Causal Inference II

Mixtape Session

Roadmap

- Introduction
  - Managing expectations
  - Introducing difference-in-differences
  - History
  - Potential outcomes
  - Identification

Introduction

- Welcome to Mixtape Sessions workshop on difference-in-differences and synthetic control (“Causal Inference II”)
- 8:00am to 5:00pm CST, 15 min breaks every hour, 1 hour lunch at noon CST
- Lecture, discussion, exercises, application

Workshop outline

Introduction to DiD basics

- Potential outcomes review
- DiD equation and estimation with OLS
- Evaluating parallel trends with falsifications, event studies
- Triple differences
- Including covariates

Workshop outline

Differential timing

- Bacon decomposition
- Aggregating group-time ATTs
- Issues and solutions with event studies
- Turning treatments on and off
- Stacked regression
- Imputation estimators
- Continuous treatments

Workshop outline

Synthetic control

- Canonical synth


I
federalism
(for the natural experiments)

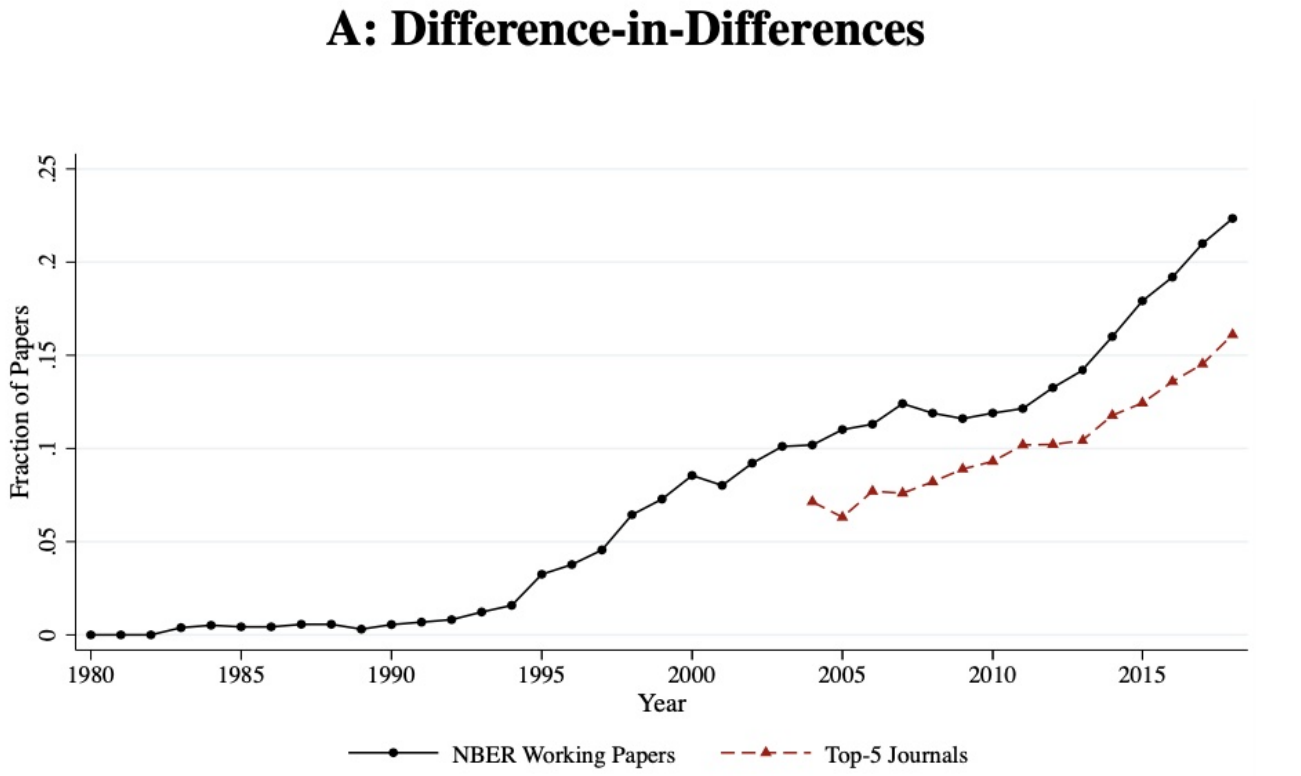
Natural experiments

“A good way to do econometrics is to look for good natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for us.” – Daniel McFadden (Nobel Laureate recipient with Heckman 1992)

What is difference-in-differences (DiD)

- DiD is a very old, relatively straightforward, intuitive research design
- A group of units are assigned some treatment and then compared to a group of units that weren't
- One of the most widely used quasi-experimental methods in economics and even used in industry
- Mostly associated with "big shocks" happening in space over time

Figure: Currie, et al. (2020)



Why an entire workshop on DiD?

