# Causal Inference II

## Mixtape Session

# Roadmap

- Background
  - Introduction
  - Potential outcomes
- Identification and Estimation
  - Parallel trends
  - Estimation with OLS specification
  - Inference
- Parallel trends violations
  - How parallel trends can get violated
  - Types of evidence
  - Triple difference

# Introduction

- Welcome to a day long workshop on difference-in-differences
- Lecture, discussion, exercises, application
- Scott Cunningham, Baylor University, Department of Economics

# NYU Workshop outline

Part 1: Introduction to DiD basics

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

# NYU Workshop outline

Part 2: Differential timing

- Fixed effects estimator and strict exogeneity
- Bacon decomposition of TWFE specification
- Callaway and Sant'Anna solution
- Sun and Abraham solution to the event study

# NYU Workshop outline

Part 3: Synthetic Control

- Abadie, et al. (2003; 2010) original synthetic control
- Ben-Michael, et al. (2021) augmented synthetic control

# Natural experiments

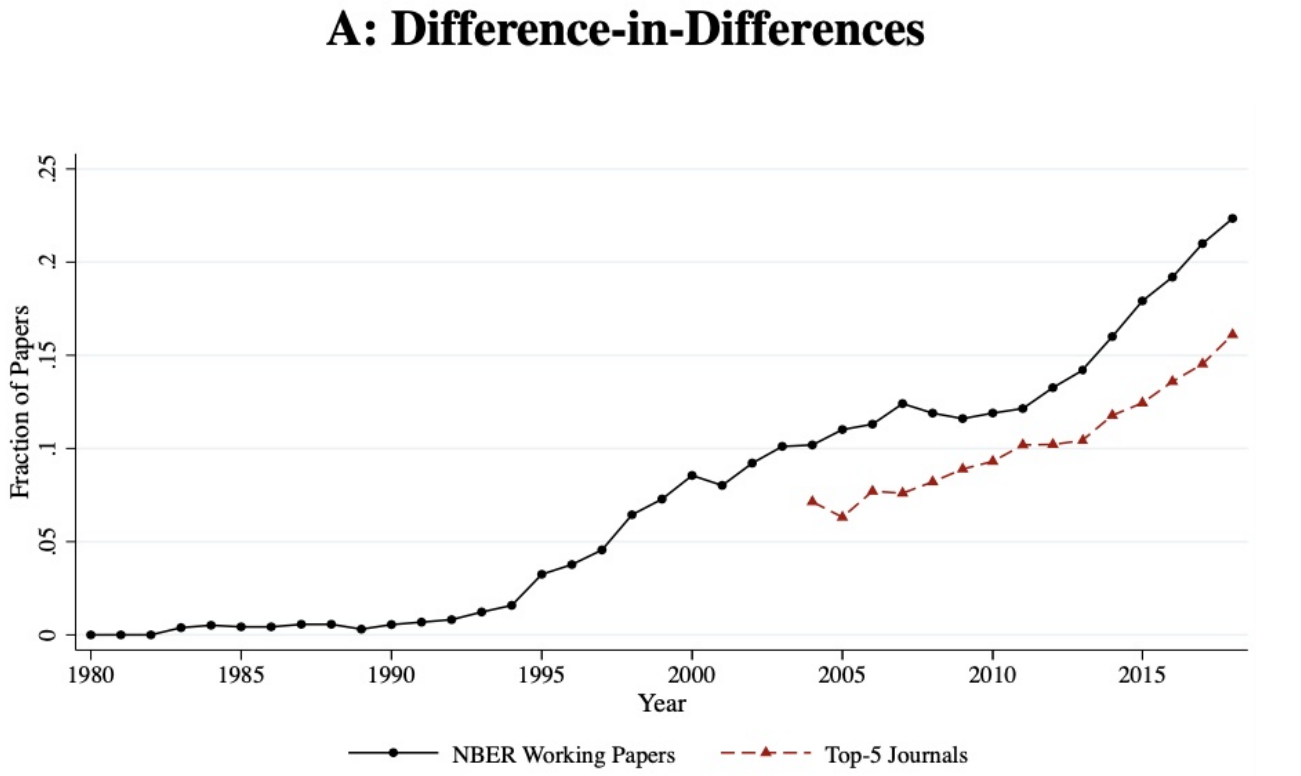
*"A good way to do econometrics is to look for good natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for us." – Daniel McFadden (Nobel Laureate recipient with Heckman 1992)*


I
federalism
(for the natural experiments)

# What is difference-in-differences (DiD)

- DiD is a very old, relatively straightforward, intuitive research design often used with “natural experiment” methodologies
- One of the most widely used quasi-experimental methods in economics and even used in industry
- Basic idea: a group of units receive some treatment and then compared to a group of units that do not

*Figure:* Currie, et al. (2020)



# Origins in Economics and Public Health

- In economics, David Card and Orley Ashenfelter are often associated with its origin – used in the 1970s and 1980s, with mixed success, to study job training programs
- Their dissatisfaction with it led to a call for more randomized controlled trials because, as we will see, it is not able to solve every kind of problem
- But it predates economics by 150 years when two health scientists (separately) used it to prove disease transmission mechanisms

