Causal Inference II

Mixtape Session

Roadmap

Introduction

- Introducing myself: Scott Cunningham (Baylor)
- Welcome to Mixtape Sessions workshop on advanced difference-in-differences and synthetic control
- 09:00am to 18:00pm, 15 min breaks every hour, 1 hour lunch
- Lecture, discussion, exercises, application

What my pedagogy is like

- Long days that don't feel long because it's high energy, with regular breaks including lunch
- Move between the econometrics, history of thought, videos, applications, code, spreadsheets, exercises
- Ask questions at any point; I'll do my best to answer them

Class goals

Pedagogical goal is to break down the procedures into plain English, rebuilding it into something you can and want to use, but also:

1. **Confidence**: You will feel like you have a good enough understanding of diff-in-diff and synthetic control, both in its basics and some more contemporary issues, so that by the end of the week it a very intuitive, friendly, and useful tool
2. **Comprehension**: You will have learned a lot both conceptually and in the specifics, particularly with regards to issues around identification and estimation in the diff-in-diff and synth context
3. **Competency**: You will have more knowledge of programming syntax in Stata and R so that later you can apply this in your own work

Day 1 outline

Introduction to DiD basics

- Potential outcomes review and the ATT parameter
- DiD equation (“four averages and three differences”), parallel trends and estimation with OLS
- Covariates
- TWFE Pathologies in static and dynamic specifications (“event study”)
- Solutions: CS, SA, dCdH, Imputation

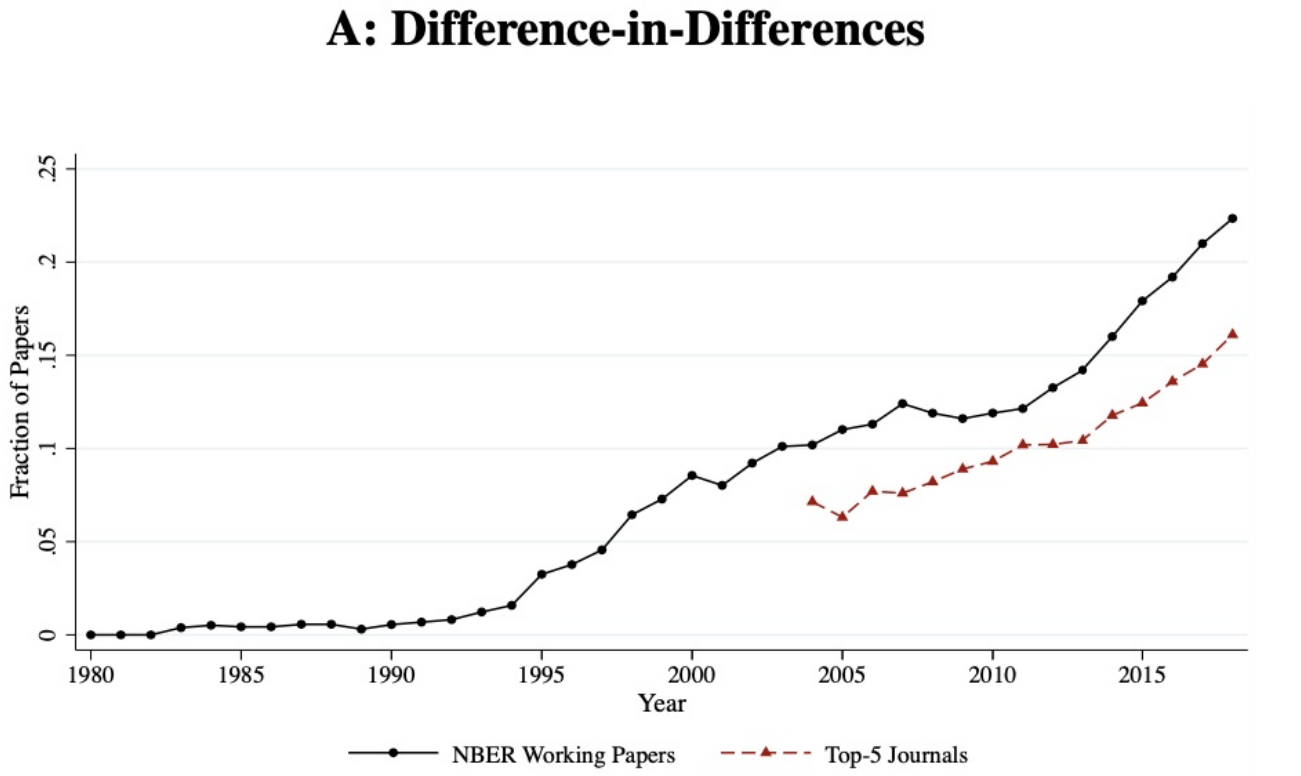
Day 3 outline

- Canonical synth (Abadie papers)
- Augmented synth (Ben-Michael, et al)
- Synthetic Difference-in-Differences

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 increasingly in industry
- Mostly associated with "big shocks" happening in space over time

Figure: Currie, et al. (2020)



Difference-in-differences and empirical crisis in labor economics

- Empirical crisis in empirical labor back in the 1970s (26:31 to 32:00) https://youtu.be/1soLdywFb_Q?t=1579
- Orley Ashenfelter graduated from Princeton in the 1970s, takes a job in Washington DC and begins studying “job trainings programs” where he develops the difference-in-differences design

Explaining diff-in-diff

- Most of us grew up on diff-in-diff being a regression
- And so did Orley – but listen to how the constraints of communicating results led to a new explanation (2:06 to 3:30)

https://youtu.be/WnB3EJ8K7lg?t=126

Equivalence

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

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

- Orley claims that the OLS estimator of δ and the "four averages and three subtractions" are the same thing numerically
- And they are – they are numerically *identical*
- And under a particular assumption, they are also unbiased estimates of an aggregate causal parameter
- But to see this we need new notation – potential outcomes

Potential outcomes notation

- Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 \text{ if in job training program } t \\ 0 \text{ if not in job training program at time } t \end{cases}$$

where i indexes an individual observation, such as a person

Potential outcomes notation

- Potential outcomes:

$$Y_{i,t}^{j} = \begin{cases} 1\text{: wages at time } t \text{ if trained} \\ 0\text{: wages at time } t \text{ if not trained} \end{cases}$$

where j indexes a counterfactual state of the world

Treatment effect definitions

Individual treatment effect

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

Missing data problem: I don't know my own counterfactual

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:

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

This is one of the most important policy parameters, if not the most important, and coincidentally it's also the parameter you get with diff-in-diff (even with heterogeneity)

Potential outcomes vs data

- ATT is expressed in terms of potential outcomes, but we do not use potential outcomes for estimation; we use data
- Potential outcomes are unknown and *hypothetical* possibilities describing states of the world but our data are realized outcomes, or "data", that actually occurred
- Potential outcomes become realized under treatment assignment

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

- Depending on how the treatment is assigned really dictates whether correlations reveal causal effects or bias

Steps of a project

1. Convert research question into causal parameter – for DiD that is the ATT *and only the ATT*
2. Deduce beliefs needed to estimate that causal parameter with data – ?
3. Create a calculator that will use data and estimate the causal parameter – ?

Most of us skipped (1) and maybe even (2) and instead simply “ran regressions” and cross our fingers that that coefficient is causal, but is it? And why is it? And what is it?

DiD equation

Orley's "four averages and three subtractions", or what Bacon will call the 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.

Does $\widehat{\delta}$ equal the ATT? If so when? If not why not?

Potential outcomes and the switching equation

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

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

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".

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"

Steps of a project

1. Convert research question into causal parameter – for DiD that is the ATT
2. Deduce beliefs needed to estimate that causal parameter with data – Parallel trends, No Anticipation, SUTVA
3. Create a calculator that will use data and estimate the causal parameter – Four averages and three subtractions

Don’t use Treated controls

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

What if our control group was treated in both periods. Replace expectations with potential outcomes and rewrite using the “add zero” trick we did. How is this similar to what we did before? Is parallel trends enough?

Don’t use Treated controls

Switching equation (notice the 1’s in the comparison group)

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

Don’t use Treated controls

$$\begin{aligned}\widehat{\delta} &= \Big(E[Y_k^1|Post] - E[Y_k^0|Pre]\Big) - \Big(E[Y_U^1|Post] - E[Y_U^1|Pre]\Big) \\ &\quad + E[Y_k^0|Post] - E[Y(0)_k|Post]\end{aligned}$$

Now let’s rearrange and do our trick and see if this becomes the DID.

Don’t use Treated controls

$$\widehat{\delta} = \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{ATT} + \underbrace{\left(E[Y_k^0|Post] - E[Y_k^0|Pre]\right) - \left(E[Y_U^1|Post] - E[Y_U^1|Pre]\right)}_{\text{Not parallel trends}}$$

Ah. So that’s the problem with using treated units as controls – the DiD equation isn’t collapsing to ATT+PT. So what is it collapsing to?

Don’t use Treated controls

Let’s add these zeroes:

$$E[Y_U^0|Post] - E[Y_U^0|Post] = 0$$
$$E[Y_U^0|Pre] - E[Y_U^0|Pre] = 0$$

$$\begin{aligned}
\widehat{\delta} \quad = \quad & \underbrace{E[Y_k^1|Post] - E[Y_k^0|Post]}_{ATT} \\
& + \underbrace{\Big(E[Y_k^0|Post] - E[Y_k^0|Pre] \Big) - \Big(E[Y_U^0|Post] - E[Y_U^0|Pre] \Big)}_{\text{Non parallel trends bias}} \\
& + \Big(E[Y_U^1|Post] - E[Y_U^0|Post] \Big) - \Big(E[Y_U^1|Pre] - E[Y_U^0|Pre] \Big)
\end{aligned}$$

Don’t use Treated controls

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

Don’t use Treated controls

We can simplify this:

$$\widehat{\delta} = ATT + PT - \Delta ATT$$

Make a distinction between the real DiD and the counterfeit DiD. Real DiD only makes assumptions about $E[Y^0]$, but counterfeit DiD make assumptions about $E[Y^0]$ *and* treatment effects. And yes you find this in Goodman-Bacon (2021) too.

But think about it – do any of us really know how these policies work? These production functions are obscure.

Don’t use Treated controls

Work together:

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

So to summarize:

1. Control group is never treated (this would apply to spillovers)
2. Treatment status at baseline is the same treatment status as that of controls treatment status

What is parallel trends?

- Parallel trends assumes away the selection bias associated with comparisons
- The assumption is thought to be more plausible than simply assuming simple comparisons held equal
  $E[Y^0|D=0] = E[Y^0|D=1]$
- But it is still a strong assumption, and differs from the assumptions have in the RCT which though also untestable, is nearly guaranteed by randomization
- Most of the hard part of the work involves the old fashioned detective work and the work of making good arguments with good exhibits (tables and figures)

Understanding parallel trends through worksheets

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 ($[Y_k^0|Pre]$)

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

Roadmap

- Introduction
  - Managing expectations
  - Introducing difference-in-differences
  - Potential outcomes
  - Identification and Estimation
- Including Covariates
  - Inverse probability weighting
  - Outcome Regression and Double Robust
  - Lalonde lab
- Differential timing
  - Introduction
- Twoway fixed effects
  - Estimator
  - Applications
- TWFE Pathologies
  - Potential outcomes

OLS and covariates

- Four averages and three subtractions is numerically identical to OLS
- So all you need is parallel trends, NA and SUTVA for either one
- OLS is also easy because you can include covariates
- But as it turns out (and we will look later) OLS with covariates has additional assumptions under the hood

Covariates and violations

- There is an assumption called “unconfoundedness”

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

- It means that within the dimensions of X (e.g., Asian males aged 45), D is assigned to units independent of their potential outcomes or any combination of them (e.g., treatment effects)
- It’s the basis for running regressions with covariates in order to recover aggregate causal parameters outside of the experiment but it claims that with the inclusion of the covariates, you have isolated a randomized experiment
- We usually motivate this assumption in diff-in-diff, too, but it is technically not what is going on

Why covariates?

- The inclusion of covariates in diff-in-diff models is not about trying to find random variation in the treatment within values of the dimension of X
- It is based on the claim that the inclusion of covariates is necessary to re-establish parallel trends
- This is itself different than how covariates will be used in synthetic control, too

Correcting the missingness problem

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

We were always missing Y^0 values for the treatment group units, but parallel trends allowed us to impute it using the change in $[Y^0]|D=0]$ as a guide

But if that trend is not a good guide, then we cannot.

Conditional parallel trends

The DiD equation yields:

$$\begin{aligned}\widehat{\delta} &= \Big(E[Y_k|Post] - E[Y_k|Pre]\Big) - \Big(E[Y_U|Post] - E[Y_U|Pre]\Big) \\ &= \text{ATT} + \text{Non-parallel trends bias}\end{aligned}$$

If we believe that conditional on covariates, parallel trends holds, but only within values of X, then there are methods we can use that incorporate covariates into the DiD equation and unbiasedness returns

The inclusion of covariates has particular regression specifications, plus there are alternative methods too, and we will review them

Three covariate DiD papers

Three papers (though sometimes you see others) about covariate adjustment in DiD:

1. Abadie (2005) on semiparametric DiD – reweights the comparison group part of the DID equation using a propensity score based on X
2. Heckman, Ichimura and Todd (1997) on outcome regression uses baseline X and control group only to impute the missing counterfactual Y^0 for treatment group units in a DiD equation
3. Sant'Anna and Zhou (2020) is double robust which means the method does both of these at the same time so that you don't have to choose between them

We will discuss both of them and then compare their performance with the more straightforward fixed effects model

Semiparametric DiD

Abadie (2005) proposed a model that simply reweights the control group in the DiD equation using a particular specification ("semiparametric") of the propensity score on pretreatment covariates

1. Calculate each unit's "after minus before" (DiD equation)
2. Estimate the conditional probability of treatment based on baseline covariates (propensity score estimation)
3. Weight the comparison group's DiD equation with the propensity score

Remember – ATT is only missing Y^0 for treatment, so we only have to apply weights to the comparison group units

Novel elements of time in Abadie's model

- There is only one treatment group so therefore there is only one relevant treatment date, t
- The period prior to treatment is called the baseline, or b, period and it is when treated units were not treated
- X_b are "baseline" covariates meaning the value of X in the pre-treatment period for either the treated or comparison group units
- Propensity scores are estimated off the b period *only*
- Abadie "throws away" covariates after treatment because this is all about re-establishing parallel trends which is a *baseline* concept recall

Assumptions

Three main assumptions

1. Conditional parallel trends

$$E[Y_t^0 - Y_b^0 | D = 1, X_b] - E[Y_t^0 - Y_t^0 | D = 0, X_b]$$

2. Common support

$$Pr(D = 1) > 0; Pr(D = 1|X) < 1$$

3. Propensity score model is properly specified

Propensity scores as dimension reduction

- Propensity scores are ways of dealing with a conditioning set X that has large dimensions
- Dimensions are not the same as covariates – if you have continuous X, then it has infinite dimensions
- Common support means that *within* all combinations of the covariates (e.g., white male 47yo versus whites, males, age) there are units in treatment and control

Common support example

Think of common support like “exact matches” but on the propensity score

I’m a white male 47 years old with a PhD; can I find a white male 47 years old without a PhD

If I can, that’s common support; if I cannot that’s off support

Propensity scores as dimension reduction

- Propensity score theorem (Rosenbaum and Rubin 1983) showed that if you need X to satisfy some assumption, the propensity score will satisfy too
- Propensity scores essentially transform your large dimensional problem into a single scalar called the propensity score, which is the conditional probability of treatment (conditional on X)
- But we need to estimate the propensity score because we don't usually know it (only an experimentalist "knows" the true propensity score)

