

### 3. Difference-in-Differences

PhD Applied Methods

Duncan Webb  
NovaSBE

Spring 2026

## Quick IV review: Compliance types are **fixed** attributes

Recall: an individual's compliance type is defined by the pair of **potential treatments**:

	$T_i(0) = 0$	$T_i(0) = 1$
$T_i(1) = 0$	Never-taker	Defier
$T_i(1) = 1$	<b>Complier</b>	Always-taker

- This is a **fixed characteristic** of each person — does not depend on realized  $Z_i$
- Avoid “ $T_i = 1$  when  $Z_i = 1$ ” — this is a *realized outcome*, not a *type*
- Correct:  $T_i(1) = 1$  and  $T_i(0) = 0 \implies$  Complier

Also: when computing type shares from a joint distribution table, **condition**:

$$P(T=1 \mid Z=1) = \frac{P(T=1, Z=1)}{P(Z=1)} \neq P(T=1, Z=1)$$

# Defining a parameter $\neq$ identifying it from data

## A common pattern in causal inference:

- It is often easy to *define* the causal parameter you want
- It is much harder to show you can *identify* it from observable data

**Example:** Suppose you want to estimate a treatment spillover effect on the untreated.

- You might try comparing untreated individuals in treated vs. control groups
- But **who are the untreated?** In each group, the composition differs:
  - In a treated group: untreated = never-takers only (compliers took up treatment)
  - In a control group: untreated = never-takers + compliers
- So the naive comparison confounds the spillover with a **composition effect**

**Lesson:** Always ask: *who is in each comparison group?* If the type composition differs, your estimator may not identify your target parameter — even if it is well-defined.

## Different instruments $\implies$ different IV estimates

Angrist & Imbens (1995) show that with **multi-valued treatment**, IV identifies a weighted average of margin-specific LATEs – and the **weights depend on the instrument**.

**Concrete example:** Two instruments for years of schooling

Instrument	Compliers	Margin
Compulsory schooling laws	Students pushed from dropping out to staying	8 $\rightarrow$ 9, 9 $\rightarrow$ 10 years
College proximity	Students induced to attend college	12 $\rightarrow$ 13, 13 $\rightarrow$ 16 years

- Both are valid instruments for “the effect of education on earnings”
- But they identify effects for **different complier populations** at **different margins**
- Different IV estimates  $\neq$  one instrument is invalid

## Different instruments $\Rightarrow$ different IV estimates (cont.)

### Why does this happen?

- With heterogeneous treatment effects, the return to an extra year of schooling differs across individuals and across margins (e.g.,  $8 \rightarrow 9$  vs.  $15 \rightarrow 16$ )
- Each instrument affects a *different set of margins* — so it weights different LATEs
- The IV estimate is a **weighted average** of these margin-specific effects, with weights determined by how much the instrument shifts treatment at each margin

**Implication:** IV estimates are *instrument-dependent*. When comparing IV results across studies, always ask: who are the compliers, and which treatment margin is being affected?

# Why difference-in-differences?

## So far in this course:

- **RCTs** give us clean causal effects — but often infeasible or unethical
- **Instrumental variables** give us causal effects without randomizing treatment directly — but good instruments are hard to find, and IV only identifies the **LATE**, not the ATE

## Today: **Difference-in-Differences (DiD)**

- Applicable when there is both **temporal** and **cross-sectional** variation in treatment exposure
- Identifies causal effects under a **stronger assumption** (parallel trends) — but one that is often plausible in policy settings
- One of the most widely used methods in applied economics

# The fundamental problem

Simple comparison of treated vs. untreated gives us:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \quad (1)$$

But this is **not** the causal effect! Why not?

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \quad (2)$$

$$= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{Selection bias}} \quad (3)$$

**Selection bias:** treated and untreated groups differ systematically, even in the absence of treatment

# Enter: Difference-in-Differences

- **Difference-in-Differences (DiD)** is one of the most widely used methods in applied econometrics
- **Key insight:** Even if treated and control groups differ in levels, we can still identify causal effects if they share common trends
- Requires observing both groups **before and after** treatment
  - Use the change in the control group to construct the counterfactual for the treated group
- **Today's goal:** Understand the mechanics, assumptions, and extensions of DiD



# Outline

1. Motivation
2. Theory: Common trends and graphical intuition
3. Multiple time periods, pre-trends, and regression
4. Threats to identification
5. Triple differences
6. Staggered treatment timing
7. Synthetic DiD

# The naive before-after estimator

One approach: compare the treated group before and after treatment

Suppose we have:

- Period 1 (before treatment):  $t = 1$
- Period 2 (after treatment):  $t = 2$
- Treatment happens between periods 1 and 2

**Naive before-after** estimator:

$$\hat{\tau}^{BA} = \mathbb{E}[Y_{i2}|D_i = 1] - \mathbb{E}[Y_{i1}|D_i = 1] \quad (4)$$

**Question:** What assumption is needed for this to identify the ATT?

# The naive before-after estimator

Recall the ATT is:

$$\tau_2^{ATT} = \mathbb{E}[Y_{i,2}(1) - Y_{i,2}(0)|D_i = 1] \quad (5)$$

The before-after estimator gives us:

$$\hat{\tau}^{BA} = \mathbb{E}[Y_{i2}|D_i = 1] - \mathbb{E}[Y_{i1}|D_i = 1] \quad (6)$$

$$= \mathbb{E}[Y_{i,2}(1)|D_i = 1] - \mathbb{E}[Y_{i,1}(0)|D_i = 1] \quad (7)$$

This equals the ATT if and only if:

$$\mathbb{E}[Y_{i,2}(0)|D_i = 1] = \mathbb{E}[Y_{i,1}(0)|D_i = 1] \quad (8)$$

**Interpretation:** The treated group's outcome (absent treatment) would have been the same in both periods

## Problems with before-after

The assumption  $\mathbb{E}[Y_{i,2}(0)|D_i = 1] = \mathbb{E}[Y_{i,1}(0)|D_i = 1]$  is very strong!

It rules out:

- **General time trends:** economic cycles, inflation, technological progress
- **Life-cycle effects:** aging, experience accumulation, depreciation
- **Seasonality:** quarterly or monthly patterns
- **Mean reversion:** regression to the mean

**Example:** NJ minimum wage increase in 1992

- Employment in fast-food rises from 20.4 to 21.0
- If economy is booming  $\implies$  underestimates negative effect (or overestimates positive effect)
- If economy is in recession  $\implies$  overestimates negative effect

# The common trends assumption

**Key insight:** We can relax the before-after assumption by using a control group

Instead of assuming treated outcomes are constant over time, we assume:

## Common Trends (Parallel Trends) Assumption:

In the absence of treatment, the average outcomes for the treated and control groups would have evolved in parallel

Formally:

$$\mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 1] = \mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0] \quad (9)$$

# Common trends: Intuition

The common trends assumption says:

- Treated and control groups can differ in **levels**
  - $\mathbb{E}[Y_{i,1}(0)|D_i = 1] \neq \mathbb{E}[Y_{i,1}(0)|D_i = 0]$  ✓
- But they must have the same **change over time** (absent treatment)
  - $\mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 1] = \mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0]$
- This allows for:
  - Permanent differences between groups
  - Common time shocks that affect everyone
- This rules out:
  - Group-specific time trends
  - Differential exposure to time-varying shocks

## Equivalence: Two ways to state parallel trends

The parallel trends assumption can be stated in two equivalent ways:

### 1. Changes are equal:

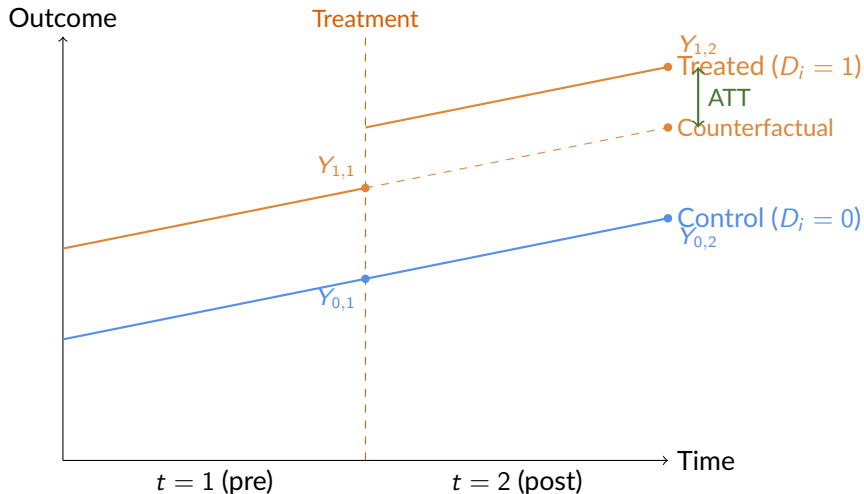
$$\mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 1] = \mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0] \quad (10)$$

