Modern Difference-in-Differences (DiD) Approaches

Callaway & Sant'Anna (2020)

Justin Eloriaga

October 31, 2025

Emory University

Roadmap

Motivation

Data Setup

Benchmark: TWFE

Modern DiD: Callaway & Sant'Anna (2020)

Estimation

Aggregating the group-time effects

Adding Covariates

Final Thoughts

Motivation

Why Difference-in-Differences?

- We often want to know: **what was the effect of a policy?**
- But we only observe outcomes *with* the policy, not the counterfactual.
- DiD compares (i) people/places that get treated to (ii) people/places that don't, **before and after** the policy.
- Classic setup:
  - If treated and control move *similarly* before the policy,
  - then any extra change for the treated group *after* the policy is a good candidate for the treatment effect.

Overview of the Data

- Uses **U.S. minimum wage data** (2001–2007) as in Callaway (2022).
- Some states/counties raise the minimum wage earlier, others later, some never.
- We compare teen employment outcomes.
- We start from the “old” DiD (two-way fixed effects) as a benchmark.
- Then we move to **modern** DiD: Callaway and Sant’Anna (2020) (CS, 2020).

Data Setup

Loading the data (Python part)

Variables

- Outcome $=$ `lemp` (log employment), group G = first year treated, id, year, minimum wage vars.

Intuition
Each county i is followed over time t. Some counties adopt higher minimum wages earlier (treated earlier), some later, some never. This is a textbook **staggered adoption** setting.

Panel view of the data

- This plot is only for **understanding** the timing and overlap. Clearly, **not everyone is treated at the same time.**

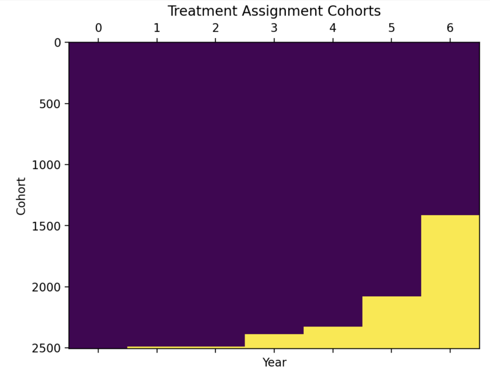

Figure 1: Who is treated in which year.

Benchmark: TWFE

The "old" way: Two-Way Fixed Effects (TWFE)

- Classic DiD regression:

$$Y_{it} = \alpha_i + \lambda_t + \beta D_{it} + \varepsilon_{it},$$

  where:
  - α_i: unit fixed effects (control for time-invariant differences across counties)
  - λ_t: time fixed effects (common shocks)
  - D_{it}: indicator that county i is treated at time t
- We "still" run this even though we know it can be problematic
- **Why?** Because it gives a baseline to compare to modern methods.

What's the problem with TWFE in staggered adoption?

- When treatment happens at **different times**, some treated units get used as controls for other treated units.
- If treatment effects are **not the same** across groups or over time, TWFE can put **negative weights** on some comparisons.
- In other words...
  - We meant to compare treated to **never** treated.
  - But with staggered timing, we accidentally compare treated to **already-treated** groups.
  - That can push the estimated effect up or down in weird ways.
- So we want something that **respects the timing**.

Modern DiD: Callaway & Sant'Anna (2020)

Key idea of Callaway & Sant'Anna (2020)

- Instead of forcing one regression to explain everything,
- they estimate **group-time average treatment effects**:

$$ATT(g, t) \quad = \text{effect for units first treated in } g\text{, evaluated at time } t.$$

- Example: effect for counties treated in 2004, measured in 2006.
- This respects:
  - **when** you got treated (your group g)
  - **when** we are measuring the effect (t)

Assumptions

1. No anticipation
Counties don't change behavior *before* they actually get the higher minimum wage.

2. Parallel trends (within groups)
If a county treated in 2005 had *not* been treated, its outcome would have moved over time like the comparison group we chose.

- CS let us pick **which** comparison group to use (never-treated vs not-yet-treated).
- That's powerful in staggered adoption.

Two main comparison strategies in CS (2020)

1. Compare treated group to **never-treated** units.
2. Compare treated group to **not-yet-treated** units (i.e. people who will be treated later, so they look more similar).

Why this matters
Using units that are "closer" in time to treatment often gives **better counterfactuals**, especially in policy diffusion settings.