Common support and the propensity score

- Exact matches mean you have people who are identical on covariate values in both treatment and control
- Common support and the propensity score means you have people nearly identical on their probability of treatment
- I am 47yo white male with a PhD with a propensity score of 0.75, but you are an Asian female 27yo without a PhD and have a propensity score of 0.75
- Same idea, but for this to work, we need to have “matches” like that (just on the propensity score)

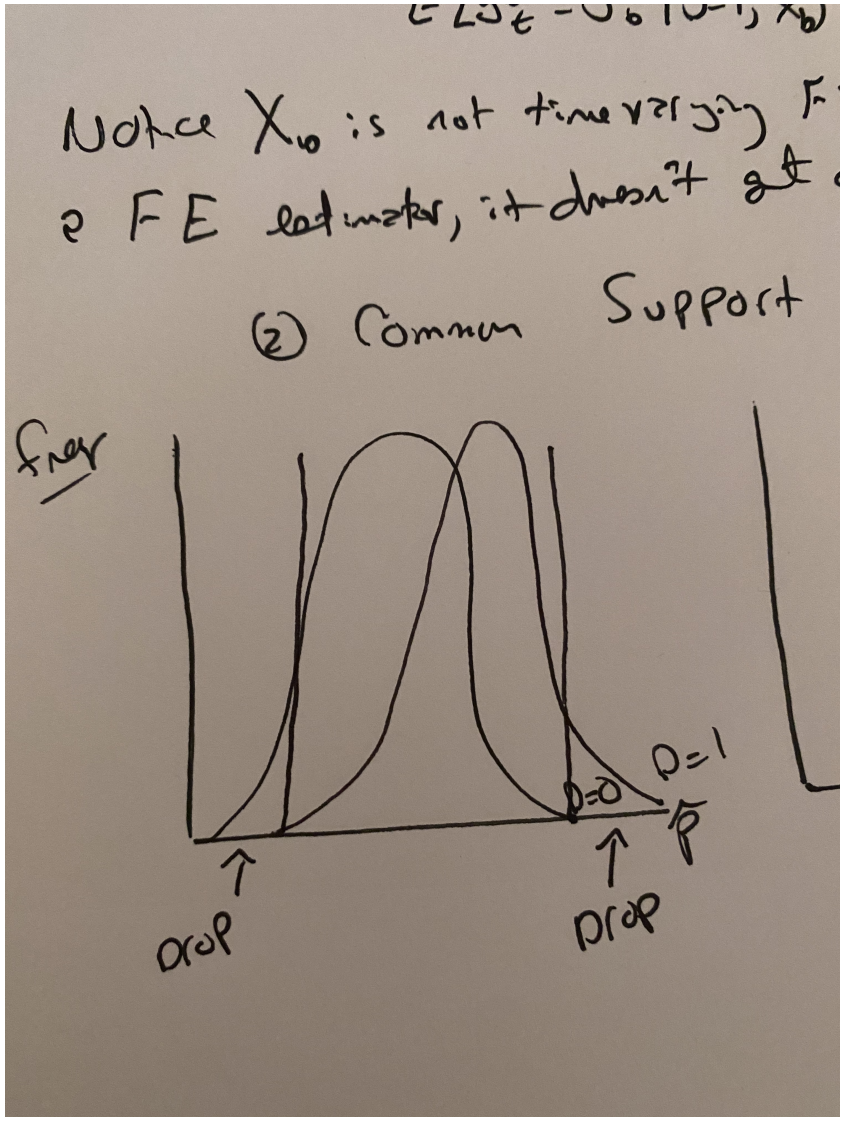
How do these work together?

Since we are identifying the ATT, and the ATT is missing Y^0 for the treated group, we are using the control group Y^0 in its place

Under conditional parallel trends and common support, some of the comparison group units are recovering the parallel trends because of their X values creating projections that in their differences perfectly aligned in expectation with the missing $\Delta E[Y^0|D=1]$

But we have to have all three for it to work

Visualizing propensity score to get common support



Definition and estimation

Defining the ATT parameter of interest

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

Abadie's inverse probability weighting (IPW) estimator

$$E\left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{D_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)}\right]$$

The first is our causal parameter; the second is our reweighted DiD equation that *estimates* our causal parameter, but we need to estimate that propensity score

Abadie's IPW estimator

Look closely; what happens mathematically when you substitute $D = 1$ vs $D = 0$?

$$E\left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{D_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)}\right]$$

The reweighting with the propensity only happens to the comparison group's first differences – not the treatment groups! Why? Because it's the Y^0 that is missing, not the Y^1

Propensity scores

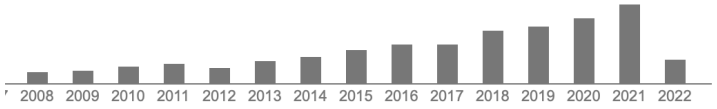
- It's common to hear people say that we don't know the propensity score; we can only estimate it. Same here – we approximate it with regressions
- Paper is titled "Semi-parametric DiD" because Abadie imposes structure on the polynomials used to construct the propensity score ("series logit")

Abadie 2005 influence

Alberto Abadie

Semiparametric difference-in-differences estimators

Authors	Alberto Abadie
Publication date	2005/1/1
Journal	The Review of Economic Studies
Volume	72
Issue	1
Pages	1-19
Publisher	Wiley-Blackwell
Description	The difference-in-differences (DID) estimator is one of the most popular tools for applied research in economics to evaluate the effects of public interventions and other treatments of interest on some relevant outcome variables. However, it is well known that the DID estimator is based on strong identifying assumptions. In particular, the conventional DID estimator requires that, in the absence of the treatment, the average outcomes for the treated and control groups would have followed parallel paths over time. This assumption may be implausible if pre-treatment characteristics that are thought to be associated with the dynamics of the outcome variable are unbalanced between the treated and the untreated. That would be the case, for example, if selection for treatment is influenced by individual-transitory shocks on past outcomes (Ashenfelter's dip). This article considers the case in which differences in observed …
Total citations	Cited by 2330



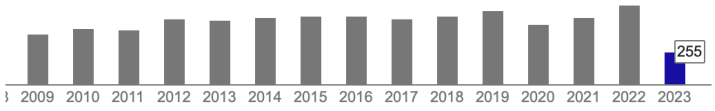
Scholar articles
Semiparametric difference-in-differences estimators
A Abadie - The Review of Economic Studies, 2005
Cited by 2330 Related articles All 12 versions

Abadie (2005) is his fourth most cited paper

Outcome Regression Paper

Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme

Authors	James J Heckman, Hidehiko Ichimura, Petra E Todd
Publication date	1997/10/1
Journal	The review of economic studies
Volume	64
Issue	4
Pages	605-654
Publisher	Wiley-Blackwell
Description	This paper considers whether it is possible to devise a nonexperimental procedure for evaluating a prototypical job training programme. Using rich nonexperimental data, we examine the performance of a two-stage evaluation methodology that (a) estimates the probability that a person participates in a programme and (b) uses the estimated probability in extensions of the classical method of matching. We decompose the conventional measure of programme evaluation bias into several components and find that bias due to selection on unobservables, commonly called selection bias in econometrics, is empirically less important than other components, although it is still a sizeable fraction of the estimated programme impact. Matching methods applied to comparison groups located in the same labour markets as participants and administered the same questionnaire eliminate much of the bias as conventionally …
Total citations	Cited by 9508



Heckman, Ichimura and Todd (1997) is Petra and Hide's most cited paper and Heckman's second most cited!

Doubly Robust Paper

Doubly robust difference-in-differences estimators

Authors	Pedro HC Sant'Anna, Jun Zhao
Publication date	2020/11/1
Journal	Journal of Econometrics
Volume	219
Issue	1
Pages	101-122
Publisher	North-Holland
Description	This article proposes doubly robust estimators for the average treatment effect on the treated (ATT) in difference-in-differences (DID) research designs. In contrast to alternative DID estimators, the proposed estimators are consistent if either (but not necessarily both) a propensity score or outcome regression working models are correctly specified. We also derive the semiparametric efficiency bound for the ATT in DID designs when either panel or repeated cross-section data are available, and show that our proposed estimators attain the semiparametric efficiency bound when the working models are correctly specified. Furthermore, we quantify the potential efficiency gains of having access to panel data instead of repeated cross-section data. Finally, by paying particular attention to the estimation method used to estimate the nuisance parameters, we show that one can sometimes construct doubly robust DID …
Total citations	Cited by 398

Sant'Anna and Zhao (2020) is Pedro's second most cited paper

Doubly Robust Difference-in-differences

- DR models control for covariates twice – once using the propensity score, once using outcomes adjusted by regression – and are unbiased so long as:
  - → The regression specification for the outcome is correctly specified
  - → The propensity score specification is correctly specified
- Sant'Anna and Zhao (2020) incorporated DR into DiD by combining inverse probability weighting and outcome regression into a single DiD model
- It's in the engine of Callaway and Sant'Anna (2020) that we discuss later so it merits close study

Identification assumptions I: Data

Assumption 1: Assume panel data or repeated cross-sectional data

Handling repeated cross-sectional data is possible but assumes stationarity which is a kind of stability assumption, but I'll use panel representation.

Cross-sections will be potentially violated with changing sample compositions (e.g., the Napster example).

Identification assumptions II: Modification to parallel trends

Assumption 2: Conditional parallel trends

Counterfactual trends for the treatment group are the same as the control group for all values of X

$$E[Y_1^0 - Y_0^0 | X, D = 1] = E[Y_1^0 - Y_0^0 | X, D = 0]$$

Identification assumptions III: Common support

Assumption 3: Common support

For some $e > 0$, the probability of being in the treatment group is greater than e and the probability of being in the treatment group conditional on X is $\leq 1 - e$.

Heckman, et al doesn't use the propensity score so we need a more general expression of support

Estimating DD with Assumptions 1-3

- Assumptions 1-3 gives us a couple of options of estimating the DiD
- We can either use the outcome regression (OR) approach of Heckman, et al 1997 (will require correct model too)
- Or we can use the inverse probability weighting (IPW) approach of Abadie (2005) (will require correct model too)

Outcome regression

This is the Heckman, et al. (1997) approach where the potential outcome evolution for the treatment group is imputed with a regression based only on X_b for the control group *only*

$$\widehat{\delta}^{OR} = \overline{Y}_{1,1} - \left[\overline{Y}_{1,0} + \frac{1}{n^T} \sum_{i|D_i=1} (\widehat{\mu}_{0,1}(X_i) - \widehat{\mu}_{0,0}(X_i)) \right]$$

where $\overline{Y}$ is the sample average of Y among units in the treatment group at time t and $\widehat{\mu}(X)$ is an estimator of the true, but unknown, $m_{d,t}(X)$ which is by definition equal to $E[Y_t|D = d, X = x]$.

Outcome regression

$$\widehat{\delta}^{OR} = \overline{Y}_{1,1} - \left[\overline{Y}_{1,0} + \frac{1}{n^T} \sum_{i|D_i=1} (\widehat{\mu}_{0,1}(X_i) - \widehat{\mu}_{0,0}(X_i)) \right]$$

1. Regress changes ΔY on X among untreated groups using baseline covariates only
2. Get fitted values of the regression using all X from $D = 1$ only. Average those
3. Calculate change in this fitted Y among treated with the average fitted values

Inverse probability weighting

This is the Abadie (2005) approach where we use weighting

$$\widehat{\delta}^{ipw} = \frac{1}{E_N[D]} E\left[\frac{D - \widehat{p}(X)}{1 - \widehat{p}(X)}(Y_1 - Y_0)\right]$$

where $\widehat{p}(X)$ is an estimator for the true propensity score. Reduces the dimensionality of X into a single scalar.

These models cannot be ranked

- Outcome regression needs $\widehat{\mu}(X)$ to be correctly specified, whereas
- Inverse probability weighting needs $\widehat{p}(X)$ to be correctly specified
- It's hard to "rank" these two in practice with regards to model misspecification because each is inconsistent when their own models are misspecified

TWFE

Consider our earlier TWFE specification:

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \varepsilon_{it}$$

Just add in covariates then right?

$$Y_{it} = \alpha_1 + \alpha_2 T_t + \alpha_3 D_i + \delta(T_i \times D_t) + \theta \cdot X_{it} + \varepsilon_{it}$$

Sure! If you're willing to impose three *more* assumptions

Decomposing TWFE with covariates

TWFE places restrictions on the DGP. Previous TWFE regression under assumptions 1-3 implies the following:

$$E[Y_1^1|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X$$

Conditional parallel trends implies

$$E[Y_1^0 - Y_0^0|D=1,X] = E[Y_1^0 - Y_0^0|D=0,X]$$
$$E[Y_1^0|D=1,X] - E[Y_0^0|D=1,X] = E[Y_1^0|D=0,X] - E[Y_0^0|D=0,X]$$
$$E[Y_1^0|D=1,X] = E[Y_0^0|D=1,X] + E[Y_1^0|D=0,X] - E[Y_0^0|D=0,X]$$
$$E[Y_1^0|D=1,X] = E[Y_0|D=1,X] + E[Y_1|D=0,X] - E[Y_0|D=0,X]$$

Switching equation substitution

Last line from the switching equation. This gives us:

$$E[Y_1^0|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta X$$

Now compare this with our earlier Y^1 expression

$$E[Y_1^1|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X$$

We can define our target parameter, the ATT, now in terms of the fixed effects representation

Collecting terms

TWFE representation of our conditional expectations of the potential outcomes

$$E[Y_1^1|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X$$
$$E[Y_1^0|D=1,X] = \alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X$$

Substitute these into our target parameter

$$\begin{aligned} ATT &= E[Y_1^1|D=1,X] - E[Y_1^0|D=1,X] \\ &= (\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta_1 X) - (\alpha_1 + \alpha_2 + \alpha_3 + \theta_2 X) \\ &= \delta + (\theta_1 X - \theta_2 X) \end{aligned}$$

What if $\theta_1 X \neq \theta_2 X$?

Assumption 4: Homogeneous treatment effects in X

TWFE requires homogenous treatment effects in X (i.e., the treatment effect is the same for all X)

If X is sex, then effects are the same for males and females.

If X is continuous, like income, then the effect is the same whether someone makes $1 or $1 million.

X-specific trends

TWFE also places restrictions on covariate trends for the two groups too. Take conditional expectations of our TWFE equation.

$$
\begin{aligned}
E[Y_1|D=1] &= \alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11} \\
E[Y_0|D=1] &= \alpha_1 + \alpha_3 + \theta X_{10} \\
E[Y_1|D=0] &= \alpha_1 + \alpha_2 + \theta X_{01} \\
E[Y_0|D=0] &= \alpha_1 + \theta X_{00}
\end{aligned}
$$

X-specific trends

Now take the DiD formula:

$$\delta^{DD} = \Big((\alpha_1 + \alpha_2 + \alpha_3 + \delta + \theta X_{11}) - (\alpha_1 + \alpha_3 + \theta X_{10})\Big) - \Big((\alpha_1 + \alpha_2 + \theta X_{01}) - (\alpha_1 + \theta X_{00})\Big)$$

Eliminating terms, we get:

$$\delta^{DD} = \delta + (\theta X_{11} - \theta X_{10}) - (\theta X_{01} - \theta X_{00})$$

Second line requires that trends in X for treatment group equal trends in X for control group.

Assumption 5 and 6

We need “no X-specific trends” for the treatment group (assumption 5) and comparison group (assumption 6)

Intuition: No X-specific trends means the evolution of potential outcome Y^0 is the same regardless of X. This would mean you cannot allow rich people to be on a different trend than poor people, for instance.