- **Research advantages**: DiD is sometimes the only way we have to study large social policies
- **Good time to retool**: Recent wave of scholarship suggest model misspecification is pronounced
- **Good news**: Better understanding of our models, new tools, new programs
- **Hope**: I think we can get to the bottom of this in a way that will stick with you

Brief history of diff-in-diff

- David Card and Orley Ashenfelter are often associated with it in economics, but it goes back further to the 19th century
- Difference-in-differences (DiD) was quietly and largely unnoticed introduced in the 19th century as a way to convince skeptics in health policy arguments
- Dominant disease theory in 19th century was *miasma* – disease caused by smelly vapor
- Keep in mind – microorganisms would not be identified until much later, partly caused by poor resolution in microscopes (Freedman 2007)

Miasma I: Ignaz Semmelweis and washing hands

- 1840s, Vienna maternity wards had high postpartum infections in one wing compared to other wings
- One division had doctors and trainee doctors, but another had midwives and trainee midwives

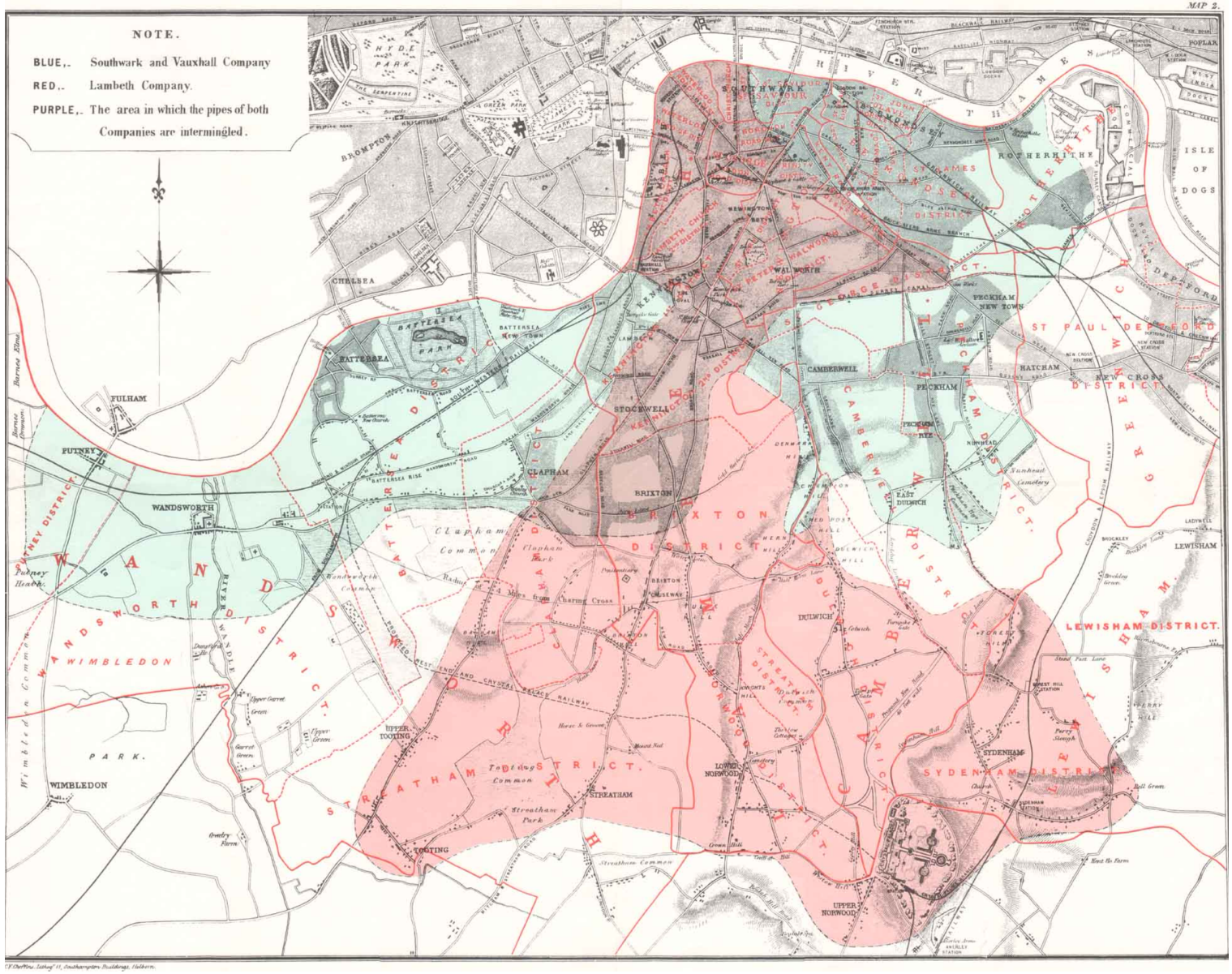
Miasma I: Ignaz Semmelweis and washing hands

