Causal Inference: Difference-in-Differences

POLI 803 Research Methods in PS

Howard Liu

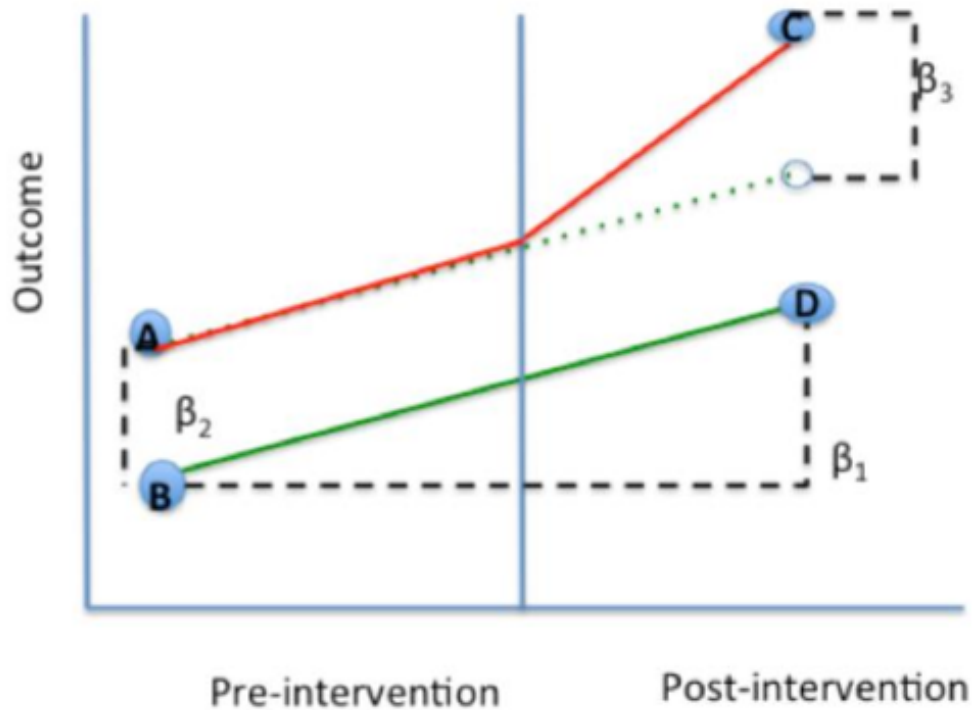
2025

Roadmap

What is difference-in-differences (DiD)

$$y = \beta_0 + \beta_1 time + \beta_2 treated + \beta_3 time * treated + \varepsilon$$

Coefficient	Calculation	Interpretation
β_0	B	Baseline average
β_1	D-B	Time trend in control group
β_2	A-B	Difference between two groups pre-intervention
β_3	(C-A)-(D-B)	Difference in changes over time



What is difference-in-differences (DiD)

- group x time: DiD is when a **group of units** are assigned some treatment and then compared to a group of units that weren't **before and after**
- One of the most widely used quasi-experimental methods in economics and increasingly in industry
- Panel data: uses panel or repeated cross section datasets, binary treatments usually, and often covariates

Example: The Cholera Death Investigation

- Mysterious rise of deaths in a London

Example: The Cholera Death Investigation

- Mysterious rise of deaths in a London
- John Snow's Cholera hypothesis: water with cholera $\rightarrow$ deaths. But how to prove it? What would you do?

Example: The Cholera Death Investigation

- Mysterious rise of deaths in a London
- John Snow's Cholera hypothesis: water with cholera $\rightarrow$ deaths. But how to prove it? What would you do?
- Collect data from two towns: one with clean water (from Lambeth or L) and one with polluted water (from Southwark and Vauxhall, or SV)
- Compare death rate before and after treatment (clean water) in these two towns
- **Goal: Show death rate is constant over time in the treatment group while the rate rises in the control group**

Example: The Cholera Death Investigation

- Mysterious rise of deaths in a London
- John Snow's Cholera hypothesis: water with cholera $\rightarrow$ deaths. But how to prove it? What would you do?
- Collect data from two towns: one with clean water (from Lambeth or L) and one with polluted water (from Southwark and Vauxhall, or SV)
- Compare death rate before and after treatment (clean water) in these two towns
- **Goal: Show death rate is constant over time in the treatment group while the rate rises in the control group**

Table 9.1: Modified Table XII (Snow 1854).