Without these six, in general TWFE will not identify ATT.

Why not both?

- Let's review the problem. What if you claim you need X for conditional parallel trends?
- You have three options:
  1. Outcome regression (Heckman, et al. 1997) – needs Assumptions 1-3
  2. Inverse probability weighting (Abadie 2005) – needs Assumptions 1-3
  3. TWFE (everybody everywhere all the time) – needs Assumptions 1-6
- Problem is 1 and 2 need the models to be correctly specified
- Doubly robust combines them to give us insurance; we now get two chances to be wrong, as opposed to just one

Double Robust DiD

$$\delta^{dr} = E\left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E\left[\frac{p(X)(1-D)}{(1-p(X))}\right]}\right)(\Delta Y - \mu_{0,\Delta}(X))\right]$$

$p(x)$: propensity score model

$\Delta Y = Y_1 - Y_0 = Y_{post} - Y_{pre}$

$\mu_{d,\Delta} = \mu_{d,1}(X) - \mu_{d,0}(X),$ where $\mu(X)$ is a model for

$m_{d,t} = E[Y_t|D = d, X = x]$

So that means $\mu_{0,\Delta}$ is just the control group's change in average Y for each $X = x$

Double Robust DiD

$$\delta^{dr} = E\left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E\left[\frac{p(X)(1-D)}{(1-p(X))}\right]}\right)(\Delta Y - \mu_{0,\Delta}(X))\right]$$

Notice how the model controls for X: you're weighting the adjusted outcomes using the propensity score

The reason you control for X twice is because you don't know which model is right. DR DiD frees you from making a choice without making you pay too much for it

Efficiency

- Authors exploit all the restrictions implied by the assumptions to construct semiparametric bounds
- This is where the influence function comes in, which those who have studied the DID code closely may have noticed
- One of the main results of the paper is that the DR DiD estimator is also DR for inference
- Let's skip to Monte Carlos

Monte Carlo details

- Compare DR with TWFE, OR and IPW
- Sample size is 1,000
- 10,000 Monte Carlo experiments
- Propensity score estimated with logit; OR estimated using linear specification

Table: Monte Carlo Simulations, DGP1, Both OR and Propensity score correct

	Bias	**RMSE**	**SE**	**Coverage**	**CI length**
TWFE	-20.9518	21.1227	2.5271	0.000	9.9061
OR	-0.0012	0.1005	0.1010	0.9500	0.3960
IPW	0.0257	2.7743	2.6636	0.9518	10.4412
DR	-0.0014	0.1059	0.1052	0.9473	0.4124

Figure 1: Monte Carlo for DID estimators, DGP1: Both pscore and OR are correctly specified

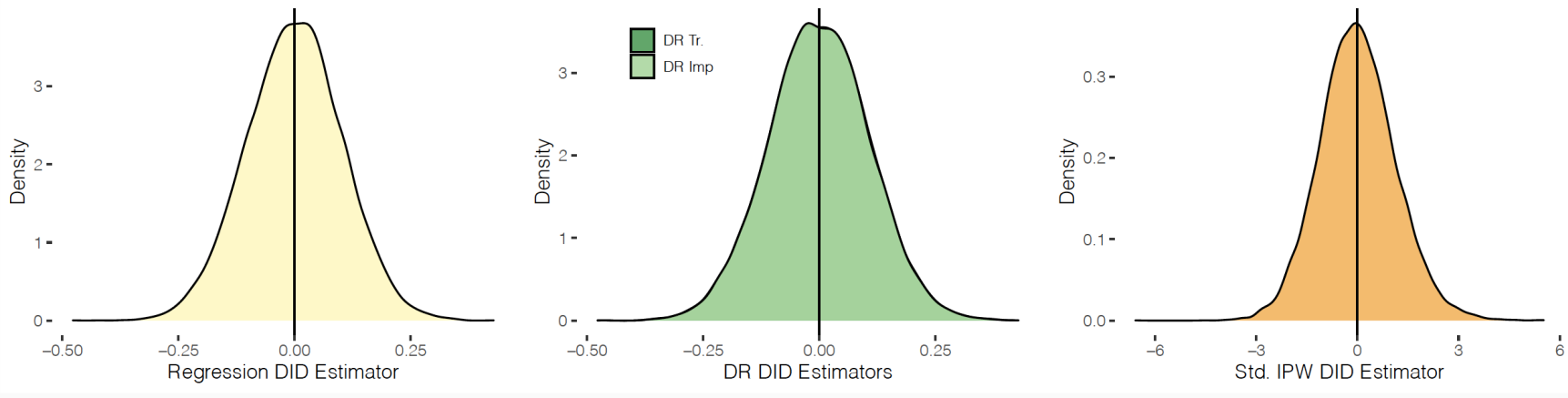
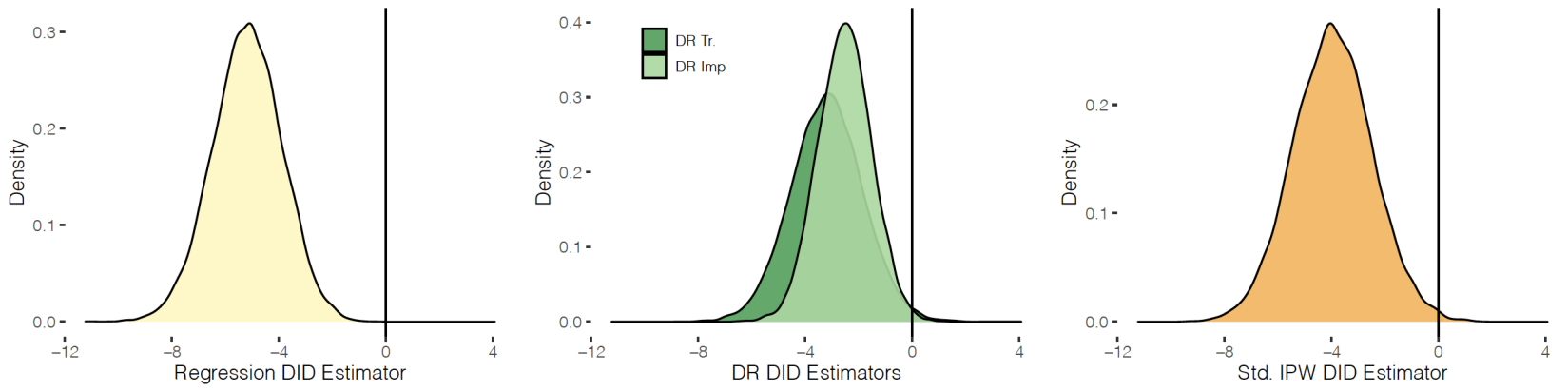


Table: Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

	Bias	**RMSE**	**SE**	**Coverage**	**CI length**
TWFE	-16.3846	16.5383	3.6268	0.000	14.2169
OR	-5.2045	5.3641	1.2890	0.0145	5.0531
IPW	-1.0846	2.6557	2.3746	0.9487	9.3084
DR	-3.1878	3.4544	1.2946	0.3076	5.0749

Figure 4: Monte Carlo for DID estimators, DGP4: Both OR and PS are misspecified



R and Stata Code

There is code in R and Stata (all DiD estimators are now beautifully arranged at a website hosted by Asjad Naqvi)

- Stata: **drdid**
- R: **drdid**

https://asjadnaqvi.github.io/DiD/docs/01_stata/

Remember – it's for 2x2 with covariates (i.e., one treatment group).

Application using real data

- Let's now use a real example with real data and see how well this does
- Famous paper in AER by Lalonde (1986), an Orley and Card student at Princeton
- Found that most program evaluation did badly, but let's revisit it with diff-in-diff

Description of NSW Job Trainings Program

The National Supported Work Demonstration (NSW), operated by Manpower Demonstration Research Corp in the mid-1970s:

- was a temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment
- was also unique in that it **randomly assigned** qualified applicants to training positions:
  - → **Treatment group**: received all the benefits of NSW program
  - → **Control group**: left to fend for themselves
- admitted AFDC females, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes

NSW Program

- Treatment group members were:
  - → guaranteed a job for 9-18 months depending on the target group and site
  - → divided into crews of 3-5 participants who worked together and met frequently with an NSW counselor to discuss grievances and performance
  - → paid for their work
- Control group members were randomized so the same
- Note: the randomization balanced observables and unobservables across the two arms, thus enabling the estimation of an ATE for the people who self-selected into the program

NSW Program

- Other details about the NSW program:
  - → Wages: NSW offered the trainees lower wage rates than they would've received on a regular job, but allowed their earnings to increase for satisfactory performance and attendance
  - → Post-treatment: after their term expired, they were forced to find regular employment
  - → Job types: varied within sites – gas station attendant, working at a printer shop – and males and females were frequently performing different kinds of work

NSW Data

- NSW data collection:
  - → MDRC collected earnings and demographic information from both treatment and control at baseline and every 9 months thereafter
  - → Conducted up to 4 post-baseline interviews
  - → Different sample sizes from study to study can be confusing, but has simple explanations

NSW Data

- Estimation:
  - → NSW was a randomized job trainings program; therefore estimating the average treatment effect is straightforward:

$$\frac{1}{N_t}\sum_{D_i=1} Y_i - \frac{1}{N_c}\sum_{D_i=0} Y_i \approx E[Y^1 - Y^0]$$

    in large samples assuming treatment selection is independent of potential outcomes (randomization) – i.e., $(Y^0, Y^1) \perp\!\!\!\perp D$.

- NSW worked: Treatment group participants' real earnings post-treatment (1978) was positive and economically meaningful – $\approx$ \$900 (LaLonde 1986) to \$1,800 (Dehejia and Wahba 2002) depending on the sample used

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review*.

LaLonde's study was **not** an evaluation of the NSW program, as that had been done, but rather an evaluation of econometric models done by:

- replacing the experimental NSW control group with non-experimental control group drawn from two nationally representative survey datasets: Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID)
- estimating the average effect using non-experimental workers as controls for the NSW trainees
- comparing his non-experimental estimates to the experimental estimates of $900

LaLonde (1986)

- LaLonde's conclusion: available econometric approaches were biased and inconsistent
  - → His estimates were way off and usually the wrong sign
  - → Conclusion was influential in policy circles and led to greater push for more experimental evaluations

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Adjusted[c] (3)	NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1978 Unadjusted (4)	NSW Treatment Earnings Less Comparison Group Earnings: Post-Training Year, 1978 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons: Without Age (6)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons: With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78: Unadjusted (8)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78: Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings (10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

[a] The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is $886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d] See Table 3 for definitions of the comparison groups.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings: Pre-Training Year, 1975 Unadjusted (2)	Pre-Training Year, 1975 Adjusted[c] (3)	Post-Training Year, 1978 Unadjusted (4)	Post-Training Year, 1978 Adjusted[c] (5)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons Without Age (6)	Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons With Age (7)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78 Unadjusted (8)	Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78 Adjusted[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings (10)
Controls	$2,063	$39	$−21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	−$15,997	−$7,624	−$15,578	−$8,067	$425	−$749	−$2,380	−$2,119	−$1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	−$4,503	−$3,669	−$4,020	−$3,482	$484	−$650	−$1,364	−$1,694	−$792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	−$509	$242	−$1,325	$629	−$552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

[a] The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is $886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d] See Table 3 for definitions of the comparison groups.

Imbalanced covariates for experimental and non-experimental samples

	All		CPS Controls $N_c = 15,992$	NSW Trainees $N_t = 297$		
covariate	mean	(s.d.)	mean	mean	t-stat	diff
Black	0.09	0.28	0.07	0.80	47.04	-0.73
Hispanic	0.07	0.26	0.07	0.94	1.47	-0.02
Age	33.07	11.04	33.2	24.63	13.37	8.6
Married	0.70	0.46	0.71	0.17	20.54	0.54
No degree	0.30	0.46	0.30	0.73	16.27	-0.43
Education	12.0	2.86	12.03	10.38	9.85	1.65
1975 Earnings	13.51	9.31	13.65	3.1	19.63	10.6
1975 Unemp	0.11	0.32	0.11	0.37	14.29	-0.26

Lab

https://github.com/Mixtape-Sessions/Causal-Inference-2/tree/main/Lab/Lalonde

Together let's do questions 1 and 2a-c

Concluding remarks

- Including covariates in a DiD design is done for reasons that are different than in regressions more generally – we are trying to address a parallel trends violation
- TWFE can only incorporate *time varying* covariates, and that places restrictions on the model, whereas other methods will not
- Doubly robust and IPW incorporate covariates through propensity scores and outcome regressions (or both) using baseline covariate means only

Roadmap

- Introduction
  - Managing expectations
  - Introducing difference-in-differences
  - Potential outcomes
  - Identification and Estimation
- Including Covariates
  - Inverse probability weighting
  - Outcome Regression and Double Robust
  - Lalonde lab
- Differential timing
  - Introduction
- Twoway fixed effects
  - Estimator
  - Applications
- TWFE Pathologies
  - Potential outcomes

Differential timing outline

We will cover some of the properties of twoway fixed effects (TWFE), some solutions and my personal opinions

1. Introduce TWFE as a panel estimator and its use in DiD
2. TWFE Pathologies in static specification
   - → Goodman-Bacon decomposition as diagnosis of the problem
   - → Aggregating group-time ATT to weaken assumptions
3. TWFE Pathologies in event study specification
   - → Sun and Abraham as both a diagnosis and a cure
   - → Comparing with Callaway and Sant'anna
4. Application, practical advice and code

Beaver dam and diff-in-diff credibility crisis

- Differential timing literature is like a stick that struck a beaver's dam
- Stick made a hole causing a leak
- Gradually that hole got larger and the leak got bigger
- Eventually the dam collapsed
- That's now

Difference-in-differences credibility crisis

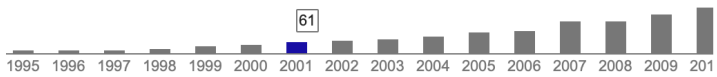
- I'll start with circa 2016 onward – several grad students and assistant professors found critical pathologies with TWFE and developed solutions
- Many simultaneous discoveries, some redundancies, and **sudden** awareness of the issues started happening around 2017, eventually became a massive thing
- Extreme meteoric rise, unusual for econometrics

Compare with LATE paper

- Compare with Imbens and Angrist 1995 LATE in *Econometrica*
- 61 annual cites the year Imbens is denied tenure at Harvard for what would later win him a Nobel Prize

Identification and estimation of local average treatment effects

Authors	Guido W Imbens, Joshua D Angrist
Publication date	1994/3/1
Journal	Econometrica: journal of the Econometric Society
Pages	467-475
Publisher	Econometric Society
Description	RANDOM ASSIGNMENT OF TREATMENT and concurrent data collection on treatment and control groups is the norm in medical evaluation research. In contrast, the use of random assignment to evaluate social programs remains controversial. Following criticism of parametric evaluation models (eg, Lalonde (1986)), econometric research has been geared towards establishing conditions that guarantee nonparametric identification of treatment effects in observational studies, ie identification without relying on functional form restrictions or distributional assumptions. The focus has been on identification of average treatment effects in a population of interest, or on the average effect for the subpopulation that is treated. The conditions required to nonparametrically identify these parameters can be restrictive, however, and the derived identification results fragile. In particular, results in Chamberlain (1986), Manski (1990 …
Total citations	Cited by 5586