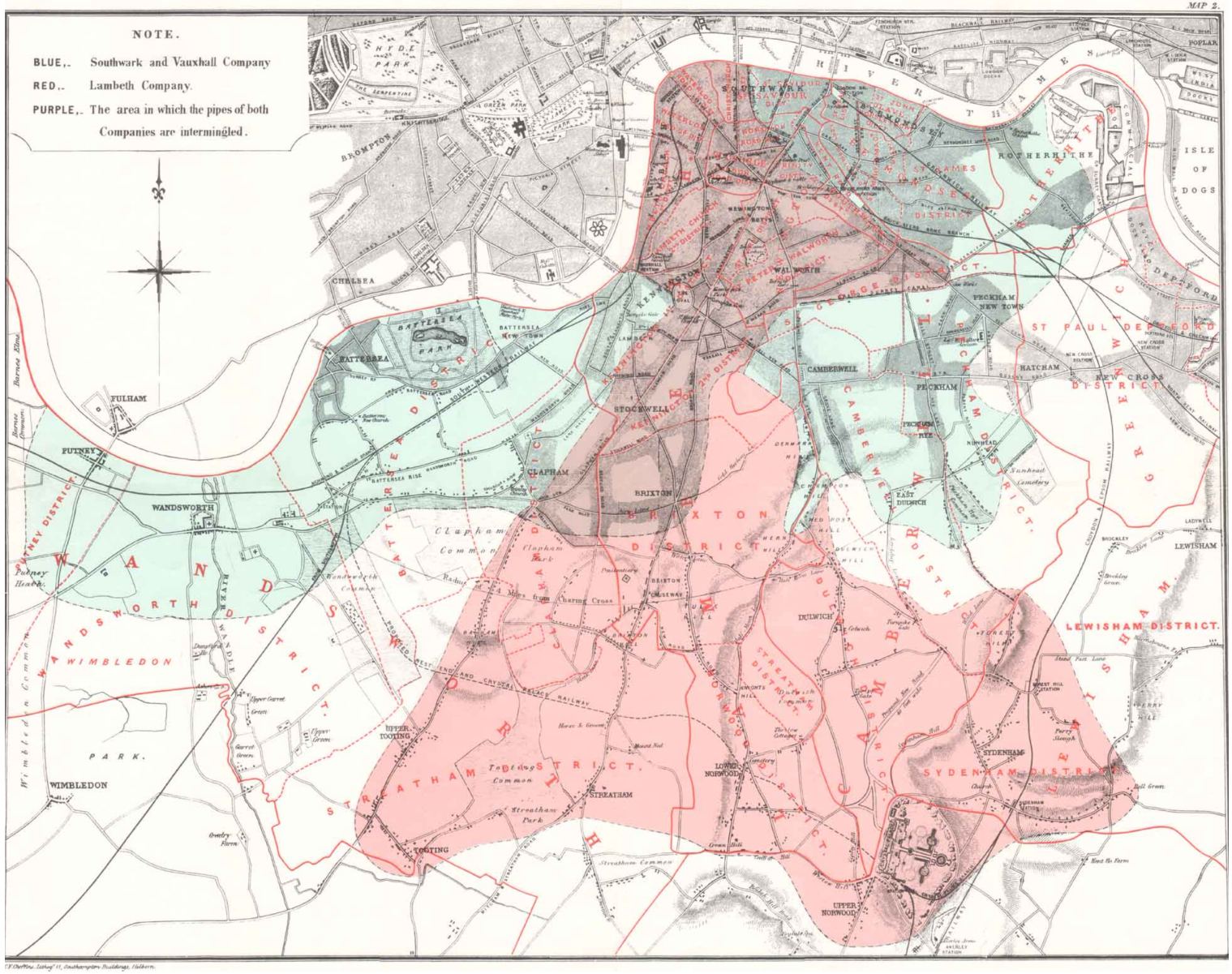
# Case I: Ignaz Semmelweis and washing hands

- 1840s, Vienna maternity wards had high postpartum mortality in wings with doctors and trainee doctors, but not in wings with midwives and trainee midwives
- Training hospitals of students had earlier moved to “anatomical” training involving cadavers for classes
- Semmelweis thinks the mortality is caused by working with cadavers and proposes in 1847 physicians wash their hands with chlorine (but not midwives)
- Comparing the two over time, he shows mortality in physician wing falls and concludes he was right (others disagree)

# Case II: John Snow and cholera

- Three major waves of cholera in the early to mid 1800s in London and people mistakenly thought the cause was “smelly air” (or *miasma*)
- John Snow argued that cholera was spread through host’s evacuations, entering Thames river, returning through water supply
- Strong evidence with maps and a novel DiD design: Lambeth water company moves its pipe between 1849 and 1854 but Southwark and Vauxhall delays

# *Figure:* Two water utility companies in London 1854



# Three ways to study this

1. Simple cross-section: compare mortality in 1854 for the two neighborhoods
2. Before and after: also called interrupted time series. Compare Lambeth in 1854 to 1849
3. Difference-in-differences: a combination of both

# 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 | $Y = SV + T_{SV}$ | $T_{SV}$ | |

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

D is the “treatment effect”. It’s the effect of moving the water on Lambeth mortality and it’s in blue because we can’t see it

# 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 | $Y = SV + T_{SV}$ | $T_{SV}$ | |

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

$T_L$ is a natural change in mortality that would have happened had they not moved the pipe upstream. It creates problems for us

# 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 | $Y = SV + T_{SV}$ | $T_{SV}$ | |

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

But, if $T_L = T_{SV}$, which is called "parallel trends", then we can identify $D$ using DiD. And that's DiD in a nutshell.

# Potential outcomes notation

We need formal notation to understand DiD and that's the potential outcomes model

- 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 realized

- Data are “realized outcomes” not “potential outcomes”
- Realized outcomes is “selected” when treatments are assigned:

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

- Example: My wages if I go to college are $Y^1$ and my wages if I don’t go to college are $Y^0$, but since I went to college ($D = 1$), my wages are $Y = Y^1$.

# Treatment effect definitions

## Individual treatment effect

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

Core building block of causal inference is the individual treatment effect.

# 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:

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

We have $E[Y^1|D=1]$ but we don't have $E[Y^0|D=1]$ so DiD imputes it using something called "parallel trends" (a strong assumption)

# Roadmap

Background
- Introduction
- Potential outcomes

Identification and Estimation
- Parallel trends
- Estimation with OLS specification
- Inference

Parallel trends violations
- How parallel trends can get violated
- Types of evidence
- Triple difference

# Steps of your causal projects

1. Define the parameter we want (“ATT”),
2. Ask what what beliefs do you need (“identification”), and
3. Build cranks that produce the correct numbers (“estimator”)

People often skip 1 and 2 and go straight to 3 and run regressions then go back and assume exogeneity (step 2), and hope that the estimates are weighted averages of individual treatment effects (1), but that is not guaranteed

Assume we are interested in the ATT. What must be true for which method to estimate it correctly?

# DiD is four averages and three differences

Let $k$ and $U$ index the treatment (Lambeth) and untreated group (Southwark and Vauxhall)

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

“Pre” (1849) and “Post” (1854) refer to when Lambeth, $k$, was treated which is why it is the same for both $k$ and $U$ groups

# Potential outcomes and the switching equation

From DiD to ATT

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

$$= \underbrace{\Big(E[Y_k^1|Post] - E[Y_k^0|Pre]\Big) - \Big(E[Y_U^0|Post] - E[Y_U^0|Pre]\Big)}_{\text{Replace with potential outcomes using switching equation}}$$

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

# Parallel trends bias

Rearrange and we get this:

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

The left hand side is our DiD estimator (i..e, four averages, three differences); the right hand side has our parameter (top) and assumption (parallel trends, bottom).

Recall from the earlier table how DiD was equal to $D + (T_L - T_{SV})$. That's this.

# 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?

# Homework

- I’ve included a simple exercise to really pin down these core ideas with simple calculations
- Please at your leisure work through this exercise

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

# OLS Specification

- Simple DiD equation (four averages, three differences) estimates ATT under parallel trends; don't need regression
- But there is an OLS specification that is numerically identical to four averages and three differences
- OLS was historically preferred because
  - → OLS estimates the ATT under parallel trends so it is valid
  - → Easy to calculate the standard errors
  - → Easy to include multiple periods which increases power and makes estimates more precise
- This specification is not appropriate under differential timing or with the inclusion of covariates

# Minimum wages

- Card and Krueger (1994) have a famous study estimating causal effect (ATT) of minimum wages on employment
- Exploited a policy change in New Jersey between February and November in mid-1990s where minimum wage was increased, but neighbor PA did not
- Using DiD, they do not find a negative effect of the minimum wage on employment which is part of its legacy today, but I mainly present it to illustrate the history and the design principles

**Binyamin Appelbaum**
@BCAppelbaum

Replying to @BCAppelbaum

The Nobel laureate James Buchanan wrote in the Wall Street Journal that Card and Krueger were undermining the credibility of economics as a discipline. He called them and their allies "a bevy of camp-following whores."

