to identify causal effects using external sources of variation that affect treatment but not outcomes directly. This is one of the most powerful—and most commonly misused—tools in the econometrician's toolkit.

Structure:

1. The problem of endogeneity and why we need IV
2. The basic IV setup and assumptions
3. Identification with IV
4. The Wald estimator
5. Two-stage least squares (2SLS)
6. The Local Average Treatment Effect (LATE) interpretation
7. Multiple instruments and overidentification
8. Weak instruments
9. Testing IV assumptions
10. Examples and applications

Solving Endogeneity with External Variation

4.1 The Problem: Endogeneity

In Part 3, we assumed we could observe all confounders. But what if we can't? This is the problem of **endogeneity**.

When CIA Fails

Recall that OLS identification requires:

$$E[u_i \mid w_i, \mathbf{x}_i] = 0$$

This fails when there exist **unobserved confounders**—variables that affect both treatment and outcome but are not in our data.

Sources of Endogeneity

1. Omitted Variables (Unobserved Confounding)

The classic problem: ability affects both education and wages, but we can't measure ability.

$$wage_i = \alpha + \tau \cdot educ_i + \gamma \cdot ability_i + \varepsilon_i$$

If we omit $ability_i$: $E[u_i \mid educ_i] \neq 0$because $Cov(educ_i, ability_i) \neq 0$.

2. Simultaneity (Reverse Causality)

Causation runs both ways:

- Does police presence reduce crime, or does crime attract police?
- Do prices affect quantity, or does quantity affect prices?

3. Measurement Error

If $educ_i^* = educ_i + v_i$where we observe $educ_i^*$with error:

- Classical measurement error biases coefficients toward zero
- Creates correlation between regressor and error

The Core Challenge

We need variation in treatment that is:

1. **Relevant:** Actually affects treatment
2. **Exogenous:** Unrelated to unobserved confounders

This is what instrumental variables provide.

4.2 The Instrumental Variables Setup

The Structural Equation

We want to estimate the causal effect τin:

$$y_i = \alpha + \tau w_i + u_i$$

where $E[u_i \mid w_i] \neq 0$due to endogeneity.

For now, I'm omitting additional covariates $\mathbf{x}_i$to focus on the core ideas. We'll add them back later.

The Instrument

An **instrumental variable** $z_i$is a variable that:

1. Affects treatment (w_i)
2. Affects outcome (y_i) **only through** its effect on treatment

The Two Key Assumptions

Assumption IV.1: Relevance (First Stage)

$$Cov(z_i, w_i) \neq 0$$

The instrument must be correlated with the endogenous variable.

Assumption IV.2: Exogeneity (Exclusion Restriction)

$$Cov(z_i, u_i) = 0$$

The instrument must be uncorrelated with the structural error.

Equivalently: $E[u_i \mid z_i] = 0$(mean independence, stronger version).

Visualizing the Assumptions

Think of a causal diagram:

u_i

↓

z_i → w_i → y_i

The instrument z_i:

- Has a direct arrow to w_i(relevance)
- Has NO direct arrow to y_i(exclusion)
- Has NO arrow from u_i(exogeneity)

If z_i had a direct effect on $y_i$or was correlated with u_i, the exclusion restriction would be violated.

4.3 Identification: The Wald Estimator

Let's start with the simplest case: a **binary instrument**.

Setup

- $z_i \in \{0,1\}$: binary instrument
- w_i: treatment (can be binary or continuous)

- y_i: outcome

The Reduced Form

The **reduced form** relates the outcome directly to the instrument:

$$y_i = \pi_0 + \pi_1 z_i + v_i$$

where $E[v_i \mid z_i] = 0$(by the exclusion restriction).

The reduced form coefficient $\pi_1$captures the **total effect** of $z_i$on y_i:

$$\pi_1 = E[y_i \mid z_i = 1] - E[y_i \mid z_i = 0]$$

The First Stage

The **first stage** relates treatment to the instrument:

$$w_i = \gamma_0 + \gamma_1 z_i + \eta_i$$

where $E[\eta_i \mid z_i] = 0$.

The first stage coefficient $\gamma_1$captures the effect of $z_i$on w_i:

$$\gamma_1 = E[w_i \mid z_i = 1] - E[w_i \mid z_i = 0]$$

The Wald Estimator

Key insight: The causal effect is the ratio of reduced form to first stage:

$$\tau = \frac{\pi_1}{\gamma_1} = \frac{E[y_i \mid z_i = 1] - E[y_i \mid z_i = 0]}{E[w_i \mid z_i = 1] - E[w_i \mid z_i = 0]}$$

This is the **Wald estimator**, named after Abraham Wald.

Proof of Identification

Step 1: Write the structural equation:

$$y_i = \alpha + \tau w_i + u_i$$

Step 2: Take expectations conditional on z_i:

$$E[y_i \mid z_i] = \alpha + \tau E[w_i \mid z_i] + E[u_i \mid z_i]$$

Step 3: By the exclusion restriction, $E[u_i \mid z_i] = 0$:

$$E[y_i \mid z_i] = \alpha + \tau E[w_i \mid z_i]$$

Step 4: Difference between $z_i = 1$and $z_i = 0$:

$$E[y_i \mid z_i = 1] - E[y_i \mid z_i = 0] = \tau \cdot [E[w_i \mid z_i = 1] - E[w_i \mid z_i = 0]]$$

Step 5: Solve for τ:

Intuition

The Wald estimator asks: “How much does the outcome change per unit change in treatment, when we use the instrument to generate the change in treatment?”

$$\tau = \frac{\text{Change in } y \text{ caused by } z}{\text{Change in } w \text{ caused by } z}$$

4.4 Sample Estimation

The Wald Estimator in Practice

$$\hat{\tau}^{Wald} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{w}_1 - \bar{w}_0}$$

where:

- \bar{y}_1 = $ mean outcome when $z_i = 1
- \bar{y}_0 = $ mean outcome when $z_i = 0
- \bar{w}_1 = $ mean treatment when $z_i = 1
- \bar{w}_0 = $ mean treatment when $z_i = 0

Example: Draft Lottery and Vietnam Service

Question: What is the effect of military service on lifetime earnings?

Problem: Volunteers differ from non-volunteers in unobserved ways (patriotism, risk tolerance, opportunity cost).

Instrument: The Vietnam-era draft lottery randomly assigned draft eligibility based on birthday.

Setup:

- y_i: lifetime earnings
- w_i: served in military (0/1)
- z_i: draft eligible based on lottery number (0/1)

Why is this a valid instrument?

- **Relevance:** Draft eligibility strongly predicts military service
- **Exclusion:** Birthday lottery is random, so it shouldn't affect earnings except through military service

Hypothetical Data:

	$z_i = 1$**(Eligible)**	$z_i = 0$**(Not Eligible)**
Mean earnings ($\bar{y}$)	$48,000	$52,000
Fraction served ($\bar{w}$)	0.26	0.07
Sample size	5,000	5,000

Wald Estimate:

$$\hat{\tau}^{Wald} = \frac{48{,}000 - 52{,}000}{0.26 - 0.07} = \frac{-4{,}000}{0.19} \approx -\$21{,}053$$

Interpretation: Military service reduced lifetime earnings by approximately $21,000.

4.5 Two-Stage Least Squares (2SLS)

The Wald estimator is limited to binary instruments. **Two-Stage Least Squares** generalizes IV to continuous instruments and multiple instruments.

The Setup

Structural equation:

$$y_i = \alpha + \tau w_i + u_i$$

First stage:

$$w_i = \gamma_0 + \gamma_1 z_i + \eta_i$$

The 2SLS Procedure

Stage 1: Regress $w_i$on $z_i$and obtain fitted values:

$$\widehat{w}_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i$$

Stage 2: Regress $y_i$on $\widehat{w}_i$:

$$y_i = \alpha + \tau\widehat{w}_i + \text{error}$$

The coefficient on $\widehat{w}_i$is the 2SLS estimate $\hat{\tau}^{2SLS}$.

Why Does This Work (most of the time)?

Key insight: $\widehat{w}_i$ contains only the variation in $w_i$that comes from z_i.

Since $z_i$is exogenous (uncorrelated with u_i), $\widehat{w}_i$is also exogenous.

Decomposition:

$$w_i = \underbrace{\widehat{w}_i}_{\text{exogenous}} + \underbrace{\hat{\eta}_i}_{\text{endogenous}}$$

The problematic correlation $Cov(w_i, u_i) \neq 0$comes from $\hat{\eta}_i$. By using only $\widehat{w}_i$, we purge the endogenous variation.

The 2SLS Formula

For the simple case with one endogenous variable and one instrument:

$$\hat{\tau}^{2SLS} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(w_i - \bar{w})} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Cov}(z_i, w_i)}$$

Note: This is exactly the Wald estimator when z_i is binary!

Matrix Formulation

For the general case, let:

- $\mathbf{y}$: $n \times 1$outcome vector
- $\mathbf{W}$: $n \times k$matrix of regressors (including endogenous variables)
- $\mathbf{Z}$: $n \times m$matrix of instruments ($m \geq k$)

The 2SLS estimator is:

$$\hat{\boldsymbol{\beta}}^{2SLS} = (\mathbf{W}'\mathbf{P}_Z\mathbf{W})^{-1}\mathbf{W}'\mathbf{P}_Z\mathbf{y}$$

where $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$is the projection matrix onto the column space of $\mathbf{Z}$.

4.6 Adding Exogenous Covariates

In practice, we usually have additional control variables.

The Model

Structural equation:

$$y_i = \alpha + \tau w_i + \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

where:

- $w_i$is endogenous: $Cov(w_i, u_i) \neq 0$
- $\mathbf{x}_i$are exogenous controls: $Cov(\mathbf{x}_i, u_i) = 0$
- $z_i$is the instrument for w_i

The Exclusion Restriction with Covariates

$$E[u_i \mid z_i, \mathbf{x}_i] = 0$$

The instrument must be uncorrelated with u_i**conditional on $\mathbf{x}_i$**.

2SLS with Covariates

Stage 1: Regress $w_i$on z_iAND $\mathbf{x}_i$:

$$w_i = \gamma_0 + \gamma_1 z_i + \mathbf{x}_i'\boldsymbol{\delta} + \eta_i$$
$$\hat{w}_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i + \mathbf{x}_i'\hat{\boldsymbol{\delta}}$$

Stage 2: Regress $y_i$on $\hat{w}_i$AND $\mathbf{x}_i$:

$$y_i = \alpha + \tau\hat{w}_i + \mathbf{x}_i'\boldsymbol{\beta} + \text{error}$$

Important Notes

1. **Include $\mathbf{x}_i$in BOTH stages.** This is crucial for consistency.
2. **Standard errors from Stage 2 are WRONG.** The OLS standard errors don't account for the fact that $\hat{w}_i$is estimated. Use IV/2SLS routines that compute correct standard errors.
3. **Exogenous variables are their own instruments.** The full instrument set is $(\mathbf{z}_i, \mathbf{x}_i)$.

4.7 Properties of 2SLS

Consistency

Theorem 4.1: Under assumptions IV.1 (relevance) and IV.2 (exogeneity), the 2SLS estimator is consistent:

$$\hat{\tau}^{2SLS} \xrightarrow{p} \tau$$

Proof (simple case):

$$\hat{\tau}^{2SLS} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Cov}(z_i, w_i)}$$

By the law of large numbers:

$$\hat{\tau}^{2SLS} \xrightarrow{p} \frac{Cov(z_i, y_i)}{Cov(z_i, w_i)}$$

From the structural equation $y_i = \alpha + \tau w_i + u_i$:

$$Cov(z_i, y_i) = \tau \cdot Cov(z_i, w_i) + Cov(z_i, u_i)$$

By exogeneity: $Cov(z_i, u_i) = 0$

Therefore: $$\text{plim } \hat{\tau}^{2SLS} = \frac{\tau \cdot Cov(z_i, w_i)}{Cov(z_i, w_i)} = \tau$$ $\square$

Asymptotic Normality

Under regularity conditions:

$$\sqrt{n}(\hat{\tau}^{2SLS} - \tau) \xrightarrow{d} N(0, V)$$

where:

$$V = \sigma_u^2 \cdot \frac{Var(z_i)}{[Cov(z_i, w_i)]^2}$$

2SLS is NOT Unbiased

Unlike OLS in the classical model, 2SLS is **biased in finite samples**. It is only consistent (unbiased asymptotically).

The finite-sample bias is approximately:

$$E[\hat{\tau}^{2SLS} - \tau] \approx \frac{\sigma_{u\eta}}{\sigma_\eta^2} \cdot \frac{1}{F+1}$$

where F is the first-stage F-statistic. This bias is toward the OLS estimate.

4.8 The Local Average Treatment Effect (LATE)

So far we've assumed a constant treatment effect. But what if effects are heterogeneous? The IV estimate has a specific interpretation.

Setup with Heterogeneous Effects

Let treatment be binary: $w_i \in \{0,1\}$.

Potential treatments:

- w_{1i}= treatment status if $z_i = 1$
- w_{0i}= treatment status if $z_i = 0$

Potential outcomes:

- y_{0i}= outcome if $w_i = 0$
- y_{1i}= outcome if $w_i = 1$

Individual treatment effect:

$$\tau_i = y_{1i} - y_{0i}$$

Compliance Types

Based on (w_{0i}, w_{1i}), we can classify individuals:

Type	w_{0i}	w_{1i}	Description
Compliers	0	1	Take treatment iff encouraged
Always-takers	1	1	Always take treatment
Never-takers	0	0	Never take treatment
Defiers	1	0	Do opposite of encouragement

The LATE Assumptions

Assumption LATE.1: Independence

$$z_i \perp\!\!\!\perp (y_{0i}, y_{1i}, w_{0i}, w_{1i})$$

The instrument is randomly assigned (or as-good-as-random).

Assumption LATE.2: Exclusion

$$y_i = y_{0i} + (y_{1i} - y_{0i}) \cdot w_i$$

The instrument affects outcomes only through treatment.

Assumption LATE.3: First Stage (Relevance)

$$E[w_{1i} - w_{0i}] \neq 0$$

The instrument affects treatment on average.

Assumption LATE.4: Monotonicity

$$w_{1i} \geq w_{0i} \text{ for all } i (\text{or } w_{1i} \leq w_{0i} \text{ for all } i)$$

No defiers exist. The instrument shifts everyone in the same direction (or not at all).

The LATE Theorem (Imbens & Angrist, 1994)

Theorem 4.2: Under assumptions LATE.1-4, the IV/Wald estimand identifies the **Local Average Treatment Effect**:

$$\tau^{IV} = \frac{E[y_i \mid z_i = 1] - E[y_i \mid z_i = 0]}{E[w_i \mid z_i = 1] - E[w_i \mid z_i = 0]} = E[y_{1i} - y_{0i} \mid w_{1i} > w_{0i}] = \text{LATE}$$

The IV estimand is the average treatment effect **for compliers only**.

Proof of the LATE Theorem

Step 1: Decompose the numerator (reduced form)

$$E[y_i \mid z_i = 1] - E[y_i \mid z_i = 0]$$

By the switching equation and independence:

$$\begin{aligned} &= E[w_{1i}y_{1i} + (1 - w_{1i})y_{0i}] - E[w_{0i}y_{1i} + (1 - w_{0i})y_{0i}] \\ &= E[(w_{1i} - w_{0i})(y_{1i} - y_{0i})] \end{aligned}$$

Step 2: Use monotonicity

With no defiers, $w_{1i} - w_{0i} \in \{0,1\}$:

- $= 1$for compliers
- $= 0$for always-takers and never-takers

Therefore:

$$E[(w_{1i} - w_{0i})(y_{1i} - y_{0i})] = E[y_{1i} - y_{0i} \mid \text{complier}] \cdot P(\text{complier})$$

Step 3: Decompose the denominator (first stage)

$$E[w_i \mid z_i = 1] - E[w_i \mid z_i = 0] = E[w_{1i} - w_{0i}] = P(\text{complier})$$

Step 4: Take the ratio

$$\tau^{IV} = \frac{E[y_{1i} - y_{0i} \mid \text{complier}] \cdot P(\text{complier})}{P(\text{complier})} = E[y_{1i} - y_{0i} \mid \text{complier}]$$ $\square$

Implications of LATE

1. **IV estimates a local effect.** It applies to compliers, not the whole population.
2. **Compliers are unobservable.** We cannot identify who they are from the data.
3. **Different instruments, different compliers.** Two valid instruments for the same treatment may give different estimates because they affect different subpopulations.

4. **External validity is limited.** LATE may not generalize to policy-relevant populations.

4.9 Example: The LATE of Education

Angrist & Krueger (1991): Quarter of Birth

Idea: Compulsory schooling laws require students to stay in school until age 16. Students born earlier in the year reach age 16 earlier and can drop out with less education.

Setup:

- y_i: log wages
- w_i: years of education
- z_i: quarter of birth (instrument)

Who are the compliers?

- People who would have dropped out at 16 with Q1 birth but stayed longer with Q4 birth
- These are the "marginal" students—those on the edge of dropping out

What LATE estimates:

- The return to education for students who are induced to stay in school by compulsory schooling laws
- NOT the return for those who would have stayed anyway (high-ability students)
- NOT the return for those who would have dropped out regardless

Policy relevance:

- If considering policies that affect marginal students, LATE is highly relevant
- If considering policies that affect all students, LATE may not generalize

4.10 Weak Instruments

A crucial practical issue: what if the instrument is only weakly correlated with treatment?