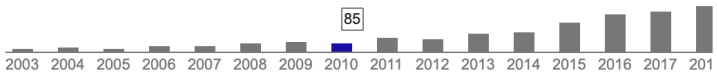


Compare with synth paper

- Athey and Imbens called synth the most important innovation in causal inference of the last two decades
- Most econometrics papers, even influential ones, show slow growth
- Something was different about diff-in-diff even before the econometricians recently shifted their attention to it

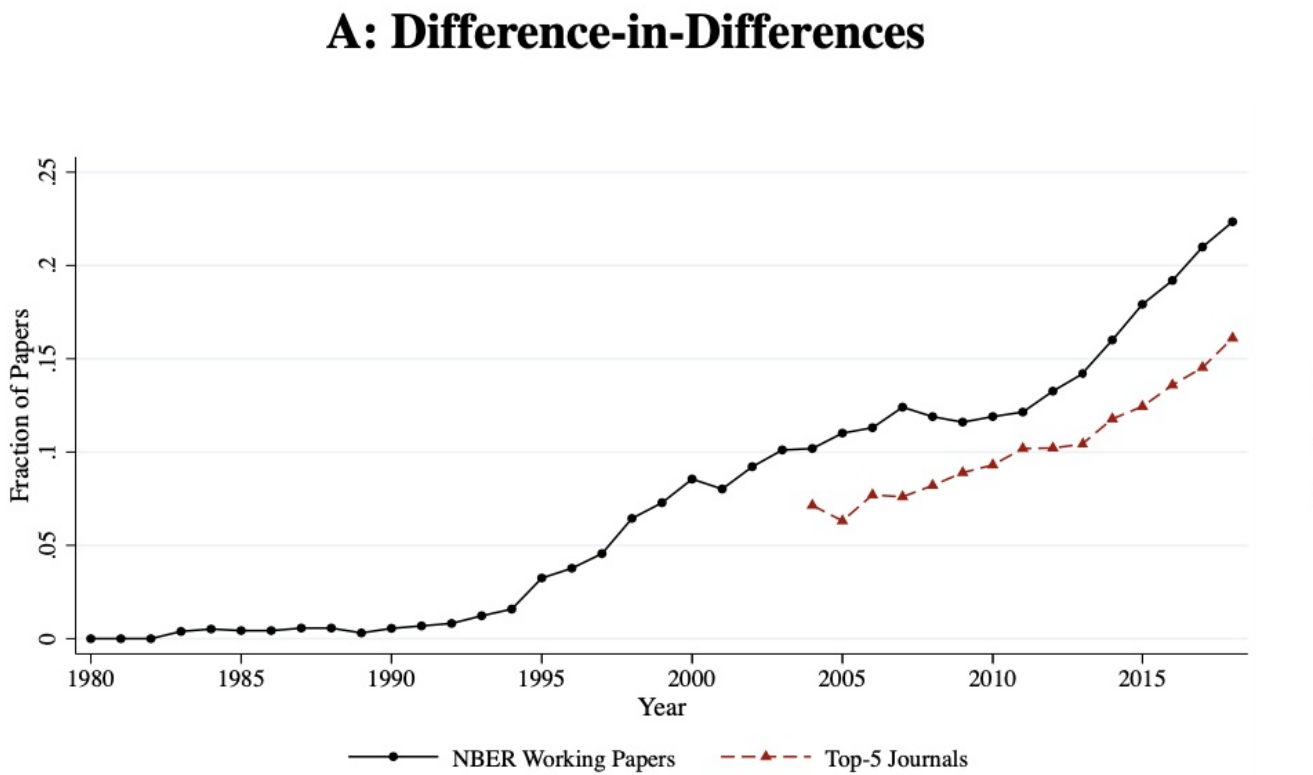
The economic costs of conflict: A case study of the Basque Country

Authors	Alberto Abadie, Javier Gardeazabal
Publication date	2003/3/1
Journal	American economic review
Volume	93
Issue	1
Pages	113-132
Publisher	American Economic Association
Description	This article investigates the economic effects of conflict, using the terrorist conflict in the Basque Country as a case study. We find that, after the outbreak of terrorism in the late 1960's, per capita GDP in the Basque Country declined about 10 percentage points relative to a synthetic control region without terrorism. In addition, we use the 1998–1999 truce as a natural experiment. We find that stocks of firms with a significant part of their business in the Basque Country showed a positive relative performance when truce became credible, and a negative relative performance at the end of the cease-fire.
Total citations	Cited by 5368



Diff-in-diff had belonged to the empiricists

Figure: Currie, et al. (2020)



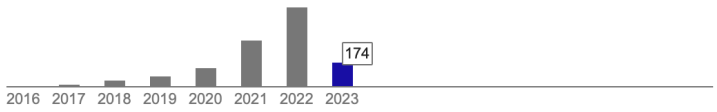
With some exception (e.g., Heckman, Ichimura and Todd 1997; Abadie 2005; Bertrand, Duflo and Mullainthan 2004), econometricians had not given it much notice

Borusyak et al

- Starts it all; written as grad students at Harvard
- Goes through many revisions, posted as working paper
- Returned to a few years ago with a third coauthor, Jahn Spiess, now R&R at Restud

Revisiting event study designs: Robust and efficient estimation

Authors	Kirill Borusyak, Xavier Jaravel, Jann Spiess
Publication date	2021/8/27
Journal	arXiv preprint arXiv:2108.12419
Description	We develop a framework for difference-in-differences designs with staggered treatment adoption and heterogeneous causal effects. We show that conventional regression-based estimators fail to provide unbiased estimates of relevant estimands absent strong restrictions on treatment-effect homogeneity. We then derive the efficient estimator addressing this challenge, which takes an intuitive "imputation" form when treatment-effect heterogeneity is unrestricted. We characterize the asymptotic behavior of the estimator, propose tools for inference, and develop tests for identifying assumptions. Extensions include time-varying controls, triple-differences, and certain non-binary treatments. We show the practical relevance of these insights in a simulation study and an application. Studying the consumption response to tax rebates in the United States, we find that the notional marginal propensity to consume is between 8 and 11 percent in the first quarter—about half as large as benchmark estimates used to calibrate macroeconomic models—and predominantly occurs in the first month after the rebate.
Total citations	Cited by 1399



“dCdH”

- First major hit (in AER), may have been in working paper in 2017 (at least 2018)
- Very thorough decomposition of the TWFE pathology, very general solution, included Stata code
- Very active and talented young team (assistant profs when this was done)

Two-way fixed effects estimators with heterogeneous treatment effects

Authors	Clément De Chaisemartin, Xavier d'Haultfoeuille
Publication date	2020/9/1
Journal	American Economic Review
Volume	110
Issue	9
Pages	2964-2996
Publisher	American Economic Association
Description	Linear regressions with period and group fixed effects are widely used to estimate treatment effects. We show that they estimate weighted sums of the average treatment effects (ATE) in each group and period, with weights that may be negative. Due to the negative weights, the linear regression coefficient may for instance be negative while all the ATEs are positive. We propose another estimator that solves this issue. In the two applications we revisit, it is significantly different from the linear regression estimator. (JEL *C21, C23, D72, J31, J51, L82*)
Total citations	Cited by 2019

Goodman-Bacon

- Arguably the most influential in terms of bringing attention to the problem (but no solution)
- Begun while grad student at Michigan, published last of the crop
- Probably Twitter network had a role as he was very active, also not an econometrician

Difference-in-differences with variation in treatment timing

Authors	Andrew Goodman-Bacon
Publication date	2021/12/1
Journal	Journal of Econometrics
Volume	225
Issue	2
Pages	254-277
Publisher	North-Holland
Description	The canonical difference-in-differences (DD) estimator contains two time periods, "pre" and "post", and two groups, "treatment" and "control". Most DD applications, however, exploit variation across groups of units that receive treatment at different times. This paper shows that the two-way fixed effects estimator equals a weighted average of all possible two-group/two-period DD estimators in the data. A causal interpretation of two-way fixed effects DD estimates requires both a parallel trends assumption and treatment effects that are constant over time. I show how to decompose the difference between two specifications, and provide a new analysis of models that include time-varying controls.
Total citations	Cited by 3307

“CS”

- Second published solution to the problem, written while assistant professors at Vanderbilt and Ole Miss,
- Pedro is a UC3M alum (2015 grad) and Brantly is a Vanderbilt grad
- Both are now coauthors with Andrew Goodman-Bacon
- Introduced new terms like group-time ATT, released very tight R code (“did”)

Difference-in-differences with multiple time periods

Authors	Brantly Callaway, Pedro HC Sant'Anna
Publication date	2021/12/1
Journal	Journal of Econometrics
Volume	225
Issue	2
Pages	200-230
Publisher	North-Holland
Description	In this article, we consider identification, estimation, and inference procedures for treatment effect parameters using Difference-in-Differences (DiD) with (i) multiple time periods, (ii) variation in treatment timing, and (iii) when the “parallel trends assumption” holds potentially only after conditioning on observed covariates. We show that a family of causal effect parameters are identified in staggered DiD setups, even if differences in observed characteristics create non-parallel outcome dynamics between groups. Our identification results allow one to use outcome regression, inverse probability weighting, or doubly-robust estimands. We also propose different aggregation schemes that can be used to highlight treatment effect heterogeneity across different dimensions as well as to summarize the overall effect of participating in the treatment. We establish the asymptotic properties of the proposed estimators and prove the …
Total citations	Cited by 2378

“SA”

- Third published solution to the problem, very similar to CS
- Focus was on decomposing the event study
- Written while grad students at MIT but Sophie Sun is now an assistant professor at CEMFI!

Estimating dynamic treatment effects in event studies with heterogeneous treatment effects

Authors	Liyang Sun, Sarah Abraham
Publication date	2021/12/1
Journal	Journal of Econometrics
Volume	225
Issue	2
Pages	175-199
Publisher	North-Holland
Description	To estimate the dynamic effects of an absorbing treatment, researchers often use two-way fixed effects regressions that include leads and lags of the treatment. We show that in settings with variation in treatment timing across units, the coefficient on a given lead or lag can be contaminated by effects from other periods, and apparent pretrends can arise solely from treatment effects heterogeneity. We propose an alternative estimator that is free of contamination, and illustrate the relative shortcomings of two-way fixed effects regressions with leads and lags through an empirical application.
Total citations	Cited by 1828

There’s even more and more coming

- Gardner, Wooldridge, John Roth, and on and on
- Too many people to name at this point
- Given the large cites, we are likely to keep seeing more on this
- Probably shifting applied practice for the better but there are some growing pains

Roadmap

- Introduction
  - Managing expectations
  - Introducing difference-in-differences
  - Potential outcomes
  - Identification and Estimation
- Including Covariates
  - Inverse probability weighting
  - Outcome Regression and Double Robust
  - Lalonde lab
- Differential timing
  - Introduction
- Twoway fixed effects
  - Estimator
  - Applications
- TWFE Pathologies
  - Potential outcomes

Twoway fixed effects

- When working with panel data, the so-called “twoway fixed effects” (TWFE) estimator was the workhorse estimator
- And from the start, it was used with diff-in-diff
- But at the start, it wasn’t staggered adoption – it was a much simpler design in which a group was treated in one year, and a comparison group wasn’t

Two OLS Models

$$Y_{ist} = \alpha_0 + \alpha_1 Treat_{is} + \alpha_2 Post_t + \delta(Treat_{is} \times Post_t) + \varepsilon_{ist} \quad (1)$$

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist} \quad (2)$$

First equation is used for simple designs when everyone is treated at once; second equation was used when different groups were treated at different times (“differential timing”)

First equation works; second one only sometimes works

Equivalence

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

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

- Orley claims that the TWFE estimator of δ and the "four averages and three subtractions" are the same thing numerically
- And they are – they are numerically *identical*
- And under a particular assumption, they are also unbiased estimates of an aggregate causal parameter
- But what if we aren't in the typical 2x2 case? Researchers used the twoway fixed effects method traditionally

Twoway fixed effects

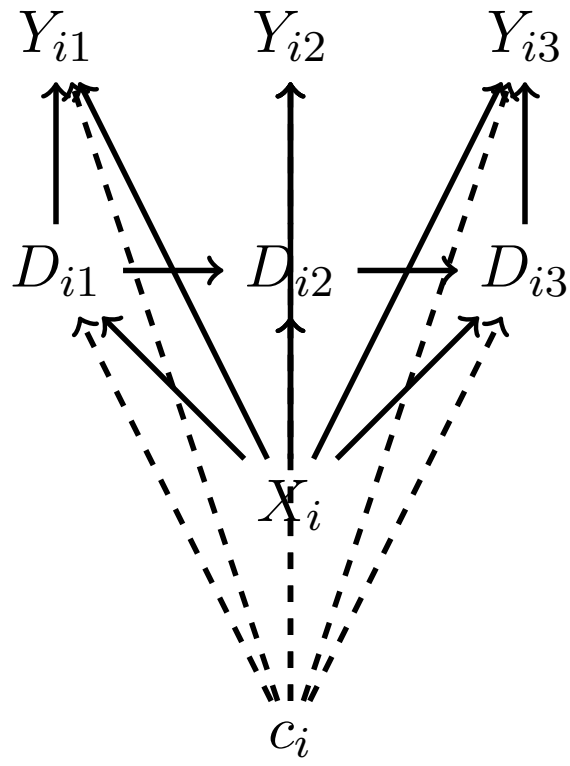
- When working with panel data, the so-called “twoway fixed effects” (TWFE) estimator is the workhorse estimator
- It was at some point adopted for difference-in-differences designs when treatments are adopted at different points in time
- It’s easy to implement, handles time-varying treatments, and has a relatively straightforward interpretation under constant treatment effects
- Turns out it’s interpretation is more complicated with heterogenous treatment effects

Types of repeating data

- Panel datasets follow the same units (individuals, firms, countries, schools, etc.) over several time periods
- Repeated cross-sections will sample a population for a given area, but not be the same people or units in that area
- Difference-in-differences accommodates either, but panel estimators are about the first

Panel estimators

- Panel estimators estimate causal effects in situations where there are unobserved factors associated with the treatment variable creating endogeneity problems
- Less about identification under parallel trends and more about modeling unobservables as unchanging over time (“time invariant”)
- Fixed effects estimation eliminate the unobserved confounder through a demeaning process while retaining the identification of the treatment parameter under constant treatment efects
- Heterogenous treatment effects are part of the problem after that



Directed acyclic graph showing when to use TWFE

When to use TWFE

- Traditionally, this was used for estimating constant treatment effects with unobserved time-invariant heterogeneity – recall the c_i was constant across all time periods
- It's a linear model, so you'll be estimating conditional mean treatment effects – if you want the median, you can't use this
- Once you enter into a world with dynamic treatment effects and differential timing, standard specifications became perverse

When not to use it

- Reverse causality: Becker predicted police reduce crime, but when you regress crime onto police, it's usually positive
  - → $\widehat{\beta}_{FE}$ inconsistent unless strict exogeneity conditional on c_i holds
    - $E[\varepsilon_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}, c_i] = 0; t = 1, 2, \ldots, T$
    - implies ε_{it} uncorrelated with past, current and future regressors
- Time-varying unobserved heterogeneity
  - → It's the time-varying unobservables you have to worry about in fixed effects
  - → Can include time-varying controls, but as always, don't condition on a collider