3:49 PM · Mar 18, 2019

179 Reply Share

Read 18 replies

# Quick comment

- Buchanan's comment gets taken out of historical context to a degree
- Empirical labor and empirical macroeconomics (e.g., Lucas Critique) had been going back to the 1970s in a bit of a "empirical crisis" much like we see sometimes today with debates about p-hacking, but theirs was more basic confusion of causality and correlation
- Consequently, the dominant paradigm in "knowing facts in economics" was theory, not empiricism
- So Buchanan's dismissiveness probably had traces of that; quality of empirical work was sub standard so people tended to not take it very seriously

# Card on that study

> *"I've subsequently stayed away from the minimum wage literature for a number of reasons. First, it cost me a lot of friends. People that I had known for many years, for instance, some of the ones I met at my first job at the University of Chicago, became very angry or disappointed. They thought that in publishing our work we were being traitors to the cause of economics as a whole."*

But let's listen to Orley's opinion about the paper's controversy at the time. https://youtu.be/bbW62axQum8

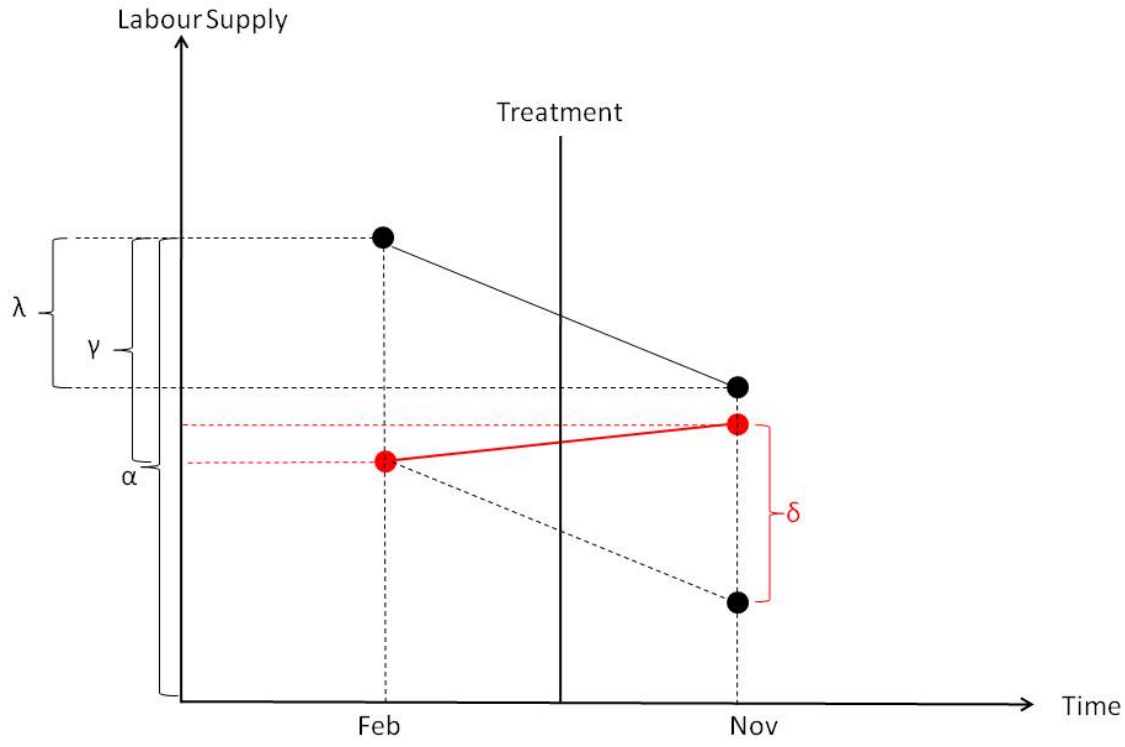
# OLS specification of the DiD equation

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

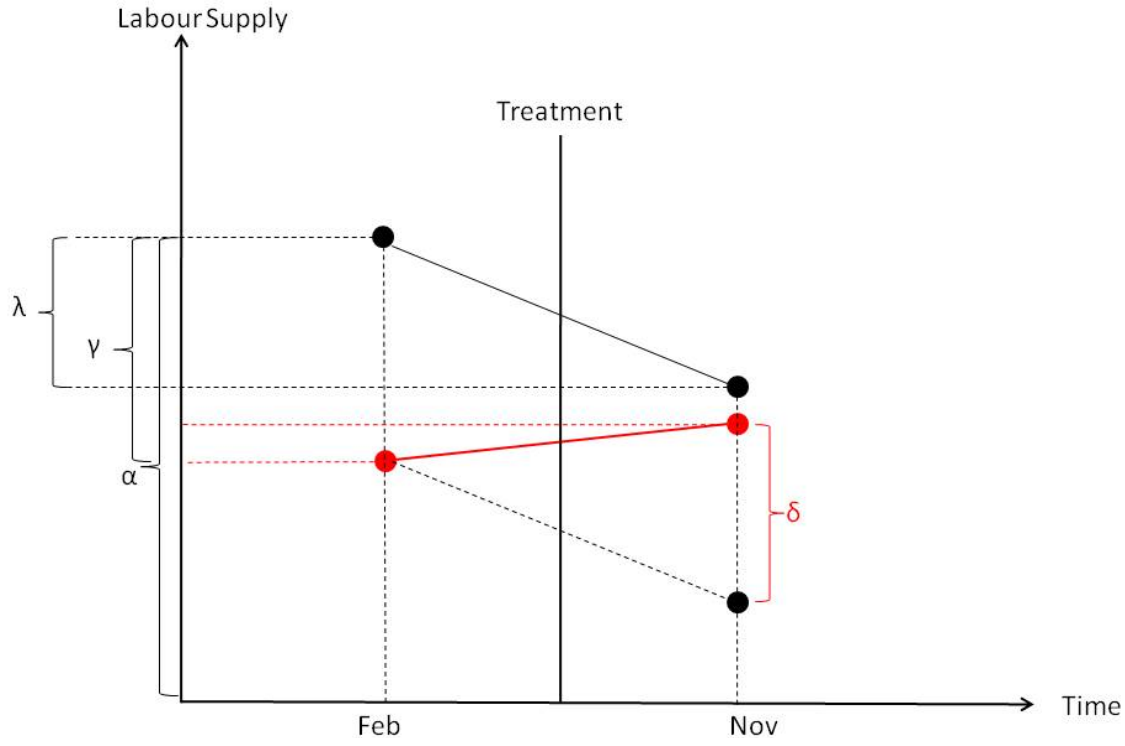
$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

  - → 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: $\alpha$
  - → PA Post: $\alpha + \lambda$
  - → NJ Pre: $\alpha + \gamma$
  - → NJ Post: $\alpha + \gamma + \lambda + \delta$
- DiD equation: (NJ Post - NJ Pre) - (PA Post - PA Pre) $= \delta$

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



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




Notice how OLS is "imputing" $E[Y^0|D=1, Post]$ for the treatment group in the post period? It is only "correct", though, if parallel trends is a good approximation

# Inference

- Bertrand, Duflo and Mullainathan (2004) show that conventional standard errors will often severely understate the standard deviation of the estimators
- Standard errors are biased downward (i.e., too small, over reject)
- They proposed three solutions, but most only use one of them (clustering)

# Inference

1. Block bootstrapping standard errors (if you analyze states the block should be the states and you would sample whole states with replacement for bootstrapping)
2. Clustering standard errors at the group level (in Stata one would simply add `, cluster(state)` to the regression equation if one analyzes state level variation)

Most people will simply cluster, but there are issues if you have too few clusters. They mention a third way but it's only a curiosity.