### 2. Selection bias is constant over time:

$$\underbrace{\mathbb{E}[Y_{i,1}(0)|D_i = 1] - \mathbb{E}[Y_{i,1}(0)|D_i = 0]}_{\text{Selection bias in } t=1} = \underbrace{\mathbb{E}[Y_{i,2}(0)|D_i = 1] - \mathbb{E}[Y_{i,2}(0)|D_i = 0]}_{\text{Selection bias in } t=2} \quad (11)$$

**Interpretation:** The "gap" between treated and control (absent treatment) stays constant

# Graphical intuition: Difference-in-Differences



**Key insight:** The change in the control group gives us the counterfactual trend for the



# The $2 \times 2$ difference-in-differences estimator

We observe four group-time averages:

	Pre-treatment ( $t = 1$ )	Post-treatment ( $t = 2$ )
Treated ( $D_i = 1$ )	$\bar{Y}_{1,1}$	$\bar{Y}_{1,2}$
Control ( $D_i = 0$ )	$\bar{Y}_{0,1}$	$\bar{Y}_{0,2}$

The **difference-in-differences estimator** is:

$$\hat{\tau}^{DiD} = (\bar{Y}_{1,2} - \bar{Y}_{1,1}) - (\bar{Y}_{0,2} - \bar{Y}_{0,1}) \quad (12)$$

- First difference: change in treated group over time
- Second difference: change in control group over time
- DiD: difference between these two changes

## Alternative formulation

The DiD estimator can equivalently be written as:

$$\hat{\tau}^{DiD} = (\bar{Y}_{1,2} - \bar{Y}_{1,1}) - (\bar{Y}_{0,2} - \bar{Y}_{0,1}) \quad (13)$$

$$= (\bar{Y}_{1,2} - \bar{Y}_{0,2}) - (\bar{Y}_{1,1} - \bar{Y}_{0,1}) \quad (14)$$

### Interpretation:

- First difference: post-treatment difference between treated and control
- Second difference: pre-treatment difference between treated and control
- DiD: how much the gap changed

This formulation makes clear that DiD **differences out** time-invariant differences between groups

# What does DiD identify?

Under the parallel trends assumption:

$$\mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 1] = \mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0] \quad (15)$$

the DiD estimator identifies the ATT:

$$\mathbb{E}[\hat{\tau}^{DiD}] = \mathbb{E}[Y_{i,2}(1) - Y_{i,2}(0)|D_i = 1] = \tau_2^{ATT} \quad (16)$$

**Proof sketch:**

$$\mathbb{E}[\hat{\tau}^{DiD}] = \mathbb{E}[Y_{i,2}(1)|D_i = 1] - \mathbb{E}[Y_{i,1}(0)|D_i = 1] \quad (17)$$

$$- (\mathbb{E}[Y_{i,2}(0)|D_i = 0] - \mathbb{E}[Y_{i,1}(0)|D_i = 0]) \quad (18)$$

## Proof (continued)

$$\mathbb{E}[\hat{\tau}^{DiD}] = \mathbb{E}[Y_{i,2}(1)|D_i = 1] - \mathbb{E}[Y_{i,1}(0)|D_i = 1] \quad (19)$$

$$- (\mathbb{E}[Y_{i,2}(0)|D_i = 0] - \mathbb{E}[Y_{i,1}(0)|D_i = 0]) \quad (20)$$

Add and subtract  $\mathbb{E}[Y_{i,2}(0)|D_i = 1]$ :

$$= \underbrace{\mathbb{E}[Y_{i,2}(1) - Y_{i,2}(0)|D_i = 1]}_{\tau_2^{ATT}} \quad (21)$$

$$+ \underbrace{\mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 1] - \mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0]}_{=0 \text{ under parallel trends}} \quad (22)$$

Under parallel trends, the second term equals zero, so:

$$\mathbb{E}[\hat{\tau}^{DiD}] = \tau_2^{ATT} \quad (23)$$

# DiD as a regression

The  $2 \times 2$  DiD estimator can be implemented via regression:

$$Y_{it} = \alpha + \gamma \cdot \mathbb{1}(D_i = 1) + \lambda \cdot \mathbb{1}(t = 2) + \beta \cdot \mathbb{1}(D_i = 1) \times \mathbb{1}(t = 2) + \varepsilon_{it} \quad (24)$$

where:

- $\mathbb{1}(D_i = 1)$ : dummy for being in treated group
- $\mathbb{1}(t = 2)$ : dummy for post-treatment period
- $\mathbb{1}(D_i = 1) \times \mathbb{1}(t = 2)$ : interaction (treatment indicator)
- $\beta$ : the DiD coefficient

**Key result:**  $\hat{\beta}_{OLS} = \hat{\tau}^{DiD}$

# Understanding the regression coefficients

$$Y_{it} = \alpha + \gamma \cdot \mathbb{1}(D_i = 1) + \lambda \cdot \mathbb{1}(t = 2) + \beta \cdot \mathbb{1}(D_i = 1) \times \mathbb{1}(t = 2) + \varepsilon_{it} \quad (25)$$

What do the parameters represent?

	$t = 1$	$t = 2$
$D_i = 0$	$\alpha$	$\alpha + \lambda$
$D_i = 1$	$\alpha + \gamma$	$\alpha + \gamma + \lambda + \beta$

- $\alpha$ : baseline outcome (control, pre-period)
- $\gamma$ : pre-treatment difference between groups
- $\lambda$ : time trend (common to both groups)
- $\beta$ : treatment effect (DiD estimator)

# Verifying $\hat{\beta} = \hat{\tau}^{DiD}$

From the regression:

$$\beta = \mathbb{E}[Y_{it}|D_i = 1, t = 2] - \mathbb{E}[Y_{it}|D_i = 0, t = 2] \quad (26)$$

$$- (\mathbb{E}[Y_{it}|D_i = 1, t = 1] - \mathbb{E}[Y_{it}|D_i = 0, t = 1]) \quad (27)$$

Rearranging:

$$\beta = (\mathbb{E}[Y_{it}|D_i = 1, t = 2] - \mathbb{E}[Y_{it}|D_i = 1, t = 1]) \quad (28)$$

$$- (\mathbb{E}[Y_{it}|D_i = 0, t = 2] - \mathbb{E}[Y_{it}|D_i = 0, t = 1]) \quad (29)$$

$$= \hat{\tau}^{DiD} \quad (30)$$

## Two-way fixed effects (TWFE) formulation

With panel data, we can rewrite the DiD regression more compactly using fixed effects:

$$Y_{it} = \alpha_i + \delta_t + \beta \cdot D_{it} + \varepsilon_{it} \quad (31)$$

where:

- $\alpha_i$ : unit fixed effects (captures  $\alpha + \gamma \cdot \mathbb{1}(D_i = 1)$  from before)
- $\delta_t$ : time fixed effects (captures  $\lambda \cdot \mathbb{1}(t = 2)$  from before)
- $D_{it} = \mathbb{1}(D_i = 1) \times \mathbb{1}(t = 2)$ : treatment indicator
- $\beta$ : treatment effect (same as before!)