Company name	1849	1854
Southwark and Vauxhall	135	147
Lambeth	85	19

Example: The Cholera Death Investigation

Compared to what? Difference in each company's differences.

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + T + D$	$T + D$	
				D
Southwark and Vauxhall	Before	$Y = SV$		
	After	$Y = SV + T$	T	

- L, SV = town fixed effects
- T = natural changes in the cholera deaths over time
- D = treatment effect (clean water)

Example: The Cholera Death Investigation

Compared to what? Difference in each company's differences.

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + T + D$	$T + D$	
				D
Southwark and Vauxhall	Before	$Y = SV$		
	After	$Y = SV + T$	T	

- L, SV = town fixed effects
- T = natural changes in the cholera deaths over time
- D = treatment effect (clean water)
- D_1: First difference (within units) removes unit fixed effects
- D_2: Second difference (betwewn units) removes temporal effects
- After D_1 and D_2 or so-called difference-in-differences (DiD), we have an unbiased treatment effect, D
- It is also why people say conceptually DiD is just a two-way fixed effects (TWFE) model

Example: The Cholera Death Investigation

Compared to what? Difference in each company's differences.

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + T + D$	$T + D$	
				D
Southwark and Vauxhall	Before	$Y = SV$		
	After	$Y = SV + T$	T	

- What are underlying problems in this approach?

Example: The Cholera Death Investigation

Compared to what? Difference in each company's differences.

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + T + D$	$T + D$	
				D
Southwark and Vauxhall	Before	$Y = SV$		
	After	$Y = SV + T$	T	

- What are underlying problems in this approach?
- Assuming there is **no unobserved temporal trends** that determines the death rates during these two time periods in these two towns E.g. People in the nearby town, Lambeth, became concerned due to the devastating news in SV and fled London.
- We call it **parallel trends** assumption

Three waves of DiD: The First Wave

- Difference-in-differences evolved in three waves from 1983 to present in economics

Three waves of DiD: The First Wave

- Difference-in-differences evolved in three waves from 1983 to present in economics
- First wave lasts from 1983 to 2011; second wave from 2011 to 2018; third wave from 2018 to present

Three waves of DiD: The First Wave

- Difference-in-differences evolved in three waves from 1983 to present in economics
- First wave lasts from 1983 to 2011; second wave from 2011 to 2018; third wave from 2018 to present
- Initially, mostly used by labor economists throughout the 1990s in the “program evaluation” area but spreads with the spread of causal inference

Three waves of DiD: The First Wave

- Difference-in-differences evolved in three waves from 1983 to present in economics
- First wave lasts from 1983 to 2011; second wave from 2011 to 2018; third wave from 2018 to present
- Initially, mostly used by labor economists throughout the 1990s in the “program evaluation” area but spreads with the spread of causal inference
- No **potential outcomes notation**, no mention of **parallel trends**, no **event studies**

Second wave grows faster than first

- Share of working papers that used diff-in-diff went from 11-12% in 2011 to 23% in 2018, or 11 percentage points

Second wave grows faster than first

- Share of working papers that used diff-in-diff went from 11-12% in 2011 to 23% in 2018, or 11 percentage points
- It had taken 24 years to reach 11% the first time, but only 7 years the second time.

Second wave grows faster than first

- Share of working papers that used diff-in-diff went from 11-12% in 2011 to 23% in 2018, or 11 percentage points
- It had taken 24 years to reach 11% the first time, but only 7 years the second time.
- 2011 marks the start of the second wave and it has very clear patterns, including the growing speed of adoption

Second wave grows faster than first

- Share of working papers that used diff-in-diff went from 11-12% in 2011 to 23% in 2018, or 11 percentage points
- It had taken 24 years to reach 11% the first time, but only 7 years the second time.
- 2011 marks the start of the second wave and it has very clear patterns, including the growing speed of adoption
- Wave 2 links **parallel trends** and **event studies** to difference-in-differences for the first time

Second wave grows faster than first

- Share of working papers that used diff-in-diff went from 11-12% in 2011 to 23% in 2018, or 11 percentage points
- It had taken 24 years to reach 11% the first time, but only 7 years the second time.
- 2011 marks the start of the second wave and it has very clear patterns, including the growing speed of adoption
- Wave 2 links **parallel trends** and **event studies** to difference-in-differences for the first time

Third wave: heterogeneity and TWFE

- Third (and current) wave has been characterized by scrutiny of the twoway fixed effects (**TWFE**) model starting around 2018