# Roadmap

- Background
  - Introduction
  - Potential outcomes
- Identification and Estimation
  - Parallel trends
  - Estimation with OLS specification
  - Inference
- Parallel trends violations
  - How parallel trends can get violated
  - Types of evidence
  - Triple difference

# Violating parallel trends exercise

- Parallel trends guides the regression's hand to correctly impute counterfactual $E[Y^0|D=1]$ using $\Delta E[Y^0|D=0]$
- OLS *always* imputes using $\Delta E[Y^0|D=0]$ but is only valid under parallel trends which means control groups matter
- To illustrate this, I've included a document (tab is "DID 2") for you to work on at your leisure

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

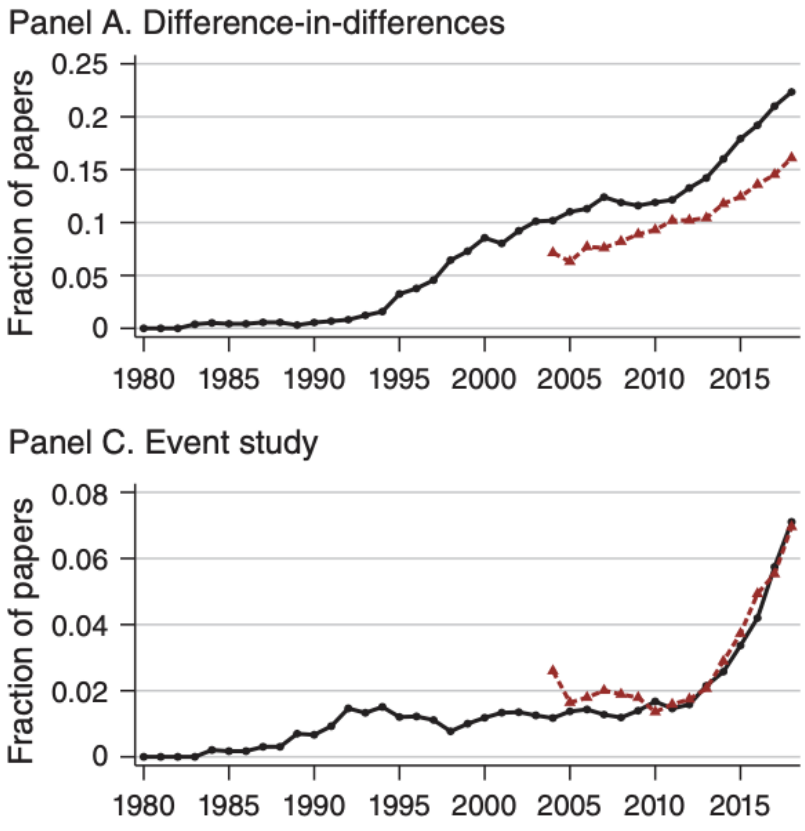
# Think like a prosecutor

- You are a prosecutor building a case before a judge and jury battling the expert defense attorney and their witnesses – what should evidence look like?
- Some example of commonly used forms of evidence can help you
- Common evidence mixes careful and informed logic about main results and mechanisms with falsifications and data visualization, primarily with the event study

# Five types of evidence

1. **Bite**: Show that the treatment impacted first order behavior before showing how it affected second order behavior
2. **Main Results**: Show the primary outcome that your project is about
3. **Mechanisms**: Can you provide evidence of a plausible pathway by which the treatment moves from first order to second order outcomes?
4. **Event studies**: A particular kind of data visualization focused on pre- and post-treatment DiD coefficients in a regression equation
5. **Placebos**: Ruling out reasonable competing theories using the same regression model on different outcomes; can include triple differences

