Lecture 07: Difference in Differences (DiD)

Applied Econometrics

Naser Amanzadeh

Graduate School of Management and Economics (GSME)
Sharif University of Technology

Spring 2025

Outline

1. Introduction
2. Simple DiD
   - Example 1: Cholera in London
   - Example 2: Minimum Wage
   - Example 3: School Construction
3. Extensions of Simple DiD
   - Event Studies
   - DDD
4. Multiple Periods and Staggered Treatment Timing
5. Violations of Parallel Trends

Table of Contents

1. Introduction
2. Simple DiD
   - Example 1: Cholera in London
   - Example 2: Minimum Wage
   - Example 3: School Construction
3. Extensions of Simple DiD
   - Event Studies
   - DDD
4. Multiple Periods and Staggered Treatment Timing
5. Violations of Parallel Trends

Introduction

- Endogeneity is a serious threat to causal identification.
- Sometimes the legal/administrative context of policy implementation provides quasi-random variation.
- Examples:
  1. Changing the water supply from a dirty to a clear source
  2. School construction program
- Are there some good exogenous variation in assigning an explanatory variable (Natural Experiment or quasi-experiment)?
  - Idea 1: Are there policy changes that apply to certain groups but not others?
  - Idea 2: Are there reasons to believe that a given policy should have a greater impact on a group of individuals?
- This lecture is on one of the strategies called Difference in Differences (DiD)

Difference-in-Differences (DiD)

- **Difference-in-differences** is one of the most popular strategies for estimating causal effects in non-experimental contexts.
    - Used in over 20% of NBER WPs (Currie et al., 2020)
- The last few years have seen an explosion of econometrics on DiD, making it hard to keep updated.
- Roth and Sant'Anna (2023), attempt to synthesize the recent literature and provide concrete recommendations for practitioners
- In this lecture, first we review the simplest DiD case or canonical model and its two extensions, Event Studies and DDD.
- Then, we focus on the **new DiD literature** loosely based on the structure in Roth and Sant'Anna (2023), focusing on staggered timing (Section 3) and violations of parallel trends (Section 4)

Comparison of Quasi-Experimental Methods

Method	**Main Assumption**	**Data Needed**	**When to Use?**
DiD	Parallel trends	Panel/repeated cross-section	Pre- and post-policy evaluations
RD	Continuity at cutoff	Data around a threshold	Rule-based treatments, e.g., policy cutoffs
Synthetic Control	Valid counterfactual	Treated and potential controls	Macro-level impacts, regional policies
IV	Exclusion restriction	Valid instrument	Non-random assignment with endogeneity

Table of Contents

1. Introduction

2. Simple DiD
   - Example 1: Cholera in London
   - Example 2: Minimum Wage
   - Example 3: School Construction

3. Extensions of Simple DiD
   - Event Studies
   - DDD

4. Multiple Periods and Staggered Treatment Timing

5. Violations of Parallel Trends

The simplest case

- We will start a description of DiD in the simplest “canonical” case
- Why? Because recent DiD lit can be viewed as relaxing various components of the canonical model while preserving others

In the canonical DiD model, we have:

- 2 periods: treatment occurs (for some units) in period 2
- Identification of the ATT from parallel trends and no anticipation
- Estimation using sample analogs, equivalent to OLS with TWFE
- A large number of independent observations (or clusters)

Canonical DiD – with math

- Panel data on Y_{it} for $t = 1, 2$ and $i = 1, ..., N$
- **Treatment timing:** Some units ($D_i = 1$) are treated in period 2; everyone else is untreated ($D_i = 0$)

Canonical DiD – with math

- Panel data on Y_{it} for $t = 1, 2$ and $i = 1, ..., N$
- **Treatment timing:** Some units ($D_i = 1$) are treated in period 2; everyone else is untreated ($D_i = 0$)
- **Potential outcomes:** Observe $Y_{i1}(0)$ and $Y_{i2}(1)$ for treated units; and $Y_{i1}(0)$ and $Y_{i2}(0)$ for comparison

Key identifying assumptions

- **Parallel trends:**

$$\mathbb{E}\left[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_{i2}(0) - Y_{i1}(0) \mid D_i = 0\right]. \qquad (1)$$

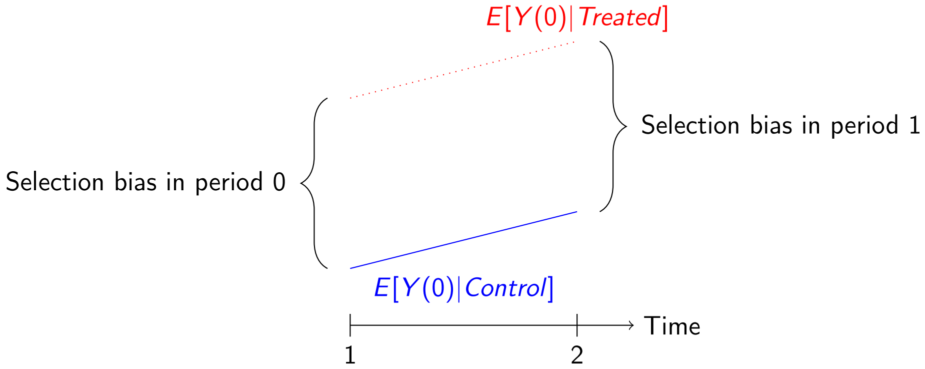
Key identifying assumptions

- **Parallel trends:**

$$\mathbb{E}\left[Y_{i2}(0) - Y_{i1}(0) \,|\, D_i = 1\right] = \mathbb{E}\left[Y_{i2}(0) - Y_{i1}(0) \,|\, D_i = 0\right]. \qquad (1)$$

- **No anticipation:** $Y_{i1}(1) = Y_{i1}(0)$
  - Intuitively, outcome in period 1 isn't affected by treatment status in period 2
  - Often left implicit in notation, but important for interpreting DiD estimand as a causal effect in period 2

Visualizing PT



Identification

- **Target parameter:** Average treatment effect on the treated (ATT) in period 2

$$\tau_{ATT} = E[Y_{i2}(1) - Y_{i2}(0) | D_i = 1]$$

Identification

- **Target parameter:** Average treatment effect on the treated (ATT) in period 2

$$\tau_{ATT} = E[Y_{i2}(1) - Y_{i2}(0) | D_i = 1]$$

- Under parallel trends and no anticipation, can show that

$$\tau_{ATT} = \underbrace{(E[Y_{i2}|D_i = 1] - E[Y_{i1}|D_i = 1])}_{\text{Change for treated}} - \underbrace{(E[Y_{i2}|D_i = 0] - E[Y_{i1}|D_i = 0])}_{\text{Change for control}},$$

a “difference-in-differences” of population means

Proof of Identification Argument

- Start with
$$E[Y_{i2} - Y_{i1}|D_i = 1] - E[Y_{i2} - Y_{i1}|D_i = 0]$$

Proof of Identification Argument

- Start with
$$E[Y_{i2} - Y_{i1}|D_i = 1] - E[Y_{i2} - Y_{i1}|D_i = 0]$$
- Apply definition of POs to obtain:
$$E[Y_{i2}(1) - Y_{i1}(1)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$

Proof of Identification Argument

- Start with
$$E[Y_{i2} - Y_{i1}|D_i = 1] - E[Y_{i2} - Y_{i1}|D_i = 0]$$
- Apply definition of POs to obtain:
$$E[Y_{i2}(1) - Y_{i1}(1)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$
- Use No Anticipation to substitute $Y_{i1}(0)$ for $Y_{i1}(1)$:
$$E[Y_{i2}(1) - Y_{i1}(0)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$

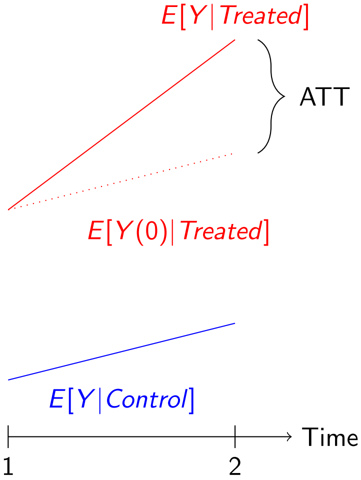
Proof of Identification Argument

- Start with
$$E[Y_{i2} - Y_{i1}|D_i = 1] - E[Y_{i2} - Y_{i1}|D_i = 0]$$
- Apply definition of POs to obtain:
$$E[Y_{i2}(1) - Y_{i1}(1)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$
- Use No Anticipation to substitute $Y_{i1}(0)$ for $Y_{i1}(1)$:
$$E[Y_{i2}(1) - Y_{i1}(0)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$
- Add and subtract $E[Y_{i2}(0)|D_i = 1]$ to obtain:
$$\begin{aligned} E[Y_{i2}(1) - Y_{i2}(0)|D_i = 1] + \\ (E[Y_{i2}(0)|D_i = 1] - E[Y_{i1}(0)|D_i = 1]) - \\ (E[Y_{i2}(0)|D_i = 0] - E[Y_{i1}(0)|D_i = 0]) \end{aligned}$$

Proof of Identification Argument

- Start with
$$E[Y_{i2} - Y_{i1}|D_i = 1] - E[Y_{i2} - Y_{i1}|D_i = 0]$$
- Apply definition of POs to obtain:
$$E[Y_{i2}(1) - Y_{i1}(1)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$
- Use No Anticipation to substitute $Y_{i1}(0)$ for $Y_{i1}(1)$:
$$E[Y_{i2}(1) - Y_{i1}(0)|D_i = 1] - E[Y_{i2}(0) - Y_{i1}(0)|D_i = 0]$$
- Add and subtract $E[Y_{i2}(0)|D_i = 1]$ to obtain:
$$\begin{aligned} &E[Y_{i2}(1) - Y_{i2}(0)|D_i = 1] + \\ &\quad (E[Y_{i2}(0)|D_i = 1] - E[Y_{i1}(0)|D_i = 1]) - \\ &\quad (E[Y_{i2}(0)|D_i = 0] - E[Y_{i1}(0)|D_i = 0]) \end{aligned}$$
- Cancel the last terms using PT to get $E[Y_{i2}(1) - Y_{i2}(0)|D_i = 1] = \tau_{ATT}$

Visualizing Identification



Estimation of Simple DiD

- The most conceptually simple estimator replaces population means with sample analogs:

$$\hat{\gamma}_{DiD} = (\bar{Y}_{12} - \bar{Y}_{11}) - (\bar{Y}_{02} - \bar{Y}_{01})$$

  where $\bar{Y}_{dt}$ is sample mean for group d in period t

Estimation of Simple DiD

- The most conceptually simple estimator replaces population means with sample analogs:

  $$\hat{\gamma}_{DiD} = (\bar{Y}_{12} - \bar{Y}_{11}) - (\bar{Y}_{02} - \bar{Y}_{01})$$

  where $\bar{Y}_{dt}$ is sample mean for group d in period t

- Difference-in-differences in a regression
  - Define two dummies

    $$T_i = \begin{cases} 0 & \text{if in group C} \\ 1 & \text{if in group T} \end{cases} \quad Post_t = \begin{cases} 0 & \text{if in pre period} \\ 1 & \text{if in post period} \end{cases}$$

  - Estimate $y_{it} = \beta_0 + \beta_1 T_i + \beta_2 Post_t + \gamma T_i \times Post_t + u_{it}$

Virtues of regression implementation

- **Inference:** Clustered standard errors are valid as number of clusters grows large
- We can add more controls

$$y_{it} = \beta_0 + \beta_1 T_i + \beta_2 Post_t + \gamma T_i \times Post_t + X_{it}\beta + u_{it}$$

- We can have unit/time fixed effects

$$y_{it} = \alpha_i + \delta_t + \gamma T_i \times Post_t + X_{it}\beta + u_{it}$$

- We can use continuous measures of treatment.