Notation

- Let y and $x \equiv (x_1, x_2, \ldots, x_k)$ be observable random variables and c be an unobservable random variable
- We are interested in the partial effects of variable x_j in the population regression function

$$E[y|x_1, x_2, \ldots, x_k, c]$$

Notation

- We observe a sample of $i = 1, 2, \ldots, N$ cross-sectional units for $t = 1, 2, \ldots, T$ time periods (a balanced panel)
  - → For each unit i, we denote the observable variables for all time periods as $\{(y_{it}, x_{it}) : t = 1, 2, \ldots, T\}$
  - → $x_{it} \equiv (x_{it1}, x_{it2}, \ldots, x_{itk})$ is a $1 \times K$ vector
- Typically assume that cross-sectional units are i.i.d. draws from the population: $\{y_i, x_i, c_i\}_{i=1}^{N} \sim i.i.d.$ (cross-sectional independence)
  - → $y_i \equiv (y_{i1}, y_{i2}, \ldots, y_{iT})'$ and $x_i \equiv (x_{i1}, x_{i2}, \ldots, x_{iT})$
  - → Consider asymptotic properties with T fixed and $N \to \infty$

Notation

Single unit:

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{pmatrix}_{T \times 1} \qquad X_i = \begin{pmatrix} X_{i,1,1} & X_{i,1,2} & X_{i,1,j} & \dots & X_{i,1,K} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{i,t,1} & X_{i,t,2} & X_{i,t,j} & \dots & X_{i,t,K} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{i,T,1} & X_{i,T,2} & X_{i,T,j} & \dots & X_{i,T,K} \end{pmatrix}_{T \times K}$$

Panel with all units:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{pmatrix}_{NT \times 1} \qquad X = \begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{pmatrix}_{NT \times K}$$

Unobserved heterogeneity

- For a randomly drawn cross-sectional unit i, the model is given by

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it},\ t = 1, 2, \ldots, T$$

  - → y_{it}: log wages i in year t
  - → x_{it} : $1 \times K$ vector of variable events for person i in year t, such as education, marriage, etc. plus an intercept
  - → β : $K \times 1$ vector of marginal effects of events
  - → c_i: sum of all time-invariant inputs known to people i (but unobserved for the researcher), e.g., ability, beauty, grit, etc., often called unobserved heterogeneity or fixed effect
  - → ε_{it}: time-varying unobserved factors, such as a recession, unknown to the farmer at the time the decision on the events x_{it} are made, sometimes called idiosyncratic error

Pooled OLS

- When we ignore the panel structure and regress y_{it} on x_{it} we get

$$y_{it} = x_{it}\beta + v_{it};\ t = 1, 2, \ldots, T$$

  with composite error $v_{it} \equiv c_i + \varepsilon_{it}$
- What happens when we regress y_{it} on x_{it} if x is correlated with c_i?
- Then x ends up correlated with v, the composite error term.
- Somehow we need to eliminate this bias, but how?

Pooled OLS

- Main assumption to obtain consistent estimates for β is:
  - → $E[v_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}] = E[v_{it}|x_{it}] = 0$ for $t = 1, 2, \ldots, T$
    - x_{it} are strictly exogenous: the composite error v_{it} in each time period is uncorrelated with the past, current and future regressors
    - But: education x_{it} likely depends on grit and ability c_i and so we have omitted variable bias and $\widehat{\beta}$ is not consistent
  - → No correlation between x_{it} and v_{it} implies no correlation between unobserved effect c_i and x_{it} for all t
    - Violations are common: whenever we omit a time-constant variable that is correlated with the regressors (heterogeneity bias)
  - → Additional problem: v_{it} are serially correlated for same i since c_i is present in each t and thus pooled OLS standard errors are invalid

Pooled OLS

- Always ask: is there a time-constant unobserved variable (c_i) that is correlated with the regressors?
- If yes, then pooled OLS is problematic
- This is how we motivate a fixed effects model: because we believe unobserved heterogeneity is the main driving force making the treatment variable endogenous

Fixed effects

- Our unobserved effects model is:

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; t = 1, 2, \ldots, T$$

- If we have data on multiple time periods, we can think of c_i as **fixed effects** to be estimated
- OLS estimation with fixed effects yields

$$(\widehat{\beta}, \widehat{c}_1, \ldots, \widehat{c}_N) = \underset{b, m_1, \ldots, m_N}{argmin} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it}b - m_i)^2$$

this amounts to including N individual dummies in regression of y_{it} on x_{it}

Fixed effects

$$(\widehat{\beta}, \widehat{c}_1, \ldots, \widehat{c}_N) = \underset{b, m_1, \ldots, m_N}{argmin} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it} b - m_i)^2$$

The first-order conditions (FOC) for this minimization problem are:

$$\sum_{i=1}^{N} \sum_{t=1}^{T} x'_{it} (y_{it} - x_{it} \widehat{\beta} - \widehat{c}_i) = 0$$

and

$$\sum_{t=1}^{T} (y_{it} - x_{it} \widehat{\beta} - \widehat{c}_i) = 0$$

for $i = 1, \ldots, N$.

Fixed effects

Therefore, for $i = 1, \ldots, N$,

$$\widehat{c}_i = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - x_{it}\widehat{\beta}) = \bar{y}_i - \bar{x}_i\widehat{\beta},$$

where

$$\bar{x}_i \equiv \frac{1}{T}\sum_{t=1}^{T} x_{it}; \bar{y}_i \equiv \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

Plug this result into the first FOC to obtain:

$$\widehat{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)'(x_{it} - \bar{x}_i)\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)'(y_{it} - \bar{y})\right)$$

$$\widehat{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}'_{it}\ddot{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{x}'_{it}\ddot{y}_{it}\right)$$

Fixed effects

Running a regression with the time-demeaned variables $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ and $\ddot{x}_{it} \equiv x_{it} - \bar{x}$ is numerically equivalent to a regression of y_{it} on x_{it} and unit specific dummy variables.

Even better, the regression with the time demeaned variables is consistent for β even when $Cov[x_{it}, c_i] \neq 0$ because time-demeaning eliminates the unobserved effects

$$
\begin{aligned}
y_{it} &= x_{it}\beta + c_i + \varepsilon_{it} \\
\bar{y}_i &= \bar{x}_i\beta + c_i + \bar{\varepsilon}_i \\
\hline
(y_{it} - \bar{y}_i) &= (x_{it} - \bar{x})\beta + (c_i - c_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \\
\ddot{y}_{it} &= \ddot{x}_{it}\beta + \ddot{\varepsilon}_{it}
\end{aligned}
$$

Fixed effects

- Identification assumptions:
  1. $E[\varepsilon_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}, c_i] = 0; t = 1, 2, \ldots, T$
     - regressors are strictly exogenous conditional on the unobserved effect
     - allows x_{it} to be arbitrarily related to c_i
  2. $rank\left(\sum_{t=1}^{T} E[\ddot{x}'_{it}\ddot{x}_{it}]\right) = K$
     - regressors vary over time for at least some i and not collinear
- Fixed effects estimator
  1. Demean and regress $\ddot{y}_{it}$ on $\ddot{x}_{it}$ (need to correct degrees of freedom)
  2. Regress y_{it} on x_{it} and unit dummies (dummy variable regression)
  3. Regress y_{it} on x_{it} with canned fixed effects routine
     - Stata: `xtreg y x, fe i(PanelID)`

Fixed effects

- Properties (under assumptions 1-2):
  - → $\widehat{\beta}_{FE}$ is consistent: $\underset{N\to\infty}{plim}\,\widehat{\beta}_{FE,N} = \beta$
  - → $\widehat{\beta}_{FE}$ is unbiased conditional on **X**

Fixed effects

- Inference:
  - → Standard errors have to be “clustered” by panel unit (e.g., farm) to allow correlation in the ε_{it}'s for the same i.
  - → Yields valid inference as long as number of clusters is reasonably large
- Typically we care about β, but unit fixed effects c_i could be of interest
  - → $\widehat{c_i}$ from dummy variable regression is unbiased but not consistent for c_i (based on fixed T and $N \to \infty$)

Application: Survey for Adult Service Providers

- From 2008-2009, I fielded a survey of Internet sex workers (685 respondents, 5% response rate)
- I asked two types of questions: static provider-specific information (e.g., age, weight) and dynamic session information over last 5 sessions
- Let's look at the panel aspect of this analysis together

Returns to risk

$$
\begin{aligned}
Y_{is} &= \beta X_i + \delta D_{is} + \gamma_{is} Z_{is} + c_i + \varepsilon_{is} \\
\ddot{Y}_{is} &= \delta \ddot{D}_{is} + \gamma_{is} \ddot{Z}_{is} + \ddot{\eta}_{is}
\end{aligned}
$$

where Y is log hourly price (i.e., gross price divided by session length in minutes times 60), D is unprotected sex with a client in session s, X are time invariant observable worker i characteristics, Z are time varying session s characteristics, and c_i is unobserved worker heterogeneity unchanging over time that is correlated with D_{is}.

Table: POLS, FE and Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers

Depvar:	**POLS**	**FE**	**Demeaned OLS**
Unprotected sex with client of any kind	0.013	0.051*	0.051*
	(0.028)	(0.028)	(0.026)
Ln(Length)	-0.308***	-0.435***	-0.435***
	(0.028)	(0.024)	(0.019)
Client was a Regular	-0.047*	-0.037**	-0.037**
	(0.028)	(0.019)	(0.017)
Age of Client	-0.001	0.002	0.002
	(0.009)	(0.007)	(0.006)
Age of Client Squared	0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)
Client Attractiveness (Scale of 1 to 10)	0.020***	0.006	0.006
	(0.007)	(0.006)	(0.005)
Second Provider Involved	0.055	0.113*	0.113*
	(0.067)	(0.060)	(0.048)
Asian Client	-0.014	-0.010	-0.010
	(0.049)	(0.034)	(0.030)
Black Client	0.092	0.027	0.027
	(0.073)	(0.042)	(0.037)
Hispanic Client	0.052	-0.062	-0.062
	(0.080)	(0.052)	(0.045)
Other Ethnicity Client	0.156**	0.142***	0.142***
	(0.068)	(0.049)	(0.045)
Met Client in Hotel	0.133***	0.052*	0.052*
	(0.029)	(0.027)	(0.024)
Gave Client a Massage	-0.134***	-0.001	-0.001
	(0.029)	(0.028)	(0.024)
Age of provider	0.003	0.000	0.000
	(0.012)	(.)	(.)

Table: POLS, FE and Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers

Depvar:	**POLS**	**FE**	**Demeaned OLS**
Body Mass Index	-0.022***	0.000	0.000
	(0.002)	(.)	(.)
Hispanic	-0.226***	0.000	0.000
	(0.082)	(.)	(.)
Black	0.028	0.000	0.000
	(0.064)	(.)	(.)
Other	-0.112	0.000	0.000
	(0.077)	(.)	(.)
Asian	0.086	0.000	0.000
	(0.158)	(.)	(.)
Imputed Years of Schooling	0.020**	0.000	0.000
	(0.010)	(.)	(.)
Cohabitating (living with a partner) but unmarried	-0.054	0.000	0.000
	(0.036)	(.)	(.)
Currently married and living with your spouse	0.005	0.000	0.000
	(0.043)	(.)	(.)
Divorced and not remarried	-0.021	0.000	0.000
	(0.038)	(.)	(.)
Married but not currently living with your spouse	-0.056	0.000	0.000
	(0.059)	(.)	(.)
N	1,028	1,028	1,028
Mean of dependent variable	5.57	5.57	0.00

Heteroskedastic robust standard errors in parenthesis clustered at the provider level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Including linear trends interacting with panel identifier

Table: Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers with provider specific trends

Depvar:	**FE w/provider trends**
Unprotected sex with client of any kind	0.004
	(0.046)
Ln(Length)	-0.450***
	(0.020)
Client was a Regular	-0.071**
	(0.023)
Age of Client	0.008
	(0.005)
Age of Client Squared	-0.000
	(0.000)
Client Attractiveness (Scale of 1 to 10)	0.003
	(0.003)
Second Provider Involved	0.126*
	(0.055)
Asian Client	-0.048***
	(0.007)
Black Client	0.017
	(0.043)
Hispanic Client	-0.015
	(0.022)
Other Ethnicity Client	0.135***
	(0.031)
Met Client in Hotel	0.073***

Concluding remarks

- This was not a review of panel econometrics; for that see Wooldridge and other excellent options
- We reviewed POLS and TWFE because they are commonly used with individual level panel data and difference-in-differences
- Their main value is how they control for unobserved heterogeneity through a simple demeaning while still incorporating time varying covariates
- Now let's discuss difference-in-differences which will at various times use the TWFE model

Difference-in-differences

Keep in mind that yesterday, we had reviewed OLS used for diff-in-diff with two groups and two time periods

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

But what if there are more than two treatment groups treated at separate times? What specification?

Difference-in-differences

- Unclear exactly when it was used, but at some point economists simply began using TWFE with state and year fixed effects and treatment dummy

$$Y_{ist} = \alpha + \delta D_{st} + \sigma_s + \tau_t + \varepsilon_{ist}$$

- The hope was that $\widehat{\delta}$ equaled a "reasonably weighted average" over all underlying treatment effects and therefore was the ATT
- Let's look at an example that is prototypical of a traditional DiD using TWFE with multiple treatment groups (five to be precise)

JHR
The Journal of Human Resources

Does Strengthening Self-Defense Law Deter Crime or Escalate Violence?
Evidence from Expansions to Castle Doctrine

Cheng Cheng

Mark Hoekstra

Abstract

From 2000 to 2010, more than 20 states passed so-called "Castle Doctrine" or "stand your ground" laws. These laws expand the legal justification for the use of lethal force in self-defense, thereby lowering the expected cost of using lethal force and increasing the expected cost of committing violent crime. This paper exploits the within-state variation in self-defense law to examine their effect on homicides and violent crime. Results indicate the laws do not deter burglary, robbery, or aggravated assault. In contrast, they lead to a statistically significant 8 percent net increase in the number of reported murders and nonnegligent manslaughters.