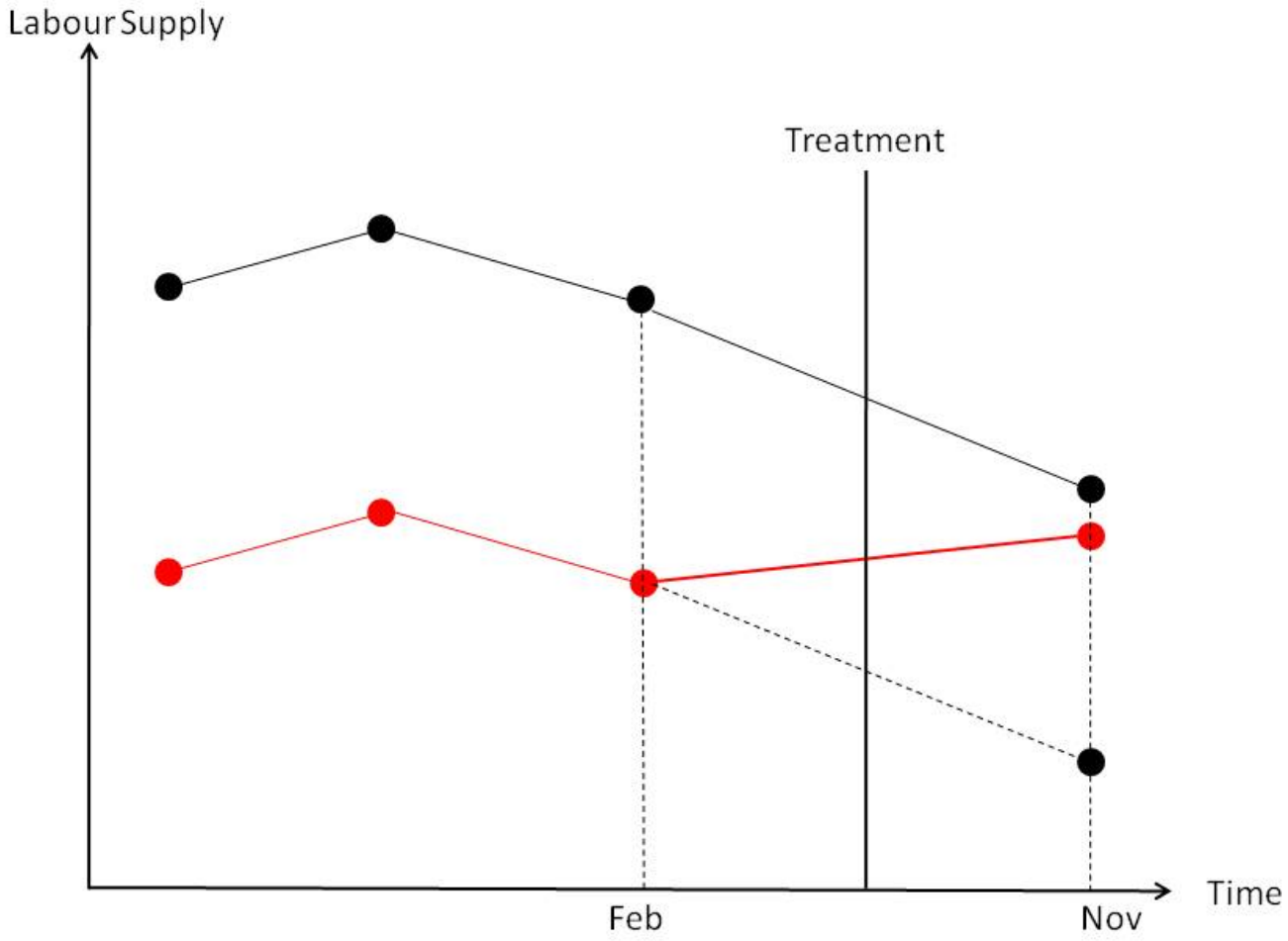
# Event studies are mandatory



# Intuition behind event studies

- We cannot verify parallel trends, but we can verify parallel *pre-trends*
- Pre-trends are a type of falsification – there should not be any effect of the treatment before the treatment occurred
- Also provides some evidence that your comparison group may satisfy parallel trends since it satisfied it earlier
- 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)

# Plot the raw data when there's only two groups



# Event study regression

- Event studies have a simple OLS specification with only one treatment group and one never-treated group

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau D_{s\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{s\tau} + \varepsilon_{ist}$$

  - → where $D$ is an interaction of the treatment group $s$ with the calendar year $\tau$
  - → Treatment occurs in year 0, no anticipation, drop baseline $t-1$
  - → Includes $q$ leads or anticipatory effects and $m$ lags or post treatment effects
- But each OLS estimate of $\mu_\tau$ and $\delta_\tau$ are just "four averages and three differences" as that's the form of the saturated regression

# Event study regression

$$Y_{its} = \alpha + \sum_{\tau=-2}^{-q} \mu_\tau D_{s\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{s\tau} + \varepsilon_{ist}$$

Typically you'll plot the coefficients and 95% CI on all leads and lags (binned or not, trimmed or not)

Under no anticipation, then you expect $\widehat{\mu}$ coefficients to be zero, which gives you confidence that parallel trends holds (but is not a guarantee, and there are still specification issues – see Jon Roth's work)

Under parallel trends, $\widehat{\delta}$ are estimates of the ATT at points in time

# 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

Get access >

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

Cite Permissions Share ▾

**Abstract**

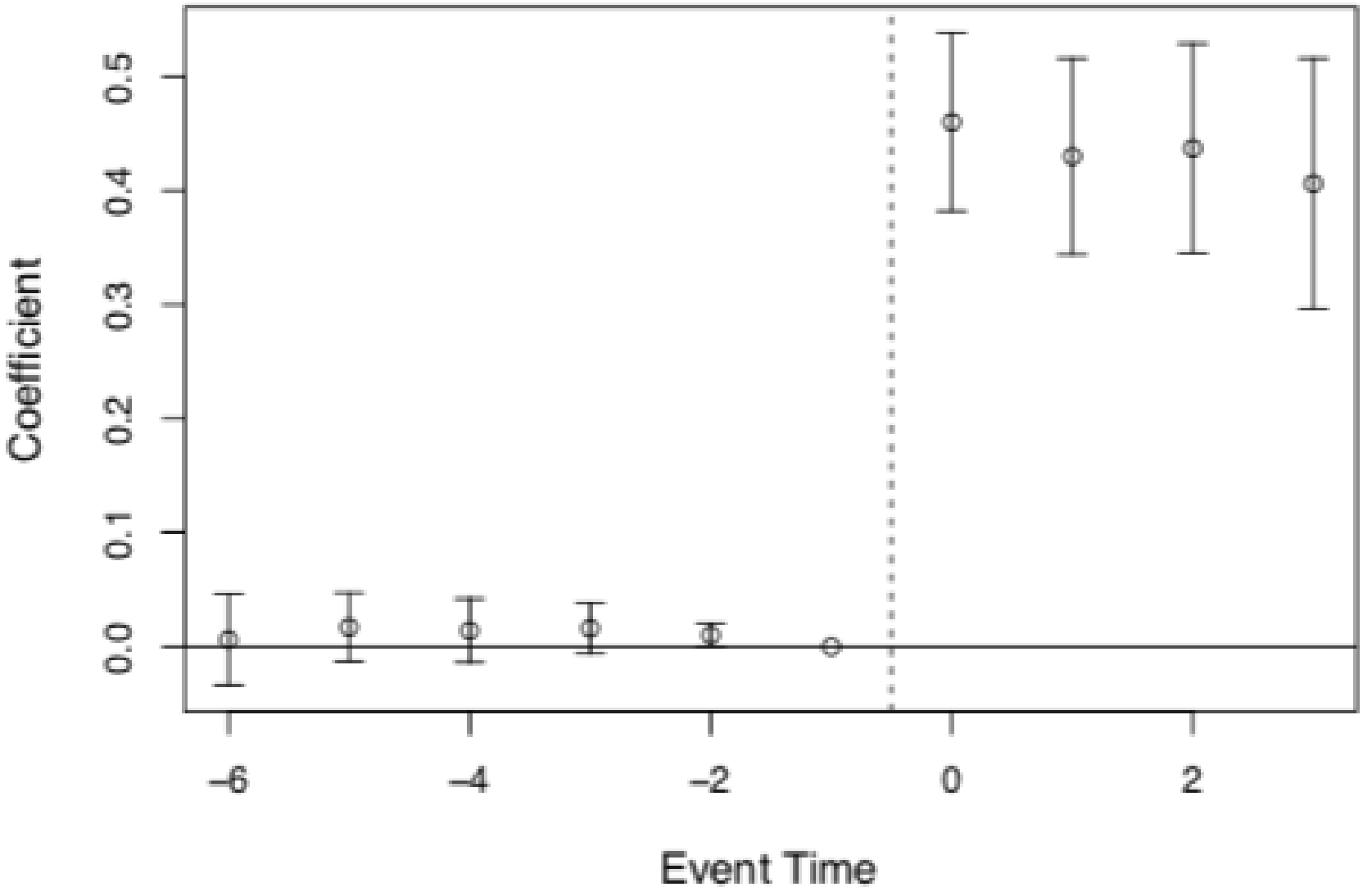
We use large-scale federal survey data linked to administrative death records to investigate the relationship between Medicaid enrollment and mortality. Our analysis compares changes in mortality for near-elderly adults in states with and without Affordable Care Act Medicaid expansions. We identify adults 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