Third wave: heterogeneity and TWFE

- Third (and current) wave has been characterized by scrutiny of the twoway fixed effects (**TWFE**) model starting around 2018
- Economists were using TWFE because it's a panel estimator and diff-in-diff could be used with panel data (or repeated cross-sections)

Third wave: heterogeneity and TWFE

- Third (and current) wave has been characterized by scrutiny of the twoway fixed effects (**TWFE**) model starting around 2018
- Economists were using TWFE because it's a panel estimator and diff-in-diff could be used with panel data (or repeated cross-sections)
- But subtle assumptions are buried in the details, or not made explicit at all, related to **heterogeneous vs constant treatment effects**
- Wave 3 links difference-in-differences with **heterogenous treatment effects**, the pathologies of twoway fixed effects (TWFE) and **begins to shift away from TWFE**
- We introduce DiD by loosly following these three waves.

Roadmap

OLS Regression

$$Y_{ist} = \alpha_0 + \alpha_1 Treat_{is} + \alpha_2 Post_t + \delta(Treat_{is} \times Post_t) + \varepsilon_{ist}$$

$$\widehat{\delta} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)}\right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)}\right)$$

- These two equations are numerically identical
- You can see it from the Cholera problem we've talked about

Introducing Potential Outcomes to DiD

- We want to know when does the DiD equation identify a causal parameter and which one (there are several)?
- We need causality concepts that can be linked to DiD if we are to answer this question
- Potential outcomes notation is the main language of modern causal inference and is rooted in the early experimental design writers like Ronald Fisher and Jerzey Neyman, as well as modern statisticians like Don Rubin

DiD equation is the 2x2

Orley Ashenfelter's "four averages and three subtractions" uses two groups, two time periods, or 2x2

$$\widehat{\delta} = \Big(E[Y_k|Post] - E[Y_k|Pre] \Big) - \Big(E[Y_U|Post] - E[Y_U|Pre] \Big)$$

k are the people in the job training program, U are the untreated people not in the program, $Post$ is after the trainees took the class, Pre is the period just before they took the class, and $E[y]$ is mean earnings.

When will $\widehat{\delta}$ equal the ATT? When will it not?

Replace with potential outcomes and add a zero

$$\widehat{\delta} = \underbrace{\Big(E[Y_k^1|Post] - E[Y_k^0|Pre]\Big) - \Big(E[Y_U^0|Post] - E[Y_U^0|Pre]\Big)}_{\text{Switching equation}}$$

$$+ \underbrace{E[Y_k^0|Post] - E[Y_k^0|Post]}_{\text{Adding zero}}$$

Parallel trends bias

$$\widehat{\delta} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{\text{ATT}} + \underbrace{\Big[E[Y_k^0|Post] - E[Y_k^0|Pre]\Big] - \Big[E[Y_U^0|Post] - E[Y_U^0|Pre]\Big]}_{\text{Non-parallel trends bias in 2x2 case}}$$

Identification through parallel trends

Parallel trends

Assume two groups, treated and comparison group, then we define parallel trends as:

$$E(\Delta Y_k^0) = E(\Delta Y_U^0)$$

In words: "The evolution of earnings for our trainees *had they not trained* is the same as the evolution of mean earnings for non-trainees".

Identification through parallel trends

Parallel trends

Assume two groups, treated and comparison group, then we define parallel trends as:

$$E(\Delta Y_k^0) = E(\Delta Y_U^0)$$

In words: "The evolution of earnings for our trainees *had they not trained* is the same as the evolution of mean earnings for non-trainees".

It's in red because parallel trends is untestable and critically important to estimation of the ATT using any method, OLS or "four averages and three subtractions"

What is and is not parallel trends?

- Parallel trends does *not* mean treatments were **randomly assigned** (though random assignment guarantees parallel trends)

What is and is not parallel trends?

- Parallel trends does *not* mean treatments were **randomly assigned** (though random assignment guarantees parallel trends)
- Parallel trends does *not* require that the groups be similar at baseline on outcomes (though random assignment guarantees that would be)

What is and is not parallel trends?

- Parallel trends does *not* mean treatments were **randomly assigned** (though random assignment guarantees parallel trends)
- Parallel trends does *not* require that the groups be similar at baseline on outcomes (though random assignment guarantees that would be)
- Parallel trends *does* require that the **comparison group** follows a ”trend” in outcomes that is approximately the same as the counterfactual trend of the treatment group (what would have had happened had the treatment not occurred) $\rightarrow$ remember the first slide?