**Equivalence:** This is just a reparameterization of the dummy variable regression

- $\alpha_i$  absorbs all time-invariant unit characteristics
- $\delta_t$  absorbs all time-varying shocks common to all units



## TWFE: Extending to $T > 2$ periods

The TWFE formulation naturally extends to multiple time periods:

$$Y_{it} = \alpha_i + \delta_t + \beta \cdot D_{it} + \varepsilon_{it}, \quad t = 1, 2, \dots, T \quad (32)$$

where now:

- $\delta_t$ : separate time fixed effect for each period  $t \in \{1, 2, \dots, T\}$
- $D_{it} = \mathbb{1}(\text{unit } i \text{ is treated at time } t)$

### Key advantages with multiple periods:

- 1 Can test parallel trends using pre-treatment data
- 2 Can study dynamic effects (how  $\beta$  changes over time since treatment)
- 3 More robust identification (not reliant on single time comparison)

**Parallel trends assumption:**  $Y_{it}(0) = \alpha_i + \delta_t + \varepsilon_{it}$  for all  $t$

## Example: Card & Krueger (1994)

**Question:** What is the effect of minimum wage on employment?

**Setting:**

- New Jersey raised minimum wage from \$4.25 to \$5.05 in April 1992
- Pennsylvania (neighboring state) did not change minimum wage
- Focus on fast-food restaurants (low-wage sector)

**Data:**

- Survey of fast-food restaurants in NJ and PA
- Before (February 1992) and after (November 1992) treatment
- Outcome: full-time equivalent (FTE) employment

## Card & Krueger: Results

	Before (Feb 1992)	After (Nov 1992)	Change
<b>NJ (treated)</b>	20.44	21.03	+0.59
<b>PA (control)</b>	23.33	21.17	-2.16
<b>Difference</b>	-2.89	-0.14	

DiD estimate:

$$\hat{\tau}^{DiD} = 0.59 - (-2.16) = 2.75 \text{ FTE workers} \quad (33)$$

**Interpretation:** Minimum wage increase led to a *relative* increase of 2.75 FTE workers in NJ restaurants (contrary to standard theory prediction)

# Outline

1. Motivation
2. Theory: Common trends and graphical intuition
3. Multiple time periods, pre-trends, and regression
4. Threats to identification
5. Triple differences
6. Staggered treatment timing
7. Synthetic DiD

## Extending to multiple time periods

So far: basic  $2 \times 2$  setup with  $t \in \{1, 2\}$

In practice, we often have:

- Multiple pre-treatment periods:  $t < t_0$
- Multiple post-treatment periods:  $t \geq t_0$
- Treatment occurs at  $t = t_0$

### Benefits of multiple time periods:

- ① Can test the parallel trends assumption using pre-treatment data
- ② Can study dynamic treatment effects (how effects evolve over time)
- ③ Can incorporate more flexible specifications

## TWFE with multiple periods

With  $T > 2$  periods, the TWFE specification becomes:

$$Y_{it} = \alpha_i + \delta_t + \beta \cdot D_{it} + \varepsilon_{it} \quad (34)$$

where now:

- $\alpha_i$ : unit fixed effects (as before)
- $\delta_t$ : time fixed effects for  $t = 1, 2, \dots, T$
- $D_{it} = \mathbb{1}(i \text{ treated at time } t)$

**Interpretation of  $\beta$ :**

- Average treatment effect across all treated units and time periods
- Assumes treatment effect is constant over time (homogeneous effects)
- We'll see later this can be problematic with staggered treatment timing

# Testing the parallel trends assumption

~~The Fundamental problem~~ The Fundamental problem: assumption about **counterfactual** outcomes

- We can never directly observe  $Y_{it}(0)$  for treated units after treatment
- So we can never definitively test whether trends would have been parallel

**But:** We can check whether trends were parallel **before** treatment!

**Pre-trends test:**

- If parallel trends holds, we should see no pre-treatment differences in trends
- If we find differential pre-trends, this casts doubt on the assumption
- Not a perfect test, but provides evidence on plausibility

## Event study specification

To test for pre-trends and examine dynamic effects, use an **event study** design:

$$Y_{it} = \alpha_i + \delta_t + \sum_{\substack{k=-K \\ k \neq -1}}^L \beta_k \cdot \mathbb{1}(t - t_0^i = k) + \varepsilon_{it} \quad (35)$$

where:

- $t_0^i$ : time when unit  $i$  is treated
- $k = t - t_0^i$ : time relative to treatment (“event time”)
- $\beta_k$ : treatment effect  $k$  periods after treatment
- $k = -1$  is the omitted reference period (normalization)

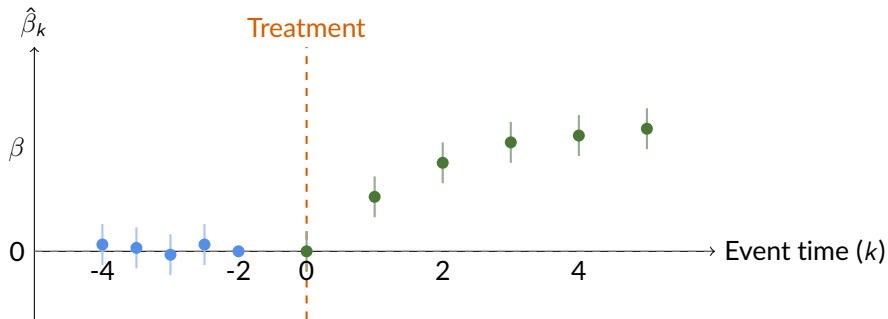
**Coefficients:**

- $\beta_k$  for  $k < 0$ : pre-treatment “effects” (should be zero if parallel trends holds)
- $\beta_k$  for  $k \geq 0$ : post-treatment effects (dynamic treatment effects)



# Event study: Graphical display

Typical event study plot:



**Good pre-trends:** Flat, close to zero before treatment (blue dots)

**Treatment effects:** Jump and evolution after treatment (green dots)

# Pre-trends: What to look for

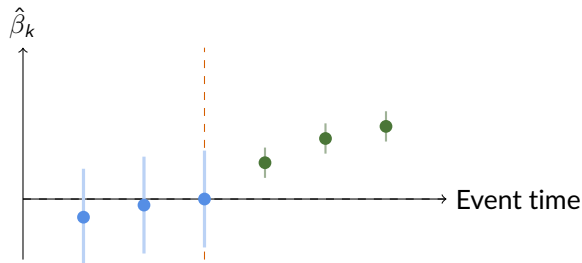
When examining pre-trends, check:

- ① **Statistical significance:** Are pre-treatment  $\hat{\beta}_k$  significantly different from zero?
  - Test individually:  $H_0 : \beta_k = 0$  for each  $k < 0$
  - Test jointly:  $H_0 : \beta_{-K} = \dots = \beta_{-2} = 0$
- ② **Economic significance:** Even if not statistically significant, are they economically large?
  - Compare magnitude of pre-trends to post-treatment effects
  - Large pre-trends (even if imprecise) are concerning
- ③ **Precision:** How precisely estimated are the pre-trends? (Roth 2022)
  - Wide confidence intervals  $\implies$  can't rule out large violations
  - Should be able to reject pre-trends as large as the treatment effect
- ④ **Number of pre-periods:** More pre-periods  $\implies$  more power to detect violations
  - With few pre-periods, tests have low power
  - Ideally want multiple pre-periods to credibly test parallel trends

## Pre-testing concern 1: Power (Roth, 2022)

**Problem:** Standard pre-trends tests have low power

Consider this event study:



**Issue:** Pre-trends not significant, but:

- Clear upward trajectory before treatment
- Confidence intervals VERY wide in pre-period

### The bias from pre-testing:

When researchers select designs based on "passing" pre-trends tests, this can induce bias:



## Ways of dealing with imprecise or differential pre-trends

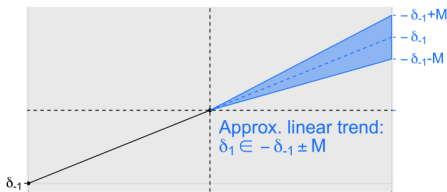
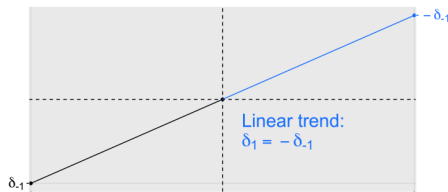
- **Report pre-trends** - report the size of the pre-trend that can be rejected at conventional levels, and discuss how this compares to the estimated treatment effect (should ideally be able to reject that pre-trend is smaller than the treatment effect)
- **Bounding estimators** (Rambachan & Roth ReStud 2023) - use information from pre-trends to bound post-trend using an assumption on smooth changes in trends over time
- **Control for linear pre-trends** - you can also just include estimates of linear differential pre-trends in your DiD regression