Case study: Castle doctrine reforms

- Cheng and Hoekstra (2013) is a good, clean example of a differential timing for us to practice on
- In 2005, Florida passed a law called Stand Your Ground that expanded self-defense protections beyond the house
- More “castle doctrine” reforms followed from 2006 to 2009

Description

Details of castle doctrine reforms

- “Duty to retreat” is removed versus castle doctrine reforms; expanded where you can use lethal force
- Presumption of reasonable fear is added
- Civil liability for those acting under the law is removed

Ambiguous predictions

Castle reforms $\rightarrow$ homicides: Increase by removing homicide penalties and increasing opportunities

- Castle doctrine expansions lowered the (expected) cost of killing someone in self-defense
- Lowering the price of lethal self-defense should increase lethal homicides

Castle reforms $\rightarrow$ homicides: decrease through deterrence

Cheng and Hoekstra's estimation model

- TWFE model

$$Y_{it} = \beta_1 D_i + \beta_2 T_t + \beta_3 (CDL_{it}) + \alpha_1 X_{it} + c_i + u_t + \varepsilon_{it}$$

- CDL is a fraction between 0 and 1 depending on the percent of the year the state has a castle doctrine law
- Preferred specifications includes "region-by-year fixed effects" (see next slide)
- Estimation with TWFE and Poisson with and without population weights
- Models will include covariates (e.g., police, imprisonment, race shares, state spending on public assistance)

Publicly available crime data

Main data: FBI Uniform Crime Reports Part 1 Offenses (2000-2010)

- Main outcomes: log homicides
- Falsification outcomes: motor vehicle theft and larceny
- Deterrence outcomes: burglary, robbery, assault

Region-by-year fixed effects

- **Parallel trends assumption**: imposed structurally with region-by-year dummies
- **Argument**: unobserved changes in crime are running “parallel” to the treatment states within region over time
- **SUTVA** and **No Anticipation**: No spillovers, no hidden variation in treatment, no behavioral change today in response to tomorrow’s law

Results – Deterrence

	OLS - Weighted by State Population						OLS - Unweighted					
	1	2	3	4	5	6	7	8	9	10	11	12
Panel A: Burglary	Log (Burglary Rate)						Log (Burglary Rate)					
Castle Doctrine Law	0.0780***	0.0290	0.0223	0.0164	0.0327*	0.0237	0.0572**	0.00961	0.00663	0.00277	0.00683	0.0207
	(0.0255)	(0.0236)	(0.0223)	(0.0247)	(0.0165)	(0.0207)	(0.0272)	(0.0291)	(0.0268)	(0.0304)	(0.0222)	(0.0259)
One Year Before Adoption of Castle Doctrine Law				-0.0201						-0.0154		
				(0.0139)						(0.0214)		
Panel B: Robbery	Log (Robbery Rate)						Log (Robbery Rate)					
Castle Doctrine Law	0.0408	0.0344	0.0262	0.0216	0.0376**	0.0515*	0.0448	0.0320	0.00839	0.00552	0.00874	0.0267
	(0.0254)	(0.0224)	(0.0229)	(0.0246)	(0.0181)	(0.0274)	(0.0331)	(0.0421)	(0.0387)	(0.0437)	(0.0339)	(0.0299)
One Year Before Adoption of Castle Doctrine Law				-0.0156						-0.0115		
				(0.0167)						(0.0283)		
Panel C: Aggravated Assault	Log (Aggravated Assault Rate)						Log (Aggravated Assault Rate)					
Castle Doctrine Law	0.0434	0.0397	0.0372	0.0362	0.0424	0.0414	0.0555	0.0698	0.0343	0.0305	0.0341	0.0317
	(0.0387)	(0.0407)	(0.0319)	(0.0349)	(0.0291)	(0.0285)	(0.0604)	(0.0630)	(0.0433)	(0.0478)	(0.0405)	(0.0380)
One Year Before Adoption of Castle Doctrine Law				-0.00343						-0.0150		
				(0.0161)						(0.0251)		
Observations	550	550	550	550	550	550	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes			Yes	Yes	Yes	Yes
Contemporaneous Crime Rates					Yes						Yes	
State-Specific Linear Time Trends						Yes						Yes

Results – Homicides

	1	2	3	4	5	6
Panel C: Homicide (Negative Binomial - Unweighted)						
Castle Doctrine Law	0.0565*	0.0734**	0.0879***	0.0783**	0.0937***	0.108***
	(0.0331)	(0.0305)	(0.0313)	(0.0355)	(0.0302)	(0.0346)
One Year Before Adoption of Castle Doctrine Law				-0.0352		
				(0.0260)		
Observations	550	550	550	550	550	550
Panel D: Log Murder Rate (OLS - Weighted)						
Castle Doctrine Law	0.0906**	0.0955**	0.0916**	0.0884**	0.0981**	0.0813
	(0.0424)	(0.0389)	(0.0382)	(0.0404)	(0.0391)	(0.0520)
One Year Before Adoption of Castle Doctrine Law				-0.0110		
				(0.0230)		
Observations	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes
Contemporaneous Crime Rates					Yes	
State-Specific Linear Time Trends						Yes

Interpretation

- Series of robustness checks (falsifications on larceny and motor vehicle theft; deterrence; many different specifications)
- Castle doctrine reforms are associated with an 8% net increase in homicide rates per year across the 21 adopting states
- Interpretation is these would not have occurred without castle doctrine reforms
- But is this robust to alternative models? Today we will check

Roadmap

- Introduction
  - Managing expectations
  - Introducing difference-in-differences
  - Potential outcomes
  - Identification and Estimation
- Including Covariates
  - Inverse probability weighting
  - Outcome Regression and Double Robust
  - Lalonde lab
- Differential timing
  - Introduction
- Twoway fixed effects
  - Estimator
  - Applications
- TWFE Pathologies
  - Potential outcomes

Twoway fixed effects

- When working with panel data, the so-called “twoway fixed effects” (TWFE) estimator was the workhorse estimator
- And from the start, it was used with diff-in-diff
- But at the start, it wasn’t staggered adoption – it was a much simpler design in which a group was treated in one year, and a comparison group wasn’t

Two OLS Models

$$Y_{ist} = \alpha_0 + \alpha_1 Treat_{is} + \alpha_2 Post_t + \delta(Treat_{is} \times Post_t) + \varepsilon_{ist} \quad (3)$$

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist} \quad (4)$$

First equation is used for simple designs when everyone is treated at once; second equation was used when different groups were treated at different times (“differential timing”)

First equation works; second one only sometimes works

Equivalence

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

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

- Orley claims that the TWFE estimator of δ and the "four averages and three subtractions" are the same thing numerically
- And they are – they are numerically *identical*
- And under a particular assumption, they are also unbiased estimates of an aggregate causal parameter
- But to see this we need new notation – potential outcomes

Discussion of estimate

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

- So that's the simple case; what about the differential timing case?
- If you estimate with OLS with differential timing, what does $\widehat{\delta}$ correspond to?
- It also corresponds to the previous "four averages and three subtractions" – but it's numerous of them, not just one

Decomposition Preview

- Andrew Goodman-Bacon decomposed $\widehat{\delta}$ and showed it is numerically identical to a weighted average of all “four averages and three subtractions”
- But, even before we get to causality there are unusual features
- TWFE model assigns its own weights which are a function of the size of a “group” and the variance of group treatment dummies

K^2 distinct DDs

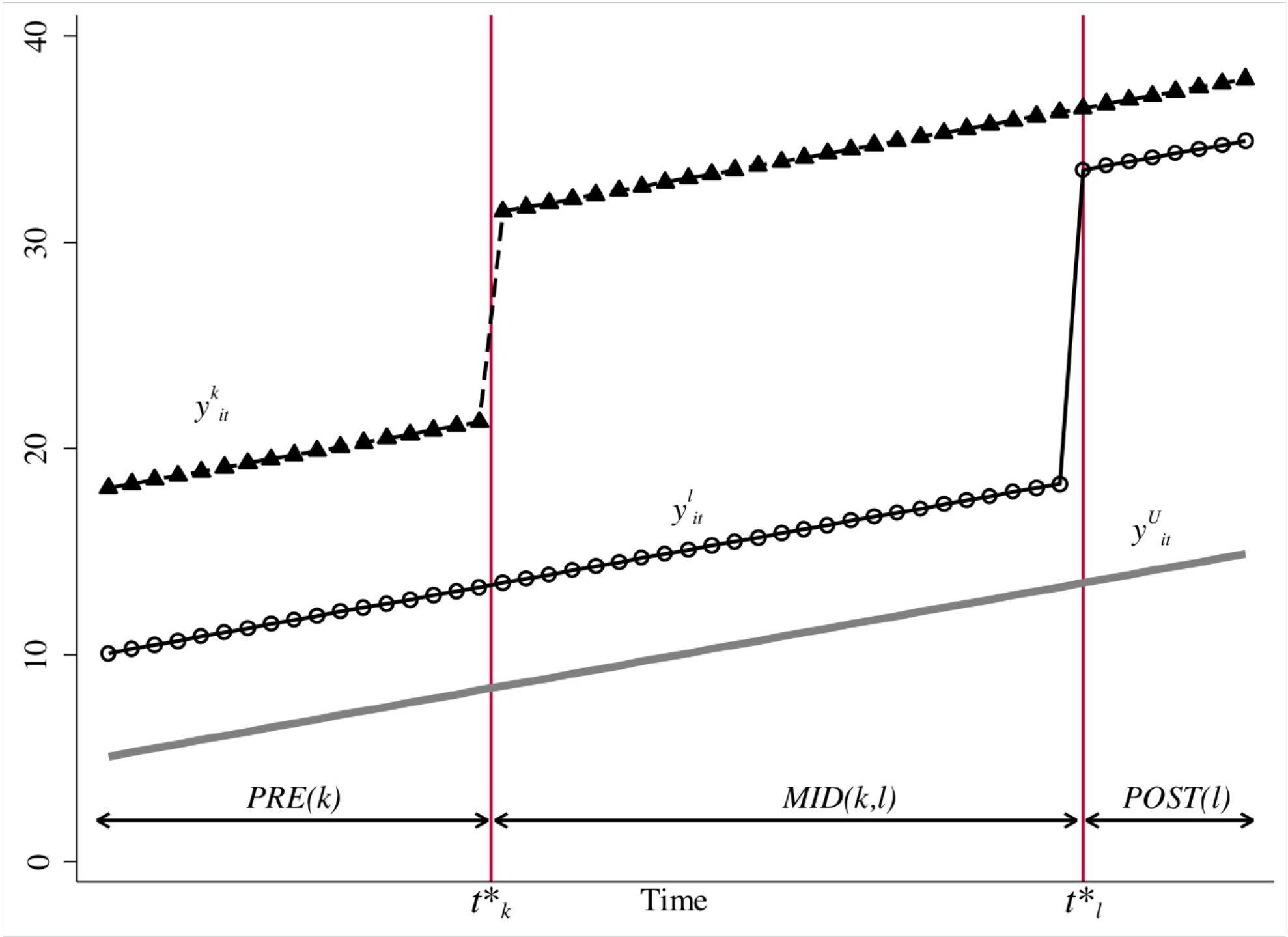
Let's look at 3 timing groups (a, b and c) and one untreated group (U). With 3 timing groups, there are 9 2x2 DDs. Here they are:

a to b	b to a	c to a
a to c	b to c	c to b
a to U	b to U	c to U

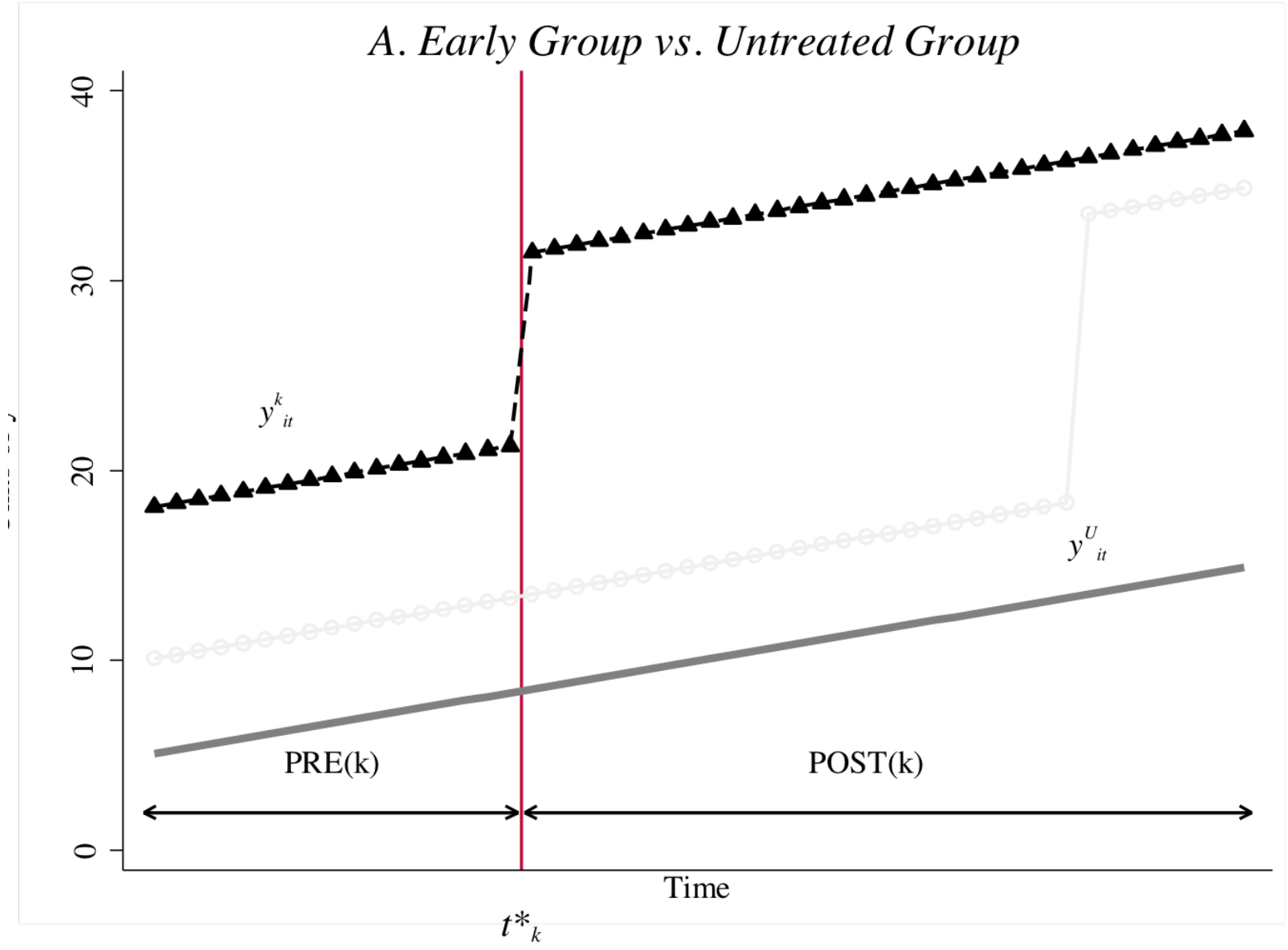
Let's return to a simpler example with only two groups – a k group treated at t_k^* and an l treated at t_l^* plus an never-treated group called the U untreated group

Terms and notation

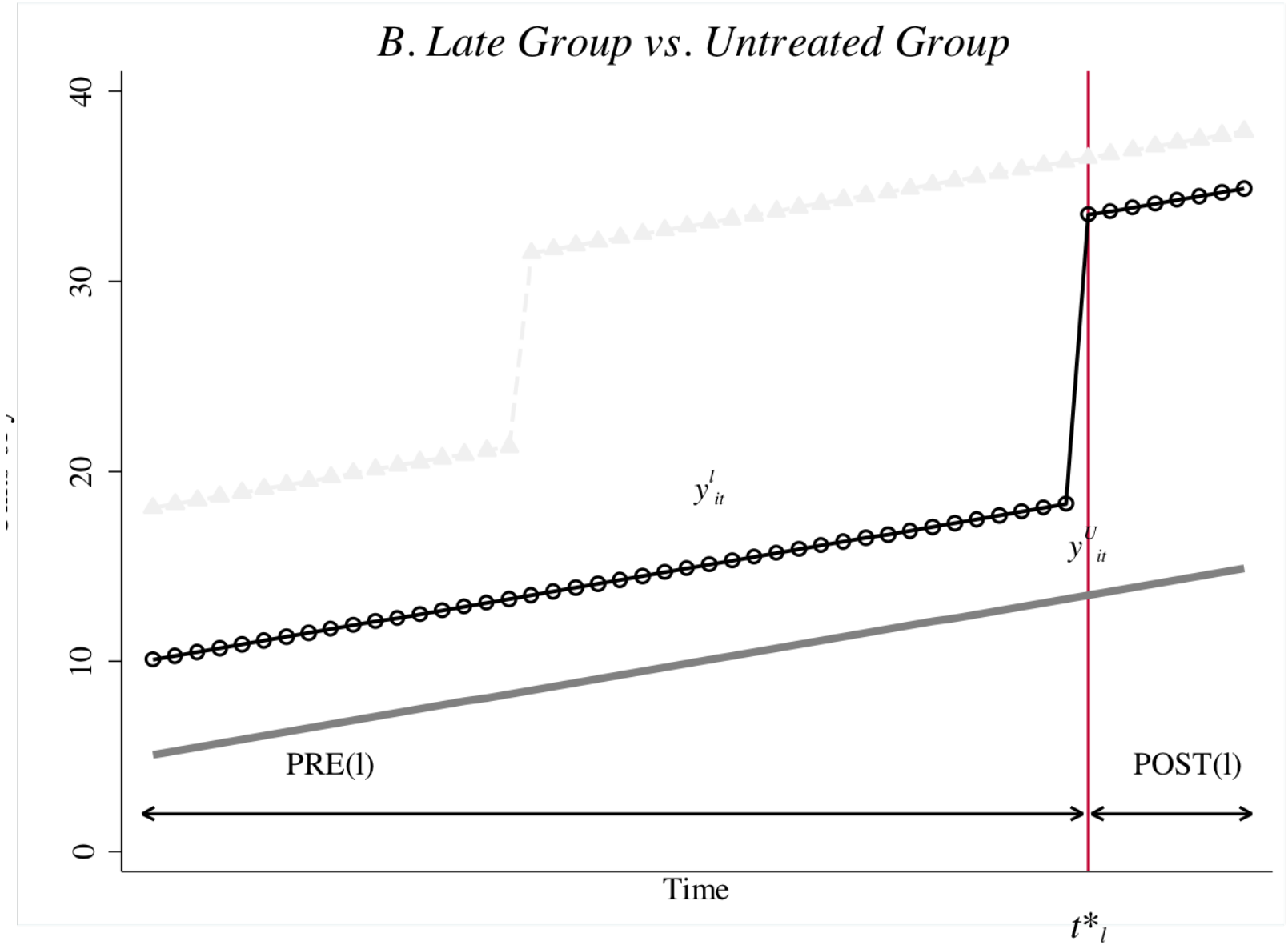
- Let there be two treatment groups (k,l) and one untreated group (U)
- k,l define the groups based on when they receive treatment (differently in time) with k receiving it earlier than l
- Denote $\overline{D}_k$ as the share of time each group spends in treatment status
- Denote $\widehat{\delta}_{jb}^{2x2}$ as the canonical 2×2 DD estimator for groups j and b where j is the treatment group and b is the comparison group