The Problem

When $Cov(z_i, w_i)$is small:

$$\hat{\tau}^{2SLS} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Cov}(z_i, w_i)}$$

- The denominator is close to zero
- Small estimation errors in the denominator lead to large errors in $\hat{\tau}$
- The estimator becomes highly variable and biased

Finite Sample Bias

With weak instruments, the 2SLS bias toward OLS can be substantial:

$$E[\hat{\tau}^{2SLS} - \tau] \approx \frac{\sigma_{u\eta}}{\sigma_\eta^2} \cdot \frac{1}{F}$$

where $F$is the first-stage F-statistic.

When F is small, bias is large!

The Staiger-Stock Rule of Thumb

Staiger & Stock (1997): If the first-stage F-statistic is less than 10, weak instrument bias is likely substantial.

$$F = \frac{(\hat{\gamma}_1)^2}{Var(\hat{\gamma}_1)} > 10$$

This is a minimum threshold. Many researchers now advocate for $F > 104$(Stock & Yogo critical values for 5% bias).

Detecting Weak Instruments

Test 1: First-stage F-statistic

- Report the F-statistic for the hypothesis $H_0 : \gamma_1 = 0$
- If $F < 10$, be very cautious

Test 2: Stock-Yogo Critical Values

- Compare F to critical values for desired maximum bias or size distortion
- Tables available in Stock & Yogo (2005)

Test 3: Partial R-squared

- Report the partial $R^2$of the excluded instruments
- Low partial R^2suggests weak instruments

Solutions for Weak Instruments

1. Find better instruments (preferred solution)

2. Limited Information Maximum Likelihood (LIML)

- Less biased than 2SLS with weak instruments
- Same asymptotic distribution as 2SLS with strong instruments

3. Anderson-Rubin confidence intervals

- Valid even with weak instruments
- Based on testing $H_0: \tau = \tau_0$for all τ_0
- Can be very wide but always valid

4. Identification-robust inference

- Kleibergen (2002) LM test
- Moreira (2003) conditional likelihood ratio test

4.11 Testing Instrument Validity

Can We Test the Exclusion Restriction?

Short answer: No.

The exclusion restriction $Cov(z_i, u_i) = 0$involves the unobservable u_i. It cannot be directly tested.

The exclusion restriction is maintained by assumption, justified by:

- Economic theory
- Institutional knowledge
- Research design

What CAN We Test?

1. Relevance (First Stage)

$H_0: \gamma_1 = 0$ is testable via the F-statistic.

2. Overidentifying Restrictions

When we have **more instruments than endogenous variables**, we can test whether the “extra” instruments are consistent with the included ones.

4.12 Multiple Instruments and Overidentification

The Setup

Suppose we have:

- 1 endogenous variable: w_i
- $m > 1$instruments: $z_{1i}, z_{2i}, \dots, z_{mi}$

The model is **overidentified**: we have more moment conditions than parameters.

Why Overidentification Helps

1. Efficiency: Multiple instruments can improve precision.

2. Testing: We can test whether all instruments give consistent estimates.

2SLS with Multiple Instruments

First stage: Regress $w_i$on all instruments:

$$w_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \cdots + \gamma_m z_{mi} + \eta_i$$

Second stage: Regress $y_i$on $\widehat{w}_i$:

$$y_i = \alpha + \tau \widehat{w}_i + \text{error}$$

The Overidentification Test (Sargan-Hansen J-Test)

Intuition: If all instruments are valid, they should all give similar estimates. If one instrument is invalid, it will give a different estimate.

Procedure:

1. Estimate the model by 2SLS, obtain residuals $\hat{u}_i = y_i - \hat{\alpha} - \hat{\tau} w_i$
2. Regress $\hat{u}_i$on all instruments: $\hat{u}_i = \mathbf{z}_i' \boldsymbol{\pi} + \text{error}$
3. Compute $J = n \cdot R^2$from this regression

Test statistic:

$$J \overset{d}{\to} \chi^2(m - k)$$

under H_0: all instruments are valid, where m= number of instruments, k= number of endogenous variables.

Interpretation:

- Reject H_0(large J) → at least one instrument is likely invalid
- Fail to reject → instruments are consistent with each other (but could all be invalid in the same way!)

Limitations of Overidentification Tests

1. **Cannot detect if all instruments are invalid.** The test has power only against instruments being inconsistent with each other.
2. **Requires at least one valid instrument.** If none are valid, the test tells us nothing.
3. **Low power with weak instruments.** Weak instruments make it hard to detect violations.

4.13 Examples of Instrumental Variables

Let's examine several classic IV applications.

Example 1: Returns to Schooling (Card, 1995)

Question: What is the causal effect of education on earnings?

Endogeneity: Ability, motivation, and family background affect both education and earnings.

Instrument: Geographic proximity to a four-year college.

Argument for validity:

- **Relevance:** Living near a college reduces the cost of attending, increasing education.
- **Exclusion:** Distance to college shouldn't directly affect adult earnings (conditional on observed characteristics).

Concerns:

- College location may be correlated with local economic conditions
- Families may choose to live near colleges for unobserved reasons

Example 2: Colonial Origins of Development (Acemoglu, Johnson, Robinson 2001)

Question: Do institutions cause economic development?

Endogeneity: Rich countries may be able to afford good institutions (reverse causality).

Instrument: European settler mortality rates in colonies.

Argument:

- **Relevance:** Where Europeans could settle (low mortality), they established good institutions. Where they couldn't (high mortality), they established extractive institutions.
- **Exclusion:** Historical mortality doesn't directly affect current GDP, only through persistent institutions.

This paper won widespread acclaim for its creative identification strategy.

Example 3: The Effect of Military Service (Angrist, 1990)

Question: How does military service affect lifetime earnings?

Instrument: Vietnam draft lottery.

Argument:

- **Relevance:** Draft eligibility strongly predicts military service.
- **Exclusion:** Birthday lottery is random, unrelated to potential earnings.

This is considered one of the cleanest natural experiments in economics.

Example 4: Police and Crime (Levitt, 1997)

Question: Do more police reduce crime?

Endogeneity: Cities with high crime hire more police (reverse causality).

Instrument: Electoral cycles—police hiring increases in election years.

Argument:

- **Relevance:** Mayors hire more police before elections to appear tough on crime.
- **Exclusion:** The timing of elections doesn't directly affect crime.

4.14 Practical Implementation

Stata Commands

stata

```
*Basic 2SLS

ivregress 2sls y x1 x2 (w = z), robust
```

```
* First-stage diagnostics
estat firststage

* Overidentification test (with multiple instruments)
ivregress 2sls y x1 x2 (w = z1 z2 z3), robust
estat overid

* Weak instrument robust inference
rivtest  // requires installation
```

R Commands

```
# Using the AER package
library(AER)
# Basic 2SLS
model <- ivreg(y ~ w + x1 + x2 | z + x1 + x2, data = mydata)
summary(model, diagnostics = TRUE)
# The diagnostics option reports:
# - Weak instruments test
# - Wu-Hausman test
# - Sargan test (with overidentification)
```

Reporting IV Results

A good IV paper reports:

1. **First-stage results:** Coefficient on instrument, F-statistic
2. **Reduced-form results:** Effect of instrument on outcome directly
3. **2SLS results:** The causal estimate
4. **OLS results:** For comparison (to show direction of bias)
5. **Overidentification test:** If applicable

6. **Discussion of exclusion restriction:** Why is the instrument valid?

4.15 Common Pitfalls and Criticisms

Pitfall 1: Invalid Instruments

The most common problem. Examples of potentially invalid instruments:

- **Lagged dependent variables:** Often correlated with persistent unobservables
- **Industry or geographic aggregates:** May be correlated with local shocks
- **“Clever” instruments without clear exclusion:** Ingenuity doesn’t guarantee validity

Pitfall 2: Weak Instruments

Even valid instruments can be too weak to be useful. Always report first-stage F-statistics.

Pitfall 3: Overinterpreting LATE

The IV estimate applies to compliers. Be careful about:

- Generalizing to the entire population
- Policy conclusions for non-complier populations

Pitfall 4: Too Many Instruments

With many instruments:

- Overidentification tests lose power
- 2SLS bias increases (Bekker, 1994)
- Can lead to overfitting in the first stage

Rule of thumb: Keep the number of instruments small.

The “Identification Zoo” Critique

Angrist and Pischke popularized IV methods, but critics argue:

- Many instruments are questionable
- Exclusion restrictions are often assumed, not justified
- Researchers choose instruments that give desired results

Response: IV requires careful thought and transparency. We must defend exclusion restrictions with theory and institutional knowledge.

4.16 Summary of Part 4

Key Assumptions

Relevance: $Cov(z_i, w_i) \neq 0$

Exogeneity: $Cov(z_i, u_i) = 0$

Monotonicity (for LATE): No defiers

Key Results

Wald Estimator:

$$\tau^{IV} = \frac{E[y_i \mid z_i = 1] - E[y_i \mid z_i = 0]}{E[w_i \mid z_i = 1] - E[w_i \mid z_i = 0]}$$

2SLS Formula:

$$\hat{\tau}^{2SLS} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Cov}(z_i, w_i)}$$

LATE Interpretation:

$$\tau^{IV} = E[y_{1i} - y_{0i} \mid \text{Compliers}]$$

Practical Checklist

Check	**What to Report**
First stage	Coefficient, F-statistic (should be > 10)
Exclusion restriction	Theoretical justification
LATE interpretation	Who are the compliers?
Overidentification	Sargan/Hansen J-test (if applicable)
Robustness	Alternative instruments, specifications

Exercises for Part 4

Exercise 4.1: In the draft lottery example, suppose:

- $E[y_i \mid z_i = 1] = 48{,}000$, $E[y_i \mid z_i = 0] = 52{,}000$
- $E[w_i \mid z_i = 1] = 0.26$, $E[w_i \mid z_i = 0] = 0.07$

(a) Calculate the Wald estimate of the effect of military service. (b) Calculate the share of compliers. (c) Interpret the estimate in terms of LATE.

Exercise 4.2: A researcher uses father's education as an instrument for own education in a wage regression. Critique this instrument in terms of relevance and exclusion.

Exercise 4.3: Explain why a first-stage F-statistic of 5 is concerning. What are the consequences for inference?

Exercise 4.4: You have two instruments for education: (1) distance to college, (2) tuition costs in the local area. Your overidentification test rejects. What do you conclude? What should you do?

Exercise 4.5: In a randomized encouragement design, 50% of subjects are encouraged to exercise (z=1). Among the encouraged, 40% actually exercise. Among the non-encouraged, 10% exercise. (a) What fraction of the population are compliers? (b) If encouraged have average health score 75 and non-encouraged have average 70, what is the LATE of exercise on health?

PART 5: DIFFERENCE-IN-DIFFERENCES

Instrumental variables require finding exogenous variation—often difficult. Next, **Difference-in-Differences**, which exploits variation across time and groups to identify causal effects without external instruments.

Structure:

1. Basic intuition and setup
2. The parallel trends assumption
3. The 2x2 case (two groups, two periods)
4. Regression formulation
5. Multiple periods and groups
6. Event studies
7. Staggered treatment timing
8. Testing parallel trends
9. Extensions and recent developments (Callaway-Sant'Anna, Sun-Abraham, etc.)
10. Practical implementation
11. Examples
12. Common pitfalls

Exploiting Time Variation for Causal Inference

5.1 The Core Idea

Difference-in-Differences (DiD, DD, or Diff-in-Diff) is one of the most widely used methods in applied economics. It exploits variation across **groups** and **time** to identify causal effects.

The Setup

Imagine we observe two groups over two time periods:

- **Treatment group:** Receives treatment starting in period 2
- **Control group:** Never receives treatment

We observe outcomes for both groups in both periods.

The Naive Approaches and Their Problems

Approach 1: Compare treated vs. control in period 2

$$\bar{y}_{treat,2} - \bar{y}_{control,2}$$

Problem: Groups may differ for reasons unrelated to treatment (selection bias).

Approach 2: Compare treated group before vs. after

$$\bar{y}_{treat,2} - \bar{y}_{treat,1}$$

Problem: Other things may have changed between periods (time trends, shocks).

The DiD Solution

Combine both comparisons:

$$\hat{\tau}^{DiD} = (\bar{y}_{treat,2} - \bar{y}_{treat,1}) - (\bar{y}_{control,2} - \bar{y}_{control,1})$$

Interpretation:

- First difference: How much did the treated group change?
- Second difference: Subtract how much the control group changed
- What remains: The effect of treatment (if assumptions hold)

The Key Assumption: Parallel Trends

DiD assumes that in the **absence of treatment**, both groups would have experienced the same change over time.

$$E[y_{0i,2} - y_{0i,1} \mid \text{Treated}] = E[y_{0i,2} - y_{0i,1} \mid \text{Control}]$$

This says: the trend in potential outcomes without treatment is the same for both groups.

5.2 Formal Framework: The 2×2 Case

Let’s formalize the basic setup with two groups and two periods.

Notation

- i: individual
- $t \in \{1,2\}$: time period (1 = pre-treatment, 2 = post-treatment)
- $G_i \in \{0,1\}$: group indicator (1 = treatment group, 0 = control group)
- D_{it}: treatment indicator (1 = actually treated at time t)

In the 2×2 case:

$$D_{it} = G_i \cdot \mathbf{1}[t = 2]$$

Treatment group members are treated only in period 2.

Potential Outcomes

- $y_{it}(0)$: outcome for individual $i$at time $t$without treatment
- $y_{it}(1)$: outcome for individual $i$at time $t$with treatment

Observed outcome:

$$y_{it} = D_{it} \cdot y_{it}(1) + (1 - D_{it}) \cdot y_{it}(0)$$

The Target Parameter

Average Treatment Effect on the Treated (ATT):

$$\tau^{ATT} = E[y_{i2}(1) - y_{i2}(0) \mid G_i = 1]$$

This is the average effect of treatment on the treatment group in period 2.

5.3 Identification

The Parallel Trends Assumption

Assumption DiD.1 (Parallel Trends):

$$E[y_{i2}(0) - y_{i1}(0) \mid G_i = 1] = E[y_{i2}(0) - y_{i1}(0) \mid G_i = 0]$$

In words: absent treatment, the treatment and control groups would have experienced the same change in outcomes.

No Anticipation Assumption

Assumption DiD.2 (No Anticipation):

$$y_{i1} = y_{i1}(0) \text{ for all } i$$

In period 1, no one is treated, so observed outcomes equal untreated potential outcomes.

Identification Result

Theorem 5.1: Under parallel trends and no anticipation, the ATT is identified by:

$$\tau^{ATT} = \{E[y_{i2} \mid G_i = 1] - E[y_{i1} \mid G_i = 1]\} - \{E[y_{i2} \mid G_i = 0] - E[y_{i1} \mid G_i = 0]\}$$

Proof:

The ATT is:

$$\tau^{ATT} = E[y_{i2}(1) \mid G_i = 1] - E[y_{i2}(0) \mid G_i = 1]$$

The first term is observed:

$$E[y_{i2}(1) \mid G_i = 1] = E[y_{i2} \mid G_i = 1]$$

The second term is counterfactual. We need to identify $E[y_{i2}(0) \mid G_i = 1]$.

Step 1: Add and subtract $E[y_{i1}(0) \mid G_i = 1]$:

$$E[y_{i2}(0) \mid G_i = 1] = E[y_{i1}(0) \mid G_i = 1] + \{E[y_{i2}(0) \mid G_i = 1] - E[y_{i1}(0) \mid G_i = 1]\}$$

Step 2: By no anticipation, $E[y_{i1}(0) \mid G_i = 1] = E[y_{i1} \mid G_i = 1]$.

Step 3: By parallel trends:

$$E[y_{i2}(0) - y_{i1}(0) \mid G_i = 1] = E[y_{i2}(0) - y_{i1}(0) \mid G_i = 0]$$

Step 4: For the control group, we observe untreated outcomes in both periods:

$$E[y_{i2}(0) - y_{i1}(0) \mid G_i = 0] = E[y_{i2} - y_{i1} \mid G_i = 0]$$

Step 5: Combining:

$$E[y_{i2}(0) \mid G_i = 1] = E[y_{i1} \mid G_i = 1] + E[y_{i2} - y_{i1} \mid G_i = 0]$$

Step 6: Substitute into the ATT:

$$\tau^{ATT} = E[y_{i2} \mid G_i = 1] - E[y_{i1} \mid G_i = 1] - E[y_{i2} - y_{i1} \mid G_i = 0]$$

$$= {E[y_{i2} \mid G_i = 1] - E[y_{i1} \mid G_i = 1]} - {E[y_{i2} \mid G_i = 0] - E[y_{i1} \mid G_i = 0]}$$ $\square$

5.4 The DiD Estimator

Sample Analog

The DiD estimator replaces population expectations with sample means:

$$\hat{\tau}^{DiD} = (\bar{y}_{1,2} - \bar{y}_{1,1}) - (\bar{y}_{0,2} - \bar{y}_{0,1})$$

where $\bar{y}_{g,t}$is the sample mean for group $g$in period t.

The 2×2 Table

	Period 1	**Period 2**	**Difference**
Treatment	$\bar{y}_{1,1}$	$\bar{y}_{1,2}$	$\bar{y}_{1,2} - \bar{y}_{1,1}$
Control	$\bar{y}_{0,1}$	$\bar{y}_{0,2}$	$\bar{y}_{0,2} - \bar{y}_{0,1}$
Difference	$\bar{y}_{1,1} - \bar{y}_{0,1}$	$\bar{y}_{1,2} - \bar{y}_{0,2}$	$\hat{\tau}^{DiD}$

Numerical Example: Minimum Wage and Employment

Question: Does raising the minimum wage reduce employment?