Three main DiD assumptions

- Parallel trends is the most common one and most well known
- But parallel trends is nested within a bundle of assumptions, and all of them are need for traditional difference-in-differences
- Other two lesser known assumptions are ”No anticipation” (or NA) and Stable Unit Treatment Value Assumption (SUTVA)

No Anticipation

- "No anticipation" simply means that the unit is not treated until it is treated (and that can be violated with rational forward looking agents but not always)
  - → **Example 1**: Tomorrow I win the lottery, but don't get paid yet. I decide to buy a new house today. That violates NA
  - → **Example 2**: Next year, a state lets you drive without a driver license and you know it. But you can't drive without a driver license today. This satisfies NA.

SUTVA

- Stable Unit Treatment Value Assumption (Imbens and Rubin 2015) focuses on what happens when in our analysis we are combining units (versus defining treatment effects)
    1. **No Interference**: a treated unit cannot impact a control unit such that their potential outcomes change (unstable treatment value)
    2. **No hidden variation in treatment**: When units are indexed to receive a treatment, their dose is the same as someone else with that same index
    3. **Scale**: If scaling causes interference or changes inputs in production process, then #1 or #2 are violated

Summarizing

- Lots of restrictions placed on difference-in-differences
  - → NA: you chose a baseline that is not treated
  - → SUTVA: your comparison group is never treated during the course of the calculations
  - → PT: your comparison group has a trend in $E[Y^0]$ that is the same as the counterfactual

Summarizing

- Lots of restrictions placed on difference-in-differences
  - → NA: you chose a baseline that is not treated
  - → SUTVA: your comparison group is never treated during the course of the calculations
  - → PT: your comparison group has a trend in $E[Y^0]$ that is the same as the counterfactual
- Only when you have NA and SUTVA, does DiD equal (ATT + PT)
- It's crucial to remember: DiD and ATT are not the same thing
- But in practice, NA and SUTVA are often ignored..

OLS is the to-go specification

- Simple DiD equation will identify ATT under parallel trends
- But so will a particular OLS specification (two groups and no covariates)
- OLS was historically preferred because
  - → OLS estimates the ATT under parallel trends
  - → Easy to calculate the standard errors
  - → Easy to include multiple periods
- People liked it also because of differential timing, continuous treatments, and covariates,

OLS specification of the DiD equation

- The correctly specified OLS regression is an interaction with time and group fixed effects:

$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

  - → NJ is a dummy equal to 1 if the observation is from NJ
  - → d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
  - → PA Pre: α
  - → PA Post: $\alpha + \lambda$
  - → NJ Pre: $\alpha + \gamma$
  - → NJ Post: $\alpha + \gamma + \lambda + \delta$
- DiD equation: (NJ Post - NJ Pre) - (PA Post - PA Pre) $= \delta$

Inference in DID

When dealing with clustered data, a crucial concept is the difference between correlated observations and correlated errors. While they may seem similar, they are distinct, and it’s essential to focus on the errors when clustering standard errors.

Correlated Observations

- Correlated observations occur when the observed variables themselves are correlated within a cluster.
- For instance, incomes within a specific region might be positively correlated.
- Correlated observations do not necessarily violate OLS assumptions.

Correlated Errors

- Correlated errors occur when the unobserved errors are correlated within a cluster.
- This violates the assumption of independent errors, leading to possibly biased standard errors and higher over rejection rates
- Failing to account for correlated errors can lead to misleading inference.

Serial correlation (as well as spatial) creates problems

- Bertrand, Duflo and Mullainathan (2004) show that conventional standard errors will often severely understate the standard deviation of the estimators
- They proposed three solutions, but most only use one of them (clustering)
- Clustering standard errors accounts for this within-cluster correlation and is a more conservative approach
- Clustering is typically recommended at the aggregate unit where the entire treatment occurred