## Rambachan and Roth (2023) suggestion

$$\mathbb{E}[\hat{\beta}_1] = \tau_{ATT} + \underbrace{\mathbb{E}[Y_{t,1}(0) - Y_{t,0}(0)|D_i = 1] - \mathbb{E}[Y_{t,1}(0) - Y_{t,0}(0)|D_i = 0]}_{\text{Post-period differential trend} =: \delta_1}$$

$$\mathbb{E}[\hat{\beta}_{-1}] = \underbrace{\mathbb{E}[Y_{t,-1}(0) - Y_{t,0}(0)|D_i = 1] - \mathbb{E}[Y_{t,-1}(0) - Y_{t,0}(0)|D_i = 0]}_{\text{Pre-period differential trend} =: \delta_{-1}}$$

- parallel trends assumes these  $\delta$  are zero. But pre-trends may not be zero.
  - R&R say: we can use the info from our pre-trends to bound post-trend
  - Use a smoothness assumption,  $M$ , on the second derivative. E.g. simple case:



## Standard errors in panel DiD

**Important:** With panel data, standard errors must account for correlation

**Problem:** Bertrand, Duflo & Mullainathan (2004)

- Outcomes for same unit are serially correlated over time
- $\varepsilon_{it}$  and  $\varepsilon_{it'}$  are correlated for  $t \neq t'$
- Standard OLS standard errors are severely downward biased
- Leads to massive over-rejection of null hypotheses

**Solution:** Cluster standard errors at the unit level

- Allows arbitrary correlation within units over time
- Conservative: only assumes independence across units
- In Stata: `reg Y X, cluster(unit_id)`
- In R: `lm_robust(Y ~ X, clusters = unit_id)`
- With fixed effects in R: `feols(Y ~ D | unit + time, cluster = "unit", data = df)`  
from the `fixest` package (fast and flexible clustering specification)

# Outline

1. Motivation
2. Theory: Common trends and graphical intuition
3. Multiple time periods, pre-trends, and regression
4. Threats to identification
5. Triple differences
6. Staggered treatment timing
7. Synthetic DiD



# Main threats to DiD identification

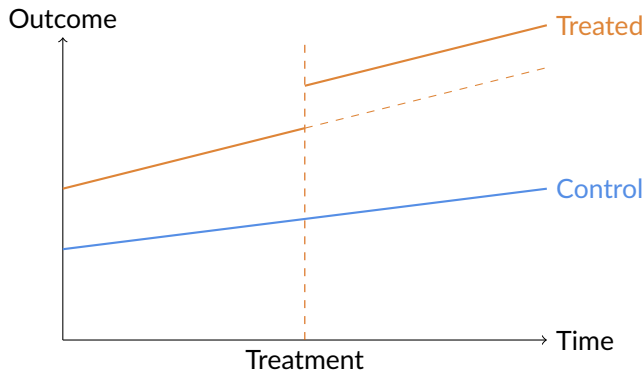
The parallel trends assumption can be violated in several ways:

- ① **Differential trends:** Treated and control groups on different trajectories
- ② **Differential shocks:** Time-varying shocks that affect groups differently
- ③ **Selection into treatment** (Ashenfelter's dip)
- ④ **Anticipation effects:** Behavioral responses before treatment
- ⑤ **Spillover effects:** Treatment affects control group
- ⑥ **Composition changes:** Different units in cross-sectional DiD
- ⑦ **Functional form:** Parallel trends in logs vs. levels

Let's discuss each in turn...

## Threat 1: Differential trends

**Problem:** Treated and control groups on systematically different trajectories



**Example:** Regions with strong economic growth more likely to get infrastructure investment

# Dealing with differential trends

## Solutions:

### 1. Group-specific linear trends:

$$Y_{it} = \alpha_i + \delta_t + \gamma_i \cdot t + \beta \cdot D_{it} + \varepsilon_{it} \quad (36)$$

- $\gamma_i$ : unit-specific linear time trend
- Allows for different slopes across units
- But: mechanically reduces post-treatment differences

### 2. Rambachan & Roth (2023) sensitivity analysis:

- Bound treatment effects under violations of parallel trends
- Assume trend violations can't be "too large"
- Provides robust confidence intervals

## Threat 2: Differential shocks

**Problem:** Time-varying shock affects treated and control groups differently

**Example:** Minimum wage study

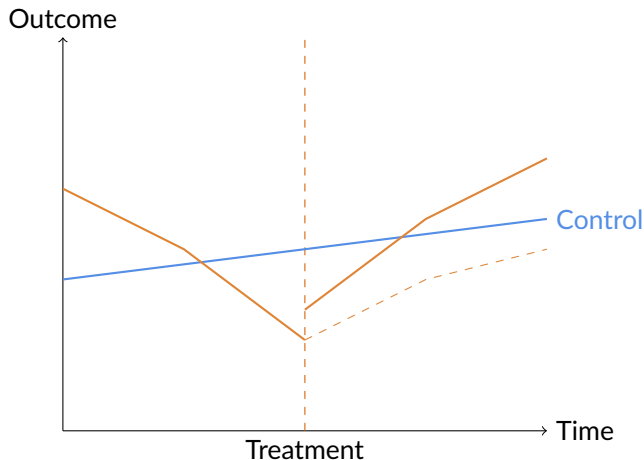
- NJ raises minimum wage; PA does not
- But NJ also experiences a state-specific recession
- Employment falls in NJ for two reasons: minimum wage + recession
- DiD incorrectly attributes recession effect to minimum wage

**Solutions:**

- Find better control groups (similar in all dimensions)
- Use multiple control groups to test robustness
- Look for placebo outcomes (unaffected by treatment)
- Triple differences (if another dimension available)

## Threat 3: Ashenfelter's dip

**Problem:** Units select into treatment precisely because they're on a downward trajectory



# Ashenfelter's dip: Evidence and solutions

## Names after Ashenfelter (1978):

- Found workers' earnings decline sharply before enrollment
- Then recover after training
- But hard to tell if recovery is due to training or mean reversion

**Critical point:** Ashenfelter's dip can be on **unobservables!**

- Even if pre-trends in observed outcomes look parallel, unobserved factors may differ
- Example: Workers enroll when motivation/health declines (unobservable)
- **Fundamentally untestable** — can't see unobservables in pre-period
- Clean pre-trends are reassuring but not definitive proof
- Threatens any setting where units select into treatment

# Solutions to Ashenfelter's dip

## Solutions:

- ① Look for the dip in pre-treatment data (event study)
  - If present in observables, likely worse in unobservables
- ② If present, focus on longer pre-treatment differences
- ③ Match treated units to controls experiencing similar pre-treatment trajectory
- ④ Use alternative control groups (e.g., future trainees)
- ⑤ **Best solution:** Find settings where treatment timing is plausibly exogenous
  - Randomization, policy changes, discontinuities
  - Removes selection-into-treatment concern

**Takeaway:** Be skeptical of DiD when treatment is chosen precisely when units need it most

## Threat 4: Anticipation effects

**Problem:** Units change behavior in anticipation of treatment

**Example:** Tax policy announced in advance

- Firms know corporate tax will increase next year
- Shift profits to current year to avoid higher future tax
- Pre-treatment profits artificially high
- DiD underestimates true revenue effect

**Detection:**

- Event study: look for effects in periods immediately before treatment
- "Leads" ( $k < 0$ ) should be zero under no anticipation

**Solutions:**

- Use earlier pre-period as baseline (before announcement)
- Model anticipation explicitly if timing is known



## Threat 5: Spillover effects

**Problem:** Treatment of one group affects outcomes in control group

**Example:** Job training program

- Treated workers become more productive
- Firms substitute away from untrained workers
- Control group employment falls
- DiD overestimates treatment effect (includes spillover)

**Solutions:**

- Choose geographically distant control groups
- Look for evidence of spillovers in untreated outcomes
- Model equilibrium effects explicitly (general equilibrium)
- Acknowledge limitation in interpretation