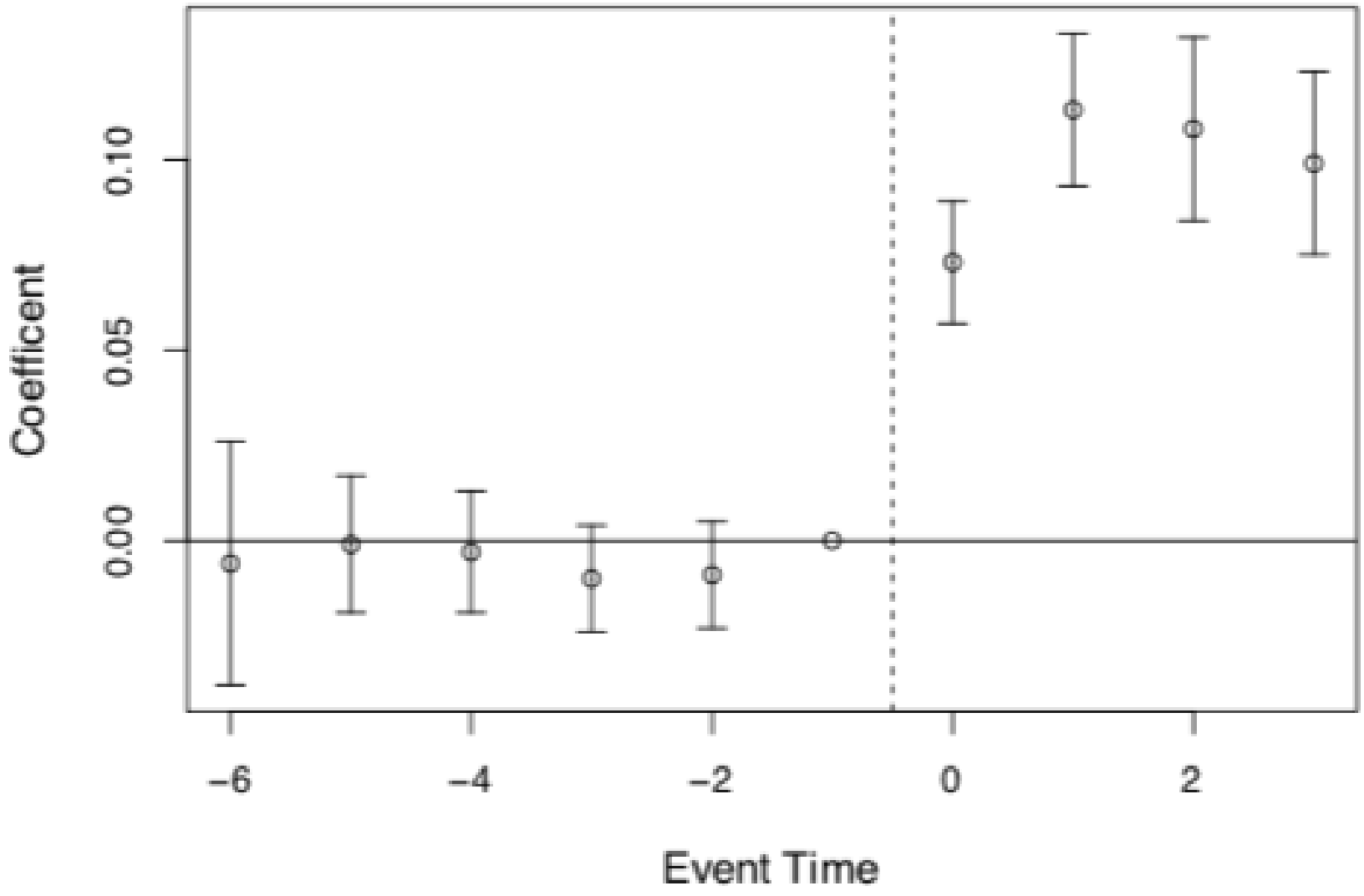
**Issue Section:** Article

# Their five types of evidence

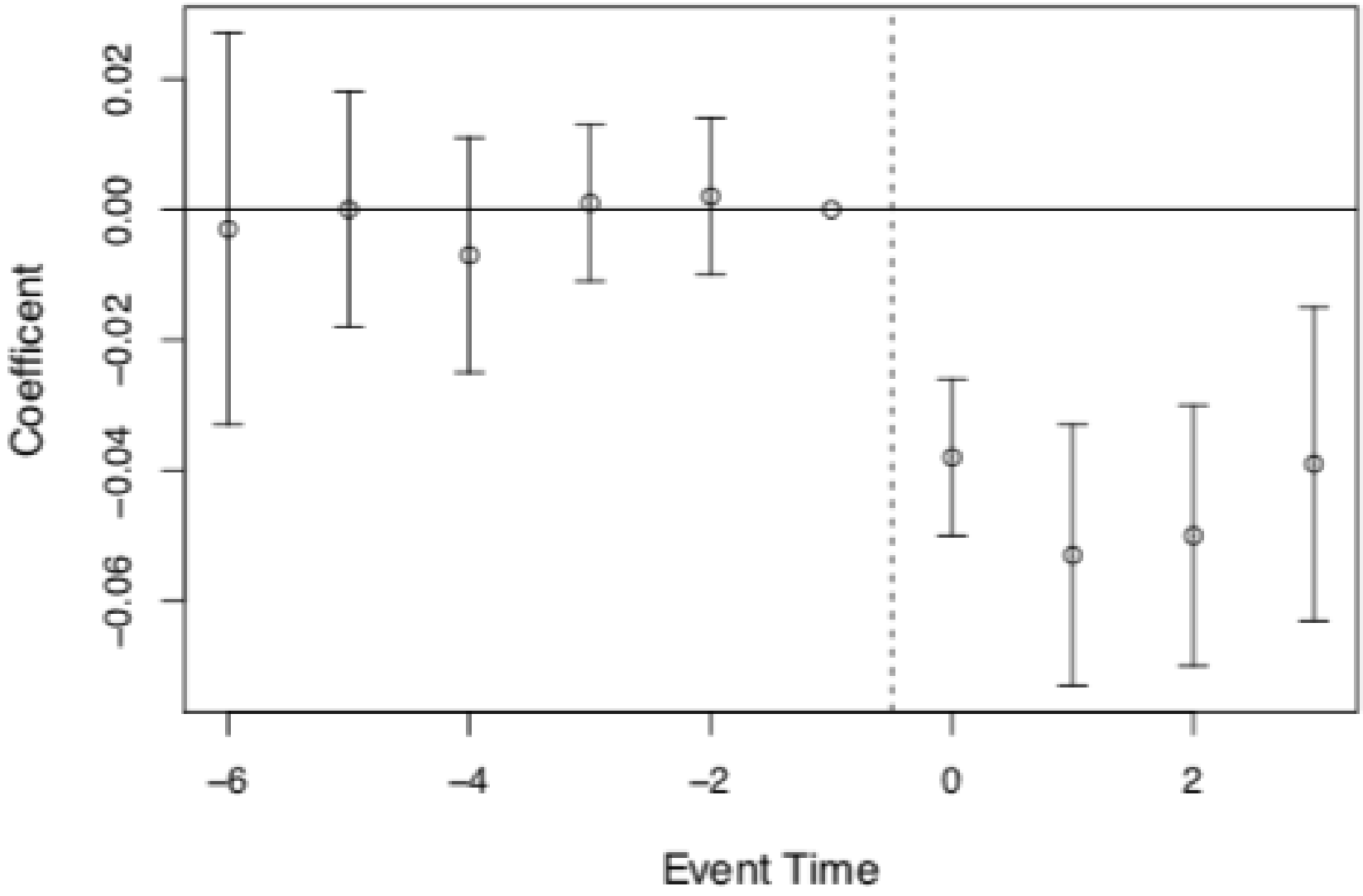
- **Bite** – show that the expansion shifted people into Medicaid and out of uninsured status
- **Main results** – show that Medicaid expansion caused near-elderly mortality to fall 0.132pp or 9.4% reduction from sample mean
- **Mechanism** – They suggest this is coming from reduced disease-related deaths which grows over time
- **Placebos** – Show that there's no effect on mortality for groups it shouldn't be affecting (people 65+)
- **Event study** – Show leads and lags on mortality