Roadmap

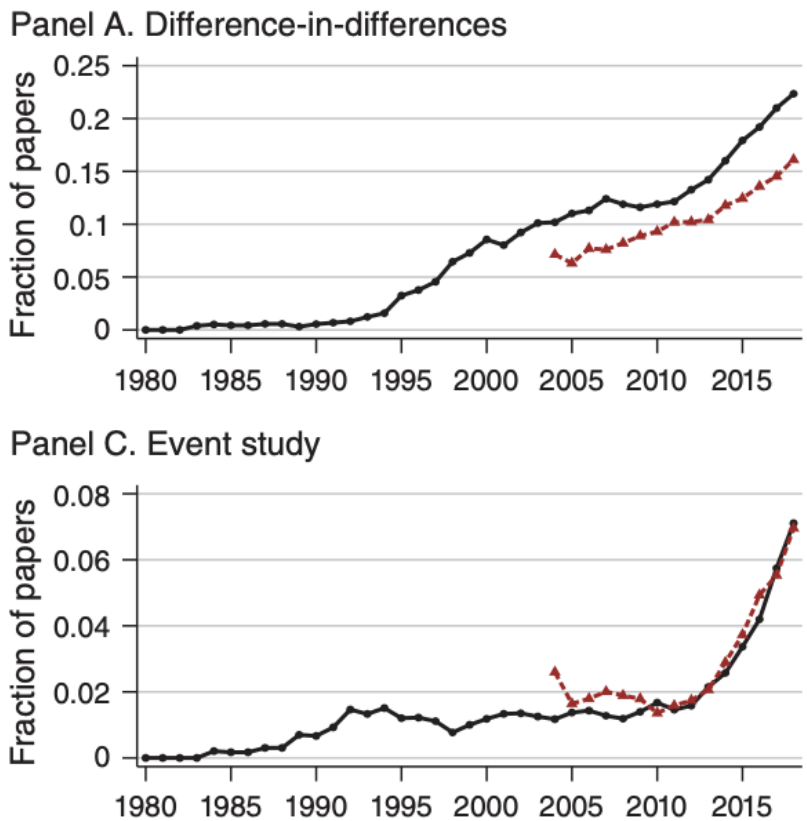
Introduction
- What is difference-in-differences (DiD)
- Three waves of DiD in Economics

Difference-in-Differences
- Potential outcomes
- Identification, Estimation and Inference

Parallel Trends Violations
- Event Studies

Staggered Treatment Difference-in-Differences

Event studies have become mandatory in DiD



Intuition behind event studies

- We cannot directly verify parallel trends, so for a long time researchers have focused on the pre-trends

Intuition behind event studies

- We cannot directly verify parallel trends, so for a long time researchers have focused on the pre-trends
- **Parallel pre-trends** not required for parallel trends and vice versa, but this is the clearest evidence we typically look for nonetheless

Intuition behind event studies

- We cannot directly verify parallel trends, so for a long time researchers have focused on the pre-trends
- **Parallel pre-trends** not required for parallel trends and vice versa, but this is the clearest evidence we typically look for nonetheless
- Think of it as a type of check for selection bias, but imperfect with false positives and false negatives