Implementation of DD as changes regression

- Consider the basic DD specification with two periods

$$y_{i2} = \beta_0 + \beta_1 T_i + \beta_2 Post_2 + \gamma T_i \times Post_2 + u_{i2}$$

$$y_{i1} = \beta_0 + \beta_1 T_i + \beta_2 Post_1 + \gamma T_i \times Post_1 + u_{i1}$$

- Subtract the second from the first equation

$$y_{i2} - y_{i1} = \beta_2 (Post_t - Post_{t'}) + \gamma \left(T_i \times Post_t - T_i \times Post_{t'} \right) + u_{i2} - u_{i1}$$

- Note that $Post_2 = 1$ and $Post_1 = 0$, therefore simplify

$$y_{i2} - y_{i1} = \beta_2 + \gamma T_i + u_{i2} - u_{i1}$$

Cholera in London

- The Difference in Difference (DiD or DD) idea was probably pioneered by physician John Snow (1855), who studied the cholera epidemics in London in the mid-nineteenth century.
- Cholera hit London three times in the early to mid 1800s causing large waves of tens of thousands of deaths
- Cholera attacked victims suddenly, with a 50% survival rate, and very painful symptoms.
- What is the source? Water or Air
- **Ideal Experiment:** Consider the ideal experiment: randomize households by coin flip to receive water from runoff (control) vs. water without runoff (treatment)

Two companies fight for customers

- Southwark and Vauxhall Waterworks Company and the Lambeth Water Company competed over some of the regions south of the Thames
- In 16 sub-districts, with a population of 300,000, they competed directly, even supplying customers side-by-side

"In many cases, a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference in the condition or occupation of the persons receiving the water of the different companies." Snow (1855) p 75

Lambeth moves its pipe

- During the 1849 epidemic, both companies drew water from Thames which was polluted with sewage and cholera
- London passes legislation requiring utility companies to move their pipes above the city
- In 1852, the Lambeth Company, a water utility company, changed supply from Hungerford Bridge
- It moved its intake pipe upstream to cleaner water and in response to legislation
- This created a natural experiment because Southwark and Vauxhall left its intake pipe in place

Would the time series estimator work?

Table: Death rates per 10000 people by time and water company

	1849	1854	Time series
Lambeth	150	10	-140

- Lambeth changed its water supply to a clean source in 1852.
- Does 10-150 = -140 show the impact of a clean water source on incidence of cholera?

$$E[Y_i^{L,Clean} \mid t = 1854] - E\left[Y_i^{L,Dirty} \mid t = 1849\right] =$$

$$\underbrace{E\left[Y_i^{L,Clean} - Y_i^{L,Dirty} \mid t = 1854\right]}_{\textbf{Treatment effect on the Treated}} +$$

$$\underbrace{\left\{E\left[Y_i^{L,Dirty} \mid t = 1854\right] - E\left[Y_i^{L,Dirty} \mid t = 1849\right]\right\}}_{\textbf{Selection bias}}$$

Would the cross-sectional estimator work?

Table: Death rates per 10000 people by time and water company

	1854
Lambeth	10
Southwark	150
Difference	-140

- Lambeth has clean water but Southwark does not.

$$
\begin{aligned}
E[Y_i^{L,Clean} \mid t = 1854] - E\left[Y_i^{S,Dirty} \mid t = 1854\right] \quad &= \\
\underbrace{E\left[Y_i^{L,Clean} - Y_i^{L,Dirty} \mid t = 1854\right]}_{\textbf{Treatment effect on the Treated}} \quad &+ \\
\underbrace{\left\{E\left[Y_i^{L,Dirty} \mid t = 1854\right] - E\left[Y_i^{S,Dirty} \mid t = 1854\right]\right\}}_{\textbf{Selection bias}} &
\end{aligned}
$$

- What is the identification assumption?

Difference-in-Difference estimator

- Snow (1855): compared changes in death rates in neighborhoods served by each company.
- Death rates per 10000 people by time and water company

Table: Death rates per 10000 people by time and water company

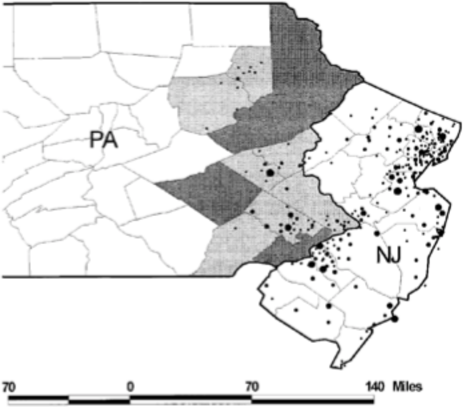
	1849	1854	Difference
Lambeth	150	10	-140
Southwark	125	150	25
Difference	-25	140	-**165**

- Would stories that invalidated time series estimates cause a problem?
- Would stories that invalidated cross sectional estimates cause a problem?

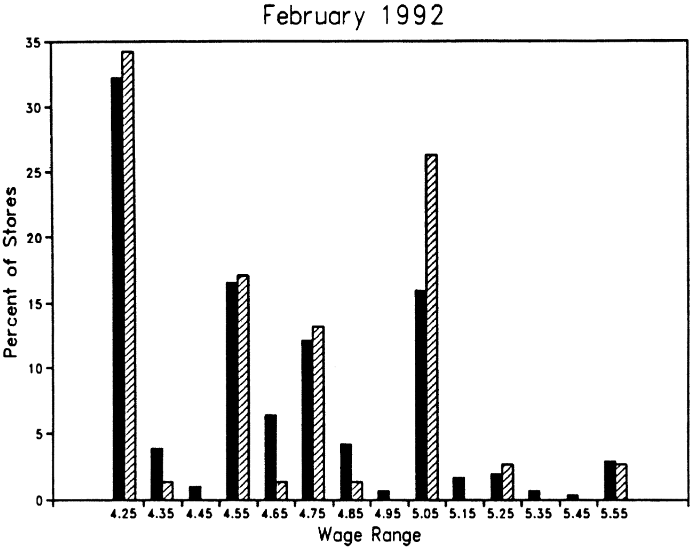
Card and Krueger (1994)

- Card and Krueger (1994) was a seminal study on the minimum wage both for the result and for the design
- Suppose you are interested in the effect of minimum wages on employment which is a classic and divisive question.
- In February 1992, New Jersey increased the state minimum wage from $4.25 to $5.05. Pennsylvania's minimum wage stayed at $4.25.
- They surveyed about 400 fast food stores both in New Jersey and Pennsylvania before and after the minimum wage increase in New Jersey!

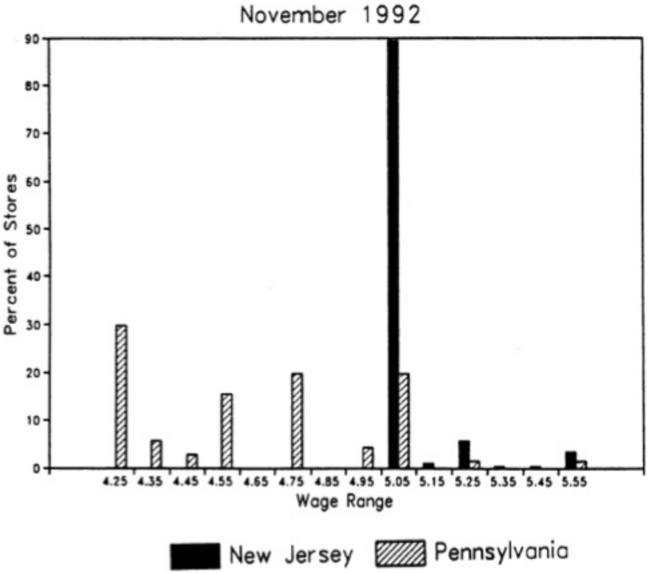
Fast Food Store Locations



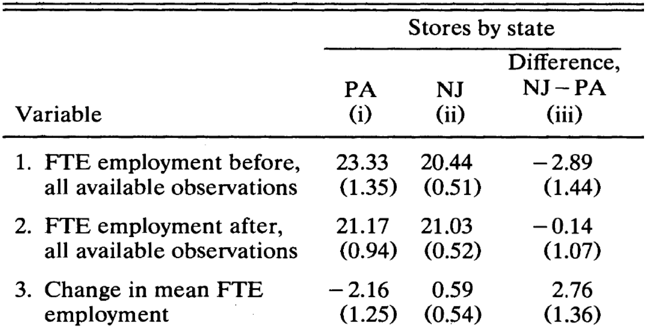
Percent of Stores with different wage ranges in Feb 1992



Percent of Stores with different wage ranges in Nov 1992



DiD estimate



	Stores by state		
Variable	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	–2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	–0.14 (1.07)
3. Change in mean FTE employment	–2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Surprisingly, employment *rose* in NJ relative to PA after the minimum wage change

Effect of School Construction

- What is the effect of school construction on educational attainment and wages?
- Think about the ideal experiment...
- An example of a difference-in-differences design
  Duflo (2001), "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment", AER.

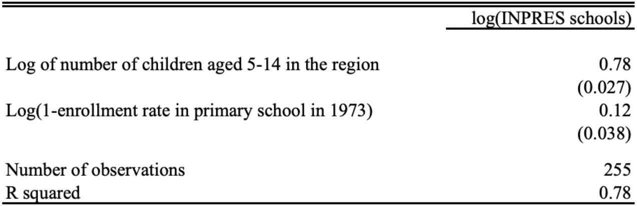
The School Construction Program: INPRES

- Oil boom: revenues for development programmes to promote equity
- Between 1973 and 1979, 61,807 new schools were built, costing 1.5% of GDP.
- Effectively 1 school for every 500 kids, the fastest construction program ever
- After school is build, government hires and pays teachers
- Stock of teachers grows by 43%

Targeting

Program designed to reach children never enrolled in school

- number of schools proportional to number of children of primary school age not enrolled in 1972
- Coefficients have the expected sign, but are <1, imperfect targeting



	log(INPRES schools)
Log of number of children aged 5-14 in the region	0.78
	(0.027)
Log(1-enrollment rate in primary school in 1973)	0.12
	(0.038)
Number of observations	255
R squared	0.78

Data

- 1995 intercensal survey, focus on men born btw 1950 and 1972
- about 153K individuals, 60K of whom are observed working for a wage
- Treatment cohorts: individuals aged 2-6 in 1974
- Control cohorts: older individuals
    - C1: aged 12-17 in 1974
    - C2: aged 18-24 in 1974
- Regions: some regions received more school construction, others less.