Estimation

Python side (conceptual)

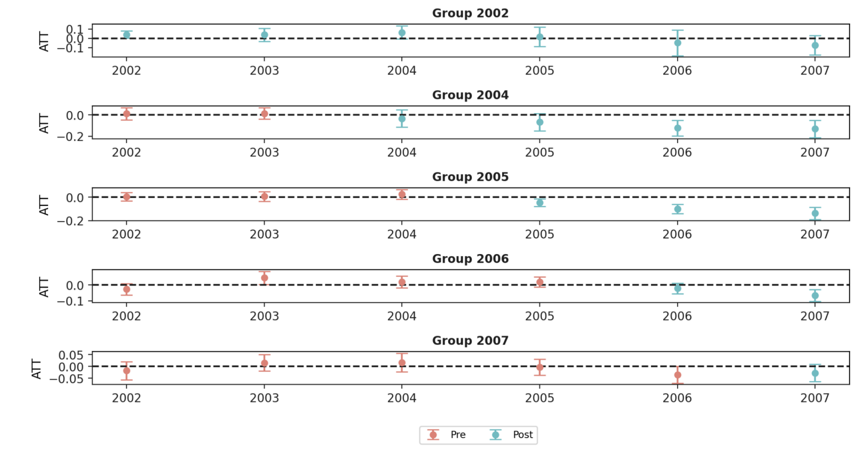

Figure 2: Event Study

Interpreting the event-study figure

- Each dot = an estimate of the effect for **a group at a time**.
- Red dots in the notebook = **pre-treatment** pseudo effects (should be near 0 if parallel trends is OK).
- If pre-period dots are flat ≈ 0, that supports the design.
- Post-period dots below 0 $\Rightarrow$ raising the min wage **reduced** teen employment.

Aggregating the group-time effects

Why aggregate?

- $ATT(g, t)$ is super detailed. Great for researchers.
- But the main policy question we want: **"So what's the effect?"**
- CS provide ways to **average** the $ATT(g, t)$ to get:
  - overall ATT (for all treated units),
  - dynamic ATT (by time since treatment),
  - group ATT (average effect for each cohort).

Simple overall average

- Take a **weighted average** of all the $ATT(g, t)$ that correspond to post-treatment periods.
- This is what the notebook calls the "simple" aggregation.
- Intuition: "On average, across all counties and years when they were treated, what was the effect?"
- In the minimum wage app, this tends to show a **negative** effect on teen employment.

```
    ATT Std. Error  [95.0%  Conf. Int.]
-0.0501     0.0074 -0.0646      -0.0356 *


---
Signif. codes: `*' confidence band does not cover 0
Control Group:  Never Treated ,
Anticipation Periods:  0
Estimation Method:  Doubly Robust
```

But: simple averages can over-weight early treated groups

- Groups treated early appear in **more** post-treatment periods.
- So they get more weight in the simple average.
- CS propose alternative aggregations (e.g. group-specific averages) that balance this out.

```
Overall summary of ATT's based on group/cohort aggregation:
    ATT Std. Error   [95.0%  Conf. Int.]
-0.0399     0.0085 -0.0567      -0.0232 *


Group Effects:
   Group  Estimate  Std. Error  [95.0% Simult.   Conf. Band
0   2002    0.0075      0.0256          -0.0428      0.0577
1   2004   -0.0888      0.0217          -0.1314     -0.0463  *
2   2005   -0.0937      0.0117          -0.1165     -0.0708  *
3   2006   -0.0439      0.0101          -0.0636     -0.0241  *
4   2007   -0.0271      0.0111          -0.0488     -0.0053  *
---
Signif. codes: `*' confidence band does not cover 0
Control Group:  Never Treated ,
Anticipation Periods:  0
Estimation Method:  Doubly Robust
```