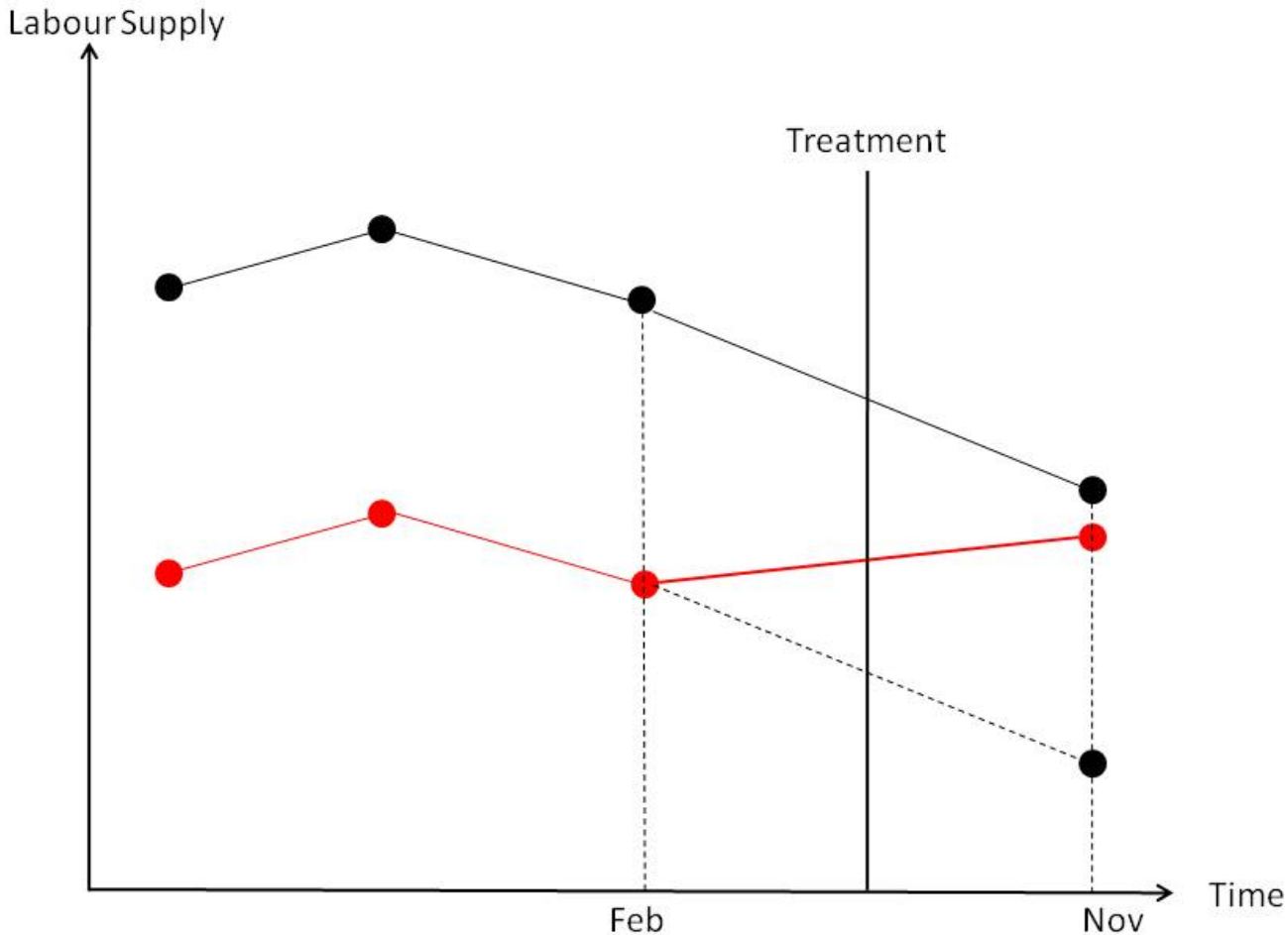
Intuition behind event studies

- We cannot directly verify parallel trends, so for a long time researchers have focused on the pre-trends
- **Parallel pre-trends** not required for parallel trends and vice versa, but this is the clearest evidence we typically look for nonetheless
- Think of it as a type of check for selection bias, but imperfect with false positives and false negatives
- Even if pre-trends are the same, one still has to worry about other policies changing at the same time (omitted variable bias is a parallel trends violation)

Creating event studies

- Originally, there were no event studies (as we saw in the First Wave)
- Economists pulled from finance and took the event study concept and changed it to suit Ashenfelter Dip reasoning
- Always presented **graphically**, but there were different ways people went about it so we will review them and make suggestions

1. Plot the raw data when there's only two groups



2. Event study regression with leads and lags

- Alternatively, present estimated coefficients from a dynamic regression specification:

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau (D_s \times \tau_t) + \sum_{\tau=0}^{m} \delta_\tau (D_s \times \tau_t) + \tau_t + D_s + \varepsilon_{ist}$$

  → With a simple 2x2, you are interacting treatment indicator **with multiple calendar year dummies**
  → Includes q **leads** or anticipatory effects and m **lags** or post treatment effects

2. Event study regression with leads and lags

- Alternatively, present estimated coefficients from a dynamic regression specification:

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau (D_s \times \tau_t) + \sum_{\tau=0}^{m} \delta_\tau (D_s \times \tau_t) + \tau_t + D_s + \varepsilon_{ist}$$

  - → With a simple 2x2, you are interacting treatment indicator **with multiple calendar year dummies**
  - → Includes q **leads** or anticipatory effects and m **lags** or post treatment effects
- Estimated $\widehat{\delta}$ coefficients are estimated ATT parameters assuming parallel trends and $\widehat{\mu}$ is part of your evidence for that
- **Note:** Dynamic regression (or dynamic treatment effects) can be used in the "canonical" version of DiD involves two periods and two groups (2x2) or staggered treatments (multi-period), which we will discuss later.