Method: DD

- Calculate DD using sample means

$$DD \quad = \quad \left[\overline{Y}^{High}_{Young} - \overline{Y}^{High}_{old}\right] - \left[\overline{Y}^{Low}_{Young} - \overline{Y}^{Low}_{old}\right]$$

- or a regression with interaction terms

$$Y_{ijk} = \alpha + \beta_1 \cdot T_k + \beta_2 \cdot P_j + \gamma \cdot T_k \times P_j + \epsilon_{ij}$$

  - Y_{ijk} outcome (schooling or wage) of individual i born in region j in year k
  - T_k equals 1 if individual belongs to young cohorts
  - P_j equals 1 if individual belongs to a region with high school construction

Results - Simple DD

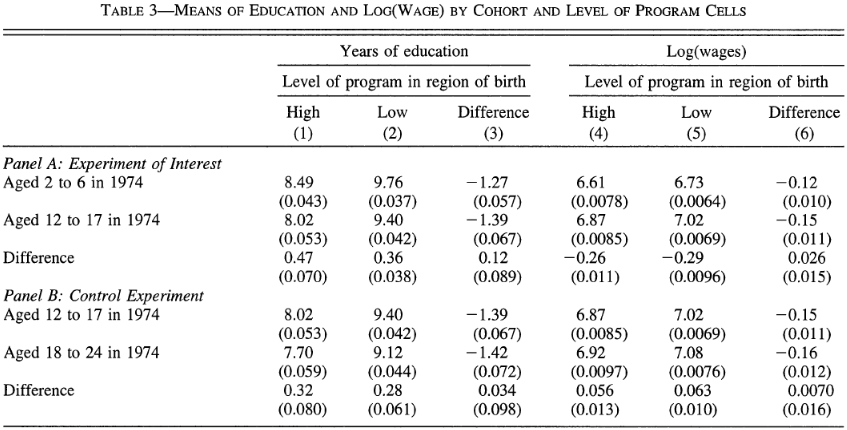


TABLE 3—MEANS OF EDUCATION AND LOG(WAGE) BY COHORT AND LEVEL OF PROGRAM CELLS

	Years of education			Log(wages)		
	Level of program in region of birth			Level of program in region of birth		
	High (1)	Low (2)	Difference (3)	High (4)	Low (5)	Difference (6)
Panel A: Experiment of Interest						
Aged 2 to 6 in 1974	8.49	9.76	−1.27	6.61	6.73	−0.12
	(0.043)	(0.037)	(0.057)	(0.0078)	(0.0064)	(0.010)
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Difference	0.47	0.36	0.12	−0.26	−0.29	0.026
	(0.070)	(0.038)	(0.089)	(0.011)	(0.0096)	(0.015)
Panel B: Control Experiment						
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Aged 18 to 24 in 1974	7.70	9.12	−1.42	6.92	7.08	−0.16
	(0.059)	(0.044)	(0.072)	(0.0097)	(0.0076)	(0.012)
Difference	0.32	0.28	0.034	0.056	0.063	0.0070
	(0.080)	(0.061)	(0.098)	(0.013)	(0.010)	(0.016)

Question of causality

- Could we claim that the INPRES increased years of schooling by 0.12?
- What are the identification assumptions?
- How does table try to check whether these assumptions hold?

Main regression specification

- Instead of a discrete measure of school construction use
  - P_j: number of schools constructed per 1000 children in region j
  - Also include other region specific variables that might have a heterogeneous effect on different cohorts

$$Y_{ijk} = \alpha_j + \beta_k + \gamma \cdot T_k \times P_j + \delta \cdot T_k \times C_j + \epsilon_{ijk}$$

  - α_j: region dummies, β_k: cohort dummies
  - C_j: number of children in 1971, enrollment rate in 1971, water and sanitation program
- $\hat{\gamma}$ gives the DD estimate.

Results - Regression DD

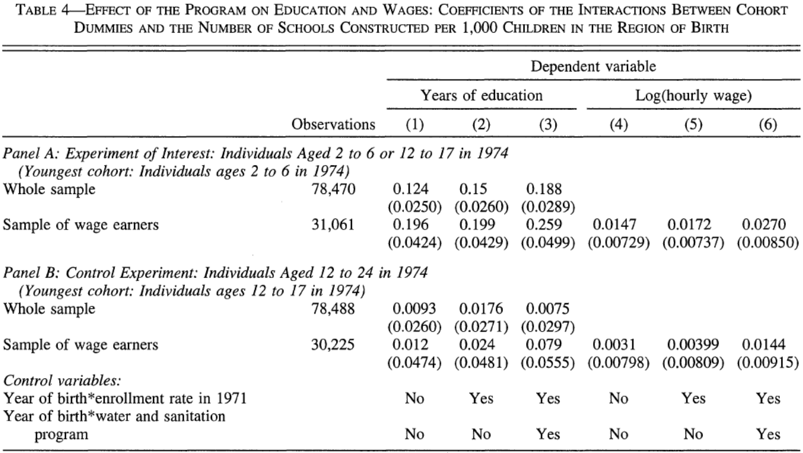


Table 4—Effect of the Program on Education and Wages: Coefficients of the Interactions Between Cohort Dummies and the Number of Schools Constructed per 1,000 Children in the Region of Birth

		Dependent variable					
		Years of education			Log(hourly wage)		
	Observations	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Experiment of Interest: Individuals Aged 2 to 6 or 12 to 17 in 1974							
(Youngest cohort: Individuals ages 2 to 6 in 1974)							
Whole sample	78,470	0.124	0.15	0.188			
		(0.0250)	(0.0260)	(0.0289)			
Sample of wage earners	31,061	0.196	0.199	0.259	0.0147	0.0172	0.0270
		(0.0424)	(0.0429)	(0.0499)	(0.00729)	(0.00737)	(0.00850)
Panel B: Control Experiment: Individuals Aged 12 to 24 in 1974							
(Youngest cohort: Individuals ages 12 to 17 in 1974)							
Whole sample	78,488	0.0093	0.0176	0.0075			
		(0.0260)	(0.0271)	(0.0297)			
Sample of wage earners	30,225	0.012	0.024	0.079	0.0031	0.00399	0.0144
		(0.0474)	(0.0481)	(0.0555)	(0.00798)	(0.00809)	(0.00915)
Control variables:							
Year of birth*enrollment rate in 1971		No	Yes	Yes	No	Yes	Yes
Year of birth*water and sanitation program		No	No	Yes	No	No	Yes

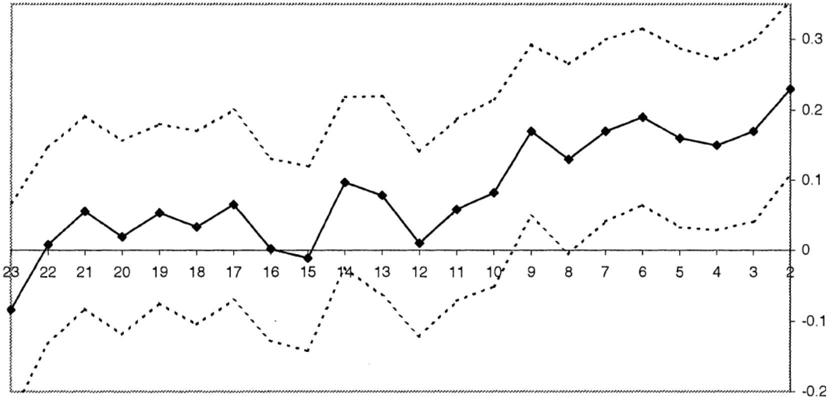
Investigating the evolution of education for all cohorts

- Instead of splitting cohorts into T and C we could include a dummy interacted with school construction measure to see whether which cohorts are affected by the program.
  - $d_{i2}, \dots, d_{i23}$: 22 dummies.
  - d_{il} equals 1 if individual i is aged l years in 1974
  - P_j: number of schools constructed per 1000 children in region j

$$
\begin{aligned}
Y_{ijk} \quad = \quad & \alpha_j + \beta_k + \\
& \sum_{l=2}^{23} \gamma_l \cdot d_{il} \times P_j + \\
& \sum_{l=2}^{23} \delta_l \cdot d_{il} \times C_j + \epsilon_{ijk}
\end{aligned}
$$

- $\hat{\gamma}_l$ gives the DD estimate of the impact of program on cohort aged l years in 1974.
- What is the pattern of coefficient that you would expect if parallel trends assumption holds?

Results - Interaction terms analysis



Age in 1974

Results - Interaction terms analysis - Wage

FIGURE 3B -- COEFFICIENTS OF THE INTERACTION AGE IN 1974* PROGRAM INTENSITY IN THE REGION OF BIRTH
IN THE WAGE AND EDUCATION EQUATIONS (SMOOTHED)

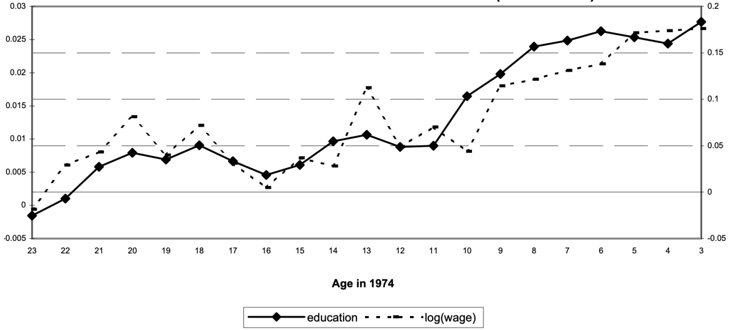


Table of Contents

1. Introduction
2. Simple DiD
    - Example 1: Cholera in London
    - Example 2: Minimum Wage
    - Example 3: School Construction
3. Extensions of Simple DiD
    - Event Studies
    - DDD
4. Multiple Periods and Staggered Treatment Timing
5. Violations of Parallel Trends

Event Study Model Specification

- **Extended DiD Model with Leads and Lags:**

$$y_{it} = \alpha_i + \delta_t + \sum_{k \neq -1} \beta_k \cdot D_i \times 1\{t = k\} + \varepsilon_{it}$$

- **Components:**
  - y_{it}: Outcome variable for unit i at time t.
  - α_i: Unit fixed effects.
  - δ_t: Time fixed effects.
  - D_i: Indicator for treated units.
  - $1\{t = k\}$: Indicator for periods relative to treatment.
  - β_k: Effect k periods from treatment.

Interpreting Event Study Coefficients

- **Pre-Treatment Periods ($k < 0$):**
  - Check for **parallel trends**.
  - Significant β_k may indicate **violation** of assumptions.
- **Post-Treatment Periods ($k \geq 0$):**
  - Assess the **timing and persistence** of treatment effects.
  - Observe if effects are **immediate**, **delayed**, or **fade over time**.
- **Plotting Coefficients:**
  - Plot $\hat{\beta}_k$ against time periods k.
  - Include **confidence intervals** to assess significance.

Example 4: Medicaid and Mortality

- Miller, S., Johnson, N., & Wherry, L. R. (2021). "Medicaid and Mortality: New Evidence from Linked Survey and Administrative Data". *The Quarterly Journal of Economics*, 136(3), 1783-1829.
- **ACA Medicaid Expansion**:
  - Largest health insurance expansion since the inception of Medicare and Medicaid
  - Over 20 million people gained coverage through the Affordable Care Act
  - Majority are low-income adults now receiving Medicaid coverage
- **Variation in State Expansion**:
  - Not all states chose to expand Medicaid eligibility
  - In 2012, the Supreme Court ruled that Medicaid expansion under the ACA was optional for states
  - Some states expanded Medicaid, while others did not (Random?)