Setup (Card & Krueger, 1994):

- Treatment group: New Jersey (raised minimum wage in April 1992)
- Control group: Eastern Pennsylvania (no change)
- Outcome: Employment at fast-food restaurants
- Period 1: February 1992 (before)
- Period 2: November 1992 (after)

Data:

	Before	**After**	**Change**
New Jersey	20.44	21.03	+0.59
Pennsylvania	23.33	21.17	-2.16
Difference	-2.89	-0.14	**+2.75**

DiD Estimate:

$$\hat{\tau}^{DiD} = 0.59 - (-2.16) = 2.75$$

Interpretation: The minimum wage increase in New Jersey **increased** employment by 2.75 FTEs per restaurant, contrary to standard theory!

5.5 Regression Formulation

The DiD estimator can be computed via OLS, which facilitates adding controls and computing standard errors.

The Basic DiD Regression

$$y_{it} = \alpha + \gamma \cdot G_i + \lambda \cdot Post_t + \tau \cdot (G_i \times Post_t) + \varepsilon_{it}$$

where:

- $G_i = 1$if individual $i$is in the treatment group
- $Post_t = 1$if $t = 2$(post-treatment period)
- $G_i \times Post_t$is the interaction term

Interpreting the Coefficients

	Period 1 ($Post = 0$)	**Period 2 ($Post = 1$)**
Control ($G = 0$)	α	$\alpha + \lambda$
Treatment ($G = 1$)	$\alpha + \gamma$	$\alpha + \gamma + \lambda + \tau$

Coefficient interpretations:

- α: baseline (control group, pre-period)
- γ: pre-existing difference between groups
- λ: common time trend
- τ: **DiD treatment effect**

Proof of Equivalence

$$\begin{aligned}\tau &= E[y_{it} \mid G = 1, Post = 1] - E[y_{it} \mid G = 1, Post = 0] \\ &-\{E[y_{it} \mid G = 0, Post = 1] - E[y_{it} \mid G = 0, Post = 0]\} \\ &= (\alpha + \gamma + \lambda + \tau) - (\alpha + \gamma) - [(\alpha + \lambda) - \alpha]\end{aligned}$$

$$= \tau$$ $\square$

Adding Covariates

$$y_{it} = \alpha + \gamma \cdot G_i + \lambda \cdot Post_t + \tau \cdot (G_i \times Post_t) + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

Covariates $\mathbf{x}_{it}$can:

1. Improve precision (reduce residual variance)
2. Make parallel trends more plausible (conditional parallel trends)

Conditional Parallel Trends:

$$E[y_{i2}(0) - y_{i1}(0) \mid G_i = 1, \mathbf{x}_i] = E[y_{i2}(0) - y_{i1}(0) \mid G_i = 0, \mathbf{x}_i]$$

Trends are parallel within cells defined by $\mathbf{x}_i$.

5.6 Panel Data and Fixed Effects

In practice, DiD is often implemented with panel data and fixed effects.

Panel Data Setup

We observe individuals $i = 1, \dots, N$over periods $t = 1, \dots, T$.

Two-Way Fixed Effects (TWFE) Model

$$y_{it} = \alpha_i + \lambda_t + \tau \cdot D_{it} + \varepsilon_{it}$$

where:

- α_i: individual fixed effects (time-invariant individual characteristics)
- λ_t: time fixed effects (common shocks affecting all individuals)
- D_{it}: treatment indicator (= 1 if individual $i$is treated at time t)
- τ: treatment effect

What Do Fixed Effects Control For?

Individual fixed effects (α_i):

- Permanent differences between individuals
- Absorbs $G_i$in the 2×2 case
- Controls for time-invariant confounders

Time fixed effects (λ_t):

- Common shocks affecting all individuals
- Absorbs $Post_t$in the 2×2 case
- Controls for aggregate trends

Equivalence in the 2×2 Case

In the 2×2 case with balanced panels:

$$\hat{\tau}^{TWFE} = \hat{\tau}^{DiD}$$

The TWFE regression gives the same estimate as the simple difference-in-differences.

Implementation in Stata

```
stata

* Method 1: Interaction regression

reg y i.treat##i.post, robust cluster(state)


* Method 2: Fixed effects with xtreg

xtset id time

xtreg y treat_post i.time, fe robust cluster(state)


* Method 3: Using reghdfe (recommended)

reghdfe y treat_post, absorb(id time) cluster(state)
```

Implementation in R

```
# Method 1: Interaction regression

lm(y ~ treat * post, data = df)


# Method 2: Fixed effects with plm

library(plm)

plm(y ~ treat_post, data = df,

  index = c(“id”, “time”), model = “within”, effect = “twoways”)


# Method 3: Using fixest (recommended)

library(fixest)

feols(y ~ treat_post | id + time, data = df, cluster = “state”)
```

5.7 Standard Errors and Clustering

Correct inference in DiD requires careful attention to standard errors.

The Problem: Serial Correlation

Outcomes for the same individual are correlated over time. Standard OLS standard errors ignore this, leading to:

- Standard errors that are too small
- Over-rejection of null hypotheses
- False positives

Bertrand, Duflo, and Mullainathan (2004)

This influential paper documented the severe under-estimation of standard errors in DiD studies.

Simulation results: With serially correlated errors, nominal 5% tests reject the null 45% of the time when there's no true effect!

Solution: Cluster Standard Errors

Cluster at the level of treatment assignment (typically state, firm, etc.):

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

where $g$indexes clusters.

Practical Guidance

1. **Cluster at the treatment level.** If treatment varies at the state level, cluster by state.
2. **Few clusters problem.** With fewer than ~40-50 clusters, cluster-robust SEs are unreliable. Consider:
   - Wild cluster bootstrap
   - Aggregation to the cluster level
   - Randomization inference
3. **Two-way clustering.** Sometimes appropriate to cluster by both entity and time:

stata

```
reghdfe y treat_post, absorb(id time) cluster(state time)
```

5.8 Testing Parallel Trends

The parallel trends assumption is crucial but **untestable** (it’s about counterfactuals). However, we can assess its plausibility.

The Logic

If parallel trends hold, treated and control groups should have similar trends in the **pre-treatment period**. We can test this.

Important caveat: Pre-treatment parallel trends don’t guarantee post-treatment parallel trends. But violations in the pre-period strongly suggest the assumption is wrong.

Visual Inspection

Plot average outcomes over time for treated and control groups:

stata

```
* Create group-time means
collapse (mean) y, by(treat time)
twoway (line y time if treat==0) (line y time if treat==1), ///
    xline(treatment_time) legend(order(1 “Control” 2 “Treatment”))
```

Look for:

- Parallel pre-trends (good)
- Diverging pre-trends (bad)
- Pre-treatment “dip” in treated group (anticipation effects, bad)

Formal Test: Pre-Treatment Placebo

Estimate “effects” in pre-treatment periods:

$$y_{it} = \alpha_i + \lambda_t + \sum_{k \neq -1} \tau_k \cdot \mathbf{1}[t - E_i = k] \cdot G_i + \varepsilon_{it}$$

where $E_i$is the treatment event time for individual i, and $k$indexes time relative to treatment.

Interpretation:

- $\tau_k$for $k < 0$: “effects” in pre-treatment periods (should be zero)
- $\tau_k$for $k \geq 0$: post-treatment effects

Testing: Joint test of $H_0 \colon \tau_{-2} = \tau_{-3} = \cdots = 0$

5.9 Event Study Designs

Event studies generalize DiD to examine dynamic treatment effects over time.

The Event Study Regression

$$y_{it} = \alpha_i + \lambda_t + \sum_{k \neq -1} \tau_k \cdot D_{it}^k + \varepsilon_{it}$$

where:

- $D_{it}^k = \mathbf{1}[t - E_i = k]$for treated units, 0 otherwise
- E_i= event time (when unit $i$first receives treatment)
- k= time relative to event ($k < 0$ is pre-event, $k \geq 0$is post-event)
- We omit $k = -1$as the reference period (normalization)

Interpreting Event Study Coefficients

- τ_{-3}, τ_{-2}: Pre-trends (should be ≈ 0 if parallel trends holds)
- τ_0: Immediate effect (at time of treatment)
- $\tau_1, \tau_2, \ldots$: Dynamic effects (may grow, shrink, or stay constant)

Binning Endpoints

With many periods, we often “bin” the endpoints:

$$y_{it} = \alpha_i + \lambda_t + \tau_{< -K} \cdot D_{it}^{< -K} + \sum_{k=-K}^{L} \tau_k \cdot D_{it}^k + \tau_{>L} \cdot D_{it}^{>L} + \varepsilon_{it}$$

where:

- $D_{it}^{< -K} = \mathbf{1}[t - E_i < -K]$: all periods more than Kbefore treatment

- $D_{it}^{>L} = \mathbf{1}[t - E_i > L]$: all periods more than Lafter treatment

Example: Event Study of Minimum Wage

```
stata
* Generate relative time indicators
gen rel_time = time - treatment_time
* Create dummies (binning at ±3)
forval k = -3/3 {
  gen D`k’ = (rel_time == `k’) & treated
}
gen D_pre = (rel_time < -3) & treated
gen D_post = (rel_time > 3) & treated
* Regression (omit k=-1 as reference)
reghdfe y D_pre D_m3 D_m2 D0 D1 D2 D3 D_post, absorb(id time) cluster(state)
* Plot coefficients
coefplot, keep(D*) vertical yline(0) xline(4.5)
```

5.10 Staggered Treatment Timing

Many real-world settings have **staggered adoption**: different units receive treatment at different times.

The Setup

- Units $i = 1, \dots, N$
- Periods $t = 1, \dots, T$
- Treatment timing: unit $i$first treated at time E_i(some units may never be treated: $E_i = \infty$)

Treatment indicator:

$$D_{it} = \mathbf{1}[t \geq E_i]$$

The Standard TWFE Approach

$$y_{it} = \alpha_i + \lambda_t + \tau \cdot D_{it} + \varepsilon_{it}$$

Researchers often estimate this and interpret $\hat{\tau}$ as "the" treatment effect.

The Problem with TWFE Under Staggered Timing

Recent econometric research has shown that TWFE can be severely biased with staggered treatment and heterogeneous effects. (for more information see SantaAnna & Callaway recent works)

Key papers:

- Goodman-Bacon (2021)
- de Chaisemartin & D'Haultfœuille (2020)
- Callaway & Sant'Anna (2021)
- Sun & Abraham (2021)
- Borusyak, Jaravel & Spiess (2024)

The Goodman-Bacon Decomposition

Goodman-Bacon (2021) showed that the TWFE estimator is a weighted average of all possible 2×2 DiD comparisons:

$$\hat{\tau}^{TWFE} = \sum_k \sum_{l \neq k} s_{kl}\, \hat{\tau}_{kl}^{2\times 2}$$

where:

- k, lindex different "cohorts" (groups treated at different times)
- $s_{kl}$are weights that depend on sample sizes and treatment timing
- $\hat{\tau}_{kl}^{2\times 2}$is the 2×2 DiD comparing cohort $k$to cohort l

The Problem: "Bad" Comparisons

Some comparisons use **already-treated units as controls**:

- Early-treated units (treated in period 2) compared to late-treated units (treated in period 5)
- In period 3, early-treated are "controls" for late-treated

If treatment effects are dynamic (change over time), this is problematic:

- Early-treated units are experiencing ongoing treatment effects
- They're not valid counterfactuals for untreated potential outcomes
- The comparison estimates: (late effect) - (earlier effect), which can even be negative!

When Is TWFE Okay?

TWFE gives a reasonable estimate when:

1. Treatment effects are **homogeneous** (same for all units and over time)
2. There are **never-treated units** that serve as clean controls
3. Treatment timing is **not too staggered**

Diagnostic: Goodman-Bacon Decomposition

Examine which comparisons drive the TWFE estimate:

stata

```
* Install bacondecomp
ssc install bacondecomp


* Run decomposition
bacondecomp y D, ddetail
```

This shows:

- The weight on each type of comparison
- The estimate from each type of comparison
- How much "bad" comparisons contribute to the overall estimate

5.11 Modern DiD Methods for Staggered Designs

Recent methods address the problems with TWFE.

Method 1: Callaway & Sant'Anna (2021)

Approach: Estimate group-time specific effects $ATT(g,t)$for each cohort g(treatment timing) and time t.

Key features:

- Uses only clean comparisons (never-treated or not-yet-treated as controls)
- Allows for heterogeneous effects
- Can aggregate to summary measures

Implementation:

```
# R
library(did)
out <- att_gt(yname = “y”, tname = “time”, idname = “id”,
          gname = “first_treat”, data = df)
summary(out)
aggte(out, type = “dynamic”)  # event study
aggte(out, type = “simple”)   # overall average
```

stata

```
* Stata
ssc install csdid
csdid y, ivar(id) time(time) gvar(first_treat)
csdid_estat simple
csdid_estat event
```

Method 2: Sun & Abraham (2021)

Approach: Interaction-weighted estimator that corrects for contamination in event studies.

Key insight: Standard event study coefficients are contaminated by effects from other periods. Sun & Abraham propose using cohort-specific effects and then aggregating.

Implementation:

stata

```
* Stata
ssc install eventstudyinteract
eventstudyinteract y L*event F*event, cohort(first_treat) ///
```

```
  control_cohort(never_treat) absorb(id time) cluster(state)
```

Method 3: Borusyak, Jaravel & Spiess (2024) — Imputation

Approach: Impute counterfactual outcomes for treated observations using untreated observations, then compare actual to imputed.

Key features:

- Efficient (uses all clean comparisons)
- Flexible (allows for covariates, different parallel trends assumptions)
- Provides valid event studies

Implementation:

```
stata
* Stata
ssc install did_imputation
did_imputation y id time first_treat, horizons(0/5) pretrends(3)

# R
library(didimputation)
did_imputation(data = df, yname = “y”, gname = “first_treat”,
           tname = “time”, idname = “id”)
```

Method 4: de Chaisemartin & D’Haultfœuille (2020)

Approach: Identify conditions under which TWFE is valid, and provide alternative estimators when it’s not.

Implementation:

```
stata
ssc install did_multiplegt
did_multiplegt y id time D, robust_dynamic dynamic(5) placebo(3) cluster(state)
```

Comparison of Methods

Method	Strengths	Limitations
Callaway-Sant'Anna	Flexible, clear framework	Can be imprecise with many cohorts
Sun-Abraham	Easy event study	Requires cohort×time interactions
Imputation (BJS)	Efficient, handles covariates well	Requires correct specification
de Chaisemartin-D'H	Minimal assumptions	Can be noisy

For Practical purposes

1. **Start with diagnostics:** Run Goodman-Bacon decomposition to see if "bad" comparisons are important.
2. **Use modern methods:** If there's staggered timing and potential heterogeneity, use Callaway-Sant'Anna or imputation methods.
3. **Compare results:** If TWFE and modern methods give similar answers, you can be more confident.
4. **Report event studies:** Always show dynamic effects to check for pre-trends and effect dynamics.

5.12 Threats to Validity

Threat 1: Violation of Parallel Trends

The fundamental assumption. If trends differ, DiD is biased.

Signs of trouble:

- Diverging pre-trends
- Treatment timing correlated with trends
- Different group compositions over time

Partial solutions:

- Conditional parallel trends (control for covariates)
- Group-specific time trends (strong assumptions)
- Matching on pre-treatment trends

Threat 2: Anticipation Effects

If units change behavior **before** treatment in anticipation:

- Pre-treatment outcomes are affected
- The “pre-period” baseline is contaminated

Example: Workers might reduce effort before a plant closure is announced.

Solutions:

- Use earlier pre-treatment periods
- Model anticipation explicitly
- Check for pre-trends

Threat 3: Spillovers

If treatment of some units affects outcomes of control units:

- Control group is “contaminated”
- DiD underestimates the effect

Example: Minimum wage increase in New Jersey affects employment in neighboring Pennsylvania.

Solutions:

- Use geographically distant controls
- Model spillovers explicitly
- Spatial econometrics

Threat 4: Compositional Changes

If the composition of groups changes due to treatment:

- Selection into/out of groups
- Changes in who is observed

Example: Minimum wage increase causes low-productivity restaurants to close. Surviving restaurants have higher average employment, but not because employment increased.

Solutions:

- Balanced panel (same units throughout)
- Bound the effect using worst-case assumptions
- Focus on outcomes that can’t change composition (e.g., mortality instead of employment)

Threat 5: Functional Form

DiD assumes effects are additive. With non-linear outcomes (proportions, counts), this may be problematic.

Example: If the outcome is a rate bounded between 0 and 1, additive parallel trends may not hold.

Solutions:

- Log transformation (for positive outcomes)
- Poisson regression (for counts)
- Consider ratio estimators

5.13 Extensions

Triple Differences (DDD)

When you have an additional comparison group, you can add a third difference.

Setup:

- Two groups (treatment, control)
- Two time periods (before, after)
- Two subgroups (affected, unaffected)

Model:

$$\begin{aligned}y_{ijt} &= \alpha + \beta_1 G_i + \beta_2 Post_t + \beta_3 Affected_j \\ &+\gamma_1(G_i \times Post_t) + \gamma_2(G_i \times Affected_j) + \gamma_3(Post_t \times Affected_j) \\ &+\tau(G_i \times Post_t \times Affected_j) + \varepsilon_{ijt}\end{aligned}$$

Interpretation: $\tau$is the treatment effect on the affected subgroup, differencing out:

- Treatment-control differences
- Before-after differences
- Affected-unaffected differences

Example (Gruber, 1994): Effect of mandated maternity benefits on wages of women of childbearing age.