Event study example: Medicaid and Affordable Care Act example

Volume 136, Issue 3
August 2021

< Previous Next >

Medicaid and Mortality: New Evidence From Linked Survey and Administrative Data

Sarah Miller, Norman Johnson, Laura R Wherry

The Quarterly Journal of Economics, Volume 136, Issue 3, August 2021, Pages 1783–1829,
https://doi.org/10.1093/qje/qjab004
Published: 30 January 2021

Abstract

We use large-scale federal survey data linked to administrative death records to investigate the relationship between Medicaid enrollment and mortality. Our analysis compares changes in mortality for near-elderly adults in states with and without Affordable Care Act Medicaid expansions. We identify adults most likely to benefit using survey information on socioeconomic status, citizenship status, and public program participation. We find that prior to the ACA expansions, mortality rates across expansion and nonexpansion states trended similarly, but beginning in the first year of the policy, there were significant reductions in mortality in states that opted to expand relative to nonexpanders. Individuals in expansion states experienced a 0.132 percentage point decline in annual mortality, a 9.4% reduction over the sample mean, as a result of the Medicaid expansions. The effect is driven by a reduction in disease-related deaths and grows over time. A variety of alternative specifications, methods of inference, placebo tests, and sample definitions confirm our main result.

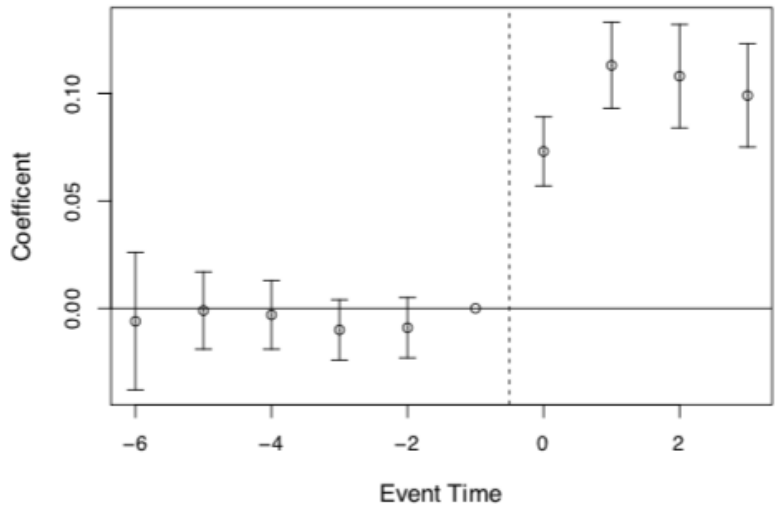
JEL: H75 - State and Local Government: Health; Education; Welfare; Public Pensions, I13 - Health Insurance, Public and Private, I18 - Government Policy; Regulation; Public Health

Issue Section: Article

Their Evidence versus Their Result

- **Bite** – they will show that the expansion shifted people into Medicaid and out of uninsured status
- **Main results** – with all of this, they will show Medicaid expansion caused near elderly mortality to fall
- **Event study** – they will lean hard on those dynamic plots