## Threat 6: Composition changes

**Problem:** In repeated cross-sections, different individuals in each period

**Example:** Regional minimum wage study

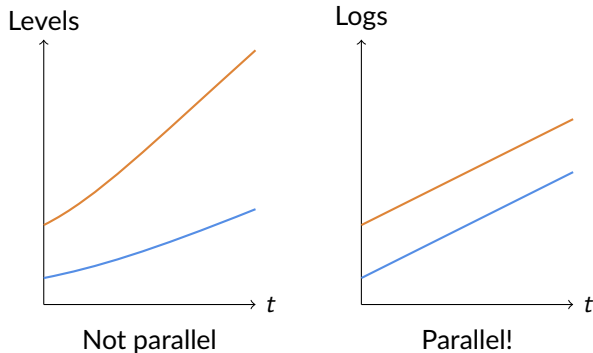
- High-wage workers migrate to treated region after treatment
- Average wage appears to increase
- But this is composition, not causal effect on incumbent workers

**Solutions:**

- ① Use panel data (follow same individuals)
- ② Test whether observable characteristics change
- ③ Control for composition using reweighting
- ④ Focus on intensive margin (hours) not extensive (employment)

## Threat 7: Functional form

**Problem:** Parallel trends may hold in one scale but not another



**Example:** Income growing at constant rates (parallel in logs, not levels)

# Functional form and DiD assumptions

**Key insight:** Every DiD design embeds a strong functional form assumption

The parallel trends assumption  $\mathbb{E}[Y_{it}(0) - Y_{it'}(0)|D_i = 1] = \mathbb{E}[Y_{it}(0) - Y_{it'}(0)|D_i = 0]$  is scale-dependent:

- If it holds in levels, it typically **does not hold** in logs
- If it holds in logs, it typically **does not hold** in levels
- Cannot be true in both scales simultaneously (unless constant trends)

**The question:** Which functional form assumption is the right one for your setting?

- **Levels:** Additive treatment effects ( $Y_{it}(1) = Y_{it}(0) + \tau$ )
- **Logs:** Proportional treatment effects ( $Y_{it}(1) = Y_{it}(0) \times (1 + \tau)$ )

# Beyond linear DiD: Change-in-changes

**Alternative:** Athey & Imbens (2006) **Changes-in-Changes** estimator

**Key idea:** Don't assume parallel trends in levels or logs. Instead:

- Allow for heterogeneous effects across the outcome distribution
- Make distributional assumptions rather than mean assumptions
- More flexible functional form

**Assumption:**

$$Y_{it}(0) = h_t(U_i) \quad (37)$$

where  $U_i$  is a time-invariant unobserved heterogeneity term, and  $h_t(\cdot)$  is a strictly increasing function that can vary over time

**Intuition:** Use the change in the distribution of control group outcomes to construct counterfactual distribution for treated group

# Why CIC is not commonly used in practice

## ① Sample requirements

- Quantile estimation is noisier than mean estimation
- Need larger samples for stable distributional estimates

## ② Still quite strong assumptions in practice

- The rank invariance assumption (individuals maintain their position in the distribution over time) is quite restrictive
- May be violated if there's genuine mobility in the outcome distribution
- Not obviously weaker than parallel trends in all applications

# Choosing the right functional form

## How to decide?

- ① **Economic theory:** Does the treatment have additive or multiplicative effects?
  - Tax policy: proportional (use logs)
  - Cash transfer: additive (use levels)
- ② **Pre-trends analysis:** Which scale shows flatter pre-trends?
  - If parallel in logs pre-treatment, assume parallel in logs post-treatment
  - But remember: not a perfect test (Roth 2022)
- ③ **Robustness:** Report results in multiple specifications
  - Levels, logs, changes-in-changes
  - If conclusions are robust, more credible
  - If sensitive, discuss why one specification is preferred
- ④ **Be explicit:** State which functional form you assume and why
  - Don't pretend it's a minor technical detail
  - Acknowledge this is a maintained assumption

# Outline

1. Motivation
2. Theory: Common trends and graphical intuition
3. Multiple time periods, pre-trends, and regression
4. Threats to identification
5. Triple differences
6. Staggered treatment timing
7. Synthetic DiD



## Triple differences (DDD)

**Motivation:** What if we're worried about differential shocks to treated vs. control?

**Idea:** Add a third dimension of differencing using an "unaffected" group

**Example:** Health insurance program for women

- Treatment: Some states introduce health insurance program for women only
- Control states: No program
- Concern: Treated states may have different macro trends (differential shocks)

**Solution:** Use men as an additional control group

- Men are not affected by the program (neither in treated nor control states)
- DiD on men captures differential macro shocks between states
- DiD on women captures differential shocks + treatment effect
- Triple difference =  $\text{DiD}_{\text{women}} - \text{DiD}_{\text{men}}$  isolates treatment effect

## Triple differences: Formula

Let:  $D_s = 1$  for treated state,  $= 0$  for control state

- $F_i = 1$  for female,  $= 0$  for male
- $t = 1$  (pre),  $t = 2$  (post)

**DiD for women:**

$$DiD_F = (\bar{Y}_{treated, female, post} - \bar{Y}_{treated, female, pre}) - (\bar{Y}_{control, female, post} - \bar{Y}_{control, female, pre}) \quad (38)$$

**DiD for men:**

$$DiD_M = (\bar{Y}_{treated, male, post} - \bar{Y}_{treated, male, pre}) - (\bar{Y}_{control, male, post} - \bar{Y}_{control, male, pre}) \quad (39)$$

**Triple difference:**

$$DDD = DiD_F - DiD_M \quad (40)$$

# Triple differences: Regression

Can implement via regression:

$$Y_{ist} = \alpha + \beta_1 D_s + \beta_2 F_i + \beta_3 Post_t \quad (41)$$

$$+ \beta_4 (D_s \times F_i) + \beta_5 (D_s \times Post_t) + \beta_6 (F_i \times Post_t) \quad (42)$$

$$+ \beta_7 (D_s \times F_i \times Post_t) + \varepsilon_{ist} \quad (43)$$

**Key coefficient:**  $\beta_7$  is the DDD estimator

**Assumption required:**

- Men and women in the same state are subject to the same differential shocks
- Common trends for men and women would have been parallel (in differences)

## Triple differences: When to use

**Advantages:** Differs out state-specific shocks that affect both genders

- More credible when worried about differential macro trends
- Provides robustness check even if not primary specification

**Disadvantages:**

- Requires finding a truly "unaffected" group
- Stronger assumptions (parallel trends for the difference-in-trends)
- Less precise (more differences = more noise)
- Spillovers to "unaffected" group would bias results

# Outline

1. Motivation
2. Theory: Common trends and graphical intuition
3. Multiple time periods, pre-trends, and regression
4. Threats to identification
5. Triple differences
- 6. Staggered treatment timing**
7. Synthetic DiD

## Moving beyond $2 \times 2$ : Staggered adoption

So far: single treatment period ( $t_0$ ), all treated units adopt simultaneously

**In reality:** Treatment often rolls out at different times

- States adopt policies in different years
- Firms receive treatment based on phased rollout
- Individuals age into eligibility at different times

**Benefits of staggered rollout:**

- ① More robust to macro shocks
  - Units treated at different times face different macro conditions
  - Differential shocks less likely to confound all comparisons
- ② Can use earlier-treated as controls for later-treated (and vice versa)

## Example: Yagan vs. Goodman-Bacon

Yagan (2015): State-level capital gains tax cuts and entrepreneurship

- All cuts happen in one year (1992)
- Control: States without cuts

**Problem:** What if 1992 is special?

- Maybe nationwide recession affects treated/control states differently
- Or tech boom affects entrepreneurship independent of taxes
- Hard to separate policy effect from concurrent macro shocks

**Better design:** Staggered rollout across years

- Some states cut in 1990, others in 1992, others in 1995...
- Macro shocks in different years unlikely to align with treatment
- More credible parallel trends assumption

# The problem with TWFE and staggered timing

**Historically:** Researchers used TWFE for staggered DiD

$$Y_{it} = \alpha_i + \delta_t + \beta \cdot D_{it} + \varepsilon_{it} \quad (44)$$

where  $D_{it} = \mathbb{1}(i \text{ has been treated by time } t)$

**Seemed reasonable:**

- $\alpha_i$  controls for unit fixed effects
- $\delta_t$  controls for common time shocks
- $\beta$  measures average treatment effect

**Recent discovery:** This doesn't work with:

- ① Staggered treatment timing, AND
- ② Heterogeneous treatment effects



# Goodman-Bacon decomposition (2 treatment times)

**Goodman-Bacon (2021):** Special case with 2 treatment times

With two treatment cohorts (early and late) plus never-treated, TWFE is a weighted average of three  $2 \times 2$  comparisons:

① **Earlier-treated vs. never-treated**

- Weight: variance share of this comparison
- Sign: positive (good comparison)

② **Later-treated vs. never-treated**

- Weight: variance share
- Sign: positive (good comparison)

③ **Later-treated vs. earlier-treated**

- Weight: variance share
- Sign: **can be negative** (forbidden comparison!)
- Earlier-treated serves as "control" for later-treated
- But earlier-treated is already experiencing treatment effects

- Consider two staggered treatments and a never-treated group
- What does the TWFE estimator estimate?
- TWFE decomposes into all possible  $2 \times 2$  comparisons

## 67 / 99

# The forbidden comparison problem

- Weighting becomes problematic if effects vary over time
- With time-varying effects, already-treated units are bad controls
- This creates **negative weights** on some treatment effects
- Goodman-Bacon decomposition reveals how much weight is problematic

Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case

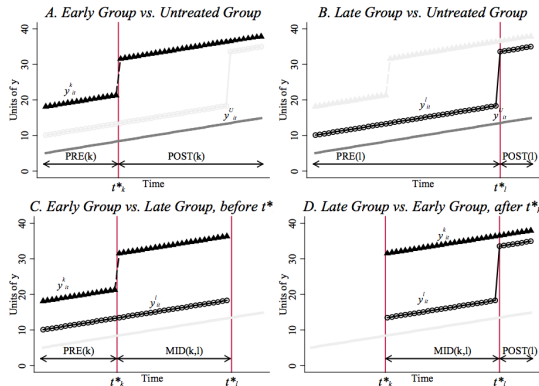
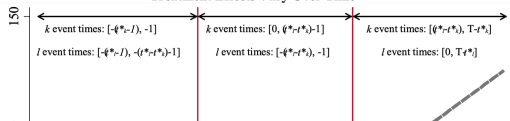


Figure 3. Difference-in-Differences Estimates with Variation in Timing Are Biased When Treatment Effects Vary Over Time



## Heterogeneous treatment effects: 3-group example

**Setup:** Three groups with parallel trends, 3 time periods ( $t = 1, 2, 3$ )

**Group E:** Treated early at  $t = T_E = 2$

- Untreated outcome:  $Y_{Et}(0) = \alpha_E + g(t)$

**Group L:** Treated late at  $t = T_L = 3 > T_E$

- Untreated outcome:  $Y_{Lt}(0) = \alpha_L + g(t)$  (parallel trends: same  $g(t)$ )

**Group C:** Never treated (control)

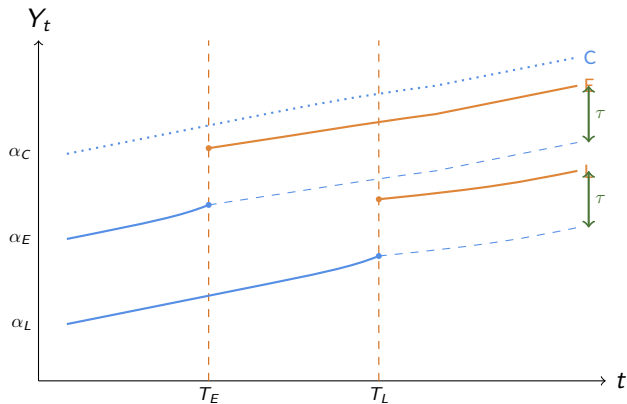
- Untreated outcome:  $Y_{Ct}(0) = \alpha_C + g(t)$  (parallel trends)

**Potential source of problems:** Treatment effects are **dynamic** (grow with exposure):

- Effect is  $\tau_l$  in first period after treatment
- Effect grows to  $\tau_h$  in second period after treatment (where  $\tau_h > \tau_l$ )
- **Problem:** Group E has been treated longer at  $t = 3$  than Group L



## With constant treatment effects: $\tau$ (no dynamics)



Constant treatment effects:  $\tau$  remains the same over time (no problem for TWFE)

## Why constant effects are not a problem

**Special case:** Suppose treatment effect is constant  $\tau$  (no dynamics)

With constant effects, Group E has effect  $\tau$  at both  $T_E$  and  $T_L$ .

**The forbidden comparison now gives:**

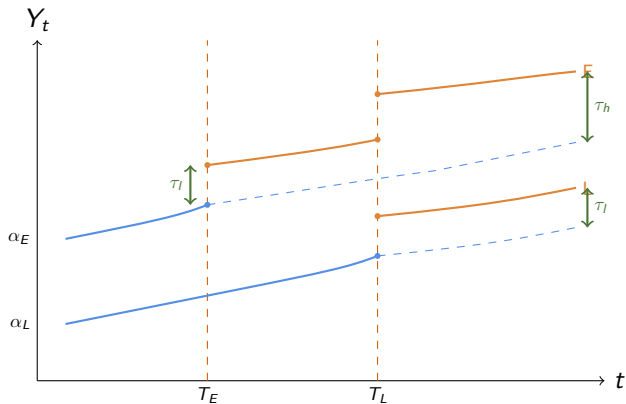
$$\begin{aligned}
 \hat{\tau}_{L \text{ vs } E}^{DID} &= [Y_{L, T_L} - Y_{L, T_E}] - [Y_{E, T_L} - Y_{E, T_E}] \\
 &= [(\alpha_L + g(T_L) + \tau) - (\alpha_L + g(T_E))] \\
 &\quad - [(\alpha_E + g(T_L) + \tau) - (\alpha_E + g(T_E) + \tau)] \\
 &= [g(T_L) - g(T_E) + \tau] - [g(T_L) - g(T_E)] \\
 &= \tau
 \end{aligned}$$

**Key insight:** With constant effects, all valid comparisons give the same answer  $\tau$

- No contamination from using already-treated units as controls
- TWFE works fine when treatment effects don't change over time



With dynamic treatment effects:  $\tau_l < \tau_h$



Dynamic treatment effects: Effects grow from  $\tau_l$  to  $\tau_h$  with exposure (problem for TWFE!)

## What TWFE estimates with dynamic effects

The TWFE estimator uses Group E as a control for Group L at  $t = T_L$

But at  $t = T_L$ :

- Group E has been treated for 2 periods  $\Rightarrow$  effect is  $\tau_h$
- Group L just got treated  $\Rightarrow$  effect is  $\tau_l$

**TWFE's implicit estimate from this comparison:**

$$\begin{aligned}
 \hat{\tau}_{L \text{ vs } E}^{TWFE} &= [Y_{L, T_L} - Y_{L, T_E}] - [Y_{E, T_L} - Y_{E, T_E}] \\
 &= [\alpha_L + g(T_L) + \tau_l - \alpha_L - g(T_E)] \\
 &\quad - [\alpha_E + g(T_L) + \tau_h - \alpha_E - g(T_E) - \tau_l] \\
 &= \tau_l - (\tau_h - \tau_l) \\
 &= 2\tau_l - \tau_h
 \end{aligned}$$

# What TWFE estimates with dynamic effects

$$\hat{\tau}_{L \text{ vs } E}^{TWFE} = 2\tau_l - \tau_h$$

**Problem:** If  $\tau_l < \tau_h$ , this can give a **negative** estimate!

- Even though treatment has a positive effect at all horizons
- TWFE uses already-treated Group E (with large effect) as "control"
- Contaminates the estimate with **heterogeneity over time**:  $2\tau_l - \tau_h$
- Really we care about the ATT or ATE  $\approx (\tau_l + \tau_h)/2$ . At the very least,  $\tau_h$  should not be counting negatively towards our estimate of treatment effects!

## General case: Many groups, many time periods

de Chaisemartin & D'Haultfoeuille (2020): Extends to general staggered timing

With many treatment cohorts and many time periods, TWFE estimates:

$$\hat{\beta}^{TWFE} = \sum_{g,t} w_{g,t} \cdot ATT_{g,t} \quad (45)$$

- $ATT_{g,t}$ : average treatment effect for cohort  $g$  at time  $t$
- $w_{g,t}$ : weight on this effect (depends on treatment variance)

**Problem:** Some weights  $w_{g,t}$  can be **negative**!

**Implication:**

- $\hat{\beta}^{TWFE}$  can be negative even if all  $ATT_{g,t} > 0$
- Or vice versa:  $\hat{\beta}^{TWFE} > 0$  even if all  $ATT_{g,t} < 0$
- Cannot interpret  $\hat{\beta}^{TWFE}$  as a meaningful average
- Their diagnostic tool shows how much negative weight in your data

# Solutions to staggered timing problem

Don't use standard TWFE with staggered timing + heterogeneous effects!