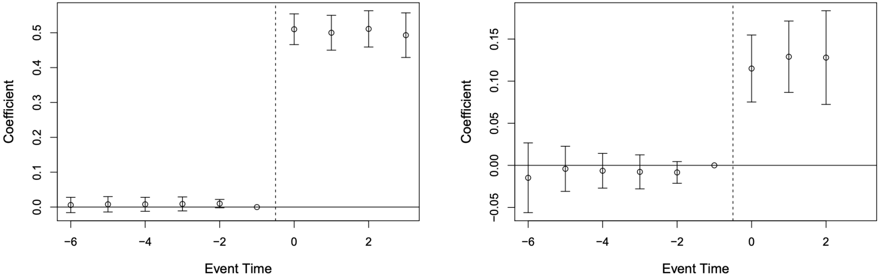
Empirical Specification

- **Method**:
  - Event study with difference-in-differences approach
- **Model**:

$$Y_{ist} = \sum_{k \neq -1} \beta_k \left(\text{Expansion}_s \times \mathbb{I}[t - t_s^* = k]\right) + \alpha_s + \delta_t + \varepsilon_{ist}$$

- **Variables**:
  - Y_{ist}: Indicator if individual i died in year t
  - Expansion_s: Medicaid expansion state indicator
  - α_s, δ_t: State and year fixed effects

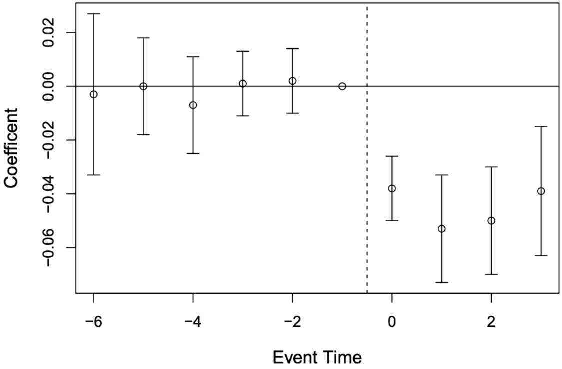
Effects on Eligibility and Enrollment



(a) Medicaid Eligibility (ACS)

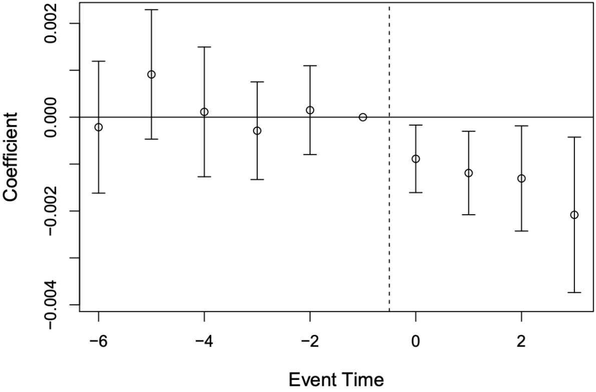
(b) Any Medicaid Enrollment in Year (CMS)

Effects on Uninsured

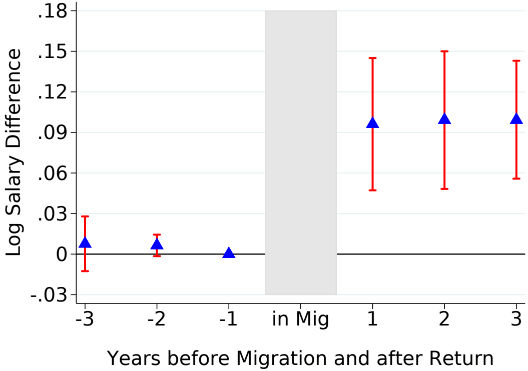


(e) Uninsured (ACS)

Effect of the ACA Medicaid Expansions on Annual Mortality



Example 5: Event Study of Migration from EM to US and Return



Migration from EM to US and Return

	Entire Sample		Mass Layoff	
	Total	Mig. Years ≥ 5	Total	Mig. Years ≥ 5
	(1)	(2)	(3)	(4)
Treatment × Post	0.0735***	0.0903***	0.0757***	0.0745*
	(0.00505)	(0.00911)	(0.0211)	(0.0379)
Observations	126934	50154	2554	820
R-Squared	0.605	0.562	0.542	0.453

Motivation for Triple Differences

- **Problem with Difference-in-Differences (DD)**:
  - DD assumes no state-specific time shocks (γ_s) or group-specific trends (δ_g).
  - If such shocks exist, DD estimates may be biased.
- **Example**:
  - Minimum wage increase in New Jersey.
  - Concern: State-specific shocks might affect employment beyond the policy change.

Triple Differences (DDD) Design

- **Adding a Third Dimension**:
  - Introduce untreated group within each state (e.g., high-wage workers).
  - Compare changes across states *and* groups over time.
- **Setup**:
  - States: Treatment (NJ) and Control (PA).
  - Groups: Low-wage (affected) and High-wage (unaffected) workers.
  - Periods: Before and After the minimum wage increase.

Triple Difference Design for Minimum Wage Example

States	Group	Period	Outcomes	D_1	D_2	D_3
NJ	Low-wage workers	After	$NJ_l + T + NJ_t + l_t + D$	$T + NJ_t +$	$(l_t - h_t) + D$	D
		Before	NJ_l	$l_t + D$		
	High-wage workers	After	$NJ_h + T + NJ_t + h_t$	$T + NJ_t + h_t$		
		Before	NJ_h			
PA	Low-wage workers	After	$PA_l + T + PA_t + l_t$	$T + PA_t + l_t$	$l_t - h_t$	
		Before	PA_l			
	High-wage workers	After	$PA_h + T + PA_t + h_t$	$T + PA_t + h_t$		
		Before	PA_h			

Estimating the Treatment Effect

- **First Differences**:
  - Compute changes over time for each group in each state.
  - Eliminates group-specific and state-specific fixed effects.
- **Second Differences (DD)**:
  - Difference between treatment and control states for low-wage workers.
  - Potentially biased due to state-specific trends (γ_s).
- **Triple Difference (DDD)**:
  - Difference-in-differences between groups and states.
  - Removes state-specific shocks and group-specific trends.
  - Isolates the treatment effect (θ) of the policy.

DDD Estimation Equation

$$
\begin{aligned}
Y_{ijt} = \alpha + \psi X_{ijt} + \beta_1 \tau_t + \beta_2 \delta_j + \beta_3 D_i \\
+ \beta_4 (\delta \times \tau)_{jt} + \beta_5 (\tau \times D)_{ti} + \beta_6 (\delta \times D)_{ij} + \beta_7 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt}
\end{aligned}
$$

- Y_{ijt}: Outcome variable for individual i, group j, at time t.
- ψX_{ijt}: Control variables for individual characteristics.
- β_7: Coefficient of interest in DDD.
- Note that one of these will be dropped due to multicollinearity, but I include them in the equation so that you can visualize all the factors used in the product of these terms.

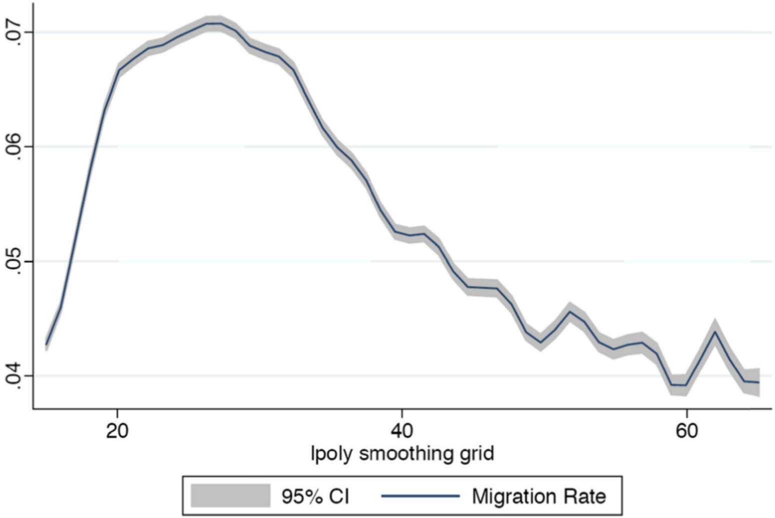
Parallel Trends Assumption

- **Assumption in DDD**:
  - Assumes that in the absence of treatment, the differences in outcomes between groups and states would have followed a parallel trend over time.
- **Relative Strength of Assumptions**:
  - All models, Simple Differences, DiD, and DDD, rely on assumptions for validity.
  - DDD assumptions are often weaker than DiD, as they account for additional variations.
  - DiD assumptions are generally weaker than those of Simple Differences.

Example 6: Natural Disasters and Youth Migration

- Baez, J., Caruso, G., Mueller, V., & Niu, C. (2017). Droughts augment youth migration in Northern Latin America and the Caribbean. Climatic change, 140, 423-435.
- They link individual-level information from multiple censuses for eight countries in the region with natural disaster indicators constructed from georeferenced climate data at the province level to measure the impact of droughts and hurricanes on internal mobility.
- They find that younger individuals are more likely to migrate in response to these disasters, especially when confronted with droughts.

Kernel density of migration by age at baseline

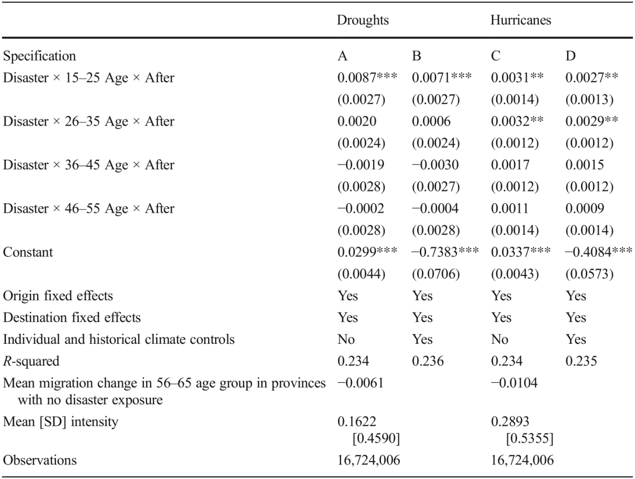


Empirical Method

$$\begin{aligned} M_{ijkat} = {} & \beta_1(Disaster_k \times Age_a \times After_t) + \beta_2(Disaster_k \times Age_a) \\ & + \beta_3(Disaster_k \times After_t) + \beta_4(Age_a \times After_t) \\ & + \theta X_{ijkt} + \alpha_j + \rho_k + \delta_t + \gamma_a + \varepsilon_{ijkat} \end{aligned}$$

- M_{ijkat}: Indicator of migration for individual i at destination province j, from origin province k, age group a, at time t.
- $Disaster_k$, Age_a, $After_t$: Interaction terms representing the impact of disasters, age, and post-disaster periods.
- θX_{ijkt}: Control variables for individual and contextual characteristics.
- α_j, ρ_k, δ_t, γ_a: FE for destination province, origin province, time, and age.
- **Identification assumption**: Differences in migration rates across age groups would be similar across provinces with high and low disaster intensity in the absence of the disaster.

Change in the probability of migration due to disasters



	Droughts		Hurricanes	
Specification	A	B	C	D
Disaster × 15–25 Age × After	0.0087***	0.0071***	0.0031**	0.0027**
	(0.0027)	(0.0027)	(0.0014)	(0.0013)
Disaster × 26–35 Age × After	0.0020	0.0006	0.0032**	0.0029**
	(0.0024)	(0.0024)	(0.0012)	(0.0012)
Disaster × 36–45 Age × After	−0.0019	−0.0030	0.0017	0.0015
	(0.0028)	(0.0027)	(0.0012)	(0.0012)
Disaster × 46–55 Age × After	−0.0002	−0.0004	0.0011	0.0009
	(0.0028)	(0.0028)	(0.0014)	(0.0014)
Constant	0.0299***	−0.7383***	0.0337***	−0.4084***
	(0.0044)	(0.0706)	(0.0043)	(0.0573)
Origin fixed effects	Yes	Yes	Yes	Yes
Destination fixed effects	Yes	Yes	Yes	Yes
Individual and historical climate controls	No	Yes	No	Yes
R-squared	0.234	0.236	0.234	0.235
Mean migration change in 56–65 age group in provinces with no disaster exposure	−0.0061		−0.0104	
Mean [SD] intensity	0.1622 [0.4590]		0.2893 [0.5355]	
Observations	16,724,006		16,724,006	

Notes: The omitted age category is 56-65. Origin province by birth year clustered standard errors presented in parentheses. ***, **, and * indicate p <0.01, p < 0.05, and p < 0.10

Characterizing the recent literature

We can group the recent innovations in DiD lit by which elements of the canonical model they relax:

- **Multiple periods and staggered treatment timing**
- **Relaxing or allowing PT to be violated**
- **Inference with a small number of clusters**

In the remaining of this lecture, we will focus on the first two.