- Treatment: States that mandated benefits

- Control: States that didn’t
- Affected: Women 20-40
- Unaffected: Men, older women

For a complete understanding of Tripple-Diff and its application; read my advisor Nishith Prakash’s and Karthik Muralidharan’s:

Muralidharan, Karthik, and Nishith Prakash. 2017. “Cycling to School: Increasing Secondary School Enrollment for Girls in India.” American Economic Journal: Applied Economics 9 (3): 321–50. DOI: 10.1257/app.20160004

YouTube Video

Or 2007 Methods Lecture, Jeffrey Wooldridge, “Difference in Differences Estimation”

Synthetic Control (it will come in detail in next part)

When we have only one or few treated units, you can construct a “synthetic” control as a weighted average of untreated units. We’ll cover this in detail in Part 6.

Changes-in-Changes

Athey & Imbens (2006) proposed a method that:

- Allows for heterogeneous effects
- Uses the entire distribution, not just means
- Requires weaker assumptions than standard DiD

Key assumption: The distribution of unobserved factors is the same across groups within quantiles.

5.14 Complete Example: The Mariel Boatlift

Background

In April 1980, Fidel Castro announced that Cubans wishing to leave could do so from the port of Mariel. About 125,000 Cubans emigrated to the US, mostly to Miami, increasing Miami’s labor force by 7% almost overnight.

Question: Did this sudden influx of low-skilled immigrants reduce wages and employment for native workers?

Card (1990) Analysis

Design:

- Treatment: Miami

- Control: Atlanta, Los Angeles, Houston, Tampa-St. Petersburg (similar cities)
- Before: 1979
- After: 1981-1985

The Data

City	Unemployment Rate 1979	Unemployment Rate 1981
Miami	5.1%	7.2%
Comparison cities	6.0%	6.8%

DiD Calculation:

$$\hat{\tau}^{DiD} = (7.2 - 5.1) - (6.8 - 6.0) = 2.1 - 0.8 = 1.3\%$$

Wait—this suggests the Mariel Boatlift **increased** unemployment in Miami! But is this statistically significant? And does it persist?

Full Results

Card examined wages and unemployment for different demographic groups over multiple years. Key findings:

- No significant effect on wages of native workers
- No significant effect on unemployment of native workers
- The labor market absorbed the shock

Controversy: Borjas (2017)

Borjas reanalyzed the data, focusing specifically on high school dropouts (most comparable to Mariel immigrants). He found wage declines of 10-30%.

The debate highlights:

- Choice of control group matters
- Choice of outcome group matters
- Parallel trends may not hold
- Small sample issues (few workers in some cells)

5.15 Summary of Part 5

Key Assumptions

Parallel Trends:

$$E[y_{i2}(0) - y_{i1}(0) \mid G_i = 1] = E[y_{i2}(0) - y_{i1}(0) \mid G_i = 0]$$

No Anticipation:

$$y_{i,pre} = y_{i,pre}(0)$$

Key Formulas

Basic DiD Estimator:

$$\hat{\tau}^{DiD} = (\bar{y}_{1,2} - \bar{y}_{1,1}) - (\bar{y}_{0,2} - \bar{y}_{0,1})$$

Regression:

$$y_{it} = \alpha + \gamma G_i + \lambda Post_t + \tau(G_i \times Post_t) + \varepsilon_{it}$$

TWFE:

$$y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \varepsilon_{it}$$

Key Insights

1. **DiD removes time-invariant confounding** through differencing.
2. **Parallel trends is crucial** and should be assessed with pre-treatment data.
3. **Cluster standard errors** at the level of treatment assignment.
4. **With staggered timing, TWFE can be biased.** Use modern methods (Callaway-Sant'Anna, imputation, etc.).
5. **Event studies** reveal dynamic effects and help assess pre-trends.

Practical Checklist

Step	What to Do
1. Plot trends	Visual check of parallel trends
2. Test pre-trends	Event study with pre-treatment coefficients
3. Choose controls	Similar units, good overlap
4. Cluster SEs	At treatment level
5. Check staggered timing	If present, use modern methods
6. Robustness	Alternative controls, specifications

Exercises for Part 5

Exercise 5.1: Consider the following data on crime rates:

	Before	After
City with new policy	45	38
Comparison city	40	36

(a) Calculate the DiD estimate. (b) Write out the regression that would give the same estimate. (c) What assumption must hold for this to be a causal estimate?

Exercise 5.2: A researcher shows that treated and control groups have parallel trends for 5 years before treatment. Does this guarantee that the parallel trends assumption holds? Explain.

Exercise 5.3: Explain why standard OLS standard errors are inappropriate for DiD. What should be done instead?

Exercise 5.4: In a staggered DiD with two cohorts (treated in 2005 and 2010) and never-treated units, the Goodman-Bacon decomposition shows that 60% of the weight comes from comparing the 2005 cohort to the 2010 cohort. Why might this be problematic?

Exercise 5.5: You're studying the effect of a state-level policy adopted by 15 states at different times. Your TWFE estimate is $\hat{\tau} = 0.05$with $SE = 0.01$. When you use Callaway-Sant'Anna, you get $\hat{\tau} = 0.12$with $SE = 0.03$. What might explain this difference? Which do you trust more?

PART 6: REGRESSION DISCONTINUITY DESIGN

What if we have only one treated unit (a single state, country, or firm)? Standard DiD requires variation across units for inference. For that we need to have idea about**: Regression Discontinuity and Synthetic Control**:

1. **Regression Discontinuity:** Exploiting arbitrary cutoffs for treatment assignment
2. **Synthetic Control:** Constructing counterfactuals from weighted combinations of controls

structure:

1. The core idea of RDD
2. Sharp RDD - setup, identification, estimation
3. Graphical analysis
4. Bandwidth selection
5. Polynomial order
6. Fuzzy RDD (combining with IV)
7. Validity tests and manipulation
8. Examples
9. Extensions (geographic, multi-cutoff, etc.)

Exploiting Assignment Thresholds for Causal Inference

6.1 The Core Idea

Regression Discontinuity Design (RDD) exploits situations where treatment is assigned based on whether a **running variable** crosses a known **cutoff**. Near the cutoff, treatment assignment is "as good as random."

The Intuition

Imagine a scholarship program that awards funding to students who score above 80 on an exam. Consider two students:

- Student A scores 79.5 → No scholarship
- Student B scores 80.5 → Gets scholarship

These students are virtually identical in ability (half a point apart), but one gets treatment and the other doesn't. By comparing outcomes for students just above and just below 80, we can estimate the causal effect of the scholarship.

Why This Works

Near the cutoff:

1. Students cannot precisely control their scores
2. Those just above and just below are comparable on all characteristics
3. The only systematic difference is treatment status
4. It's like a **local randomized experiment**

Note

RDD was introduced by Thistlethwaite & Campbell (1960) in psychology but gained prominence in economics through the work of Angrist & Lavy (1999), Hahn, Todd & Van der Klaauw (2001), and Lee & Lemieux (2010).

6.2 Setup and Notation

The Basic Elements

- y_i: outcome of interest
- x_i: **running variable** (also called forcing variable, assignment variable, or score)
- c: **cutoff** (threshold) value
- w_i: treatment indicator

Treatment Assignment Rule

In the **sharp RDD**, treatment is a deterministic function of the running variable:

$$w_i = \mathbf{1}[x_i \geq c]$$

Everyone above the cutoff is treated; everyone below is not.

Potential Outcomes

As always:

- y_{1i}: outcome if treated
- y_{0i}: outcome if not treated
- $y_i = w_i \cdot y_{1i} + (1 - w_i) \cdot y_{0i}$

The Causal Parameter of Interest

We focus on the treatment effect **at the cutoff**:

$$\tau_{RD} = E[y_{1i} - y_{0i} \mid x_i = c]$$

This is a **local** treatment effect—it applies to units at the cutoff, not the entire population.

6.3 Identification in Sharp RDD

The Key Assumption: **Continuity**

Assumption RDD.1 (Continuity):

$$E[y_{0i} \mid x_i = x] \text{ and } E[y_{1i} \mid x_i = x] \text{ are continuous in } x \text{ at } x = c$$

This says: the expected potential outcomes are smooth functions of the running variable, with no jumps at the cutoff.

What Continuity Means

In the absence of treatment, outcomes would vary smoothly with the running variable. Any **discontinuity** at the cutoff must be due to treatment.

Identification Result

Theorem 6.1: Under continuity, the RDD treatment effect is identified by:

$$\tau_{RD} = \lim_{x \downarrow c} E[y_i \mid x_i = x] - \lim_{x \uparrow c} E[y_i \mid x_i = x]$$

The treatment effect equals the **jump** in the conditional expectation function at the cutoff.

Proof

Step 1: Just above the cutoff, everyone is treated ($w_i = 1$):

$$\lim_{x \downarrow c} E[y_i \mid x_i = x] = \lim_{x \downarrow c} E[y_{1i} \mid x_i = x]$$

Step 2: Just below the cutoff, no one is treated ($w_i = 0$):

$$\lim_{x \uparrow c} E[y_i \mid x_i = x] = \lim_{x \uparrow c} E[y_{0i} \mid x_i = x]$$

Step 3: By continuity of $E[y_{1i} \mid x_i = x]$:

$$\lim_{x \downarrow c} E[y_{1i} \mid x_i = x] = E[y_{1i} \mid x_i = c]$$

Step 4: By continuity of $E[y_{0i} \mid x_i = x]$:

$$\lim_{x \uparrow c} E[y_{0i} \mid x_i = x] = E[y_{0i} \mid x_i = c]$$

Step 5: Therefore:

$$\lim_{x \downarrow c} E[y_i \mid x_i = x] - \lim_{x \uparrow c} E[y_i \mid x_i = x] = E[y_{1i} \mid x_i = c] - E[y_{0i} \mid x_i = c] = \tau_{RD}$$

6.4 Why Continuity is Plausible

The "Local Randomization" Interpretation

Near the cutoff, small differences in $x_i$determine treatment, but these differences are essentially random if:

1. **Individuals cannot precisely manipulate** their running variable
2. **There is some randomness** in the running variable

Example: Test Scores

A student aiming for a score of 80 might get 78, 79, 80, 81, or 82 due to:

- Random variation in test questions
- Having a good or bad day
- Minor differences in preparation

This randomness means students scoring 79.9 vs 80.1 are essentially exchangeable.

When Continuity Might Fail

Manipulation: If individuals can precisely control $x_i$to be just above the cutoff.

Example: A government program awards grants to districts with poverty rates above 20%. If districts can manipulate their reported poverty rates, those just above 20% may be systematically different from those just below.

6.5 Estimation: Graphical Analysis

Step 1: Plot the Raw Data

Before any formal analysis, **visualize** the data:

1. Create a scatter plot of $y_i$against x_i

2. Add a vertical line at the cutoff c
3. Look for a visible jump at the cutoff

Step 2: Binned Scatter Plot

With many observations, raw scatter plots can be hard to read. Instead:

1. Divide the running variable into bins
2. Calculate the mean of $y_i$within each bin
3. Plot bin means against bin midpoints
4. Fit separate curves on each side of the cutoff

Implementation

```
stata
* Create bins
gen bin = floor(x_i / binwidth) * binwidth + binwidth/2

* Calculate means within bins
bysort bin: egen y_mean = mean(y)

* Scatter plot with fitted lines
twoway (scatter y_mean bin if x_i < c, msymbol(circle)) ///
    (scatter y_mean bin if x_i >= c, msymbol(circle)) ///
    (lfit y x_i if x_i < c) ///
    (lfit y x_i if x_i >= c), ///
    xline(c) legend(off) ///
    xtitle("Running Variable") ytitle("Outcome")
```

What to Look For

Good RDD:

- Clear jump at the cutoff
- Smooth trends on either side

- No other discontinuities

Problematic signs:

- No visible jump (null effect or noise)
- Jump occurs away from the cutoff
- Highly non-linear relationships

6.6 Estimation: Local Linear Regression

The Problem with Global Polynomials

One approach: fit a polynomial to the entire data with a discontinuity at c:

$$y_i = \alpha + \tau w_i + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i$$

Problems:

- Results sensitive to polynomial order
- Observations far from cutoff can influence estimate at cutoff
- High-order polynomials can overfit and create spurious jumps

The Solution: Local Linear Regression

Fit **separate linear regressions** on each side of the cutoff, using only observations **near** the cutoff.

The Estimator

For a bandwidth $h > 0$, use observations with $| x_i - c | \leq h$.

On each side of the cutoff, fit:

Below cutoff ($x_i < c$):

$$y_i = \alpha_- + \beta_-(x_i - c) + \varepsilon_i$$

Above cutoff ($x_i \geq c$):

$$y_i = \alpha_+ + \beta_+(x_i - c) + \varepsilon_i$$

The RDD estimate:

$$\hat{\tau}_{RD} = \hat{\alpha}_+ - \hat{\alpha}_-$$

This is the difference in the intercepts, which equals the jump at $x = c$.

Equivalent Single Regression

This is equivalent to running:

$$y_i = \alpha + \tau w_i + \beta_1(x_i - c) + \beta_2 w_i \cdot (x_i - c) + \varepsilon_i$$

on observations with $| x_i - c | \leq h$.

Coefficients:

- $\alpha = \hat{\alpha}_-$: intercept below cutoff
- $\tau = \hat{\alpha}_+ - \hat{\alpha}_-$: **RDD treatment effect**
- $\beta_1 = \hat{\beta}_-$: slope below cutoff
- $\beta_2 = \hat{\beta}_+ - \hat{\beta}_-$: difference in slopes

Centering the Running Variable

Note that we use $(x_i - c)$rather than x_i. This **centering** ensures:

- α_- is the predicted value at $x = c$from below
- α_+ is the predicted value at $x = c$from above
- τ directly gives the discontinuity

6.7 Kernel Weighting

The Idea

Instead of using all observations within bandwidth $h$equally, give **more weight to observations closer to the cutoff**.

Weighted Local Linear Regression

Minimize:

$$\sum_{i=1}^{N} K\left(\frac{x_i - c}{h}\right)[y_i - \alpha - \tau w_i - \beta_1(x_i - c) - \beta_2 w_i(x_i - c)]^2$$

where $K(\cdot)$is a **kernel function**.

Common Kernels

Uniform (Rectangular):

$$K(u) = \frac{1}{2}\mathbf{1}[| u |\leq 1]$$

All observations within bandwidth get equal weight.

Triangular:

$$K(u) = (1-| u |)\mathbf{1}[| u |\leq 1]$$

Observations closer to cutoff get more weight; recommended by Imbens & Kalyanaraman (2012).

Epanechnikov:

$$K(u) = \frac{3}{4}(1 - u^2)\mathbf{1}[| u |\leq 1]$$

Gaussian:

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$$

Uses all data with exponentially declining weights.

Why Triangular is Often Preferred

Triangular kernels are **boundary optimal**: they minimize mean squared error at boundary points (like the cutoff). Since we're estimating at the cutoff specifically, this is appropriate for RDD.

6.8 Bandwidth Selection

The Bias-Variance Tradeoff

Smaller bandwidth h:

- Less bias (comparing more similar units)
- Higher variance (fewer observations)

Larger bandwidth h:

- More bias (comparing less similar units)
- Lower variance (more observations)

Optimal Bandwidth

The goal is to minimize **Mean Squared Error (MSE)** of the estimator:

$$MSE(\hat{\tau}) = Bias(\hat{\tau})^2 + Var(\hat{\tau})$$

Imbens-Kalyanaraman (IK) Bandwidth

Imbens & Kalyanaraman (2012) derived an optimal bandwidth:

$$h_{IK} = C \cdot \left(\frac{\sigma^2(c)}{\sum (x_i - c)^2 f(c) \cdot (m''(c))^2} \right)^{1/5} \cdot N^{-1/5}$$

where:

- $\sigma^2(c)$: conditional variance at cutoff
- $f(c)$: density of running variable at cutoff
- $m''(c)$: second derivative of conditional mean at cutoff
- C: constant depending on kernel

Key insight: Optimal bandwidth shrinks with sample size at rate $N^{-1/5}$.

Calonico-Cattaneo-Titiunik (CCT) Bandwidth

Calonico, Cattaneo & Titiunik (2014) proposed a **bias-corrected** approach:

1. Estimate the bias using local quadratic regression
2. Subtract the bias estimate
3. Construct standard errors that account for bias correction

This allows for valid inference even with MSE-optimal bandwidth.

Implementation

```
stata

* Install rdrobust package

ssc install rdrobust


* Basic RDD estimation with automatic bandwidth

rdrobust y x, c(0)
```

```
* With triangular kernel (default) and CCT bandwidth
rdrobust y x, c(0) kernel(triangular) bwselect(mserd)
```

```
# R
library(rdrobust)

# Basic estimation
rd <- rdrobust(y, x, c = 0)
summary(rd)

# With specific options
rd <- rdrobust(y, x, c = 0, kernel = “triangular”, bwselect = “mserd”)
```

6.9 Inference

Standard Errors

The local linear regression produces standard errors, but several issues require attention:

1. Heteroskedasticity: Use heteroskedasticity-robust standard errors.

2. Clustering: If observations are clustered (e.g., students within schools), cluster standard errors.

3. Bias: Standard errors don’t account for bias from using a finite bandwidth.

Robust Bias-Corrected Inference (CCT)

Calonico, Cattaneo & Titiunik (2014) propose:

1. **Bias correction:** Estimate and subtract the leading bias term
2. **Robust standard errors:** Account for estimation error in bias correction
3. **Valid confidence intervals:** Even with MSE-optimal bandwidth

This is now the **standard approach** in applied work.

Confidence Intervals

The robust 95% confidence interval is:

$$\hat{\tau}_{bc} \pm 1.96 \times SE_{robust}$$

where $\hat{\tau}_{bc}$is the bias-corrected estimate.

Bandwidth Sensitivity

Always show results for multiple bandwidths:

- Half the optimal bandwidth
- The optimal bandwidth
- Twice the optimal bandwidth

Estimates should be stable across reasonable bandwidths.