Weighted Groups

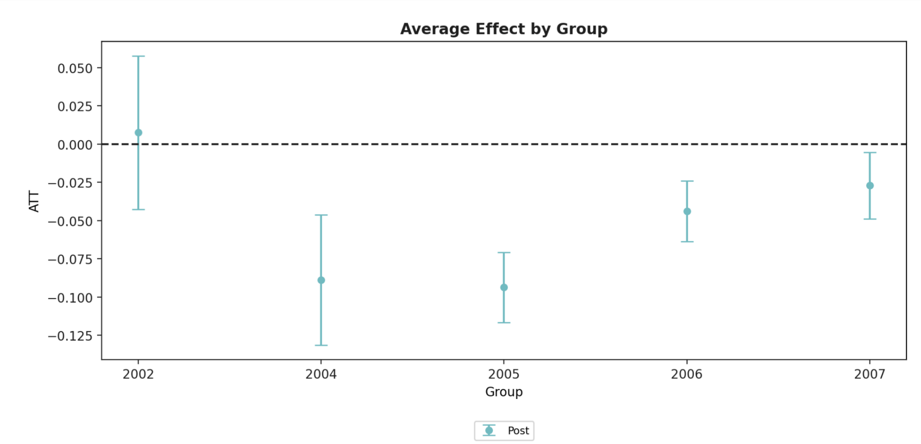

Figure 3: ATT by Groups

Group-specific effects

- That gives, for each treatment cohort (e.g. treated in 2004, 2005, 2006, 2007), its own average treatment effect.
- In the minimum wage data:
    - some cohorts show a stronger negative effect,
    - others a milder one.
- This is exactly why TWFE can go wrong: **effects differ by cohort.**

Adding Covariates

Adding covariates

- What if we control for things like population, average pay, etc.?
- In CS-style estimators you can add a formula like

$$Y \sim \text{treatment} + \text{covariates}$$

- Purpose: tighten up the parallel trends assumption by comparing **more similar** treated and control units.

```
Overall summary of ATT's based on group/cohort aggregation:
    ATT Std. Error  [95.0%  Conf. Int.]
-0.0399     0.0082 -0.0559       -0.024 *


Group Effects:
   Group  Estimate  Std. Error  [95.0% Simult.   Conf. Band
0   2002    0.0075      0.0260         -0.0435      0.0585
1   2004   -0.0888      0.0231         -0.1341     -0.0436  *
2   2005   -0.0937      0.0109         -0.1151     -0.0723  *
3   2006   -0.0439      0.0096         -0.0626     -0.0251  *
4   2007   -0.0271      0.0117         -0.0500     -0.0042  *
---
Signif. codes: `*' confidence band does not cover 0
Control Group:  Never Treated ,
Anticipation Periods:  0
Estimation Method:  Doubly Robust
```

With covariates

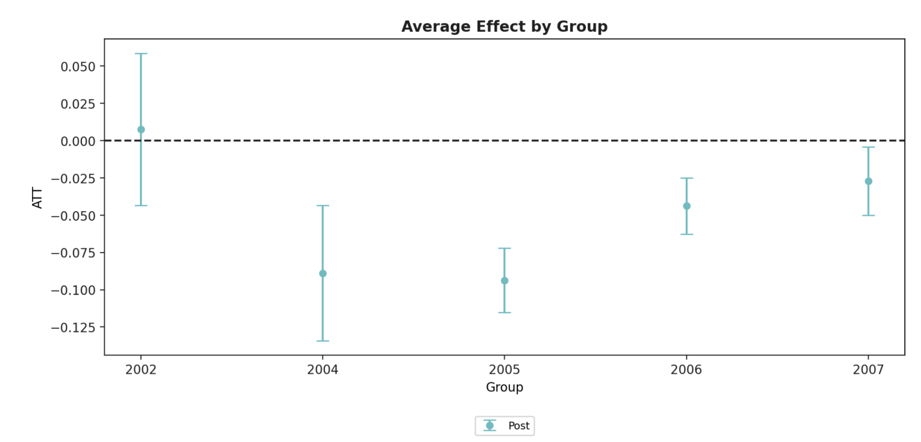

Figure 4: ATT with Coveriates

Changing the control group

- What about `control_group="notyettreated"`.
- That means: for a group treated in year g, compare them to units that are still untreated in year t (but may be treated later).
- Intuition:
  - Comparing 2005-treated counties to 2006-treated counties **in 2005** is often better than comparing to counties that will *never* raise their min wage.
- If results don't change much across these choices, that **stabilizes** the conclusion.

Final Thoughts

Economic story to tell in class

- Policy: raising the minimum wage.
- Theory: higher wages can reduce teen employment if teen labor demand is elastic.
- Evidence from modern DiD:
  - once we cleanly compare each treated cohort to the right control group,
  - we still see mostly **negative** effects on teen employment,
  - and the pattern is **similar** even when we switch to multiple comparison groups or add covariates.
- So the conclusion is **not** an artifact of messy DiD.

Main takeaways

- Staggered adoption + TWFE $\Rightarrow$ can mislead.
- Modern DiD (CS 2020) fixes this by estimating $ATT(g,t)$.
- Then we aggregate in ways that do **not** over-weight early treated groups.
- In this notebook's application (minimum wage $\rightarrow$ teen employment), the main message **survives** across choices:
  - multiple comparison groups,
  - adding covariates,
  - alternative aggregations.