Table of Contents

1 Introduction

2 Simple DiD

- Example 1: Cholera in London
- Example 2: Minimum Wage
- Example 3: School Construction

3 Extensions of Simple DiD

- Event Studies
- DDD

4 Multiple Periods and Staggered Treatment Timing

5 Violations of Parallel Trends

Staggered Timing

- Remember that in the canonical DiD model we had:
  - Two periods and a common treatment date
  - Identification from parallel trends and no anticipation
  - A large number of clusters for inference
- A very active recent literature has focused on relaxing the first assumption: **what if there are multiple periods and units adopt treatment at different times?**
- This literature typically maintains the remaining ingredients: parallel trends and many clusters

Overview of Staggered Timing Literature

1. Negative results: TWFE OLS doesn't give us what we want with treatment effect heterogeneity
2. New estimators: perform better under treatment effect heterogeneity

Staggered timing set-up

- Panel of observations for periods $t = 1, ..., T$
- Suppose units adopt a binary treatment at different dates $G_i \in \{1, ..., T\} \cup \infty$ (where $G_i = \infty$ means "never-treated")
  - Literature is now starting to consider cases with continuous treatment and treatments that turn on/off – that lit is still developing (see Section 3.4 of Roth and Sant'Anna (2023))
- Potential outcomes $Y_{it}(g)$ – depend on time and time you were first-treated
  - Example: $Y_{i,2016}(2014)$ would represent the insurance coverage in state i in 2016 if they had first expanded Medicaid in 2014.

Extending the Identifying Assumptions: Parallel Trend

- The key identifying assumptions from the canonical model are extended in a natural way
- **Parallel trends:** Intuitively, says that if treatment hadn't happened, all "adoption cohorts" would have parallel average outcomes in all periods

$$E[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g] = E[Y_{it}(\infty) - Y_{i,t-1}(\infty) | G_i = g'] \text{ for all } g, g', t$$

  - If there had been no Medicaid expansions, insurance rates would have evolved in parallel on average for all groups of states that adopted Medicaid expansion in different years, including those that never expanded Medicaid.
- Note: can impose slightly weaker versions (e.g. only require PT post-treatment, or for only treatment groups)

Extending the Identifying Assumptions: No Anticipation

- **No anticipation:** Intuitively, says that treatment has no impact before it is implemented

$$Y_{it}(g) = Y_{it}(\infty) \text{ for all } t < g$$

Interpreting the estimand of two-way fixed effects models

- Suppose we again run the regression

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it},$$

  where $D_{it} = 1[t \geq G_i]$ is a treatment indicator.

- Suppose we're willing to assume no anticipation and parallel trends across all adoption cohorts as described above

Interpreting the estimand of two-way fixed effects models

- Suppose we again run the regression

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it},$$

  where $D_{it} = 1[t \geq G_i]$ is a treatment indicator.

- Suppose we're willing to assume no anticipation and parallel trends across all adoption cohorts as described above
- Good news: if treatment effects are constant across time and units, $Y_{it}(g) - Y_{it}(\infty) \equiv \tau$ for all $t \geq g$, then $\beta = \tau$

Interpreting the estimand of two-way fixed effects models

- Suppose we again run the regression

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it},$$

  where $D_{it} = 1[t \geq G_i]$ is a treatment indicator.

- Suppose we're willing to assume no anticipation and parallel trends across all adoption cohorts as described above
- Good news: if treatment effects are constant across time and units, $Y_{it}(g) - Y_{it}(\infty) \equiv \tau$ for all $t \geq g$, then $\beta = \tau$
- Bad news: if treatment effects are heterogeneous, then β may put negative weights on treatment effects for some units and time periods
  - E.g., if treatment effect depends on time since treatment, $Y_{it}(t-r) - Y_{it}(\infty) = \tau_r$, then some τ_rs may get negative weight

Where do these negative weights come from?

- Negative weights will bias the estimator and may even lead to opposite results. ($\beta = \sum_r \omega_r \tau_r$ where the weights sum to 1 but may be negative)
- The intuition for these negative weights is that the TWFE OLS specification combines two sources of comparisons:

  1. **Clean comparisons:** DiD's between treated units (e.g., **switchers**) and **never-treated** units
  2. **Forbidden comparisons:** DiD's between two sets of already-treated units (who began treatment at different times), such as **switchers** and **always-treated** units

- These forbidden comparisons can lead to negative weights: the "control group" is already treated, so we encounter problems if their treatment effects change over time.

Some intuition for forbidden comparisions

- Consider the two-period model with three groups:
  1. **Always-treated (AT)** units: Treated in both periods
  2. **Switchers (S)**: Treated only in period 2
  3. **Never-treated (NT)** units: Never receive treatment
- With two periods, the coefficient β from the model:

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it}$$

is equivalent to the first-differenced regression:

$$\Delta Y_i = \Delta\phi + \Delta D_i\,\beta + \Delta\epsilon_i$$

Some intuition for forbidden comparisions - II

- Observe that ΔD_i equals:
  - 0 for always-treated units ($D_{i2} = D_{i1} = 1$)
  - 1 for switchers ($D_{i2} = 1,\ D_{i1} = 0$)
  - 0 for never-treated units ($D_{i2} = D_{i1} = 0$)
- **Key point:** The switchers are the only group with a change in treatment status. In the TWFE model, the control group includes both always-treated and never-treated units.

$$\bar{Y}_{\text{Control}} = \frac{N_{AT}}{N_{AT} + N_{NT}} \bar{Y}_{\text{AT}} + \frac{N_{NT}}{N_{AT} + N_{NT}} \bar{Y}_{\text{NT}}$$

Clean and forbidden comparisons in this context

- The estimated coefficient $\hat{\beta}$ captures the average difference in outcome changes between the switchers and the control group:

$$\hat{\beta} = \Delta \bar{Y}_{\text{Swithers}} - \Delta \bar{Y}_{\text{Controls}}$$

- However, the control group is a mix of:
  - **Never-treated** units (providing clean comparisons)
  - **Always-treated** units (leading to forbidden comparisons)
- Breaking it down:

$$\hat{\beta} = \underbrace{\left(\Delta \bar{Y}_{\text{S}} - \Delta \bar{Y}_{\text{NT}}\right)}_{\text{Clean Comparison}} - \frac{N_{AT}}{N_{AT} + N_{NT}} \underbrace{\left(\Delta \bar{Y}_{\text{AT}} - \Delta \bar{Y}_{\text{NT}}\right)}_{\text{Forbidden Comparison}}$$

Problem arising from forbidden comparisons

- If the treatment effect for the always-treated units changes over time (e.g., grows), the term $\left(\Delta\bar{Y}_{\text{Always-Treated}} - \Delta\bar{Y}_{\text{Never-Treated}}\right)$ captures this change.
- When the treatment effect increases over time for always-treated units, it subtracts from $\hat{\beta}$, potentially making it negative even if the true treatment effect is positive.
- This introduces bias into $\hat{\beta}$ because the always-treated units are being used as part of the control group, but they are experiencing treatment effects.

Issues with dynamic TWFE

- Sun and Abraham (2021) show that similar issues arise with dynamic TWFE specifications:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{k \neq 0} \gamma_k D_{i,t}^k + \varepsilon_{i,t},$$

  where $D_{i,t}^k = 1\{t - G_i = k\}$ are "event-time" dummies.

- Like for the static spec, γ_k may put negative weight on treatment effects after k periods for some units
- SA also show that γ_k may be "contaminated" by treatment effects at lags $k' \neq k$

Dynamic TWFE - Continued

- The results in SA suggest that interpreting the $\hat{\gamma}_k$ for $k = 1, 2, ...$ as estimates of the dynamic effects of treatment may be misleading
- These results also imply that pre-trends tests of the γ_k for $k < 0$ may be misleading – could be non-zero even if parallel trends holds, since they may be "contaminated" by post-treatment effects!

Dynamic TWFE - Continued

- The results in SA suggest that interpreting the $\hat{\gamma}_k$ for $k = 1, 2, ...$ as estimates of the dynamic effects of treatment may be misleading
- These results also imply that pre-trends tests of the γ_k for $k < 0$ may be misleading – could be non-zero even if parallel trends holds, since they may be "contaminated" by post-treatment effects!
- The issues discussed in SA arise if dynamic path of treatment effects is heterogeneous across adoption cohorts
  - Biases may be less severe than for "static" specs if dynamic patterns are similar across cohorts

New estimators for Staggered Timing

- Several new (closely-related) estimators have been proposed to try to address these negative weighting issues
- The key components of all of these are:
  1. Be precise about the target parameter (estimand) – i.e., how do we want to aggregate treatment effects across time/units
  2. Estimate the target parameter using only "clean-comparisons"

Example – Callaway and Sant'Anna (2021)

- Define $ATT(g,t)$ to be ATT in period t for units first treated at period g,

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

Example – Callaway and Sant'Anna (2021)

- Define $ATT(g,t)$ to be ATT in period t for units first treated at period g,

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- Under PT and No Anticipation, $ATT(g,t)$ is identified as

$$ATT(g,t) = \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Change for cohort g}} - \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Change for never-treated units}}$$

- Why?

Example – Callaway and Sant'Anna (2021)

- Define $ATT(g,t)$ to be ATT in period t for units first treated at period g,

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty)|G_i = g]$$

- Under PT and No Anticipation, $ATT(g,t)$ is identified as

$$ATT(g,t) = \underbrace{E[Y_{it} - Y_{i,g-1}|G_i = g]}_{\text{Change for cohort g}} - \underbrace{E[Y_{it} - Y_{i,g-1}|G_i = \infty]}_{\text{Change for never-treated units}}$$

- Why? This is a two-group two-period comparison, so the argument is the same as in the canonical case!

Proof of Identification Argument

- Start with

$$E[Y_{it} - Y_{i,g-1} | G_i = g] - E[Y_{it} - Y_{i,g-1} | G_i = \infty]$$

Proof of Identification Argument

- Start with

$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$

- Apply definition of POs to obtain:

$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

Proof of Identification Argument

- Start with

$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$

- Apply definition of POs to obtain:

$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

- Use No Anticipation to substitute $Y_{i,g-1}(\infty)$ for $Y_{i,g-1}(g)$:

$$E[Y_{it}(g) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

Proof of Identification Argument

- Start with

$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$

- Apply definition of POs to obtain:

$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

- Use No Anticipation to substitute $Y_{i,g-1}(\infty)$ for $Y_{i,g-1}(g)$:

$$E[Y_{it}(g) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

- Add and subtract $E[Y_{it}(\infty)|G_i = g]$ to obtain:

$$E[Y_{it}(g) - Y_{it}(\infty)|G_i = g] +$$
$$[E[Y_{it}(\infty) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]]$$

Proof of Identification Argument

- Start with

$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$

- Apply definition of POs to obtain:

$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

- Use No Anticipation to substitute $Y_{i,g-1}(\infty)$ for $Y_{i,g-1}(g)$:

$$E[Y_{it}(g) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

- Add and subtract $E[Y_{it}(\infty)|G_i = g]$ to obtain:

$$E[Y_{it}(g) - Y_{it}(\infty)|G_i = g] +$$
$$[E[Y_{it}(\infty) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]]$$

- Cancel the last term using PT to get
$E[Y_{it}(g) - Y_{it}(\infty)|G_i = g] = ATT(g, t)$

Example – Callaway and Sant'Anna (2020)

- Define $ATT(g,t)$ to be ATT in period t for units first treated at period g,

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

Example – Callaway and Sant'Anna (2020)

- Define $ATT(g,t)$ to be ATT in period t for units first treated at period g,

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- Under PT and No Anticipation,

$$ATT(g,t) = \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Change for cohort g}} - \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Change for never-treated}}$$

Example – Callaway and Sant'Anna (2020)

- Define $ATT(g,t)$ to be ATT in period t for units first treated at period g,

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- Under PT and No Anticipation,

$$ATT(g,t) = \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Change for cohort g}} - \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Change for never-treated}}$$

- We can then estimate this with sample analogs:

$$\widehat{ATT}(g,t) = \underbrace{\widehat{E}[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Sample change for cohort g}} - \underbrace{\widehat{E}[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Sample change for never-treated}}$$

where $\hat{E}$ denotes sample means.