Alternative estimators:

① Callaway & Sant'Anna (2021):

- Compute clean  $2 \times 2$  DiDs for each cohort-time pair
- Only use never-treated or not-yet-treated as controls
- Aggregate using explicit weights

② de Chaisemartin & D'Haultfoeuille (2020):

- Similar approach: avoid forbidden comparisons
- Provides diagnostic for negative weights in your data
- R package: `DIDmultiplegt`

③ Sun & Abraham (2021):

- Event-study approach with interaction-weighted estimator
- Clean estimates of dynamic effects by cohort

# Callaway & Sant'Anna (2021) estimator

**Basic approach:** Explicitly **exclude** any forbidden comparisons.

- ① Define cohorts by treatment timing:  $g \in \{2, 3, \dots, T, \infty\}$ 
  - $g = t$  if unit first treated at time  $t$
  - $g = \infty$  if never treated
- ② For each cohort  $g$  and time  $t \geq g$ , compute:

$$ATT(g, t) = \mathbb{E}[Y_t - Y_{g-1} | G_i = g] - \mathbb{E}[Y_t - Y_{g-1} | G_i = C_t] \quad (46)$$

where  $C_t$  is comparison group (never-treated or not-yet-treated at  $t$ )

- ③ Aggregate across cohorts and times:

$$ATT_{overall} = \sum_{g,t} w_{g,t} \cdot ATT(g, t) \quad (47)$$

with explicit, non-negative weights  $w_{g,t}$ . These could be e.g. proportional to the number of units in each cell.

**Key advantage:** Transparent about what's being compared and weighted.

## Note: Fuzzy DiD and partial treatment

**Important:** The staggered timing problem also applies to **fuzzy DiD**

**Fuzzy DiD:** When the "control" group is partially treated

- Treatment occurs in the treatment group, but also (to lesser extent) in control group
- Example: Policy rollout affects neighboring regions
- Example: Media coverage spills over to control areas

**Why this matters:**

- If control group has small treatment effect  $\tau_C > 0$
- And treatment group has larger effect  $\tau_T > \tau_C$
- Standard DiD estimates  $\tau_T - \tau_C$ , not  $\tau_T$
- This is the same issue as forbidden comparisons!
- Comparing "more treated" vs. "less treated" rather than "treated" vs. "untreated"

**Takeaway:** Be careful about control group contamination and partial treatment

# Practical recommendations

If you have staggered treatment timing:

## ① Check for heterogeneity:

- Run event study: do effects vary across cohorts or over time?
- If yes, standard TWFE is problematic

## ② Use diagnostic tools:

- `bacon` package (Goodman-Bacon decomposition)
- `DIDmultiplgt` (de Chaisemartin & D'Haultfoeuille)
- Check for negative weights in your data

## ③ Use robust estimators:

- Callaway & Sant'Anna: R package `did`
- Sun & Abraham: Stata package `eventstudyinteract`
- Report both TWFE and robust estimator for comparison

## ④ Be transparent:

- Document which comparisons are being made
- Show event studies by cohort if heterogeneity is present
- Discuss sensitivity to choice of comparison group



# Outline

1. Motivation
2. Theory: Common trends and graphical intuition
3. Multiple time periods, pre-trends, and regression
4. Threats to identification
5. Triple differences
6. Staggered treatment timing
7. Synthetic DiD

# Synthetic control methods

**Motivation:** What if parallel trends doesn't hold for all control units?

**Key idea:** Construct a **weighted combination** of control units that best matches the treated unit pre-treatment:

$$\tau = \underbrace{Y_{post}(1)}_{\text{Fully observed}} - \underbrace{\hat{Y}_{post}(0)}_{\text{Constructed}} \quad (48)$$

## Synthetic Control example (Abadie et al., 2010))

- Consider following problem: California bans smoking in 1989. What does that do to smoking?
  - Define estimand:
$$\tau_{ban, CA} = Y_{california, post}(1) - Y_{california, post}(0)$$
  - This is the effect of the California smoking ban
  - How can we get at it?
- We need a “synthetic California” as our control
- In an ideal world, the average of the other states would work – however, not clear empirically that they are a good counterfactual

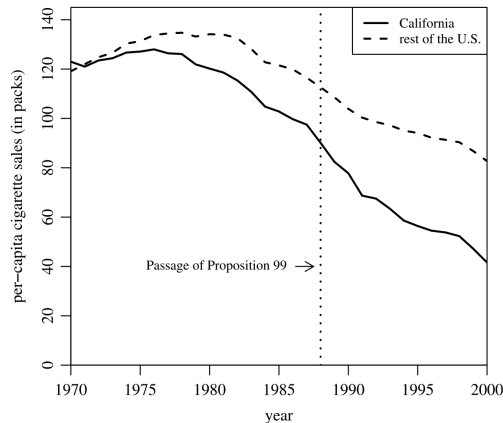


Figure 1. Trends in per-capita cigarette sales: California vs. the rest of the United States.

## Synthetic control: basic method

Method (Abadie et al., 2010):  $\sum_j \omega_j = 1$

- Estimate counterfactual untreated California using a weighted sum of other states that “look like” California (the synthetic California)

$$\hat{Y}_{\text{post,treated}}(0) = \sum_j \omega_j Y_{\text{post,control}}$$

- Select weights to make minimize the distance in terms of pre-treatment covariates:

$$\{\hat{\omega}\}_i = \arg \min_{\mathbf{W}} \|\mathbf{X}_{\text{treat}} - \mathbf{X}_{\text{control}} \mathbf{W}\|$$

- This approach can be incredibly

- This approach can be incredibly successful
- By careful construction of a synthetic control, can calculate counterfactual impacts due to policy
- Still subject to same caveats from DiD – not invariant to some transformations (e.g. log and linear)

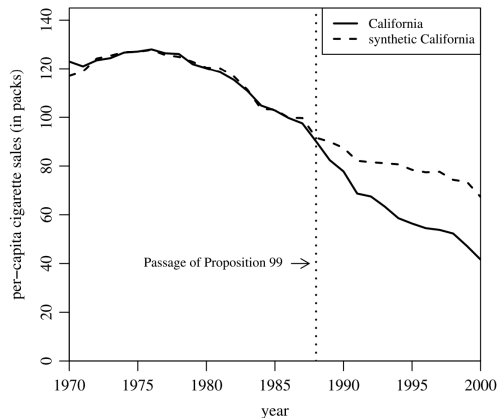
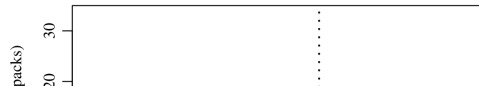


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.



**Arkhangelsky et al. (2021):** Combine unit weights and time weights

- **Unit weights** ( $\omega_i$ ): reweight controls so that their pre-outcomes match the treated units' pre-outcomes (same as synthetic control)
- **Time weights** ( $\lambda_t$ ): reweight time so that for the controls, the pre-period looks like the post period
  - This soaks up aggregate trends
  - E.g., if there was a macroeconomic recession in the post-period, want to weight pre-periods more if there was a recession in those periods

Then do a **DiD** on this reweighted setup.