(a) Medicaid Eligibility




(b) Medicaid Coverage

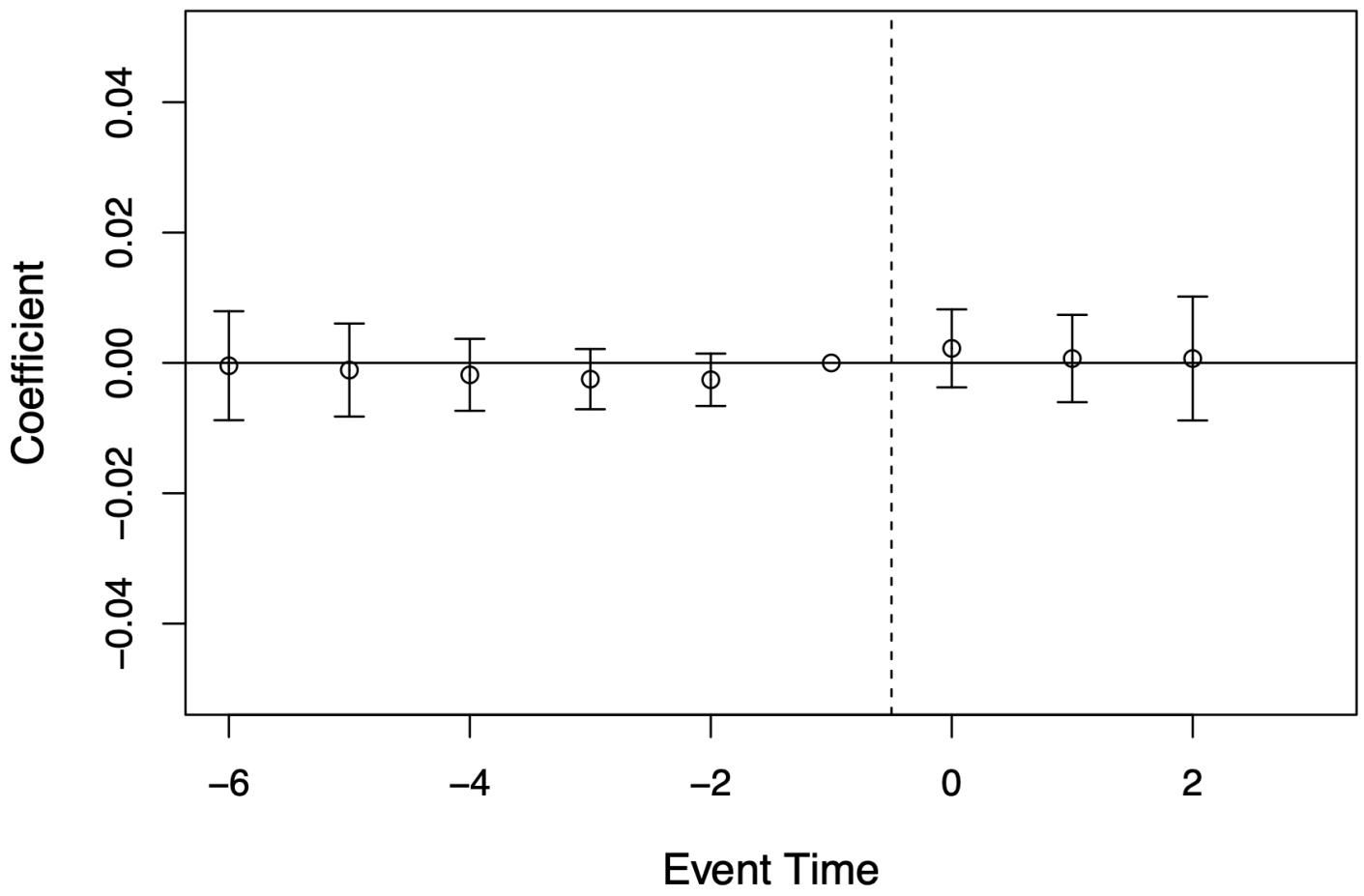



(c) Uninsured

# Quick review

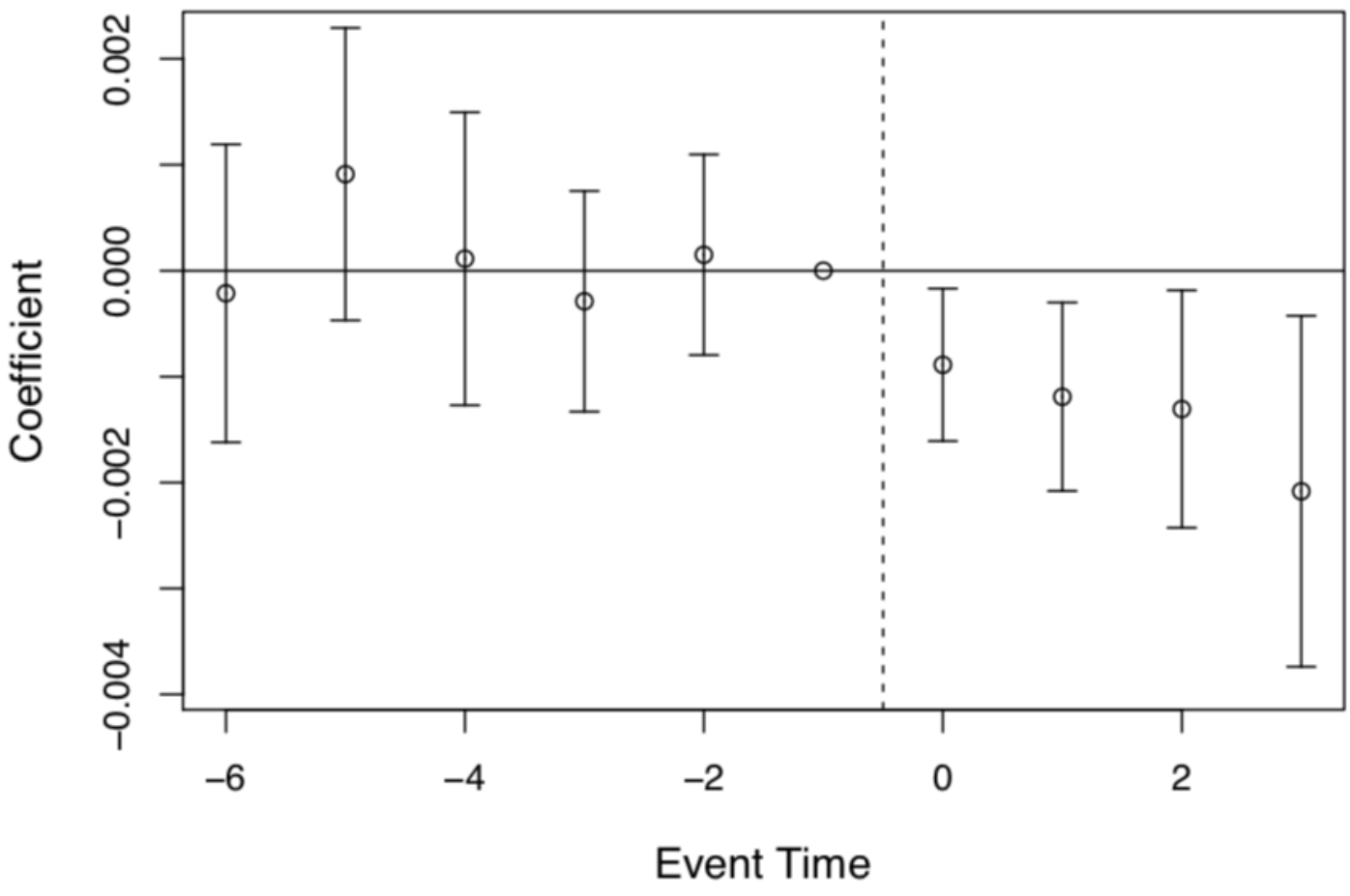
- **Bite**: Did the expansion of Medicaid put more people on Medicaid?
  1. 40pp increase in people eligible (but this was mechanical)
  2. 6-10pp increase in people on Medicaid (but maybe it was crowding out private insurance?)
  3. 4-6pp decrease in uninsured (at least some of the marginal Medicaid enrollees had been uninsured)

# 65 and older mortality placebo




**Discussion**: Why do they do this? Explain to me like I'm 5 the value of a picture like this.

# Main Results: Medicaid expansion and near-elderly mortality



# Lab

- I've provided a lab for you to deepen your understanding of the simple mechanics involved in estimating DiD and event studies
- Please go to this link `https://github.com/Mixtape-Sessions/Causal-Inference-2/tree/main/Lab/Lalonde`
- Q1:a-c vs. Q2a. Skip the covariates

# Triple differences as alternative strategy

- Very common for readers and others to request a variety of "robustness checks" from a DD design
- We saw some of these just now (e.g., falsification test using data for alternative control group, the Medicare population)
- Triple differences uses a within-state untreated group; little trickier, so let's use the table again

# DDD Example by Gruber

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES ON HOURLY WAGES

| Location/year | Before law change | After law change | Time difference for location |
|---|---|---|---|
| A. *Treatment Individuals: Married Women, 20–40 Years Old:* | | | |
| Experimental states | 1.547<br>(0.012)<br>[1,400] | 1.513<br>(0.012)<br>[1,496] | −0.034<br>(0.017) |
| Nonexperimental states | 1.369<br>(0.010)<br>[1,480] | 1.397<br>(0.010)<br>[1,640] | 0.028<br>(0.014) |
| Location difference at a point in time: | 0.178<br>(0.016) | 0.116<br>(0.015) | |
| Difference-in-difference: | −0.062<br>(0.022) | | |
| B. *Control Group: Over 40 and Single Males 20–40:* | | | |
| Experimental states | 1.759<br>(0.007)<br>[5,624] | 1.748<br>(0.007)<br>[5,407] | −0.011<br>(0.010) |
| Nonexperimental states | 1.630<br>(0.007)<br>[4,959] | 1.627<br>(0.007)<br>[4,928] | −0.003<br>(0.010) |