40
30
20
10
0
y^{k}_{it}
y^{l}_{it}
y^{U}_{it}
PRE(k)
MID(k,l)
POST(l)
t^{*}_{k}
Time
t^{*}_{l}

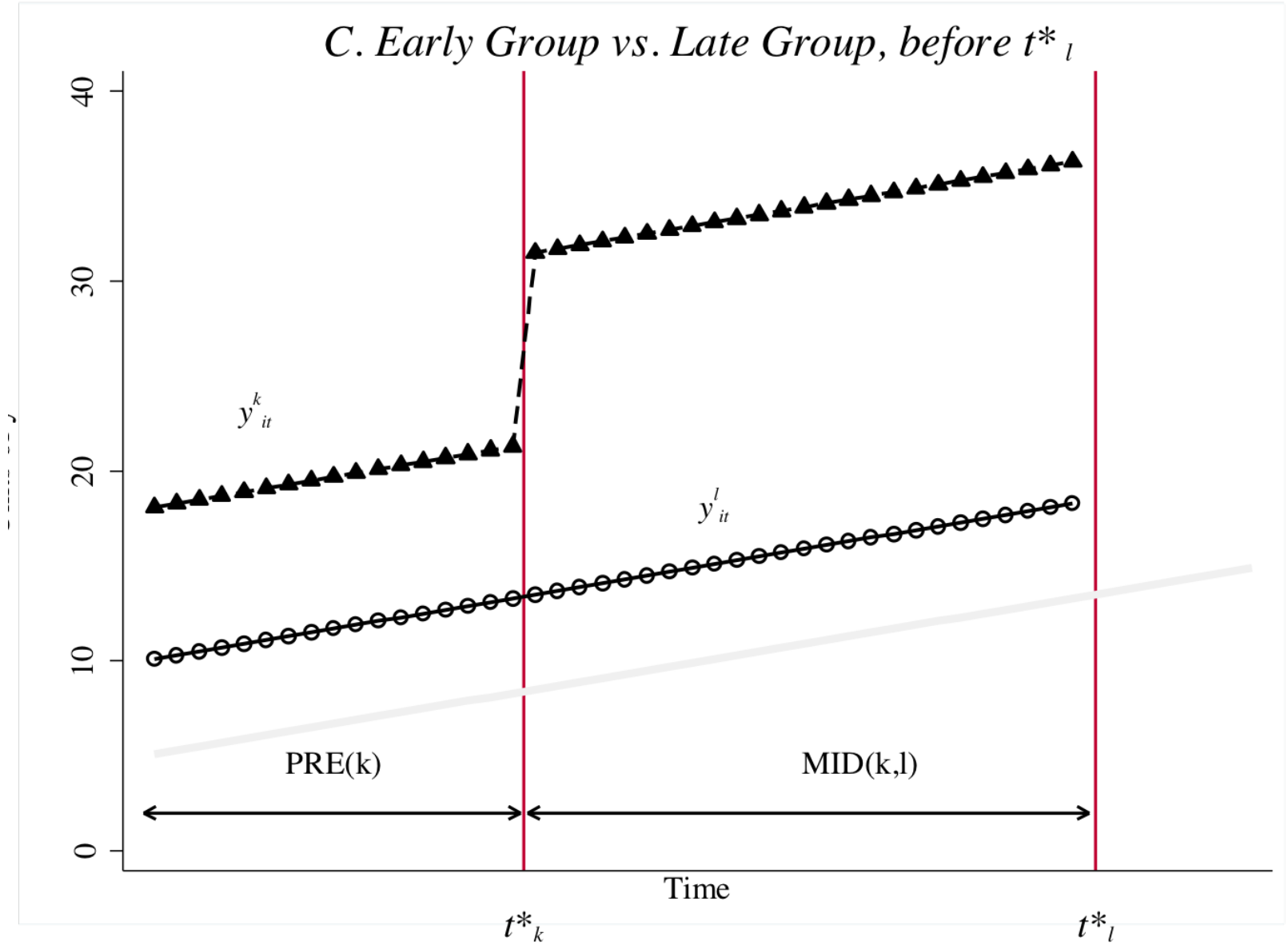
$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)} \right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)} \right)$$



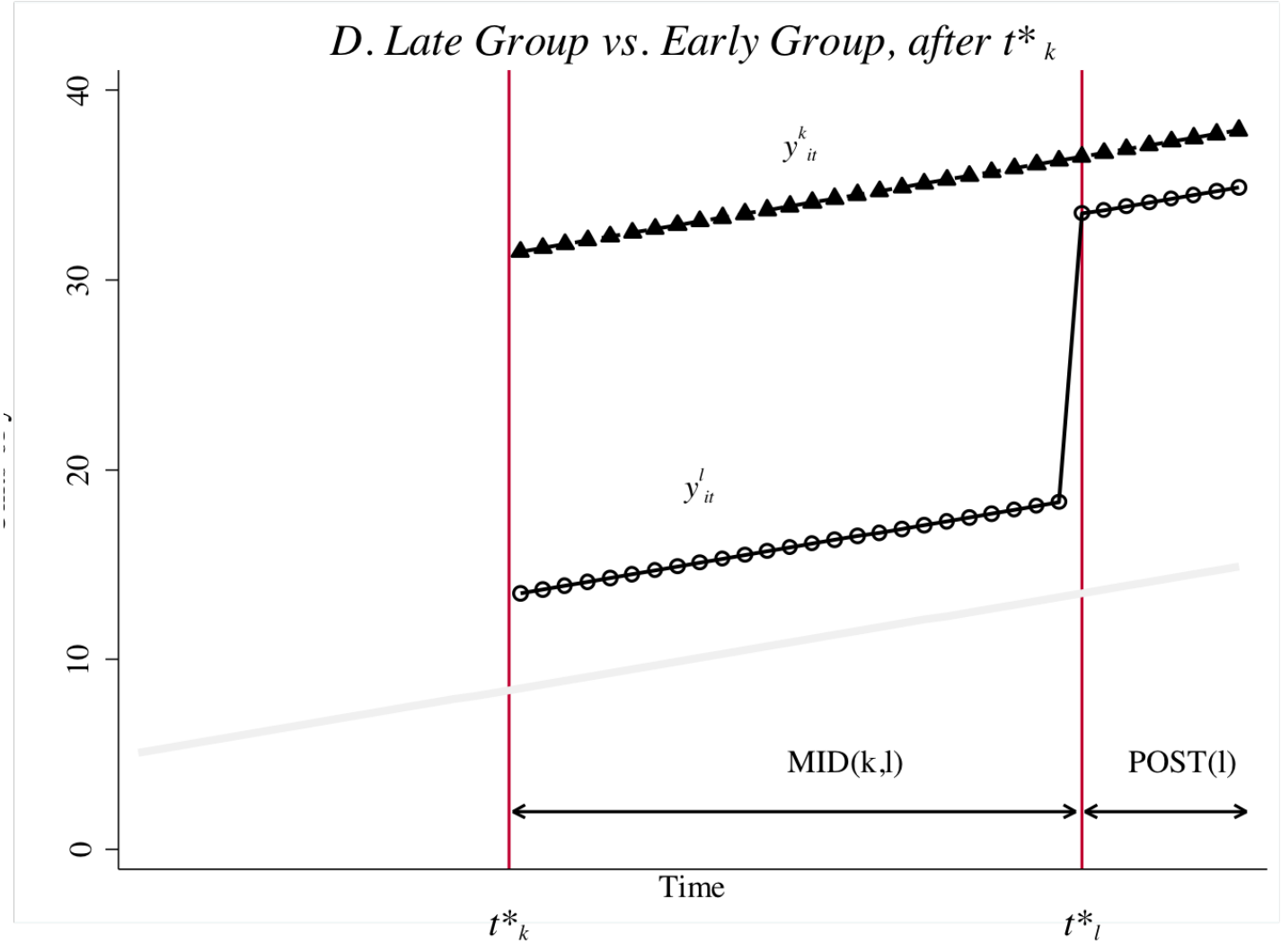
$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_{l}^{post(l)} - \overline{y}_{l}^{pre(l)}\right) - \left(\overline{y}_{U}^{post(l)} - \overline{y}_{U}^{pre(l)}\right)$$



$$\delta_{kl}^{2x2,k} = \left(\overline{y}_k^{MID(k,l)} - \overline{y}_k^{Pre(k,l)} \right) - \left(\overline{y}_l^{MID(k,l)} - \overline{y}_l^{PRE(k,l)} \right)$$



$$\delta_{lk}^{2x2,l} = \left(\overline{y}_l^{POST(k,l)} - \overline{y}_l^{MID(k,l)} \right) - \left(\overline{y}_k^{POST(k,l)} - \overline{y}_k^{MID(k,l)} \right)$$



Bacon decomposition

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

TWFE estimate of $\widehat{\delta}$ is equal to a weighted average over all group 2x2 (of which there are 4 in this example)

$$\widehat{\delta}^{TWFE} = \sum_{k \neq U} s_{kU} \widehat{\delta}^{2x2}_{kU} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}^{2x2,k}_{kl} + (1 - \mu_{kl}) \widehat{\delta}^{2x2,l}_{lk} \right]$$

where that first 2x2 combines the k compared to U and the l to U (combined to make the equation shorter)

Third, the Weights

$$
\begin{aligned}
s_{ku} &= \frac{n_k n_u \overline{D}_k (1 - \overline{D}_k)}{\widehat{Var}(\tilde{D}_{it})} \\
s_{kl} &= \frac{n_k n_l (\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l))}{\widehat{Var}(\tilde{D}_{it})} \\
\mu_{kl} &= \frac{1 - \overline{D}_k}{1 - (\overline{D}_k - \overline{D}_l)}
\end{aligned}
$$

where n refer to sample sizes, $\overline{D}_k(1 - \overline{D}_k)$ $(\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l))$ expressions refer to variance of treatment, and the final equation is the same for two timing groups.

Weights discussion

- Two things to note:
  - → More units in a group, the bigger its 2x2 weight is
  - → Group treatment variance weights up or down a group's 2x2
- Think about what causes the treatment variance to be as big as possible. Let's think about the s_{ku} weights.
  - → $\overline{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
  - → $\overline{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
  - → $\overline{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$
  - → $\overline{D} = 0.6$. Then $0.6 \times 0.4 = 0.24$
- This means the weight on treatment variance is maximized for *groups treated in middle of the panel*

More weights discussion

- But what about the “treated on treated” weights (i.e., $\overline{D}_k - \overline{D}_l$)
- Same principle as before - when the difference between treatment variance is close to 0.5, those 2x2s are given the greatest weight
- For instance, say $t_k^* = 0.15$ and $t_l^* = 0.67$. Then $\overline{D}_k - \overline{D}_l = 0.52$. And thus $0.52 \times 0.48 = 0.2496$.

Summarizing TWFE centralities

- Groups in the middle of the panel weight up their respective 2x2s via the variance weighting
- Decomposition highlights the strange role of panel length when using TWFE
- Different choices about panel length change both the 2x2 and the weights based on variance of treatment

Back to TWFE

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

- So we know that the estimate is a weighted average over all "four averages and three subtractions" but is that good or bad?
- It's good if it's unbiased; it's bad if it isn't, and the decomposition doesn't tell us which unless we replace realized outcomes with potential outcomes
- Bacon shows that TWFE estimate of δ needs two assumptions for unbiasedness:
  1. variance weighted parallel trends are zero and
  2. no dynamic treatment effects (not the case with 2x2)
- Under those assumptions, TWFE estimator estimates the variance weighted ATT as a weighted average of all possible ATTs (not just weighted average of DiDs)

Moving from 2x2s to causal effects and bias terms

Let's start breaking down these estimators into their corresponding estimation objects expressed in causal effects and biases

$$
\begin{aligned}
\widehat{\delta}_{kU}^{2x2} &= ATT_k Post + \Delta Y_k^0(Post(k), Pre(k)) - \Delta Y_U^0(Post(k), Pre) \\
\widehat{\delta}_{kl}^{2x2} &= ATT_k(MID) + \Delta Y_k^0(MID, Pre) - \Delta Y_l^0(MID, Pre)
\end{aligned}
$$

These look the same because you're always comparing the treated unit with an untreated unit (though in the second case it's just that they haven't been treated *yet*).

The dangerous 2x2

But what about the 2x2 that compared the late groups to the already-treated earlier groups? With a lot of substitutions we get:

$$\widehat{\delta}_{lk}^{2x2} = ATT_{l,Post(l)} + \underbrace{\Delta Y_l^0(Post(l), MID) - \Delta Y_k^0(Post(l), MID)}_{\text{Parallel trends bias}} - \underbrace{(ATT_k(Post) - ATT_k(Mid))}_{\text{Heterogeneity bias!}}$$

Substitute all this stuff into the decomposition formula

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}^{2x2}_{kU} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}^{2x2,k}_{kl} + (1 - \mu_{kl}) \widehat{\delta}^{2x2,l}_{kl} \right]$$

where we will make these substitutions

$$\begin{aligned}
\widehat{\delta}^{2x2}_{kU} &= ATT_k(Post) + \Delta Y^0_l(Post, Pre) - \Delta Y^0_U(Post, Pre) \\
\widehat{\delta}^{2x2,k}_{kl} &= ATT_k(Mid) + \Delta Y^0_l(Mid, Pre) - \Delta Y^0_l(Mid, Pre) \\
\widehat{\delta}^{2x2,l}_{lk} &= ATT_l Post(l) + \Delta Y^0_l(Post(l), MID) - \Delta Y^0_k(Post(l), MID) \\
&\quad -(ATT_k(Post) - ATT_k(Mid))
\end{aligned}$$

Notice all those potential sources of biases!