Aggregation schemes

- If have a large number of observations and relatively few groups/periods, can report $\widehat{ATT}(g,t)$'s directly.
- If there are many groups/periods, the $\widehat{ATT}(g,t)$ may be very imprecisely estimated and/or too numerous to report concisely

Aggregation schemes

- In these cases, it is often desirable to report sensible averages of the $\widehat{ATT}(g,t)$'s.

Aggregation schemes

- In these cases, it is often desirable to report sensible averages of the $\widehat{ATT}(g,t)$'s.
- One of the most useful is to report event-study parameters which aggregate $\widehat{ATT}(g,t)$'s at a particular lag since treatment
  - E.g. $\hat{\theta}_k = \sum_g \omega_g \widehat{ATT}(g, t+k)$ aggregates effects for cohorts in the kth period after treatment
  - The weights ω_g could be chosen to weight different cohorts equally, or in terms of their relative frequencies in the treated population.
  - Can also construct for $k < 0$ to estimate "pre-trends"

Aggregation schemes

- In these cases, it is often desirable to report sensible averages of the $\widehat{ATT}(g,t)$'s.

- One of the most useful is to report event-study parameters which aggregate $\widehat{ATT}(g,t)$'s at a particular lag since treatment
  - E.g. $\hat{\theta}_k = \sum_g \omega_g \widehat{ATT}(g, t+k)$ aggregates effects for cohorts in the kth period after treatment
  - The weights ω_g could be chosen to weight different cohorts equally, or in terms of their relative frequencies in the treated population.
  - Can also construct for $k < 0$ to estimate "pre-trends"

- C&S discuss other sensible aggregations too – e.g., if interested in whether treatment effects differ across good/bad economies, may want to "calendar averages" that pool the $\widehat{ATT}(t,g)$ for the same year

Comparisons of new estimators

- Callaway and Sant'Anna also propose an analogous estimator using *not-yet-treated* rather than never-treated units.
- Sun and Abraham (2021) propose a similar estimator but with different comparison groups (e.g. using last-to-be treated rather than not-yet-treated)
- Borusyak et al. (2024), Wooldridge (2022), Gardner (2022) propose "imputation" estimators that estimate the counterfactual $\hat{Y}_{it}(0)$ using a TWFE model that is fit using only pre-treatment data

Borusyak et al. (2024)

- They fit a TWFE using observations only for units and time periods that are not-yet-treated. ($Y_{it}(\infty) = \alpha_i + \lambda_t + \epsilon_{it}$)
- They then infer the never-treated potential outcome for each treated unit using the predicted value from this regression ($\widehat{Y}_{it}(\infty)$).
- This provides an estimate of the treatment effect for each treated unit, $Yit - \widehat{Y}_{it}(\infty)$, and these individual-level estimates can be aggregated to form estimates of summary parameters like the ATT(g, t)
- Main difference from C & S is that this uses more pre-treatment periods, not just period $g - 1$
- This can sometimes be more efficient (if outcome not too serially correlated), but also relies on a stronger PT assumption that may be more susceptible to bias

Conclusions on Staggered Timing I

- While conventional TWFE specifications make sensible comparisons of treated and untreated units in the canonical two-period DiD setting, in the staggered case they typically make "*forbidden comparisons*" between already-treated units.

- As a result, treatment effects for some units and time periods may receive negative weights in the TWFE estimand if we have heterogeneous treatment effect ("*Wrong sign* for ATT in extreme cases).

- If the researcher is not willing to impose assumptions on treatment effect heterogeneity, the most direct remedy for this problem is to use the heterogeneity robust DiD estimators that explicitly specify the comparisons to be made between treatment and control groups (e.g. not-yet-treated or never-treated units), as well as the desired weights in the target parameter.

Conclusions on Staggered Timing II

- Using a TWFE specification may be justified for efficiency reasons if one is confident that treatment effects are homogeneous, but researchers are often unwilling to restrict treatment effect heterogeneity.

- Usually, the various heterogeneity-robust DiD estimators typically (although not always) produce similar answers. The first-order consideration is therefore to use an approach that makes clear what the target parameter is and which groups are being compared for identification.

- And while more complex to express in regression format, they can be viewed as simple aggregations of comparisons of group means. Once researchers gain experience using the newer heterogeneity-robust DiD methods, they will not seem so scary!

- Thankfully, there are now statistical packages that make implementing (and comparing) the results from these estimators straightforward in practice (See Table 2 of Roth and Sant'Anna (2023))

Statistical packages for recent DiD methods

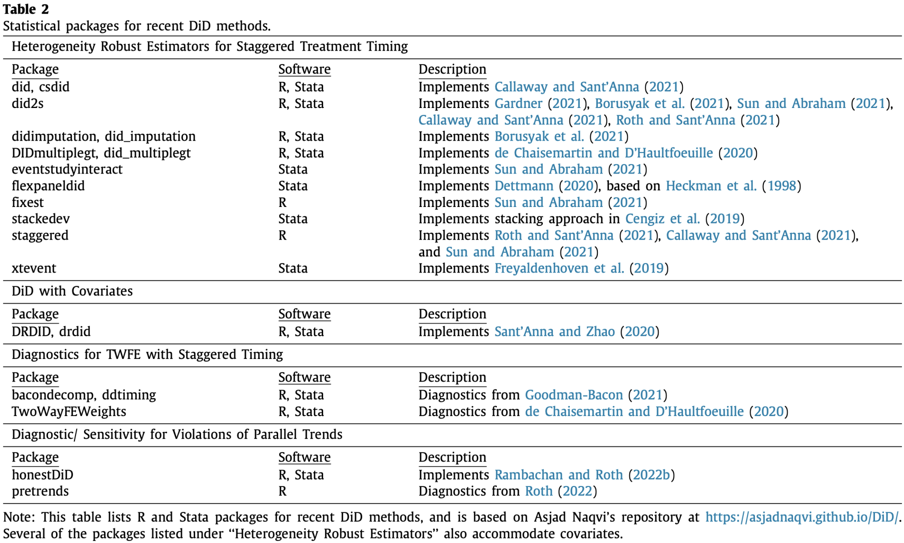


Table 2
Statistical packages for recent DiD methods.

Heterogeneity Robust Estimators for Staggered Treatment Timing

Package	Software	Description
did, csdid	R, Stata	Implements Callaway and Sant'Anna (2021)
did2s	R, Stata	Implements Gardner (2021), Borusyak et al. (2021), Sun and Abraham (2021), Callaway and Sant'Anna (2021), Roth and Sant'Anna (2021)
didimputation, did_imputation	R, Stata	Implements Borusyak et al. (2021)
DIDmultiplegt, did_multiplegt	R, Stata	Implements de Chaisemartin and D'Haultfoeuille (2020)
eventstudyinteract	Stata	Implements Sun and Abraham (2021)
flexpaneldid	Stata	Implements Dettmann (2020), based on Heckman et al. (1998)
fixest	R	Implements Sun and Abraham (2021)
stackedev	Stata	Implements stacking approach in Cengiz et al. (2019)
staggered	R	Implements Roth and Sant'Anna (2021), Callaway and Sant'Anna (2021), and Sun and Abraham (2021)
xtevent	Stata	Implements Freyaldenhoven et al. (2019)

DiD with Covariates

Package	Software	Description
DRDID, drdid	R, Stata	Implements Sant'Anna and Zhao (2020)

Diagnostics for TWFE with Staggered Timing

Package	Software	Description
bacondecomp, ddtiming	R, Stata	Diagnostics from Goodman-Bacon (2021)
TwoWayFEWeights	R, Stata	Diagnostics from de Chaisemartin and D'Haultfoeuille (2020)

Diagnostic/ Sensitivity for Violations of Parallel Trends

Package	Software	Description
honestDiD	R, Stata	Implements Rambachan and Roth (2022b)
pretrends	R	Diagnostics from Roth (2022)

Note: This table lists R and Stata packages for recent DiD methods, and is based on Asjad Naqvi's repository at https://asjadnaqvi.github.io/DiD/. Several of the packages listed under "Heterogeneity Robust Estimators" also accommodate covariates.

Table of Contents

1. Introduction

2. Simple DiD
   - Example 1: Cholera in London
   - Example 2: Minimum Wage
   - Example 3: School Construction

3. Extensions of Simple DiD
   - Event Studies
   - DDD

4. Multiple Periods and Staggered Treatment Timing

5. Violations of Parallel Trends

Violations of PT

- Remember that in the canonical DiD model, we had:
  - Two periods and a common treatment date
  - Identification from parallel trends and no anticipation
  - A large number of clusters for inference
- A second literature has focused on relaxing the second assumption: **what if parallel trends may be violated?**
- The ideas from this literature apply even if there is non-staggered timing, although as we'll see, many of the tools can be applied with staggered timing as well. Large clusters is maintained throughout.

Violations of parallel trends

- Three substrands of this literature:
  - **Parallel trends only conditional on covariates**
  - **Testing for violations of (conditional) parallel trends**
  - **Sensitivity analysis and bounding exercises**
- I will focus on the latter two

Why might we be skeptical of PT?

- Recall PT requires the selection bias to be constant over time.
  Why might we be skeptical of this?

Why might we be skeptical of PT?

- Recall PT requires the selection bias to be constant over time. Why might we be skeptical of this?
- There might be different confounding factors in period 1 as in period 0
  - E.g. states that pass a minimum wage increase might also change unemployment insurance at the same time
  - Then UI is a confound in period 1 but not in period 0

Why might we be skeptical of PT?

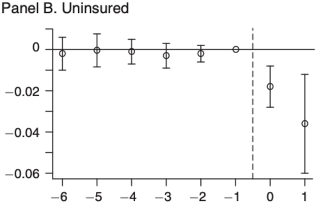
- Recall PT requires the selection bias to be constant over time.
  Why might we be skeptical of this?
- There might be different confounding factors in period 1 as in period 0
  - E.g. states that pass a minimum wage increase might also change unemployment insurance at the same time
  - Then UI is a confound in period 1 but not in period 0
- The same confounding factors may have different effects on the outcome in different time periods
  - Suppose people who enroll in a job training program are more motivated to find a job
  - Motivation might matter more in a bad economy than in a good economy

Why might we be skeptical of PT? Part 2

- Another reason to be skeptical of parallel trends is that its validity will often be **functional form** dependent
- Consider an example:
  - In period 0, all control units have outcome 10; all treated units have outcome 5.
  - In period 1, all control units have outcome 15.
  - If treatment hadn't occurred, would treated units' outcome have increased by 5 also (PT in levels)?
  - Or would they have increased by 50% ($\sim$ PT in logs)?

Pre-trends to the rescue...