6.10 Validity Tests

Test 1: Continuity of Covariates

Idea: If units just above and below the cutoff are comparable, pre-determined covariates should be continuous at the cutoff.

Implementation: Run RDD with each covariate as the outcome. None should show a discontinuity.

$$\hat{\tau}_{RD}^{x_k} = \lim_{x \downarrow c} E[x_{ki} \mid x_i = x] - \lim_{x \uparrow c} E[x_{ki} \mid x_i = x] \approx 0$$

Example: For the scholarship RDD, check continuity of:

- Family income
- Parental education
- High school quality
- Race/ethnicity

If these show discontinuities, the design is compromised.

Test 2: Density Test (McCrary Test)

Idea: If individuals can manipulate the running variable, we'll see "bunching" just above the cutoff (more people than expected).

Implementation: Test whether the density of $x_i$is continuous at c.

McCrary (2008) test:

1. Estimate density separately on each side of cutoff using local polynomial
2. Test whether there’s a discontinuity

```
stata
* McCrary test
DCdensity x, breakpoint(c) generate(Xj Yj r0 fhat se_fhat)


# R - using rddensity package
library(rddensity)
dens_test <- rddensity(x, c = 0)
summary(dens_test)
rdplotdensity(dens_test, x)
```

Interpretation:

- No discontinuity in density → supports validity
- Jump up in density above cutoff → suggests manipulation (invalid)

Test 3: Placebo Cutoffs

Idea: There should be no treatment effect at points other than the true cutoff.

Implementation:

1. Estimate RDD at “placebo” cutoffs (e.g., $c - 5$, $c + 5$)
2. These estimates should be approximately zero

Test 4: Placebo Outcomes

Idea: Treatment shouldn’t affect outcomes that it cannot plausibly influence.

Example: A scholarship in 2010 shouldn’t affect test scores from 2005.

6.11 Complete Example: Maimonides’ Rule

Background (Angrist & Lavy, 1999)

A 12th-century Talmudic scholar, Maimonides, established that class size should not exceed 40 students. Israeli public schools follow this rule:

- Enrollment 1-40: 1 class
- Enrollment 41-80: 2 classes
- Enrollment 81-120: 3 classes

This creates discontinuities in class size at enrollment thresholds.

The Design

Running variable: School enrollment (x_i) **Cutoffs:** 41, 81, 121, etc. **Treatment:** Smaller class size (due to splitting) **Outcome:** Student test scores

Expected Class Size Function

$$\text{Expected class size} = \frac{\text{Enrollment}}{\text{int}\left(\frac{\text{Enrollment} - 1}{40}\right) + 1}$$

At enrollment = 40: class size = 40 At enrollment = 41: class size = 20.5 (average of two classes)

Identification

At enrollment = 41, there’s a sharp drop in class size. If students in schools with 40 vs 41 enrollment are otherwise similar, comparing their outcomes identifies the effect of class size.

Results

Angrist & Lavy found:

- Reducing class size by 10 students increases test scores by about 0.2-0.3 standard deviations
- Effects are larger for disadvantaged students

Visual Evidence

The paper shows compelling graphs:

1. Class size drops sharply at multiples of 40
2. Test scores jump up at these same points
3. Pre-determined covariates are smooth at the cutoffs

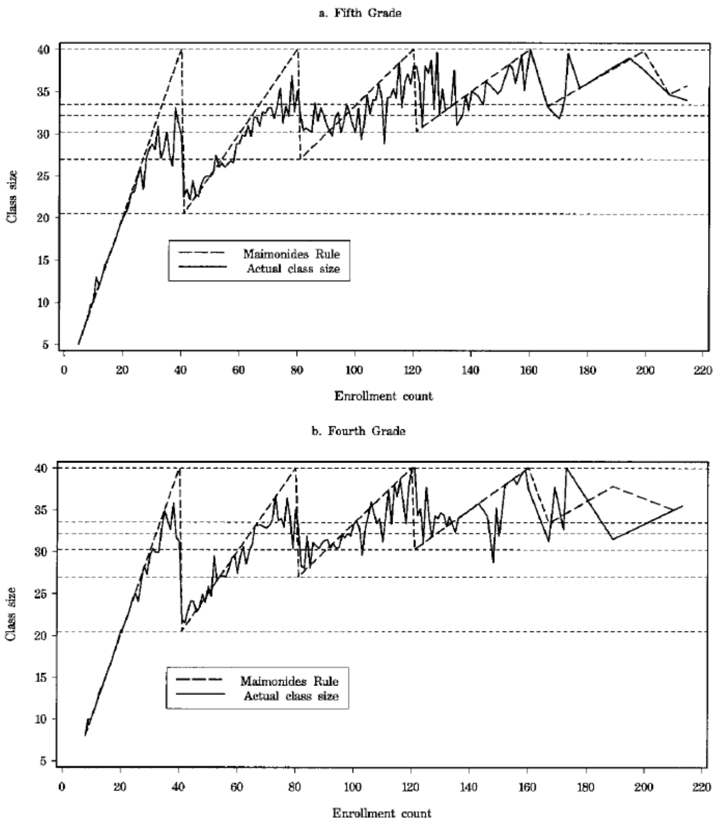

FIGURE I
Class Size in 1991 by Initial Enrollment Count, Actual Average Size and as Predicted by Maimonides' Rule

6.12 Fuzzy RDD

When Assignment Isn't Perfect

In many settings, crossing the cutoff doesn't perfectly determine treatment:

- Scoring above the threshold **increases probability** of treatment
- But some above don't get treated, and some below do get treated

This is **Fuzzy RDD**.

Formal Setup

Sharp RDD:

$$w_i = \mathbf{1}[x_i \geq c]$$

Fuzzy RDD:

$$P(w_i = 1 \mid x_i = x) \text{ has a discontinuity at } x = c$$

The probability of treatment jumps at the cutoff, but from neither 0 to 1 nor 1 to 0.

The Identification Problem

We observe a jump in both:

1. Treatment probability: $\lim_{x\downarrow c} P(w = 1 \mid x) - \lim_{x\uparrow c} P(w = 1 \mid x) = \pi$
2. Outcome: $\lim_{x\downarrow c} E[y \mid x] - \lim_{x\uparrow c} E[y \mid x]$

But the outcome jump reflects both:

- The effect of treatment
- The fact that not everyone is treated/untreated

The Solution: IV/Wald Estimator

Fuzzy RDD is **IV at the cutoff**:

$$\tau_{FRD} = \frac{\lim\limits_{x\downarrow c} E[y_i \mid x_i = x] - \lim\limits_{x\uparrow c} E[y_i \mid x_i = x]}{\lim\limits_{x\downarrow c} E[w_i \mid x_i = x] - \lim\limits_{x\uparrow c} E[w_i \mid x_i = x]}$$
$$= \frac{\text{Jump in outcome at cutoff}}{\text{Jump in treatment probability at cutoff}}$$

This is exactly the Wald estimator from Part 4, applied locally at the cutoff!

Instrument and LATE Interpretation

Instrument: $z_i = \mathbf{1}[x_i \geq c]$(crossing the cutoff)

Compliers: Those who get treated when $x_i \geq c$but wouldn't if $x_i < c$

LATE: The fuzzy RDD estimate is the treatment effect for compliers at the cutoff:

$$\tau_{FRD} = E[y_{1i} - y_{0i} \mid x_i = c, \text{complier}]$$

Estimation

Two-Stage Least Squares:

First stage:

$$w_i = \gamma + \pi \cdot \mathbf{1}[x_i \geq c] + \delta_1(x_i - c) + \delta_2 \cdot \mathbf{1}[x_i \geq c] \cdot (x_i - c) + \eta_i$$

Second stage:

$$y_i = \alpha + \tau \widehat{w}_i + \beta_1(x_i - c) + \beta_2 \cdot \mathbf{1}[x_i \geq c] \cdot (x_i - c) + \varepsilon_i$$

```
stata
* Fuzzy RDD using rdrobust
rdrobust y x, c(0) fuzzy(w)
```

Example: Financial Aid Eligibility

Students with GPA ≥ 3.0 are eligible for financial aid, but:

- Not all eligible students apply
- Some ineligible students get aid through appeals

Sharp: Eligibility jumps from 0 to 1 at GPA = 3.0 **Fuzzy:** Actual aid receipt jumps from, say, 20% to 70% at GPA = 3.0

Fuzzy RDD estimates the effect of aid for students who receive it because they crossed the threshold.

6.13 Practical Issues

Issue 1: Choosing the Polynomial Order

Local linear (order 1): Generally preferred

- Robust to boundary bias
- Well-understood theoretical properties
- Recommended by Gelman & Imbens (2019)

Local quadratic (order 2): Sometimes useful

- Reduces bias when true function is highly curved
- Higher variance
- Use for bias correction (CCT approach)

Higher-order polynomials: Avoid

- Gelman & Imbens (2019) show they can produce misleading results
- Sensitive to observations far from cutoff
- Can create spurious discontinuities

Issue 2: Discrete Running Variables

When x_i takes only a few values (e.g., age in years, test scores in integers):

Problems:

- Can't observe outcomes arbitrarily close to cutoff
- Clustering of observations at specific values
- Bias from discretization

Solutions:

- Use all unique values as "mass points"
- Cluster standard errors by running variable value
- Lee & Card (2008) provide methods for discrete x_i

Issue 3: Multiple Cutoffs

Sometimes the same running variable has multiple cutoffs (e.g., different benefit levels at different income thresholds).

Options:

1. Analyze each cutoff separately
2. Pool cutoffs by normalizing running variable: $\tilde{x}_i = x_i - c_j$where $c_j$is the relevant cutoff
3. Cattaneo et al. (2016) provide methods for multi-cutoff RDD

Issue 4: Covariates

Should you include covariates in RDD?

If the design is valid, covariates are not necessary for identification (treatment is locally random). However, they can:

- Improve precision
- Serve as a robustness check (results shouldn't change much)

Warning: If results are highly sensitive to covariate inclusion, the design may be compromised.

6.14 Geographic and Spatial RDD

The Idea

Treatment is assigned based on geographic location: crossing a boundary determines treatment.

Examples:

- Minimum wage differences across state borders
- School district boundaries
- Electoral district boundaries
- International borders

The Design

Running variable: Distance to boundary (negative = control side, positive = treatment side)

Cutoff: $c = 0$(the boundary itself)

Key assumption: Locations just across the boundary from each other are comparable.

Challenges

1. Multi-dimensional running variable:

- Location is two-dimensional (latitude, longitude)
- Need to reduce to distance from boundary

2. Sorting:

- People may choose to live on one side of boundary
- More problematic than manipulation in standard RDD

3. Boundary effects:

- Other things may change at the boundary (culture, amenities)
- Exclusion restriction harder to defend

Example: Dell (2010)

Question: Do colonial-era labor institutions affect development today?

Setting: Peru/Bolivia, where the "Mita" system (forced labor in mines) affected some areas but not others.

Running variable: Distance to Mita boundary

Finding: Areas subject to Mita have substantially lower household consumption today, 200+ years later.

6.15 Regression Kink Design

The Idea

Sometimes treatment intensity changes at a threshold, but treatment status doesn't. The **slope** of the relationship between running variable and treatment changes—a "kink" rather than a jump.

Setup

Example: Income tax rates increase from 25% to 35% at $50,000 income.

- Everyone pays taxes (no discontinuity in treatment status)
- But the marginal tax rate changes (discontinuity in slope)

Identification

If the relationship between running variable and outcome has a kink at the threshold, and this kink is caused by the policy kink, we can estimate causal effects.

$$\tau_{RK} = \frac{\lim_{x \downarrow c} \frac{dE[y \mid x]}{dx} - \lim_{x \uparrow c} \frac{dE[y \mid x]}{dx}}{\lim_{x \downarrow c} \frac{dE[w \mid x]}{dx} - \lim_{x \uparrow c} \frac{dE[w \mid x]}{dx}}$$

The treatment effect is the ratio of the kink in the outcome to the kink in treatment.

Implementation

stata

```
* Using rdrobust with deriv option

rdrobust y x, c(0) deriv(1)
```

6.16 RDD vs. Other Methods

Comparison with Randomized Experiments

Aspect	**RCT**	**RDD**
Internal validity	High (by design)	High (if valid)
External validity	Depends on sample	Limited (local to cutoff)
Required sample	Moderate	Large (need density at cutoff)
Feasibility	Often low	Higher (uses existing rules)

Comparison with DiD

Aspect	DiD	RDD
Variation	Across time and groups	Across running variable
Key assumption	Parallel trends	Continuity
Testability	Pre-trends	Manipulation, covariate balance
Locality	Group-level	Cutoff-local

Comparison with IV

Aspect	IV	RDD
Instrument	External variable	Cutoff crossing
Key assumption	Exclusion restriction	Continuity
Compliers	Those affected by instrument	Those at cutoff
Locality	Population of compliers	Neighborhood of cutoff

When to Use RDD

RDD is ideal when:

1. Treatment assignment follows a clear cutoff rule
2. Individuals cannot precisely manipulate the running variable
3. You care about effects near the cutoff (or are willing to extrapolate)
4. You have sufficient data near the cutoff

6.17 Limitations and Critiques

Limitation 1: External Validity

RDD estimates the effect **at the cutoff**. This may not generalize to:

- Individuals far from the cutoff
- Different cutoff values
- Different populations

Partial solution: Combine RDD estimates from multiple cutoffs or conduct extrapolation analysis.

Limitation 2: Bandwidth Sensitivity

Results can be sensitive to bandwidth choice. Always:

- Report results for multiple bandwidths
- Use data-driven bandwidth selection
- Be suspicious if results only appear for specific bandwidths

Limitation 3: Specification Sensitivity

Results can depend on:

- Polynomial order
- Kernel choice
- Covariate inclusion

Robust results should be stable across reasonable specifications.

Limitation 4: Manipulation

If individuals can manipulate the running variable, RDD fails. Even partial manipulation by a subset can bias results.

Limitation 5: Precision

RDD is "data hungry." We need many observations near the cutoff for precise estimates. This is especially problematic with:

- Discrete running variables
- Running variables with low density at cutoff
- Small overall samples

6.18 Reporting Standards

What to Include in an RDD Paper

Following Lee & Lemieux (2010) and current best practices:

1. Graphical Evidence

- Scatter plot or binned means of outcome vs. running variable
- Separate fitted curves on each side
- Clear visual of discontinuity

2. Main Estimates

- Local linear regression with optimal bandwidth
- Bias-corrected estimates with robust standard errors
- Bandwidth used

3. Robustness

- Results with different bandwidths (0.5h, h, 1.5h, 2h)
- Results with different polynomial orders
- With and without covariates

4. Validity Tests

- Density test (McCrary or rddensity)
- Covariate balance tests
- Placebo cutoffs (if appropriate)

5. Sensitivity Analysis

- What happens if we move the cutoff slightly?
- Sensitivity to outliers near cutoff

6.19 Summary of Part 6

Key Assumptions

Sharp RDD:

$$E[y_{0i} \mid x_i = x] \text{ and } E[y_{1i} \mid x_i = x] \text{ continuous at } x = c$$

Fuzzy RDD (additional):

$$P(w_i = 1 \mid x_i = x) \text{ has a discontinuity at } x = c$$

Plus monotonicity (no defiers).

Key Formulas

Sharp RDD:

$$\tau_{RD} = \lim_{x \downarrow c} E[y_i \mid x_i = x] - \lim_{x \uparrow c} E[y_i \mid x_i = x]$$

Fuzzy RDD:

$$\tau_{FRD} = \frac{\lim_{x \downarrow c} E[y_i \mid x_i = x] - \lim_{x \uparrow c} E[y_i \mid x_i = x]}{\lim_{x \downarrow c} E[w_i \mid x_i = x] - \lim_{x \uparrow c} E[w_i \mid x_i = x]}$$

Local Linear Regression:

$$y_i = \alpha + \tau w_i + \beta_1 (x_i - c) + \beta_2 w_i (x_i - c) + \varepsilon_i$$

Key Insights

1. **RDD exploits discontinuities** in treatment assignment at known cutoffs.
2. **Identification is local:** Effects apply to units at the cutoff.
3. **Continuity is key:** If potential outcomes are smooth, any jump in observed outcomes is causal.
4. **Manipulation is the main threat:** Test with density analysis.
5. **Use local linear regression** with data-driven bandwidths and bias-corrected inference.
6. **Fuzzy RDD is IV:** When crossing the cutoff doesn't perfectly determine treatment.

Practical Checklist

Step	Action
1. Understand the rule	What determines treatment assignment?

Step	Action
2. Visualize	Plot outcome vs. running variable
3. Test validity	Density test, covariate balance
4. Estimate	Local linear with CCT bandwidth
5. Robustness	Multiple bandwidths, specifications
6. Report	Graphs, estimates, validity tests

Exercises for Part 6

Exercise 6.1: A university admits students with entrance exam scores ≥ 500. You want to estimate the effect of attending this university on earnings.

(a) What is the running variable? The cutoff? (b) Write the continuity assumption in this context. (c) What validity tests would you conduct?

Exercise 6.2: In the setup of Exercise 6.1, suppose 90% of students scoring above 500 attend, and 30% of students scoring below 500 also attend (through appeals).

(a) Is this sharp or fuzzy RDD? (b) Write the fuzzy RDD estimator. (c) What is the LATE interpretation?

Exercise 6.3: A researcher estimates an RDD with bandwidths of 5, 10, and 20. The estimates are:

- $h = 5$: $\hat{\tau} = 2.1$(SE = 1.8)
- $h = 10$: $\hat{\tau} = 3.5$(SE = 0.9)
- $h = 20$: $\hat{\tau} = 5.2$(SE = 0.5)

Comment on these results. What might explain the pattern?