- Ignaz Semmelweis notes the difference in 1841 when hospitals moved to “anatomical” training involving cadavers (Pamela Jakeila lecture notes on DiD)
- New training happens to one but not the other and Semmelweis thinks the mortality is caused by working with cadavers
- Proposes washing hands with chlorine in 1847 in the midwives’ wing and uses a DiD design of pre and post

Miasma II: John Snow and cholera

- Three major waves of cholera in the early to mid 1800s in London
- John Snow believed cholera was spread through the Thames water supply which contradicted dominant theory about “dirty air” transmission
- Grand experiment: Lambeth moves its pipe between 1849 and 1854; Southwark and Vauxhall delay
- He can evaluate the effect in three ways (one of which is DiD)

Figure: Two water utility companies in London 1854



1) Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

Company	**Cholera mortality**
Lambeth	$Y = L + D$
Southwark and Vauxhall	$Y = SV$

$$\widehat{\delta}_{cs} = D + (L - SV)$$

What is L and SV?

1) Simple cross-sectional design

Table: Lambeth and Southwark and Vauxhall, 1854

Company	**Cholera mortality**
Lambeth	$Y = L + D$
Southwark and Vauxhall	$Y = SV$

$$\widehat{\delta}_{cs} = D + (L - SV)$$

This is biased if L $\neq SV$ (selection bias). Give an example when we're pretty sure they are equal.

2) Interrupted time series design

Table: Lambeth, 1849 and 1854

Company	Time	Cholera mortality
Lambeth	1849	$Y = L$
	1854	$Y = L + (T + D)$

$$\widehat{\delta}_{its} = D + T$$

What is required for this estimator to be unbiased?

3) Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

Companies	**Time**	**Outcome**	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + T_L + D$	$T_L + D$	
				D
Southwark and Vauxhall	Before	$Y = SV$		
	After	$Y = SV + T_{SV}$	T_{SV}	

$$\widehat{\delta}_{did} = D + (T_L - T_{SV})$$

How do we calculate T_{SV}?

3) Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

Companies	**Time**	**Outcome**	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + T_L + D$	$T_L + D$	
				D
Southwark and Vauxhall	Before	$Y = SV$		
	After	$Y = SV + T_{SV}$	T_{SV}	

$$\widehat{\delta}_{did} = D + (T_L - T_{SV})$$

How do we calculate T_L?

3) Difference-in-differences

Table: Lambeth and Southwark and Vauxhall, 1849 and 1854

Companies	**Time**	**Outcome**	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + T_L + D$	$T_L + D$	
				D
Southwark and Vauxhall	Before	$Y = SV$		
	After	$Y = SV + T_{SV}$	T_{SV}	

$$\widehat{\delta}_{did} = D + (T_L - T_{SV})$$

This second term is called “parallel trends”

Potential outcomes review

- DiD really can't be understood without committing to useful causality notation
- Standard language is the potential outcomes model, sometimes called the Rubin-Neyman model
- Potential outcomes are thought experiments about worlds that never existed, but which *could have*
- Important we are all on the same page, so I'm going to review this

Potential outcomes notation

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 \text{ if pipe inlet is upstream at time } t \\ 0 \text{ if pipe inlet is downstream at time } t \end{cases}$$

where i indexes an individual observation, such as a person

Potential outcomes notation

- Potential outcomes:

$$Y_{i,t}^{j} = \begin{cases} 1\text{: health if drank from upstream at time } t \\ 0\text{: health if drank from downstream at time } t \end{cases}$$

where j indexes a counterfactual state of the world

Potential vs observed

- A potential outcome Y^1 and an observed outcome Y are distinct
- Potential outcomes are *hypothetical* possibilities describing states of the world but historical outcomes actually occurred
- Potential outcomes become observed outcomes when treatments are assigned (the "switching equation")

$$Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$$

- I'll often drop the subscripts to reduce clutter, but note the key point here is one's own realized outcome at some point in time is based on one's own treatment assignment at that time