- Luckily, in most DiD applications we have several periods before anyone was treated
- We can test whether the groups were moving in parallel prior to the treatment
    - If so, then assumption that confounding factors are stable seems more plausible
    - If not, then it's relatively implausible that would have magically started moving in parallel after treatment date
- Testing for pre-trends provides a natural plausibility check on the parallel trends assumption



- Carey, Miller, and Wherry (2020) do a DiD comparing states who expanded Medicaid in 2014 to states that didn't.
- Report results from "event-study" regression:

$$Y_{its} = \phi_t + \lambda_s + \sum_{r \neq -1} (D_i \times 1[t = 2014 + r]) \cdot \beta_r + \epsilon_{it}$$

where Y_{its} is insurance for person i in year t in state s, and $D_i = 1$ if in an expansion state.

- Testing for pre-existing trends is a very natural way to assess the plausibility of the PT assumption
- But it also has several *limitations*, highlighted in recent work (Freyaldenhoven et al., 2019; Kahn-Lang and Lang, 2020; Bilinski and Hatfield, 2018; Roth, 2022)
- The remainder of this section will focus on these issues, as well as some solutions.
- We will focus mainly on these two papers:
  - Roth (2022 AER:I, “Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends")
  - Rambachan and Roth (2023 RESTUD, “A More Credible Approach to Parallel Trends”)

Overview of Limitations

- Parallel pre-trends doesn't necessarily imply parallel (counterfactual) post-treatment trends
  - If other policies change at the same time as the one of interest — e.g. min wage and UI reform together — can produce parallel pre-trends but non-parallel post-trends
  - Likewise, could be that treated/control groups are differentially exposed to recessions, but there is only a recession in the post-treatment period

Overview of Limitations

- Parallel pre-trends doesn't necessarily imply parallel (counterfactual) post-treatment trends
  - If other policies change at the same time as the one of interest — e.g. min wage and UI reform together — can produce parallel pre-trends but non-parallel post-trends
  - Likewise, could be that treated/control groups are differentially exposed to recessions, but there is only a recession in the post-treatment period
- **Low power:** even if pre-trends are non-zero, we may fail to detect it statistically

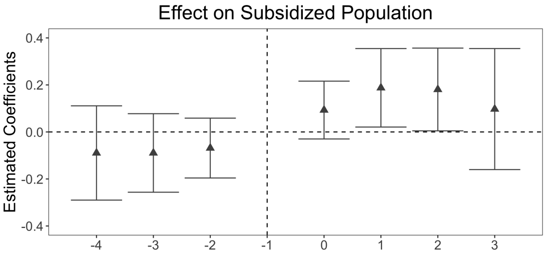
Overview of Limitations

- Parallel pre-trends doesn't necessarily imply parallel (counterfactual) post-treatment trends
  - If other policies change at the same time as the one of interest — e.g. min wage and UI reform together — can produce parallel pre-trends but non-parallel post-trends
  - Likewise, could be that treated/control groups are differentially exposed to recessions, but there is only a recession in the post-treatment period
- **Low power:** even if pre-trends are non-zero, we may fail to detect it statistically
- **Pre-testing issues:** if we only analyze cases without statistically significant pre-trends, this introduces a form of selection bias (which can make things worse)

Overview of Limitations

- Parallel pre-trends doesn't necessarily imply parallel (counterfactual) post-treatment trends
    - If other policies change at the same time as the one of interest — e.g. min wage and UI reform together — can produce parallel pre-trends but non-parallel post-trends
    - Likewise, could be that treated/control groups are differentially exposed to recessions, but there is only a recession in the post-treatment period
- **Low power:** even if pre-trends are non-zero, we may fail to detect it statistically
- **Pre-testing issues:** if we only analyze cases without statistically significant pre-trends, this introduces a form of selection bias (which can make things worse)
- If we fail the pre-test, what next? May still want to write a paper (especially if violation is "small")

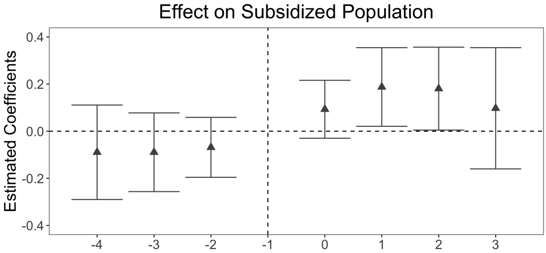
Issue 1 - Low Power



- He & Wang (2017) study impacts of placing college grads as village officials in China
- Use an "event-study" approach comparing treated and untreated villages

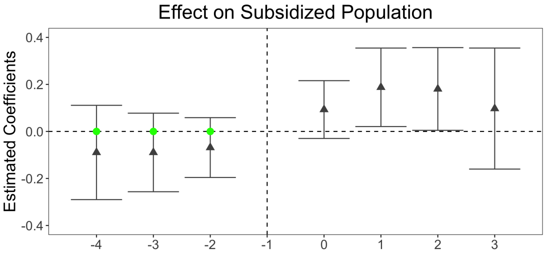
$$Y_{it} = \sum_{k \neq -1} D_{it}^{k} \beta_k + \alpha_i + \phi_t + \epsilon_{it}$$

Issue 1 - Low Power



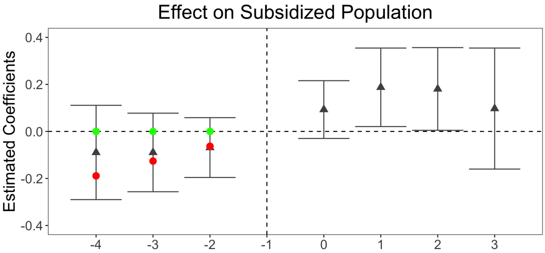
"The estimated coefficients on the leads of treatment ... are statistically indifferent from 0. ... We conclude that the pretreatment trends in the outcomes in both groups of villages are similar, and villages without CGVOs can serve as a suitable control group for villages with CGVOs in the treatment period." (He and Wang, 2017)

Issue 1 - Low Power



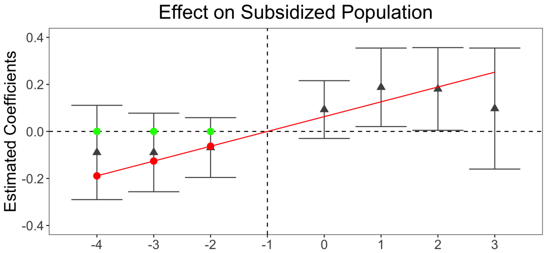
- P-value for $H_0 : \beta_{pre} =$ green dots (no pre-trend): 0.81

Issue 1 - Low Power



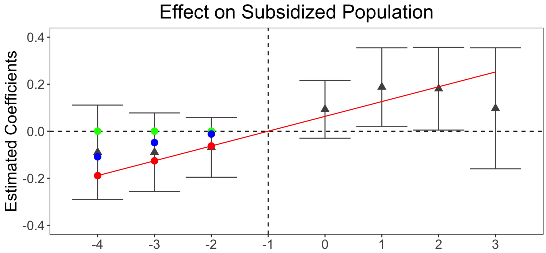
- P-value for $H_0 : \beta_{pre} =$ green dots (no pre-trend): 0.81
- P-value for $H_0 : \beta_{pre} =$ red dots: 0.81

Issue 1 - Low Power



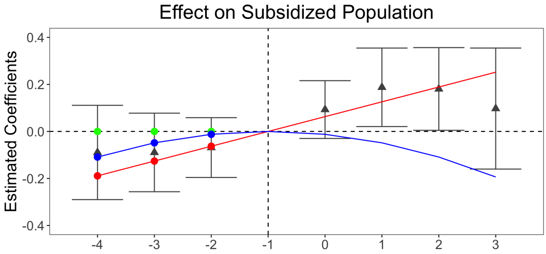
- P-value for $H_0 : \beta_{pre} =$ green dots (no pre-trend): 0.81
- P-value for $H_0 : \beta_{pre} =$ red dots: 0.81

Issue 1 - Low Power



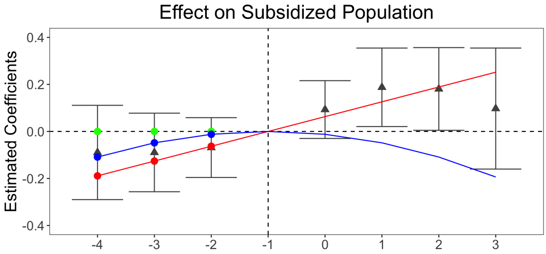
- P-value for $H_0 : \beta_{pre} =$ green dots (no pre-trend): 0.81
- P-value for $H_0 : \beta_{pre} =$ red dots: 0.81
- P-value for $H_0 : \beta_{pre} =$ blue dots: 0.81

Issue 1 - Low Power



- P-value for $H_0 : \beta_{pre} =$ green dots (no pre-trend): 0.81
- P-value for $H_0 : \beta_{pre} =$ red dots: 0.81
- P-value for $H_0 : \beta_{pre} =$ blue dots: 0.81

Issue 1 - Low Power



- P-value for $H_0 : \beta_{pre} =$ green dots (no pre-trend): 0.81
- P-value for $H_0 : \beta_{pre} =$ red dots: 0.81
- P-value for $H_0 : \beta_{pre} =$ blue dots: 0.81
- We can't reject zero pre-trend, but we also can't reject pre-trends that under smooth extrapolations to the post-treatment period would produce substantial bias

More systematic evidence

- Roth (2022): simulations calibrated to papers published in *AER*, *AEJ: Applied*, and *AEJ: Policy* between 2014 and mid-2018
  - 70 total papers contain an event-study plot; focus on 12 w/available data
- Evaluate properties of standard estimates/CIs under linear violations of parallel trends against which conventional tests have limited power (50 or 80%):
  1. Bias often of magnitude similar to estimated treatment effect
  2. Confidence intervals substantially undercover in many cases
  3. Distortions from pre-testing can further exacerbate these issues

Issue 2 - Distortions from Pre-testing

- When parallel trends is violated, we will sometimes fail to find a significant pre-trend
- But the draws of data where this happens are a **selected sample**. This is known as *pre-test bias*.
- Analyzing this selected sample introduces additional statistical issues, and can make things worse!