Exercise 6.4: You conduct a McCrary density test and find a statistically significant jump in the density just above the cutoff. What does this mean? Can you still use RDD?

Exercise 6.5: Explain why Gelman & Imbens (2019) recommend against high-order polynomial specifications in RDD.

PART 7: SYNTHETIC CONTROL METHOD

Constructing Counterfactuals for Case Studies

7.1 Motivation

The Problem with Traditional Methods

Consider estimating the effect of a policy implemented in a **single unit**:

- California's tobacco control program (Proposition 99)
- German reunification on West German GDP
- Brexit on UK economic outcomes
- A merger on a single firm's performance

Traditional methods struggle here:

- ***DiD*:** Requires choosing control units, but which ones? Any single control may be a poor match.
- ***RDD*:** Requires a running variable with a cutoff—often doesn't exist for policy interventions.
- ***Randomization inference*:** With $N = 1$treated unit, classical inference fails.

The Synthetic Control Idea

Instead of choosing a single control unit (or arbitrary group), **construct** a synthetic version of the treated unit as a weighted average of untreated units.

The synthetic control is designed to match the treated unit's characteristics **before treatment**, providing a credible counterfactual for what would have happened without treatment.

Key Innovation

Abadie & Gardeazabal (2003) and Abadie, Diamond & Hainmueller (2010, 2015) formalized this approach:

1. **Data-driven control selection:** Weights are chosen to minimize pre-treatment differences
2. **Transparent:** The weights show exactly how the counterfactual is constructed
3. **Valid inference:** Placebo tests provide a framework for statistical inference

7.2 Setup and Notation

Setting

We observe $J + 1$units over Ttime periods:

- Unit 1: the **treated unit** (receives treatment at time $T_0 + 1$)
- Units $2, \dots, J + 1$: the **donor pool** (potential controls, never treated)

Time periods:

- $t = 1, \dots, T_0$: **pre-treatment periods**
- $t = T_0 + 1, \dots, T$: **post-treatment periods**

Potential Outcomes

For unit $j$at time t:

- y_{jt}^N: potential outcome **without** treatment (N for “no intervention”)
- y_{jt}^I: potential outcome **with** treatment (I for “intervention”)

Treatment Indicator

$$D_{jt} = \begin{cases} 1 & \text{if } j = 1 \text{ and } t > T_0 \\ 0 & \text{otherwise} \end{cases}$$

Only unit 1 is treated, and only after period T_0.

Observed Outcome

$$y_{jt} = y_{jt}^N + (y_{jt}^I - y_{jt}^N) \cdot D_{jt}$$

For control units: $y_{jt} = y_{jt}^N$(always observe untreated outcome) For treated unit before T_0: $y_{1t} = y_{1t}^N$For treated unit after T_0: $y_{1t} = y_{1t}^I$

The Target Parameter

We want to estimate the treatment effect on the treated unit:

$$\tau_{1t} = y_{1t}^{I} - y_{1t}^{N} \text{for } t > T_0$$

We observe $y_{1t}^{I} = y_{1t}$, but we need to estimate y_{1t}^{N}—what would have happened to unit 1 without treatment.

7.3 The Synthetic Control Estimator

The Core Idea

Construct a **synthetic control** as a weighted average of donor units:

$$\hat{y}_{1t}^{N} = \sum_{j=2}^{J+1} w_j \, y_{jt}$$

where $w_j$are weights satisfying:

- $w_j \geq 0$for all j(non-negative weights)
- $\sum_{j=2}^{J+1} w_j = 1$(weights sum to one)

The synthetic control is a **convex combination** of donor units.

Choosing Weights

Weights are chosen so that the synthetic control **matches the treated unit in the pre-treatment period**.

Let $\mathbf{X}_1$be a $(k \times 1)$vector of pre-treatment characteristics for the treated unit:

- Pre-treatment outcomes: $y_{1,1}, y_{1,2}, \ldots, y_{1,T_0}$(or averages)
- Predictor variables: $z_{1,1}, z_{1,2}, \ldots$(covariates that predict the outcome)

Let $\mathbf{X}_0$be a $(k \times J)$matrix of the same characteristics for donor units.

The Optimization Problem

Choose weights $\mathbf{w} = (w_2, \ldots, w_{J+1})'$to minimize:

$$\| \mathbf{X}_1 - \mathbf{X}_0\mathbf{w} \|_V = \sqrt{(\mathbf{X}_1 - \mathbf{X}_0\mathbf{w})'\mathbf{V}(\mathbf{X}_1 - \mathbf{X}_0\mathbf{w})}$$

subject to:

$$w_j \geq 0 \text{ for all } j, \sum_{j=2}^{J+1} w_j = 1$$

where $\mathbf{V}$ is a $(k \times k)$positive semi-definite matrix of weights on predictors.

What is $\mathbf{V}$?

The matrix $\mathbf{V}$ determines the **relative importance** of different predictors in constructing synthetic control.

Options for choosing $\mathbf{V}$:

1. **Identity matrix:** All predictors weighted equally
2. **Inverse variance:** Weight by $1/Var(x_k)$to standardize
3. **Data-driven:** Choose $\mathbf{V}$ to minimize pre-treatment prediction error (nested optimization)

The data-driven approach is most common:

$$\mathbf{V}^* = \arg \min_{\mathbf{V}} \sum_{t=1}^{T_0} \left(y_{1t} - \sum_{j=2}^{J+1} w_j^* (\mathbf{V}) y_{jt} \right)^2$$

where $\mathbf{w}^*(\mathbf{V})$solves the inner optimization given $\mathbf{V}$.

The Treatment Effect Estimate

Once we have optimal weights $\mathbf{w}^*$, the estimated treatment effect is:

$$\hat{\tau}_{1t} = y_{1t} - \sum_{j=2}^{J+1} w_j^* y_{jt} = y_{1t} - \hat{y}_{1t}^N \text{for } t > T_0$$

This is the **gap** between the treated unit and its synthetic control.

7.4 Identification

When Does Synthetic Control Work?

The method relies on several key conditions:

Condition 1: Good Pre-Treatment Fit

The synthetic control must closely match the treated unit in the pre-treatment period:

$$\sum_{j=2}^{J+1} w_j^* y_{jt} \approx y_{1t} \text{for } t = 1, \dots, T_0$$

If pre-treatment fit is poor, the counterfactual is unreliable.

Condition 2: Convex Hull

The treated unit must lie within the **convex hull** of donor units in terms of pre-treatment characteristics.

Why? Weights are constrained to be non-negative and sum to one. If the treated unit is “outside” the donor pool, no convex combination can match it.

Example: If California has the highest pre-treatment smoking rate among all states, no weighted average of other states can match it.

Condition 3: No Interference

Treatment of unit 1 should not affect outcomes of donor units (SUTVA).

Potential violation: If California's tobacco program reduces tobacco company revenues, this might affect tobacco advertising in other states.

Condition 4: No Anticipation

The treated unit doesn't change behavior before treatment in anticipation.

Formal Identification Result

Theorem 7.1 (Abadie, Diamond & Hainmueller, 2010):

Suppose outcomes follow a factor model:

$$y_{jt}^{N} = \delta_t + \boldsymbol{\theta}_t' \mathbf{z}_j + \boldsymbol{\lambda}_t' \boldsymbol{\mu}_j + \varepsilon_{jt}$$

where:

- δ_t: common time effects
- $\mathbf{z}_j$: observed covariates
- $\boldsymbol{\mu}_j$: unobserved unit factors
- $\boldsymbol{\lambda}_t$: time-varying factor loadings
- ε_{jt}: idiosyncratic shocks

If weights $\mathbf{w}^*$satisfy:

$$\sum_{j=2}^{J+1} w_j^* \, y_{jt} = y_{1t} \text{ for } t = 1, \dots, T_0$$
$$\sum_{j=2}^{J+1} w_j^* \, \mathbf{z}_j = \mathbf{z}_1$$

Then (under regularity conditions) the bias of $\hat{\tau}_{1t}$converges to zero as $T_0 \to \infty$.

Intuition

If the synthetic control matches the treated unit on:

1. Pre-treatment outcomes
2. Observed predictors

Then it also implicitly matches on unobserved factors $\boldsymbol{\mu}_1$, because these factors affect pre-treatment outcomes.

7.5 Inference: Placebo Tests

The Challenge

With only one treated unit, classical inference is impossible:

- No sampling variation across treated units
- Standard errors from regression are meaningless

The Solution: Permutation Inference

Idea: Apply the synthetic control method to units that were **not actually treated** (placebo treatments). If the effect for the truly treated unit is unusually large compared to placebo effects, we have evidence of a real effect.

In-Space Placebo Test

1. For each donor unit $j = 2, \ldots, J + 1$:
    - Pretend unit $j$was treated at time T_0
    - Construct a synthetic control for unit $j$using remaining donors
    - Estimate "placebo effect" $\hat{\tau}_{jt}^{placebo}$
2. Compare the true effect $\hat{\tau}_{1t}$to the distribution of placebo effects
3. If $\hat{\tau}_{1t}$is in the tail of the placebo distribution, the effect is "significant"

Visual Representation

Plot the gap (treated minus synthetic) for:

- The actual treated unit (bold line)
- Each placebo unit (gray lines)

If the treated unit's gap is much larger than all placebo gaps, the effect is credible.

The Ratio of Post/Pre MSPE

To account for pre-treatment fit quality, compute:

$$\text{Ratio}_j = \frac{\text{MSPE}_{j,post}}{\text{MSPE}_{j,pre}}$$

where:

$$\text{MSPE}_{j,pre} = \frac{1}{T_0}\sum_{t=1}^{T_0}(y_{jt} - \hat{y}_{jt}^{synth})^2$$
$$\text{MSPE}_{j,post} = \frac{1}{T - T_0}\sum_{t=T_0+1}^{T}(y_{jt} - \hat{y}_{jt}^{synth})^2$$

Interpretation: A high ratio means the gap increased substantially after treatment relative to pre-treatment fit.

P-Value Calculation

$$p\text{-value} = \frac{\#\{j\text{: Ratio}_j \geq \text{Ratio}_1\}}{J + 1}$$

The p-value is the fraction of units (including the treated) with a ratio at least as large as the treated unit's ratio.

In-Time Placebo Test

Apply treatment at a **fake time** before the actual treatment:

1. Pretend treatment occurred at $T_0' < T_0$
2. Estimate synthetic control using data up to T_0'
3. Check for a "gap" between $T_0'$and T_0

If there's a gap before the actual treatment, this suggests:

- Poor pre-treatment fit
- The synthetic control doesn't capture relevant trends
- Something other than treatment caused the divergence

7.6 Complete Example: California's Proposition 99

Background

In 1988, California passed Proposition 99, a comprehensive tobacco control program including:

- 25-cent tax increase on cigarettes
- Anti-smoking media campaigns
- Tobacco education programs

Question: Did Proposition 99 reduce cigarette consumption in California?

Data

- **Treated unit:** California
- **Donor pool:** 38 other US states (excluding states with similar programs)
- **Outcome:** Per capita cigarette sales (packs per person)
- **Pre-treatment:** 1970-1988
- **Post-treatment:** 1989-2000

Predictors Used

- Average cigarette sales: 1975, 1980, 1988
- Log GDP per capita
- Percent age 15-24
- Retail price of cigarettes
- Beer consumption per capita

Results: Synthetic California

The optimal weights produce a synthetic California that is primarily a combination of:

State	Weight
Colorado	0.164
Connecticut	0.069
Montana	0.199
Nevada	0.234
New Mexico	0.055
Utah	0.279
All others	0.000

Pre-Treatment Fit

Variable	California	Synthetic California
Cigarette sales 1975	127.1	127.0
Cigarette sales 1980	120.2	120.4
Cigarette sales 1988	90.1	91.4
Log GDP per capita	10.08	9.89
Percent age 15-24	17.4%	17.4%
Retail price	89.4	89.4
Beer consumption	24.3	24.1

The synthetic California closely matches actual California on all predictors.

Estimated Effect

After 1988:

- California’s cigarette consumption declined faster than synthetic California
- By 2000, the gap was approximately 26 packs per capita

- This represents about a 25% reduction relative to synthetic California

Placebo Tests

When the method is applied to each control state:

- California shows the largest gap
- The post/pre MSPE ratio for California exceeds all but one control state
- Implied p-value: approximately 0.026 (2/38)

Interpretation

Proposition 99 caused a substantial and statistically significant reduction in cigarette consumption in California—approximately 25% by 2000.

7.7 Implementation

Stata

```
stata
* Install synth package
ssc install synth

* Basic syntax
synth outcome predictor1 predictor2 ... predictorK, ///
    trunit(treated_unit_id) trperiod(treatment_period) ///
    counit(donor_unit_ids) fig

* Example: California tobacco study
use smoking.dta, clear

synth cigsale beer(1984(1)1988) lnincome retprice age15to24 ///
    cigsale(1975) cigsale(1980) cigsale(1988), ///
    trunit(3) trperiod(1989) fig
```

```
* Save gaps for placebo analysis
matrix gaps = e(Y_treated) - e(Y_synthetic)
```

R

```
# Install Synth package
install.packages(“Synth”)
library(Synth)

# Prepare data
dataprep_out <- dataprep(
  foo = smoking_data,
  predictors = c(“beer”, “lnincome”, “retprice”, “age15to24”),
  predictors.op = “mean”,
  time.predictors.prior = 1980:1988,
  special.predictors = list(
    list(“cigsale”, 1975, “mean”),
    list(“cigsale”, 1980, “mean”),
    list(“cigsale”, 1988, “mean”)
  ),
  dependent = “cigsale”,
  unit.variable = “state_id”,
  unit.names.variable = “state”,
  time.variable = “year”,
  treatment.identifier = 3,  # California
  controls.identifier = c(1:2, 4:39),
  time.optimize.ssr = 1970:1988,
  time.plot = 1970:2000
)

# Run synthetic control
synth_out <- synth(data.prep.obj = dataprep_out)

# View results
synth.tables <- synth.tab(dataprep.res = dataprep_out,
                          synth.res = synth_out)
print(synth.tables)

# Plot
path.plot(synth.res = synth_out, dataprep.res = dataprep_out)
gaps.plot(synth.res = synth_out, dataprep.res = dataprep_out)
```

R: Tidysynth Package (More User-Friendly)

```
library(tidysynth)

smoking_synth <- smoking_data %>%
  synthetic_control(
    outcome = cigsale,
    unit = state,
    time = year,
    i_unit = “California”,
    i_time = 1988,
    generate_placebos = TRUE
  ) %>%
  generate_predictor(
    time_window = 1980:1988,
    beer = mean(beer),
    lnincome = mean(lnincome),
    retprice = mean(retprice),
    age15to24 = mean(age15to24)
  ) %>%
  generate_predictor(time_window = 1975, cigsale_1975 = cigsale) %>%
  generate_predictor(time_window = 1980, cigsale_1980 = cigsale) %>%
  generate_predictor(time_window = 1988, cigsale_1988 = cigsale) %>%
  generate_weights(optimization_window = 1970:1988) %>%
  generate_control()

# Plot with placebos
smoking_synth %>% plot_placebos()

# Get p-value
smoking_synth %>% grab_significance()
```

7.8 Practical Considerations

Consideration 1: Donor Pool Selection

Include units that:

- Are similar to the treated unit
- Were not affected by the treatment (no spillovers)
- Did not receive similar treatments

Exclude units that:

- Received similar policies
- Are clearly incomparable
- Have missing data for key periods

Note: Don't cherry-pick donors to get desired results. Pre-specify the donor pool before seeing results.

Consideration 2: Predictor Selection

Good predictors:

- Pre-treatment outcomes (most important!)
- Variables that predict the outcome
- Variables that differ between treated and potential controls

Rules of thumb:

- Include several pre-treatment outcome values (not just the average)
- Don't include too many predictors (overfitting risk)
- Focus on economic fundamentals

Consideration 3: Pre-Treatment Fit

Critical: If pre-treatment fit is poor, the synthetic control is unreliable.

Diagnostics:

- Compare predictor values (Table)
- Plot pre-treatment outcomes (Graph)
- Report pre-treatment RMSPE

What counts as "good" fit?

- No formal threshold
- Visual inspection: lines should be close
- Pre-treatment RMSPE should be small relative to outcome scale

Consideration 4: Sparsity of Weights

Synthetic control often produces **sparse** weights—only a few donors get positive weight.

This is a feature, not a bug:

- Interpretable: “California is like 28% Utah + 23% Nevada + ...”
- Avoids extrapolation beyond the data
- But: many donors with zero weight reduces effective sample for inference

Consideration 5: Interpolation Bias

If the treated unit is extreme on some dimension, the synthetic control may interpolate poorly.

Example: If California has the highest GDP per capita, a weighted average of other states will underestimate California’s counterfactual GDP.

Solution: Check that treated unit is not at the boundary of the covariate distribution.

7.9 Extensions

Extension 1: Multiple Treated Units

When several units receive treatment:

Option A: Apply synthetic control separately to each treated unit, then average effects.

Option B: Use **generalized synthetic control** (Xu 2017), which combines synthetic control with interactive fixed effects.

Extension 2: Augmented Synthetic Control (ASCM)

Ben-Michael, Feller & Rothstein (2021) propose combining synthetic control with outcome regression:

$$\hat{\tau}_{1t}^{aug} = \left(y_{1t} - \sum_j w_j\, y_{jt}\right) - \left(\hat{m}(\mathbf{X}_1) - \sum_j w_j\, \hat{m}(\mathbf{X}_j)\right)$$

where $\hat{m}(\cdot)$is an outcome model estimated on control units.