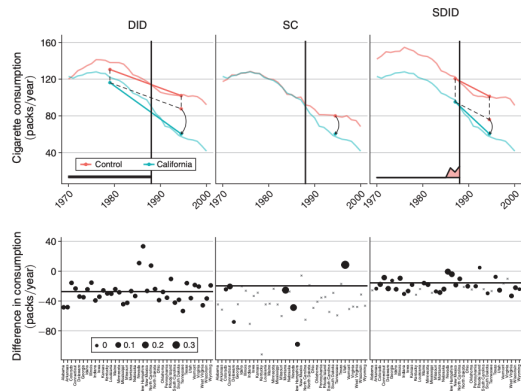


FIGURE 1. A COMPARISON BETWEEN DID, SC, AND SDID ESTIMATES FOR THE EFFECT OF CALIFORNIA PROPOSITION 99 ON PER-CAPITA ANNUAL CIGARETTE CONSUMPTION (IN PACKS/YEAR)

# Summary: The DiD toolkit

**Basic DiD (2x2 design):** Parallel trends assumption:  $\mathbb{E}[Y_{it}(0) - Y_{it'}(0)|D_i = 1] = \mathbb{E}[Y_{it}(0) - Y_{it'}(0)|D_i = 0]$

- Identifies ATT under parallel trends
- Can implement via regression:  $Y_{it} = \alpha_i + \delta_t + \beta \cdot D_{it} + \varepsilon_{it}$

## Key threats:

- Differential trends, differential shocks, Ashenfelter's dip
- Anticipation, spillovers, composition, functional form

## Extensions:

- Triple differences for additional robustness
- Event studies for dynamic effects (with caution on pre-trends)

## Summary: Recent developments

### Standard TWFE timing: negative weights with heterogeneous effects

- “Forbidden comparisons”: later-treated vs. already-treated
- **Solutions:** Callaway & Sant’Anna (2021), de Chaisemartin & D’Haultfoeuille (2020), Sun & Abraham (2021)

### Synthetic control methods:

- When parallel trends may not hold for all controls
- Construct weighted combination matching pre-treatment characteristics
- Transparent, data-driven approach to control group construction
- Best for few treated units with rich pre-treatment data

**Key takeaway:** Choice of method depends on your setting, data structure, and assumptions you’re willing to make



# Practical advice

## ① Always visualize your data:

- Plot trends for treatment and control groups
- Show event studies (but interpret pre-trends carefully)
- Make the parallel trends assumption transparent

## ② Be honest about threats:

- Discuss potential violations of identifying assumptions
- Show robustness checks (functional form, sample restrictions, etc.)
- Consider alternative explanations

## ③ With staggered timing:

- Check for heterogeneity across cohorts/time
- Use diagnostic tools (Goodman-Bacon decomposition)
- Report both TWFE and robust estimators

## ④ Document your choices:

- Which comparison groups are being used
- How you handle standard errors (clustering level)
- Sensitivity to key decisions

# Thank you!

Questions?

## Generalized panel setup

Consider a panel with  $T$  time periods and  $N + 1$  units. Intervention  $D_{it}$  at time  $T_0$  for one unit (unit  $i = 0$ ).

Let  $\mathbf{Y}_{a,b}$  denote outcomes for  $a \in \{\text{treated, control}\}$  and  $b \in \{\text{pre, post}\}$ :

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_{t,\text{post}} & \mathbf{Y}_{c,\text{post}} \\ \mathbf{Y}_{t,\text{pre}} & \mathbf{Y}_{c,\text{pre}} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_{t,\text{post}}(1) & \mathbf{Y}_{c,\text{post}}(0) \\ \mathbf{Y}_{t,\text{pre}}(0) & \mathbf{Y}_{c,\text{pre}}(0) \end{pmatrix}$$

**Key insight:** We need to estimate  $\mathbf{Y}_{t,\text{post}}(0)$ , the counterfactual for the treated unit(s) in the post period.

# Synthetic DiD: The estimator

**Standard DiD:**

$$(\hat{\alpha}, \hat{\gamma}, \hat{\tau}) = \arg \min \sum_{i,t} (Y_{it} - \alpha_i - \gamma_t - D_{it}\tau)^2$$

**Synthetic Control:**

$$(\hat{\gamma}, \hat{\tau}) = \arg \min \sum_{i,t} (Y_{it} - \gamma_t - D_{it}\tau)^2 \hat{\omega}_i$$

where  $\hat{\omega}_i$  chosen to match pre-treatment characteristics

**Synthetic DiD:**

$$(\hat{\alpha}, \hat{\gamma}, \hat{\tau}) = \arg \min \sum_{i,t} (Y_{it} - \alpha_i - \gamma_t - D_{it}\tau)^2 \hat{\omega}_i \hat{\lambda}_t$$

where both  $\hat{\omega}_i$  (unit weights) and  $\hat{\lambda}_t$  (time weights) are data-driven

# Generalized estimator form

Consider estimators of the form:

$$\hat{Y}_{t,\text{post}}(0) = \mu + \sum_{j \in \text{controls}} \omega_j Y_{j,T}$$

## Components:

- $\mu$ : Constant allowing for level differences (common in DiD)
- $\omega_j$ : Weights that vary across control units
  - Simple average would be standard DiD
  - Different weights allow more flexibility

**Question:** How should we choose the weights  $\omega_j$ ?

# Synthetic control weight restrictions

Abadie, Diamond, Hainmueller (2010) impose three restrictions:

- ①  $\mu = 0$  (no intercept)
- ②  $\sum_j \omega_j = 1$  (weights sum to one)
- ③  $\omega_j \geq 0 \forall j$  (non-negative weights)

## Interpretation:

- These create a counterfactual whose outcomes are **within the convex hull** of control units
- Treated unit is a weighted average of a subset of control states
- More transparent than allowing negative weights or extrapolation

## Formal weight estimation

Weights  $\omega_j$  are chosen by minimizing distance between covariates in pre-period:

$$\{\hat{\omega}_j\}_j = \arg \min_{\mathbf{W}} \|\mathbf{X}_{\text{treat}} - \mathbf{X}_{\text{control}} \mathbf{W}\|$$

subject to  $\sum_j \omega_j = 1$  and  $\omega_j \geq 0$ .

**Crucial feature:**  $\mathbf{X}$  can include:

- Lagged outcomes:  $Y_{i,t-1}, Y_{i,t-2}, \dots$
- Time-invariant covariates: demographics, geography, etc.
- Time-varying covariates

Re-envision the panel:

- **Observed outcomes:**  $\mathbf{Y}_{t,\text{post}}(1), \mathbf{Y}_{c,\text{post}}(0)$
- **Covariates/predictors:**  $\mathbf{Y}_{t,\text{pre}}(0), \mathbf{Y}_{c,\text{pre}}(0), \mathbf{X}_t, \mathbf{X}_c$

## Inference with synthetic control

**Challenge:** With only one treated unit, standard large-sample asymptotics don't apply.

**Standard approach:** Placebo/permutation tests

- Apply synthetic control method to **each potential control unit**
- Compute “placebo effects” for untreated units
- Compare actual treatment effect to distribution of placebo effects
- Similar to randomization inference

**Interpretation:**

- If treatment effect is large relative to placebos  $\Rightarrow$  evidence of real effect
- If treatment effect is in middle of distribution  $\Rightarrow$  could be noise



## Staggered adoption with synthetic DiD

**Issue:** Staggered adoption isn't as natural for synthetic control

- How can we adapt it?

**Solution** (following Callaway & Sant'Anna approach):

- Split up adoption timings by cohort
- Estimate synthetic DiD separately for each  $(g, t)$  pair
  - $g$  = adoption cohort
  - $t$  = time period
- Aggregate cohort-time effects

**Advantage:** Allows for heterogeneous treatment effects across cohorts and time while maintaining synthetic control benefits

# Practical considerations and skepticism

## Why limited adoption despite being “cool”?

### Challenges:

- Strong structural assumptions
  - Not clear we have good tests yet
  - Pre-trends in DiD felt more testable/transparent
- Researcher degrees of freedom
  - Choice of covariates to match on
  - Which control units to include
  - How to weight different matching variables
  - True in DiD too, but perhaps less transparent?

### Alternative interpretation:

- Maybe DiD is equally problematic, but we're not aware of it
- If we accept DiD is sensitive to functional form, then ML methods that construct counterfactuals are natural

## Practical recommendations

• **Ideal:** Single treatment event (“big bang”)

### When to use synthetic control:

- Get a good synthetic control for treatment unit
- If no good match exists in pre-period, stop (or adjust)
- **Better approach:** Ben-Michael, Feller & Rothstein (2021) adjust for imperfect pre-match

### When to use synthetic DiD:

- Promising generalization for multiple treated units
- Key challenge: Convince readers why this works better than traditional DiD
- Recommendation: Show results with **both** DiD and synthetic DiD

### Software packages:

- `augsynth`: Augmented synthetic control (Ben-Michael et al.)
- `synthdid`: Synthetic DiD (Arkhangelsky et al.)
- `tidysynth`: User-friendly synthetic control