Medicaid and Affordable Care Act example

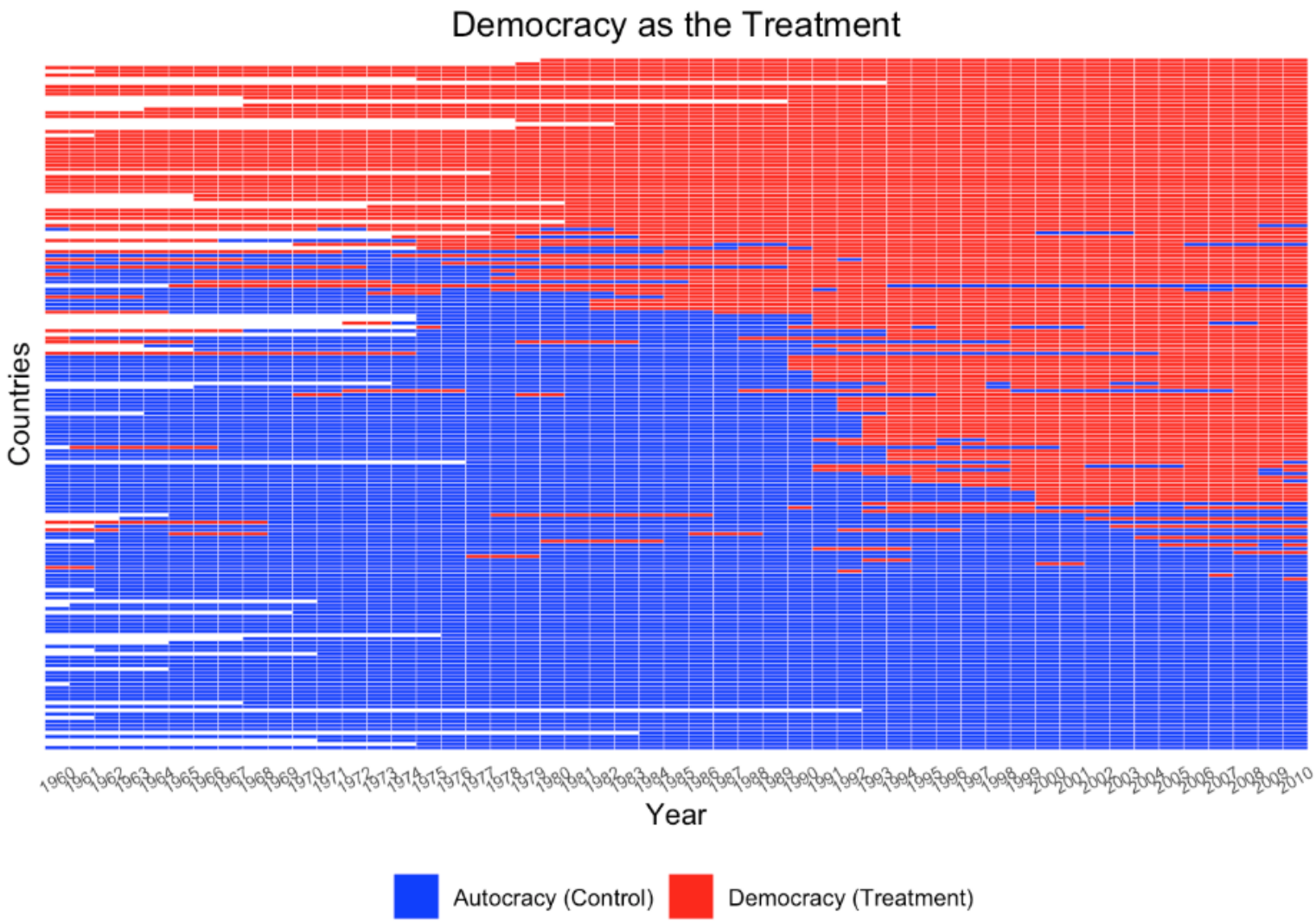


(b) Medicaid Coverage

- The reference period (the dashline): the last year/period before the treatment kick in
- Leads: we want to see mostly insignificant DiD coefficients (evidence for no pre-trend/placebo test)
- Lags: the effect may become ”immediately evident” or ”gradually evident”

Roadmap

Staggered Treatment



- Our lab exercise in the week of matching

Staggered Treatment

- Can we still use the OLS regression with TWFE?

Staggered Treatment

- Can we still use the OLS regression with TWFE?
- We can’t because things are more complex

Staggered Treatment

- Can we still use the OLS regression with TWFE?
- We can't because things are more complex
- Treated groups: always treated (t_{1-n}), later treated (t_{k-n})
- Control groups: never treated, not-yet-treated, previously treated (and got reversed), **which control groups to use?**

Staggered Treatment

- Can we still use the OLS regression with TWFE?
- We can't because things are more complex
- Treated groups: always treated (t_{1-n}), later treated (t_{k-n})
- Control groups: never treated, not-yet-treated, previously treated (and got reversed), **which control groups to use?**
- Parallel pre-trend assumptions are a mess: don't and shouldn't have a consistent pre-trends in the control group $\rightarrow$ standard DiD, TWFE approach failed (third wave of DiD)

Estimiation Strategies for Staggered Treatment

- An active field of research $\rightarrow$ no consenus for now
- Solution: estimate treatment effects for each cohert (that are treated in the same pattern in time) and then average the effect with some weights $\rightarrow$ like PanelMatch by Imai et al.
- The debate now is how we weigh them (groups treated in the middle verse at the end)

Estimiation Strategies for Staggered Treatment

- An active field of research → no consenus for now
- Solution: estimate treatment effects for each cohert (that are treated in the same pattern in time) and then average the effect with some weights → like PanelMatch by Imai et al.
- The debate now is how we weigh them (groups treated in the middle verse at the end)
- Two most recent approaches:
  - → San and Abraham (2020): Only use the ”last cohort” (not the ”not-yet-treated”) in the control groups
  - → Once a group is treated, it is no longer used as a control for others who are treated later.
- Egami and Yamauchi (2023): Use double DID that finds the optimal weights for DiD estimator
  - → optimal weights for combining cohort-specific effects.