Benefits:

- Reduces bias when synthetic control fit is imperfect
- Provides valid confidence intervals
- Robust to either good SC fit or good outcome model

Implementation in R:

```
library(augsynth)
# Augmented synthetic control
ascm <- augsynth(
  cigsale ~ treated,
  unit = state,
  time = year,
  data = smoking_data,
  progfunc = "Ridge",  # outcome model
  scm = TRUE           # include synthetic control
)

summary(ascm)
plot(ascm)
```

Extension 3: Penalized Synthetic Control

Abadie & L'Hour (2021) propose adding a penalty for pairwise distance between treated and donor units:

$$\min_{\mathbf{w}} \; \| \mathbf{X}_1 - \mathbf{X}_0 \mathbf{w} \|^2 + \lambda \sum_{j=2}^{J+1} w_j \| \mathbf{X}_1 - \mathbf{X}_j \|^2$$

This penalizes placing weight on dissimilar donors, improving extrapolation properties.

Extension 4: Matrix Completion Methods

Athey et al. (2021) propose viewing the problem as **matrix completion**:

- The outcome matrix has missing entries (treated unit, post-treatment)
- Use matrix factorization techniques to impute missing values
- Provides valid confidence intervals

Extension 5: Synthetic Difference-in-Differences

Arkhangelsky et al. (2021) combine synthetic control with difference-in-differences:

$$\hat{\tau}^{SDID} = \sum_{t>T_0} \lambda_t \left[\left(y_{1t} - \sum\nolimits_j w_j \, y_{jt}\right) - \left(y_{1,pre} - \sum\nolimits_j w_j \, y_{j,pre}\right)\right]$$

where weights are chosen for both units (w_j) and time periods (λ_t).

Benefits:

- Combines strengths of DiD and SC
- Valid inference with many treated units
- Robust to some violations of parallel trends

Implementation in R:

```
library(synthdid)
# Synthetic DiD
setup <- panel.matrices(smoking_data)
sdid <- synthdid_estimate(setup$Y, setup$N0, setup$T0)
summary(sdid)
plot(sdid)
```

7.10 Comparison with Other Methods

Synthetic Control vs. DiD

Aspect	DiD	Synthetic Control
Control selection	Researcher chooses	Data-driven weights
Number of treated	Any	Typically few/one
Assumption	Parallel trends	Factor model / good fit
Transparency	Less (implicit equal weights)	More (explicit weights)
Pre-trends test	Yes	Pre-treatment fit

Synthetic Control vs. Matching

Aspect	Matching	Synthetic Control
Unit of analysis	Individual	Aggregate
Number matched	Many	One synthetic unit
Weights	Binary (0/1) or limited	Continuous [0,1]
Time dimension	Often cross-sectional	Panel required

When to Use Synthetic Control

Ideal settings:

- Small number of treated units (1-10)
- Aggregate panel data (states, countries, firms)
- Long pre-treatment period
- Clear treatment date
- Good potential controls available

Less ideal:

- Many treated units (use DiD or SDID)
- Short pre-treatment period
- Outcome is very noisy
- Treated unit is extreme/unusual

7.11 Limitations and Critiques

Limitation 1: Pre-Treatment Fit

If the synthetic control cannot match the treated unit pre-treatment, the counterfactual is unreliable.

Problem: There’s no formal test for “good enough” fit. Judgment is required.

Limitation 2: Extrapolation

Synthetic control stays within the convex hull of donors. But this may not be the right counterfactual if the treated unit is unusual.

Limitation 3: Finite-Sample Inference

Placebo tests have limited power with few donor units:

- With 39 donors, minimum p-value is 1/40 = 0.025
- Cannot detect effects significant at 1% level

Limitation 4: Model Dependence

Results can be sensitive to:

- Choice of predictors
- Choice of $\mathbf{V}$matrix
- Donor pool composition

Best practice: Report sensitivity to these choices.

Limitation 5: No Negative Weights

Constraining weights to be non-negative prevents extrapolation but can limit fit.

Example: If treated unit has the highest value of some predictor, a non-negative weighted average of donors must underestimate it.

The Ferman & Pinto Critique (2021)

When treatment is correlated with unobserved factors:

- Synthetic control may not adequately control for these factors
- Bias can persist even with perfect pre-treatment fit
- The factor model assumptions may not hold

7.12 Second Example: German Reunification

Background

In 1990, East and West Germany reunified after the fall of the Berlin Wall.

Question: What was the economic effect of reunification on West Germany?

Challenge

This seems difficult to study because:

- Only one "treated" unit (West Germany)
- Reunification was a unique historical event
- No obvious control country

Synthetic Control Approach (Abadie, Diamond & Hainmueller, 2015)

Treated unit: West Germany **Donor pool:** OECD countries not affected by reunification **Outcome:** GDP per capita **Pre-treatment:** 1960-1990 **Post-treatment:** 1990-2003

Synthetic West Germany

The synthetic control is constructed from:

Country	Weight
Austria	0.42
United States	0.22
Japan	0.16
Switzerland	0.11
Netherlands	0.09
Others	0.00

Results

- Pre-1990: Synthetic Germany tracks actual West Germany very closely
- Post-1990: West Germany's GDP falls below synthetic Germany
- By 2003: Gap of approximately $1,500 per capita (about 6%)

Interpretation: Reunification reduced West German GDP per capita by approximately 6%—a substantial negative effect.

Placebo Tests

Applying the method to other OECD countries:

- West Germany shows one of the largest post-1990 gaps
- Implied p-value: approximately 0.07

Economic Interpretation

The costs came from:

- Massive fiscal transfers to East Germany
- Integration costs and economic disruption
- Currency union at unfavorable rates

7.13 Reporting Standards

What to Include in a Synthetic Control Paper

1. Donor Pool Justification

- Which units were included/excluded and why
- Pre-specified or post-hoc?

2. Predictor Selection

- What predictors were used
- Justification for choices
- Sensitivity to alternatives

3. Pre-Treatment Fit

- Table comparing treated and synthetic on predictors
- Graph of pre-treatment outcomes
- Pre-treatment RMSPE

4. Weights

- Table showing donor weights
- Which donors contribute most

5. Results

- Graph of treated vs. synthetic over full period
- Gap plot (treated minus synthetic)
- Effect estimates (point and by year)

6. Placebo Tests

- In-space placebos: all donors
- Graph showing all gaps
- P-value calculation
- Ratio of post/pre MSPE

7. Robustness

- Different predictor sets
- Different donor pools
- Different time windows

- Leave-one-out (dropping each major donor)

7.14 Summary of Part 7

Key Concepts

Synthetic Control: A weighted average of untreated units designed to match the treated unit pre-treatment.

Identification: Based on matching pre-treatment outcomes and predictors, which implicitly matches unobserved factors.

Inference: Placebo tests comparing treated unit's gap to placebo gaps.

Key Formulas

Synthetic Control:

$$\hat{y}_{1t}^{N} = \sum_{j=2}^{J+1} w_j^* \, y_{jt}$$

Treatment Effect:

$$\hat{\tau}_{1t} = y_{1t} - \sum_{j=2}^{J+1} w_j^* \, y_{jt}$$

Weight Selection:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \; \| \mathbf{X}_1 - \mathbf{X}_0\mathbf{w} \|_V$$

subject to $w_j \geq 0, \sum w_j = 1$

P-value (Permutation):

$$p = \frac{\#\{j\colon \text{Ratio}_j \geq \text{Ratio}_1\}}{J+1}$$

Key Insights

1. **Data-driven control:** Weights are chosen to match treated unit, not arbitrarily.
2. **Transparency:** The composition of synthetic control is explicit.
3. **Local identification:** Works by matching on pre-treatment outcomes, which implicitly matches unobserved factors.
4. **Placebo inference:** Permutation tests provide valid p-values.
5. **Pre-treatment fit is crucial:** Poor fit = unreliable counterfactual.

Practical Checklist

Step	**Action**
1. Define donor pool	Similar, unaffected units
2. Choose predictors	Pre-treatment outcomes + covariates
3. Estimate weights	Using synth/tidysynth/augsynth
4. Check pre-treatment fit	Table + graph
5. Estimate effects	Gap between treated and synthetic
6. Conduct placebo tests	In-space and in-time
7. Robustness checks	Alternative specifications

Exercises for Part 7

Exercise 7.1: Explain why the synthetic control method requires non-negative weights that sum to one. What would happen if we allowed negative weights?

Exercise 7.2: A researcher applies synthetic control to study the effect of a policy in State A. The synthetic control assigns weight 0.95 to State B and 0.05 to State C. All other states have zero weight. Is this problematic? Why or why not?

Exercise 7.3: In a synthetic control study with 20 donor states, what is the minimum possible p-value from the permutation test? How does this limit inference?

Exercise 7.4: You're studying the effect of a corporate merger on firm performance. You have data on 50 similar firms that didn't merge. (a) How would you construct the donor pool? (b) What predictors would you use? (c) What validity checks would you perform?

Exercise 7.5: The pre-treatment RMSPE for your treated unit is 5.2, while the average pre-treatment RMSPE across placebo units is 3.1. Does this affect your interpretation of the results? How?

Exercise 7.6: Compare and contrast synthetic control with difference-in-differences. When would you prefer one over the other?

PART 8: ADVANCED TOPICS AND RESEARCH FRONTIERS

explore:

1. **Combining methods:** How to use multiple identification strategies
2. **Sensitivity analysis:** Formal approaches to assessing robustness
3. **Machine learning for causal inference:** LASSO, random forests, and causal forests
4. **Heterogeneous treatment effects:** Moving beyond average effects
5. **Mediation and mechanisms:** Understanding how treatments work
6. **Research design principles:** How to choose and defend your approach

structure:

1. Combining multiple identification strategies
2. Sensitivity analysis (Rosenbaum bounds, Oster's approach, etc.)
3. Machine learning for causal inference (LASSO for variable selection, causal forests, etc.)
4. Heterogeneous treatment effects
5. Mediation analysis
6. Bounds and partial identification

Synthesis, Extensions, and Modern Methods

8.1 Introduction: The State of the Art

We've now covered the foundational methods of causal inference:

Method	**Key Assumption**	**Identifies**
RCT	Random assignment	ATE
Selection on Observables	Conditional independence	ATE/ATT
Instrumental Variables	Exclusion restriction	LATE
Difference-in-Differences	Parallel trends	ATT
Regression Discontinuity	Continuity	Local ATE
Synthetic Control	Factor model / good fit	ATT for treated unit

In this final part, I tried doing:

- How to combine these methods
- How to assess robustness formally
- Modern machine learning approaches
- Heterogeneous treatment effects
- Partial identification and bounds
- Best practices for credible research

8.2 Combining Identification Strategies

Why Combine Methods?

Each method has different assumptions. If multiple methods give similar answers under different assumptions, our confidence in the results increases.

Strategy 1: Same Question, Different Methods

Apply multiple methods to the same causal question.

Example: Returns to Education

Method	**Estimate**	**Assumption**
OLS	10%	Selection on observables
IV (quarter of birth)	8%	QOB affects education, not wages directly
IV (college proximity)	9%	Distance affects education, not wages directly
Twins fixed effects	7%	Within-twin variation is exogenous
RDD (college admission)	11%	Continuity at admission cutoff

If estimates cluster around 7-11%, we're more confident the true effect is in this range.

Strategy 2: Bounding with Different Assumptions

Use methods with different biases to bound the true effect.

Example: If OLS is biased upward (positive selection) and some IV is biased downward (weak instruments), the true effect may lie between them.

Strategy 3: Triangulation

Use different data sources, time periods, or populations.

Example: Minimum Wage Effects

- Card & Krueger (1994): NJ vs. PA, +2.75 employees
- Neumark & Wascher (2000): Same setting, different data, -2.0 employees
- Cengiz et al. (2019): Bunching estimator, near-zero effect
- Harasztosi & Lindner (2019): Hungarian data, small negative effect

The debate continues, but triangulation reveals where estimates agree and disagree.

Strategy 4: Nested Designs

Embed one method within another.

Example: RDD + DiD

If there's a cutoff AND time variation:

$$y_{ist} = \alpha + \tau \cdot \mathbf{1}[x_i \geq c] \cdot Post_t + f(x_i) + \gamma_s + \lambda_t + \varepsilon_{ist}$$

This uses both the discontinuity AND the before-after comparison, allowing each to address threats to the other.

8.3 Sensitivity Analysis: How Robust Are Your Results?

The Core Question

Every causal estimate relies on untestable assumptions. Sensitivity analysis asks: **How much would these assumptions need to be violated to overturn our conclusions?**

Approach 1: Rosenbaum Bounds (Selection on Observables)

For matched/propensity score studies, Rosenbaum (2002) developed bounds for unobserved confounding.

Setup: Let Γmeasure the degree of unobserved confounding:

$$\frac{1}{\Gamma} \leq \frac{P(w_i = 1 \mid \mathbf{x}_i, u_i)/P(w_i = 0 \mid \mathbf{x}_i, u_i)}{P(w_j = 1 \mid \mathbf{x}_j, u_j)/P(w_j = 0 \mid \mathbf{x}_j, u_j)} \leq \Gamma$$

for matched pairs $i, j$with $\mathbf{x}_i = \mathbf{x}_j$but potentially different unobserved u.

Interpretation:

- $\Gamma = 1$: No unobserved confounding (randomized experiment)
- $\Gamma = 2$: Unobserved factors could double odds of treatment
- $\Gamma = 3$: Could triple odds

Procedure:

1. Start with $\Gamma = 1$(CIA holds)
2. Gradually increase Γ
3. Find the value where significance disappears

Reporting: “Results are robust to unobserved confounding that increases treatment odds by up to $\Gamma = 2.3$.”

Implementation in R:

```
library(rbounds)

# After matching

psens(match_object, Gamma = 2, GammaInc = 0.1)
```

Approach 2: Oster (2019) Bounds for OLS

Emily Oster developed a practical sensitivity analysis for regression.

Key insight: Observe how the coefficient changes when you add controls. Use this to infer what would happen if you could add unobserved controls.

Setup:

- $\tilde{\tau}$: coefficient from regression without controls, $\tilde{R}^2$
- $\hat{\tau}$: coefficient from regression with controls, R^2
- τ^*: hypothetical coefficient with all controls (observed + unobserved), R^2_{max}

The Key Equation:

Assuming proportional selection (δ):

$$\tau^* \approx \hat{\tau} - \delta \cdot (\tilde{\tau} - \hat{\tau}) \cdot \frac{R^2_{max} - R^2}{R^2 - \tilde{R}^2}$$

Parameters:

- δ: ratio of selection on unobservables to selection on observables
- R^2_{max}: maximum R-squared if all confounders were included

Common benchmarks:

- $\delta = 1$: unobservables as important as observables
- $R^2_{max} = 1$: outcome fully determined (conservative)
- $R^2_{max} = 1.3 \times R^2$: modest increase (Oster's recommendation)

Reporting: "To explain away the effect, unobserved confounding would need to be $\delta =$ 3.2times as important as observed confounding."

Implementation in Stata:

```
stata
* Install psacalc
ssc install psacalc
* Basic usage
psacalc tau treatment, mcontrol(controls) rmax(0.9) delta(1)
```

Approach 3: Coefficient Stability (Altonji, Elder & Taber, 2005)

Compare coefficients across specifications:

$$\frac{\hat{\tau}_{full}}{\hat{\tau}_{restricted} - \hat{\tau}_{full}}$$

Interpretation: How many times larger would the effect of unobservables need to be (relative to observables) to explain away the result?

Approach 4: Placebo Tests

Test for "effects" where none should exist:

***Placebo outcomes*:** Treatment shouldn't affect outcomes determined before treatment.

***Placebo treatments*:** Fake treatments (wrong time, wrong threshold) shouldn't show effects.

***Placebo samples*:** Effect shouldn't appear in populations not affected by treatment.

Approach 5: Leave-One-Out Analysis

Systematically drop:

- Individual observations (influential outliers?)
- Time periods (results driven by specific years?)
- Subgroups (results driven by specific populations?)
- Control variables (results sensitive to specific controls?)

8.4 Machine Learning for Causal Inference

The Promise and the Challenge

Machine learning excels at **prediction** but causal inference requires more:

- Prediction: $E[y \mid \mathbf{x}]$
- Causal inference: $E[y^1 - y^0 \mid \mathbf{x}]$

ML can help with causal inference when used carefully:

1. Covariate selection
2. Propensity score estimation
3. Outcome modeling
4. Heterogeneous treatment effect estimation

LASSO for Covariate Selection

The Problem: With many potential controls, how do we choose which to include?

LASSO (Least Absolute Shrinkage and Selection Operator):

$$\hat{\boldsymbol{\beta}}^{LASSO} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p} \mid \beta_j \mid$$

The L_1penalty shrinks coefficients toward zero and sets some exactly to zero (variable selection).

For Causal Inference: Use LASSO to select controls, then run OLS with selected controls and the treatment variable.

Warning: Naive LASSO can introduce bias. Use **double selection** (Belloni, Chernozhukov & Hansen, 2014):

1. LASSO of $y$on **x**: select controls $\hat{S}_y$
2. LASSO of $w$on **x**: select controls $\hat{S}_w$
3. OLS of $y$on $w$and $\hat{S}_y \cup \hat{S}_w$

This ensures we include variables that predict treatment OR outcome.

Implementation in R:

```
library(hdm)

# Double selection LASSO

lasso_fit <- rlassoEffect(x = X, y = Y, d = W, method = "double selection")

summary(lasso_fit)
```

Double/Debiased Machine Learning (Chernozhukov et al., 2018)

The Framework:

Use ML to estimate nuisance functions, then use these to estimate causal effects.