| Location difference at a point in time: | 0.129<br>(0.010) | 0.121<br>(0.010) | |
| Difference-in-difference: | −0.008:<br>(0.014) | | |
| **DDD:** | **−0.054**<br>**(0.026)** | | |

*Table:* Difference-in-Difference-in-Differences numerical example

| States | Group | Period | Outcomes | $D_1$ | $D_2$ | $D_3$ |
|---|---|---|---|---|---|---|
| Experimental states | Married women, 20-40yo | After | $NJ + T + NJ_t + l_t + D$ | $T + NJ_t + l_t + D$ | | |
| | | Before | $NJ$ | | | |
| | Older 40, Single men 20-40yo | After | $NJ + T + NJ_t + s_t$ | | $D + l_t - s_t$ | |
| | | Before | $NJ$ | $T + NJ_t + s_t$ | | |
| Non-experimental states | Married women, 20-40yo | After | $PA + T + PA_t + l_t$ | | | $D$ |
| | | Before | $PA$ | $T + PA_t + l_t$ | | |
| | Older 40, Single men 20-40yo | After | $PA + T + PA_t + s_t$ | | $l_t - s_t$ | |
| | | Before | $PA$ | $T + PA_t + s_t$ | | |

## What is our identifying assumption?

**Answer:** $l_t - s_t$ is the same for both experimental and non-experimental states. This is "change in inequality between two groups hourly wages" from pre to post. It's a new parallel trend assumption.

# DDD in Regression

$$\begin{aligned} Y_{ijt} &= \alpha + \beta_2 \tau_t + \beta_3 \delta_j + \beta_4 D_i + \beta_5 (\delta \times \tau)_{jt} \\ &\quad + \beta_6 (\tau \times D)_{ti} + \beta_7 (\delta \times D)_{ij} + \beta_8 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt} \end{aligned}$$

- Your panel is now a group $j$ state $i$ (e.g., AR high wage worker 1991, AR high wage worker 1992, etc.)
- Assume we drop $\tau_t$ but I just want to show it to you for now.
- If the placebo DD is non-zero, it might be difficult to convince the reviewer that the DDD removed all the bias

# Concluding remarks

- So we hopefully see a few of the key elements of DiD
  - → Remember: the DiD equation and ATT equation are distinct concepts and definitions
  - → DiD designs can be implemented with OLS specifications that calculate differences in means
  - → Parallel pre-trends and parallel trends are not the same thing – the first is testable, the latter is not testable
  - → Event studies are mandatory but pre-trends are smoking guns, but can mislead nonetheless
- Now we want to move into the fixed effects work