Treatment effect definitions

Individual treatment effect

The individual treatment effect, δ_i, equals $Y_i^1 - Y_i^0$

Individual causal effects cannot be calculated because one of the two needed potential outcomes will always be missing. Epistemologically "unknowable" in some important but difficult to define way.

Conditional Average Treatment Effects

Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

$$\begin{aligned} E[\delta|D=1] &= E[Y^1 - Y^0|D=1] \\ &= E[Y^1|D=1] - E[Y^0|D=1] \end{aligned}$$

Again that “epistemological” uncertainty. We can estimate the ATT, but never be sure due to missing potential outcomes for the treated group

Identification without randomization

- We may be unable to randomize – not because we lack the imagination, but because we lack the permission
- If we cannot randomize, then how does DiD identify a treatment effect, and which treatment effect?
- DiD identifies the ATT, and since we are missing Y^0 for treated group, we will restrict counterfactual Y^0 in expectation
- One of the main advantages of DiD is the hope that we can identify the ATT *without* randomization

DiD equation

I call this the DiD equation, but Goodman-Bacon calls it the “2x2”; I’ll use his k and U notation for treated and untreated groups

$$\widehat{\delta}_{kU}^{2x2} = \Big(E[Y_k|Post] - E[Y_k|Pre] \Big) - \Big(E[Y_U|Post] - E[Y_U|Pre] \Big)$$

k index people with Lambeth, U index people with Southwark and Vauxhall, $Post$ is after Lambeth moved pipe upstream, Pre before Lambeth moved its pipe (baseline), and $E[y]$ mean cholera mortality.

DiD equation

"Pre" and "Post" refer to when Lambeth, k, was treated which is why it is the same for both k and U groups

If we had more than one treatment group, then "Pre" and "Post" no longer are defined for all units

Potential outcomes and the switching equation

$$\widehat{\delta}_{kU}^{2x2} = \underbrace{\Big(E[Y_k^1|Post] - E[Y_k^0|Pre]\Big) - \Big(E[Y_U^0|Post] - E[Y_U^0|Pre]\Big)}_{\text{Switching equation}}$$

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

Parallel trends bias

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

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 cholera mortality for Lambeth *had it kept its pipe downstream* is the same as the evolution of cholera mortality for Southwark and Vauxhall".

It's in red so you know it's a nontrivial assumption. But why? Can't we just check?

What is “The Science”?

- You’ve probably heard people say RCT is the gold standard for causal inference. But why?
- Because randomization gives *near certainty* that selection bias won’t exist
- Don Rubin commented once, “we know how the science works”
- DiD **does not** use “the science” though

Identification with Independence ("the Science")

Independence assumption

Treatment is independent of potential outcomes

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

RCTs use *independence* to estimate causal effects; DiD does not

Independence

Independence allows us to write down conditional expected potential outcome equations (i.e., impute) like

$$E[Y^0|D=1] = E[Y^0|D=0]$$

In the simple comparison in means, $\widehat{\delta}_{cs} = D + (L - SV)$, independence implies $L = SV$ (no selection bias).

Parallel trends is hard

- **There is no guaranteed "science" in parallel trends**
- IMO, this makes DiD a "hard" design – *because* it doesn't rely on randomization (it relies on parallel trends), there are no "slam dunks" and evidence tends to be multi-layered
- Before we move into regression, let's go through a simple exercise to really pin down these core ideas with simple calculations

https://docs.google.com/spreadsheets/d/1onabpc14JdrGo6NFv0zCWo-nuWDLLV2L1qNogDT9SBw/edit?usp=sharing

No Anticipation

- Additional assumption is “no anticipation” – poorly named as it doesn’t require literally no anticipation
- No anticipation means that the treatment effect happens only at the time that the treatment occurs or after, but not before
  - → **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.
- We need NA because we are comparing to a baseline period and it needs to not be treated

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
- Shifts from defining treatment effects to estimating them, which means being careful about who is the control group, how you define treatments and what questions can and cannot be answered with this method