For ATE with selection on observables:

1. **Estimate propensity score:** $\hat{p}(\mathbf{x}_i) = \hat{E}[w_i \mid \mathbf{x}_i]$using ML
2. **Estimate outcome regressions:** $\hat{\mu}_0(\mathbf{x}_i), \hat{\mu}_1(\mathbf{x}_i)$using ML
3. **Compute doubly robust estimator:**

$$\hat{\tau}^{DML} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(\mathbf{x}_i) - \hat{\mu}_0(\mathbf{x}_i) + \frac{w_i(y_i - \hat{\mu}_1(\mathbf{x}_i))}{\hat{p}(\mathbf{x}_i)} - \frac{(1 - w_i)(y_i - \hat{\mu}_0(\mathbf{x}_i))}{1 - \hat{p}(\mathbf{x}_i)}\right]$$

Cross-fitting: To avoid overfitting bias:

1. Split sample into K folds
2. For each fold, estimate nuisance functions on other folds
3. Compute estimator on held-out fold
4. Average across folds

Implementation in R:

```
library(DoubleML)
# Setup
dml_data <- DoubleMLData$new(data, y_col = “Y”, d_cols = “W”, x_cols = X_names)


# Choose ML methods
ml_g <- lrn(“regr.ranger”)  # Random forest for outcome
ml_m <- lrn(“classif.ranger”)  # Random forest for propensity


# Double ML
dml_plr <- DoubleMLPLR$new(dml_data, ml_g, ml_m)
dml_plr$fit()
dml_plr$summary()
```

Causal Forests (Wager & Athey, 2018)

Goal: Estimate **heterogeneous treatment effects** $\tau(\mathbf{x}) = E[y^1 - y^0 \mid \mathbf{x}]$.

The Idea: Adapt random forests to estimate treatment effects rather than outcomes.

Procedure:

1. Grow trees that split to maximize heterogeneity in treatment effects

2. At each leaf, estimate local treatment effect
3. Average across trees (forest)
4. Use “honest” estimation: split sample for tree construction vs. effect estimation

Key Innovation: Honesty

Standard random forests use the same data to:

- Decide where to split
- Estimate outcomes in leaves

This causes overfitting. Causal forests use separate samples for each task.

Implementation in R:

```
library(grf)

# Causal forest

cf <- causal_forest(X, Y, W)

# Estimate treatment effects

tau_hat <- predict(cf)$predictions

# Average treatment effect

average_treatment_effect(cf)

# Variable importance

variable_importance(cf)
```

When to Use ML for Causal Inference

Good applications:

- High-dimensional covariates (many potential controls)
- Complex functional forms
- Heterogeneous treatment effects
- Prediction as an intermediate step

Limitations:

- ML doesn’t solve fundamental identification problems

- Black-box methods reduce interpretability
- Requires large samples
- Can be computationally intensive

8.5 Heterogeneous Treatment Effects

Why Heterogeneity Matters

Average effects hide important variation:

- A drug may help some patients but harm others
- A policy may benefit the rich but hurt the poor
- Understanding heterogeneity is crucial for targeting

Notation

Conditional Average Treatment Effect (CATE):

$$\tau(\mathbf{x}) = E[y_i^1 - y_i^0 \mid \mathbf{x}_i = \mathbf{x}]$$

The treatment effect for individuals with characteristics **x**.

Approach 1: Subgroup Analysis

Estimate effects separately for subgroups:

$$y_i = \alpha + \tau_1 w_i \cdot \mathbf{1}[x_i = 1] + \tau_2 w_i \cdot \mathbf{1}[x_i = 2] + \cdots + \varepsilon_i$$

Problems:

- Multiple testing (many subgroups → spurious findings)
- Low power in small subgroups
- Which subgroups to examine?

Solution: Pre-specify subgroups or adjust for multiple testing.

Approach 2: Interaction Terms

$$y_i = \alpha + \tau w_i + \beta x_i + \gamma(w_i \times x_i) + \varepsilon_i$$

Interpretation: γmeasures how the treatment effect varies with x_i.

Limitations:

- Assumes linear interaction
- Hard to interpret with many covariates
- Can't capture complex heterogeneity

Approach 3: Causal Forests (ML Approach)

As described above, causal forests estimate $\tau(\mathbf{x})$flexibly.

Advantages:

- No need to pre-specify interactions
- Captures complex heterogeneity
- Provides variable importance for heterogeneity

Typical workflow:

```
library(grf)
# Estimate causal forest
cf <- causal_forest(X, Y, W, num.trees = 2000)
# Individual treatment effects
tau_hat <- predict(cf)$predictions
# Find important variables for heterogeneity
var_imp <- variable_importance(cf)
print(var_imp)
# Best linear projection onto selected variables
```

```
blp <- best_linear_projection(cf, X[, c(“age”, “income”)])

print(blp)

# Rank-weighted average treatment effect (compare high vs low predicted effect)

rate <- rank_average_treatment_effect(cf, tau_hat)

print(rate)
```

Approach 4: Generic ML (Kennedy, 2020)

DR-Learner: Use any ML method with a doubly robust transformation.

Steps:

1. Estimate propensity score $\hat{p}(\mathbf{x})$
2. Estimate outcome regressions $\hat{\mu}_0(\mathbf{x}), \hat{\mu}_1(\mathbf{x})$
3. Compute pseudo-outcome: $$\tilde{Y}_i = \hat{\mu}_1(\mathbf{x}_i) - \hat{\mu}_0(\mathbf{x}_i) + \frac{w_i(y_i - \hat{\mu}_1(\mathbf{x}_i))}{\hat{p}(\mathbf{x}_i)} - \frac{(1-w_i)(y_i - \hat{\mu}_0(\mathbf{x}_i))}{1 - \hat{p}(\mathbf{x}_i)}$$
4. Regress $\tilde{Y}_i$ on $\mathbf{x}_i$ using any ML method

Reporting Heterogeneous Effects

Good practices:

5. Report average effect first
6. Pre-specify hypotheses about heterogeneity
7. Use appropriate multiple testing corrections
8. Visualize CATE distribution
9. Validate findings in holdout sample

8.6 Mediation Analysis: Mechanisms

The Question

Once we establish that $W$ causes $Y$, we often want to know **how**:

- Through what mechanism does education raise earnings?
- Why does the drug improve health outcomes?

The Mediation Framework

W → M → Y

W ------→ Y

- **Direct effect:** $W \rightarrow Y$(not through M)
- **Indirect effect:** $W \rightarrow M \rightarrow Y$(through M)
- **Total effect:** Direct + Indirect

Potential Outcomes Notation

Let:

- $M_i(w)$: potential mediator when treatment is w
- $Y_i(w, m)$: potential outcome when treatment is $w$and mediator is m

Natural Direct Effect (NDE):

$$NDE = E[Y_i(1, M_i(0)) - Y_i(0, M_i(0))]$$

Effect of treatment holding mediator at its “natural” value under control.

Natural Indirect Effect (NIE):

$$NIE = E[Y_i(1, M_i(1)) - Y_i(1, M_i(0))]$$

Effect of changing mediator from control to treatment value, holding treatment at 1.

Total Effect:

$$TE = E[Y_i(1, M_i(1)) - Y_i(0, M_i(0))] = NDE + NIE$$

The Traditional Approach: Baron & Kenny (1986)

Step 1: Estimate total effect:

$$y_i = \alpha_1 + \tau w_i + \varepsilon_{1i}$$

Step 2: Estimate effect of treatment on mediator:

$$m_i = \alpha_2 + \beta w_i + \varepsilon_{2i}$$

Step 3: Estimate effect of mediator controlling for treatment:

$$y_i = \alpha_3 + \tau' w_i + \gamma m_i + \varepsilon_{3i}$$

Interpretation:

- τ' = direct effect
- $\beta \times \gamma = $ indirect effect (product of coefficients)
- $\tau = \tau' + \beta\gamma$(total effect decomposition)

Problems with Traditional Approach

1. **The mediator is endogenous!** Even if $W$is randomly assigned, $M$is not.
2. **Omitted confounders of $M \to Y$:** Variables affecting both mediator and outcome bias γ.
3. **Post-treatment confounding:** Controlling for Mcan induce bias if there are post-treatment confounders.

Modern Causal Mediation Analysis

Required assumptions (Imai, Keele & Yamamoto, 2010):

1. **Sequential ignorability:**
   - $\{Y_i(w', m), M_i(w)\} \perp\perp W_i \mid \mathbf{X}_i$(treatment is as-good-as-random given **X**)
   - $Y_i(w', m) \perp\perp M_i \mid W_i = w, \mathbf{X}_i$ (mediator is as-good-as-random given treatment and **X**)
2. The second assumption is **very strong**: no unmeasured confounders of $M \to Y$ relationship.

Implementation in R:

```
library(mediation)
# Step 1: Model for mediator
med_model <- lm(M ~ W + X1 + X2, data = df)
# Step 2: Model for outcome
out_model <- lm(Y ~ W + M + X1 + X2, data = df)
# Step 3: Mediation analysis
med_results <- mediate(med_model, out_model,
                 treat = “W”, mediator = “M”,
                 boot = TRUE, sims = 1000)
summary(med_results)
```

Sensitivity Analysis for Mediation

Since sequential ignorability is often implausible, sensitivity analysis is crucial:

```
# Sensitivity analysis
sens_results <- medsens(med_results, rho.by = 0.1)
summary(sens_results)
plot(sens_results)
```

This shows how results change as we vary the correlation between mediator and outcome errors.

Alternative: Controlled Direct Effects

If we can **experimentally manipulate** the mediator, we can estimate the **Controlled Direct Effect (CDE)**:

$$CDE(m) = E[Y_i(1, m) - Y_i(0, m)]$$

This requires manipulating both $W$and M, which is feasible in some experimental settings.

8.7 Partial Identification and Bounds

When Point Identification Fails

Sometimes our assumptions aren't strong enough for point identification—we can't pinpoint the exact causal effect. But we can often **bound** it.

Manski Bounds: The Worst Case

Manski (1990) derived bounds under minimal assumptions.

Setup: Binary treatment, outcome bounded in $[y_L, y_U]$.

Observed: $E[Y \mid W = 1]$and $E[Y \mid W = 0]$

Unobserved: $E[Y^0 \mid W = 1]$and $E[Y^1 \mid W = 0]$

Worst-case bounds:

$$E[Y^1] \in [E[Y \mid W = 1] \cdot P(W = 1) + y_L \cdot P(W = 0), E[Y \mid W = 1] \cdot P(W = 1) + y_U \cdot P(W = 0)]$$
$$E[Y^0] \in [E[Y \mid W = 0] \cdot P(W = 0) + y_L \cdot P(W = 1), E[Y \mid W = 0] \cdot P(W = 0) + y_U \cdot P(W = 1)]$$

ATE bounds:

$$ATE \in [\text{lower bound for } E[Y^1] - \text{upper bound for } E[Y^0], \text{upper bound for } E[Y^1] - \text{lower bound for } E[Y^0]]$$

These bounds are often **very wide**—uninformative without additional assumptions.

Tightening Bounds with Assumptions

Monotone Treatment Response (MTR): Assume treatment never hurts (or never helps): $Y_i^1 \geq Y_i^0$for all i.

Monotone Treatment Selection (MTS): Assume selection is positive: $E[Y^0 \mid W = 1] \geq E[Y^0 \mid W = 0]$.

Monotone Instrumental Variable (MIV): Assume the instrument has a monotone relationship with potential outcomes.

Each assumption tightens the bounds.

Example: Returns to Education with Bounds

Suppose we're agnostic about selection:

- OLS gives 10% return (but assumes no selection)
- What if we only assume earnings are bounded $[0, \$500, 000]$?

The Manski bounds might be $[-5\%, +25\%]$—consistent with education helping, hurting, or having no effect!

Adding monotonicity (education never hurts): bounds might tighten to $[0\%, +15\%]$.

Lee (2009) Bounds for Sample Selection

When treatment affects sample selection (attrition, censoring):

Example: Job training program. Some people in both groups don't report earnings (unemployed).

Problem: If training affects employment, comparing observed earnings is biased.

Lee bounds: Trim the group with higher observation rate to match the lower rate, using best-case and worst-case assumptions.

Implementation in Stata:

```
stata

* Lee bounds

leebounds outcome treatment, select(employed)
```

8.8 Regression Discontinuity: Advanced Topics

RDD with Multiple Cutoffs and Scores

Multiple cutoffs on one score:

- Pool data, normalize running variable: $\tilde{x}_{ic} = x_i - c$ for cutoff c
- Estimate single effect (if effects are homogeneous) or separate effects

Multiple running variables:

- Geographic RDD: latitude and longitude
- Need to define treatment as function of both variables

Extrapolating from the Cutoff

RDD identifies effects only at the cutoff. To generalize:

Assumption: Constant effects

$$\tau(x) = \tau \text{ for all } x$$

Under this assumption, the local effect equals the global effect.

Testing: If effects vary with x, estimate effects at different bandwidths and check stability.

Parametric extrapolation: Estimate $\tau(x)$ as a function of the running variable:

$$y_i = \alpha + \tau(x_i) w_i + f(x_i) + \varepsilon_i$$

where $\tau(x) = \tau_0 + \tau_1 x$.

RDD with Discrete Running Variables

When x takes few values (test scores, age in years):

Challenge: Can't observe outcomes arbitrarily close to cutoff.

Solutions:

1. Use mass points directly

2. Cluster standard errors by running variable value
3. Cattaneo, Idrobo & Titiunik (2020) methods

Donut RDD

If manipulation occurs very close to the cutoff, exclude observations near the cutoff:

$$\text{Use only observations with } |x_i - c| > \epsilon$$

This sacrifices precision but may reduce manipulation bias.

8.9 Difference-in-Differences: Advanced Topics

A. Doubly Robust DiD (Sant'Anna & Zhao, 2020)

Combine outcome regression and propensity score weighting:

$$\hat{\tau}^{DR} = \frac{1}{n_1} \sum_{i: G_i=1} \left[(y_{i,post} - y_{i,pre}) - \hat{m}(\mathbf{x}_i)\right]$$

$$- \frac{1}{n} \sum_{i: G_i=0} \frac{\hat{p}(\mathbf{x}_i)}{1-\hat{p}(\mathbf{x}_i)} \cdot \frac{n_1/n}{P(G=0)} \left[(y_{i,post} - y_{i,pre}) - \hat{m}(\mathbf{x}_i)\right]$$

where $\hat{m}(\mathbf{x})$ is the conditional mean of Δy for controls.

Advantage: Consistent if either propensity score or outcome model is correct.

B. Continuous Treatment DiD

When treatment intensity varies:

$$y_{it} = \alpha_i + \lambda_t + \tau \cdot D_{it} \cdot Dose_i + \varepsilon_{it}$$

Challenge: Units selecting higher doses may differ systematically.

Solutions:

- IV for dose (if available)
- Bounding approaches
- Careful interpretation as dose-response

C. DiD with Multiple Time Periods and Groups

Recent methods for staggered designs also handle:

- Multiple treatment cohorts
- Treatment turning on and off
- Heterogeneous treatment timing

did package in R handles many of these cases.

8.10 Instrumental Variables: Advanced Topics

A. Weak IV Robust Methods

Anderson-Rubin (AR) test:

Test $H_0: \tau = \tau_0$ by testing whether z is uncorrelated with $y - \tau_0 w$.

$$AR(\tau_0) = \frac{(y - \tau_0 w)' P_z (y - \tau_0 w) / k}{(y - \tau_0 w)' M_z (y - \tau_0 w) / (n-k)}$$

Under H_0: $AR \sim F(k, n-k)$

Confidence interval: Invert the test—all τ_0 values not rejected.

Advantage: Valid even with weak instruments.

Disadvantage: Can be wide or even unbounded.

B. Many Instruments

With many instruments, 2SLS is biased toward OLS.

Solutions:

- LIML (Limited Information Maximum Likelihood)
- JIVE (Jackknife IV Estimator)
- Regularized IV (LASSO-IV)

C. Shift-Share (Bartik) Instruments

Common in labor and trade economics:

$$z_i = \sum_k s_{ik} g_k$$

where s_{ik} = share of industry k in location i (historical), g_k = national growth of industry k.

Interpretation: Local exposure to national shocks.

Recent work (Goldsmith-Pinkham, Sorkin & Swift, 2020; Borusyak, Hull & Jaravel, 2022):

- Identification comes from either share or shock exogeneity
- Need to be clear about which assumption is maintained
- Appropriate standard errors depend on the source of variation

The "Five-Slide Rule"

Can you explain your paper in five slides?

1. Question and why it matters
2. Ideal experiment
3. Your identification strategy
4. Main results
5. Robustness and implications

The Core Message

Causal inference requires both:

1. **Identification:** A credible source of exogenous variation
2. **Estimation:** Appropriate statistical methods

Neither alone is sufficient. A perfect estimator can't overcome bad identification, and good identification can be ruined by bad estimation.

Method Summary Table

Method	Variation	Key Assumption	Estimand	Best For
RCT	Randomization	Random assignment	ATE	When feasible
Matching/IPW	Cross-sectional	CIA	ATE/ATT	Rich observables
IV/2SLS	Instrument	Exclusion	LATE	Clear instrument
DiD	Time × Group	Parallel trends	ATT	Policy changes
RDD	Running variable	Continuity	Local	Clear cutoffs
Synthetic Control	Time series	Good fit	Unit effect	Few treated

Key references for further reading:

- Angrist & Pischke, *Mostly Harmless Econometrics*
- Wooldridge, *Econometric Analysis of Cross Section and Panel Data*
- Imbens & Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences*
- Cunningham, *Causal Inference: The Mixtape*
- Huntington-Klein, *The Effect*