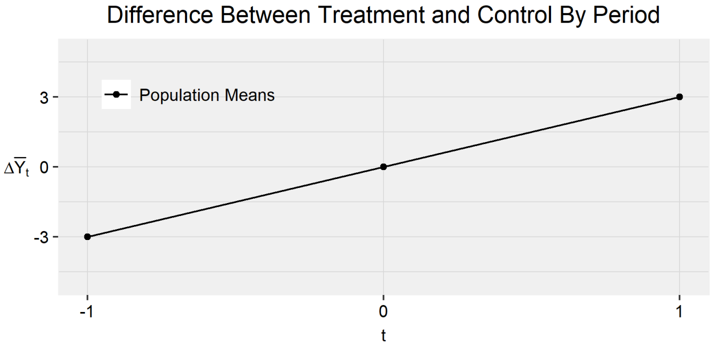
Stylized Three-Period DiD Example

- Consider a 3-period model ($t = -1, 0, 1$) where treatment occurs in last period

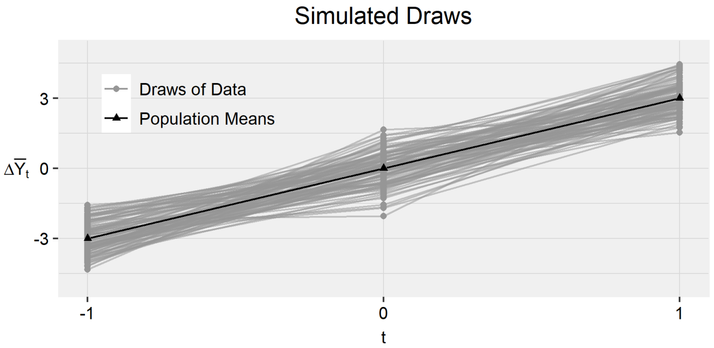
- No causal effect of treatment: $Y_{it}(0) = Y_{it}(1)$ in all periods

- In population, treatment group is on a linear trend relative to the control group with slope δ
  - Control group mean in period t: $E[Y_{it}(0) \mid \text{Control group}] = 0$
  - Treatment group mean in period t: $E[Y_{it}(0) \mid \text{Treated group}] = \delta \cdot t$

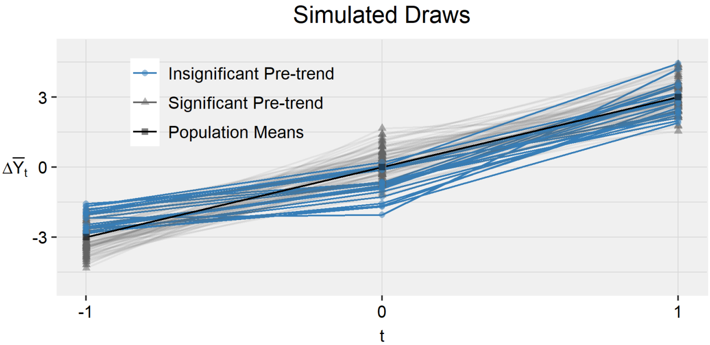
- Simulate from this model with Y_{it} equal to the group mean plus independent normal errors



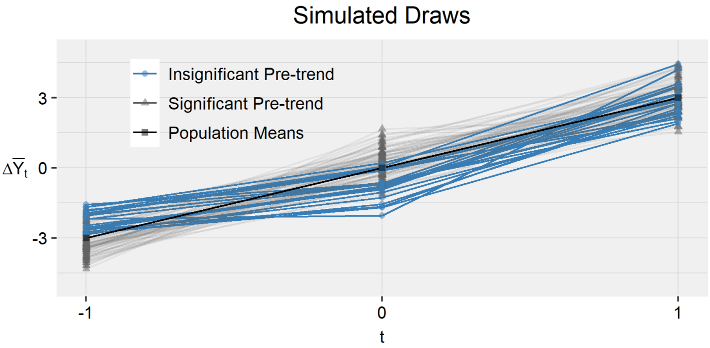
- Example: In population, there is a linear difference in trend with slope 3



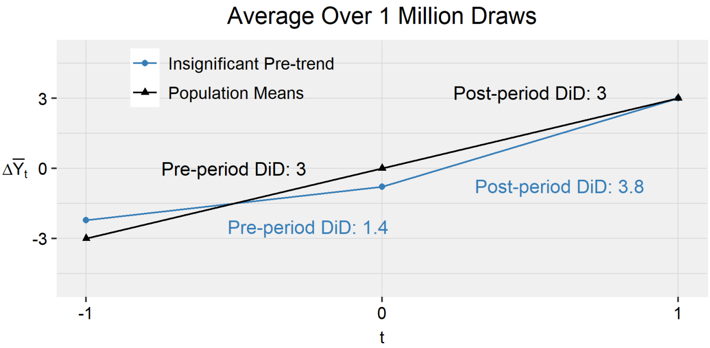
- Example: In population, there is a linear difference in trend with slope 3
- In actual draws of data, there will be noise around this line



- Example: In population, there is a linear difference in trend with slope 3
- In some of the draws of the data, highlighted in blue, the difference between period -1 and 0 will be insignificant



- In some of the draws of the data, highlighted in blue, the difference between period -1 and 0 will be insignificant
- In the insignificant draws, we tend to underestimate the difference between treatment and control at $t = 0$



- In the insignificant draws, we tend to underestimate the difference between treatment and control at $t = 0$
- As a result, the DiD between period 0 and 1 tends to be particularly large when we get an insignificant pre-trend

To Summarize

What are the Limitations of Pre-trends Testing?

1. Low Power – May not find significant pre-trend even if PT is violated
2. Pre-testing Issues – Selection bias from only analyzing cases with insignificant pre-trend
3. If reject pre-trends test, what comes next?

To Summarize

What are the Limitations of Pre-trends Testing?

1. Low Power – May not find significant pre-trend even if PT is violated
2. Pre-testing Issues – Selection bias from only analyzing cases with insignificant pre-trend
3. If reject pre-trends test, what comes next?

What Can We Do About It?

1. Diagnostics of power and distortions from pre-testing (Roth, 2022, “Pre-Test with Caution..., AER, Insight, 2022"). See `pretrends` package.
2. Formal sensitivity analysis that avoids pre-testing (Rambachan and Roth, 2023, ReStud, “A More Credible Approach...”). See `HonestDiD` package.

“A More Credible Approach to Parallel Trends”

- The intuition motivating pre-trends testing is that the pre-trends are informative about counterfactual post-treatment trends
- Formalize this by imposing the restriction that the counterfactual difference in trends can’t be “too different” than the pre-trend
- This allows us to bound the treatment effect and obtain uniformly valid (“honest”) confidence sets under the imposed restrictions
- Enables **sensitivity analysis:** How different would the counterfactual trend have to be from the pre-trends to negate a conclusion (e.g. a positive effect)?

Restrictions on Violations of PT

- Consider the 3-period model ($t = -1, 0, 1$) where treatment occurs in last period
- Let δ_1 be the violation of PT:

$$\delta_1 = \mathbb{E}\left[Y_{i,t=1}(0) - Y_{i,t=0}(0) \,|\, D_i = 1\right] - \mathbb{E}\left[Y_{i,t=1}(0) - Y_{i,t=0}(0) \,|\, D_i = 0\right]$$

- We don't directly identify δ_1, but we do identify its pre-treatment analog, δ_{-1}:

$$\delta_{-1} = \mathbb{E}\left[Y_{i,t=-1}(0) - Y_{i,t=0}(0) \,|\, D_i = 1\right] - \mathbb{E}\left[Y_{i,t=-1}(0) - Y_{i,t=0}(0) \,|\, D_i = 0\right]$$

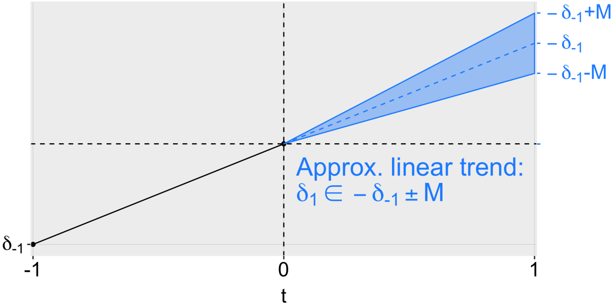
- Key idea: restrict possible values of δ_1 given δ_{-1}
  Intuitively, counterfactual trend can't be too different from pre-trend

Examples of Restrictions on δ

- **Bounds on relative magnitudes:** Require that $|\delta_1| \leq \bar{M}|\delta_{-1}|$

Examples of Restrictions on δ

- **Bounds on relative magnitudes:** Require that $|\delta_1| \leq \bar{M}|\delta_{-1}|$
- **Smoothness restriction:** Bound how far δ_1 can deviate from a linear extrapolation of the pre-trend: $\delta_1 \in [-\delta_{-1} - M, -\delta_1 + M]$



Robust confidence intervals

- In the paper, they develop confidence intervals for the treatment effect of interest under the assumptions on δ discussed above
- The CIs account for the fact that we don't observe the true (population) pre-trend δ_{pre}, only our estimate $\widehat{\beta}_{pre}$.
- The robust CIs tend to be wider the larger are the confidence intervals on the pre-trends — intuitive, since if we know less about the pre-trends, we should have more uncertainty
- This contrasts with pre-trends tests, where you're less likely to reject the null that $\beta_{pre} = 0$ when the SEs are larger!

Benzarti & Carloni (2019)

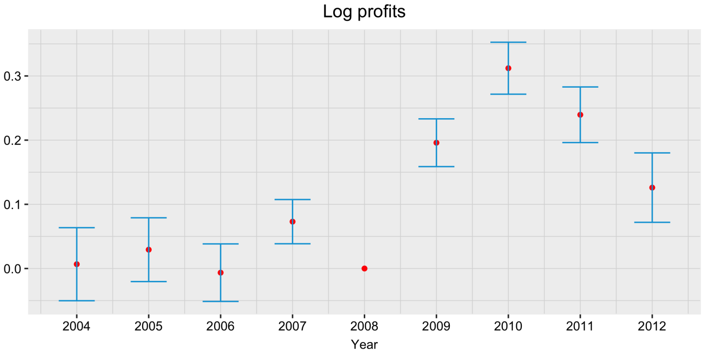
- BC study the incidence of a cut in the value-added tax on sit-down restaurants in France. France reduced the VAT on restaurants from 19.6 to 5.5 percent in July of 2009.
- BC analyze the impact of this change using a difference-in-differences design comparing restaurants to a control group of other market services firms

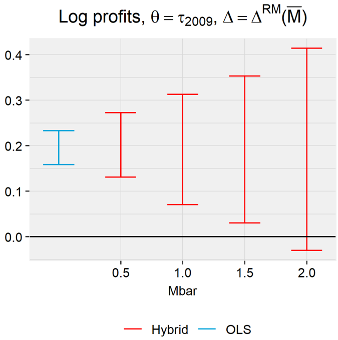
$$Y_{irt} = \sum_{s=2004}^{2012} \beta_s \times 1[t = s] \times D_{ir} + \phi_i + \lambda_t + \epsilon_{irt}, \tag{2}$$

  - Y_{irt} = outcome of interest for firm i in region r
  - D_{ir} = indicator if firm i in region r is a restaurant
  - Φ_i, λ_t = firm and year FEs

- Outcomes of interest include firm profits, prices, wage bill & employment. We focus on impact on profits in first year after reform.

Event-study coefficients for log profits





- "Breakdown" $\bar{M}$ for null effect is ~ 2
- Can rule out a null effect unless allow for violations of PT 2x larger than the max in pre-period

Wrapping Up

- Tests of pre-trends are intuitive but not a panacea!
- Roth (2021) and Rambachan and Roth (2021) provide tools for diagnostics and sensitivity analysis
- It's important to incorporate *context-specific* knowledge when using these tools.
- **Think about how parallel trends may be violated in your context!** This puts the "econ" back into "econometrics"

References I

Bilinski, A. and Hatfield, L. A. (2018). Seeking evidence of absence: Reconsidering tests of model assumptions. *arXiv:1805.03273 [stat]*.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event-study designs: robust and efficient estimation. *Review of Economic Studies*, page rdae007.

Currie, J., Kleven, H., and Zwiers, E. (2020). Technology and big data are changing economics: Mining text to track methods. In *AEA Papers and Proceedings*, volume 110, pages 42–48. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Freyaldenhoven, S., Hansen, C., and Shapiro, J. M. (2019). Pre-event Trends in the Panel Event-Study Design. *American Economic Review*, 109(9):3307–3338.

Gardner, J. (2022). Two-stage differences in differences. *arXiv preprint arXiv:2207.05943*.

References II

Kahn-Lang, A. and Lang, K. (2020). The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications. *Journal of Business & Economic Statistics*, 38(3):613–620.

Roth, J. (2022). Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends. *American Economic Review: Insights*, 4(3):305–322.

Roth, J. and Sant'Anna, P. H. (2023). Efficient estimation for staggered rollout designs. *Journal of Political Economy Microeconomics*, 1(4):669–709.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics*, 225(2):175–199.

Wooldridge, J. (2022). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators; 2021. *Available at SSRN*.