Potential Outcome Notation

$$p\,lim\,\widehat{\delta}^{TWFE}_{n\to\infty} \quad = \quad VWATT + VWPT - \Delta ATT$$

- Notice the number of assumptions needed *even* to estimate this very strange weighted ATT (which is a function of how you drew the panel in the first place).
- With dynamics, it attenuates the estimate (bias) and can even reverse sign depending on the magnitudes of what is otherwise effects in the sign in a reinforcing direction!
- Model can flip signs (does not satisfy a “no sign flip property”)

Simulated data

- 1000 firms, 40 states, 25 firms per states, 1980 to 2009 or 30 years, 30,000 observations, four groups
- I'll impose "unit level parallel trends", which is much stronger than we need (we only need average parallel trends)
- Also no anticipation of treatment effects until treatment occurs but does *not* guarantee homogenous treatment effects
- Two types of situations: constant versus dynamic treatment effects

Constant vs Dynamic Treatment Effects

Calendar Time	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1981	0	0	0	0
1982	0	0	0	0
1983	0	0	0	0
1984	0	0	0	0
1985	0	0	0	0
1986	10	0	0	0
1987	10	0	0	0
1988	10	0	0	0
1989	10	0	0	0
1990	10	0	0	0
1991	10	0	0	0
1992	10	8	0	0
1993	10	8	0	0
1994	10	8	0	0
1995	10	8	0	0
1996	10	8	0	0
1997	10	8	0	0
1998	10	8	6	0
1999	10	8	6	0
2000	10	8	6	0
2001	10	8	6	0
2002	10	8	6	0

Calendar Time	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1981	0	0	0	0
1982	0	0	0	0
1983	0	0	0	0
1984	0	0	0	0
1985	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0

Group-time ATT

Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			

- Heterogenous treatment effects across time and across groups
- Cells are called “group-time ATT” (Callaway and Sant’anna 2020) or “cohort ATT” (Sun and Abraham 2020)
- ATT is weighted average of all cells and $+82$ with uniform weights $1/60$

Estimation

Estimate the following equation using OLS:

$$Y_{ist} = \alpha_i + \gamma_t + \delta D_{it} + \varepsilon_{ist}$$

Table: Estimating ATT with different models

	Truth	**(TWFE)**	**(CS)**	**(SA)**	**(BJS)**
$\widehat{ATT}$	82	-6.69***			

The sign flipped. Why? Because of *extreme* dynamics (i.e., $-\Delta ATT$)

Bacon decomposition

Table: Bacon Decomposition (TWFE $= -6.69$)

DD Comparison	**Weight**	**Avg DD Est**
Earlier T vs. Later C	0.500	51.800
Later T vs. Earlier C	0.500	-65.180

T $=$ Treatment; C$=$ Comparison

$(0.5 * 51.8) + (0.5 * -65.180) = -6.69$

While large weight on the “late to early 2x2” is *suggestive* of an issue, these would appear even if we had constant treatment effects

Roadmap

- Introduction
  - Managing expectations
  - Introducing difference-in-differences
  - Potential outcomes
  - Identification and Estimation
- Including Covariates
  - Inverse probability weighting
  - Outcome Regression and Double Robust
  - Lalonde lab
- Differential timing
  - Introduction
- Twoway fixed effects
  - Estimator
  - Applications
- TWFE Pathologies
  - Potential outcomes

Callaway and Sant'Anna 2020

CS is a DiD estimator used for estimating and then summarizing smaller ATT parameters under differential timing and conditional parallel trends into more policy relevant ATT parameters (either dynamic or static)

Difference-in-differences with multiple time periods

Authors	Brantly Callaway, Pedro HC Sant'Anna
Publication date	2021/12/1
Journal	Journal of Econometrics
Volume	225
Issue	2
Pages	200-230
Publisher	North-Holland
Description	In this article, we consider identification, estimation, and inference procedures for treatment effect parameters using Difference-in-Differences (DiD) with (i) multiple time periods, (ii) variation in treatment timing, and (iii) when the "parallel trends assumption" holds potentially only after conditioning on observed covariates. We show that a family of causal effect parameters are identified in staggered DiD setups, even if differences in observed characteristics create non-parallel outcome dynamics between groups. Our identification results allow one to use outcome regression, inverse probability weighting, or doubly-robust estimands. We also propose different aggregation schemes that can be used to highlight treatment effect heterogeneity across different dimensions as well as to summarize the overall effect of participating in the treatment. We establish the asymptotic properties of the proposed estimators and prove the …
Total citations	Cited by 2378

When is CS used

Just some examples of when you'd want to consider it:

1. When treatment effects differ depending on when it was adopted
2. When treatment effects change over time
3. When shortrun treatment effects are different than longrun effects
4. When treatment effect dynamics differ if people are first treated in a recession relative to expansion years

CS estimates the ATT by identifying smaller causal effects and aggregating them using non-negative weights

Group-time ATT

Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			

Each cell contains that group's ATT(g,t)

$$ATT(g,t) = E[Y_t^1 - Y_t^0 | G_g = 1]$$

CS identifies all feasible ATT(g,t)

Group-time ATT

Group-time ATT is the ATT for a specific group and time

- Groups are basically cohorts of units treated at the same time
- Group-time ATT estimates are simple (weighted) differences in means
- Does not directly restrict heterogeneity with respect to observed covariates, timing or the evolution of treatment effects over time
- Allows us ways to choose our aggregations
- Inference is the bootstrap

Notation

- T periods going from $t = 1, \ldots, T$
- Units are either treated ($D_t = 1$) or untreated ($D_t = 0$) but once treated cannot revert to untreated state
- G_g signifies a group and is binary. Equals one if individual units are treated at time period t.
- C is also binary and indicates a control group unit equalling one if "never treated" (can be relaxed though to "not yet treated")
  - $\rightarrow$ Recall the problem with TWFE on using treatment units as controls
- Generalized propensity score enters into the estimator as a weight:

$$\widehat{p(X)} = Pr(G_g = 1 | X, G_g + C = 1)$$

Assumptions

Assumption 1: Sampling is iid (panel data, but repeated cross-sections are possible)

Assumption 2: Conditional parallel trends (for either never treated or not yet treated)

$$E[Y_t^0 - Y_{t-1}^0 | X, G_g = 1] = [Y_t^0 - Y_{t-1}^0 | X, C = 1]$$

Assumption 3: Irreversible treatment

Assumption 4: Common support (propensity score)

Assumption 5: Limited treatment anticipation (i.e., treatment effects are zero pre-treatment)

CS Estimator (the IPW version)

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E\left[\frac{\hat{p}(X)C}{1-\hat{p}(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

This is the inverse probability weighting estimator. Alternatively, there is an outcome regression approach and a doubly robust. Sant'Anna recommends DR. CS uses the never-treated or the not-yet-treated as controls but never the already-treated

Aggregated vs single year/group ATT

- The method they propose is really just identifying very narrow ATT per group time.
- But we are often interested in more aggregate parameters, like the ATT across all groups and all times
- They present two alternative methods for building “interesting parameters”
- Inference from a bootstrap

Group-time ATT

Truth				
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			
Feasible ATT	68.3333333			

CS estimates				
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1981	-0.0548	0.0191	0.0578	0
1986	10.0258	-0.0128	-0.0382	0
1987	20.0439	0.0349	-0.0105	0
1988	30.0028	-0.0516	-0.0055	0
1989	40.0201	0.0257	0.0313	0
1990	50.0249	0.0285	-0.0284	0
1991	60.0172	-0.0395	0.0335	0
1992	69.9961	8.013	0	0
1993	80.0155	16.0117	0.0105	0
1994	89.9912	24.0149	0.0185	0
1995	99.9757	32.0219	-0.0505	0
1996	110.0465	40.0186	0.0344	0
1997	120.0222	48.0338	-0.0101	0
1998	129.9164	56.0051	6.027	0
1999	139.9235	63.9884	11.969	0
2000	150.0087	71.9924	18.0152	0
2001	159.9702	80.0152	23.9656	0
2002	169.9857	88.0745	29.9757	0
2003	179.981	96.0161	36.013	0
2004				
2005				
2006				
2007				
2008				
2009				
Total ATT	n/a			
Feasible ATT	68.33718056			

Question: Why didn't CS estimate all ATT(g,t)? What is "feasible ATT"?

Reporting results

Table: Estimating ATT using only pre-2004 data

	(Truth)	(TWFE)	(CS)	(SA)	(BJS)
$\widehat{Feasible\ ATT}$	68.33	26.81 ***	68.34***		

TWFE is no longer negative, interestingly, once we eliminate the last group (giving us a never-treated group), but is still suffering from attenuation bias.

Event study and differential timing

- Sometimes we care about a simple summary, and sometimes we care about separating it out in time and sometimes in even more interesting ways
- Event studies with one treatment group and one untreated group were relatively straightforward
- Interact treatment group with calendar date to get a series of leads and lags
- But when there are more than one treatment group, specification challenges emerge

Differential timing complicates plotting sample averages

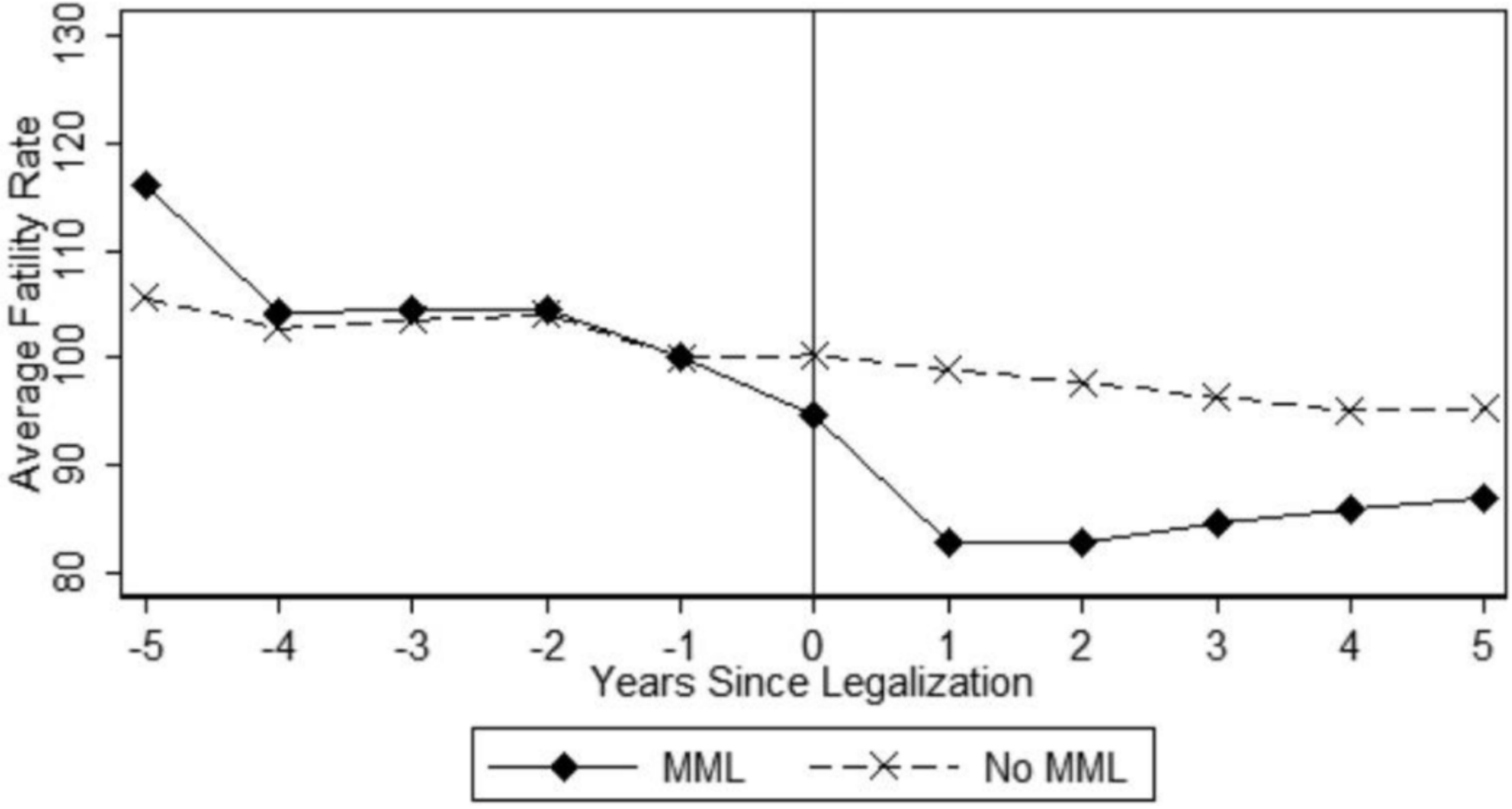


Figure: Anderson, et al. (2013) display of raw traffic fatality rates for re-centered treatment states and control states with randomized treatment dates

Replicated from a project of mine

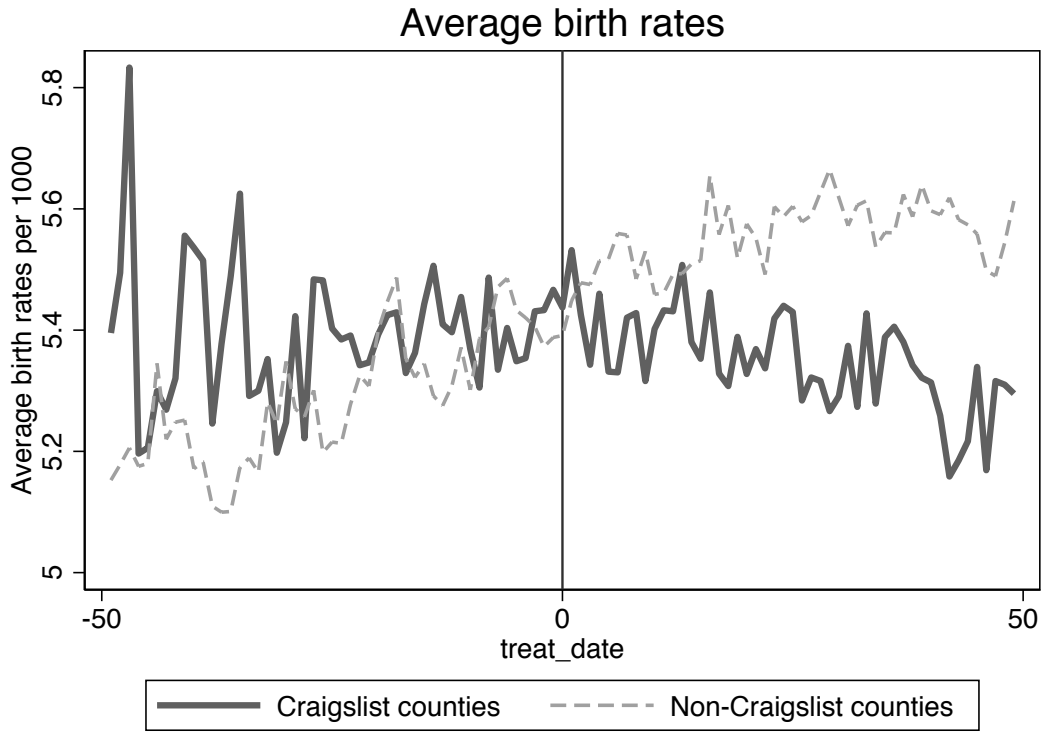
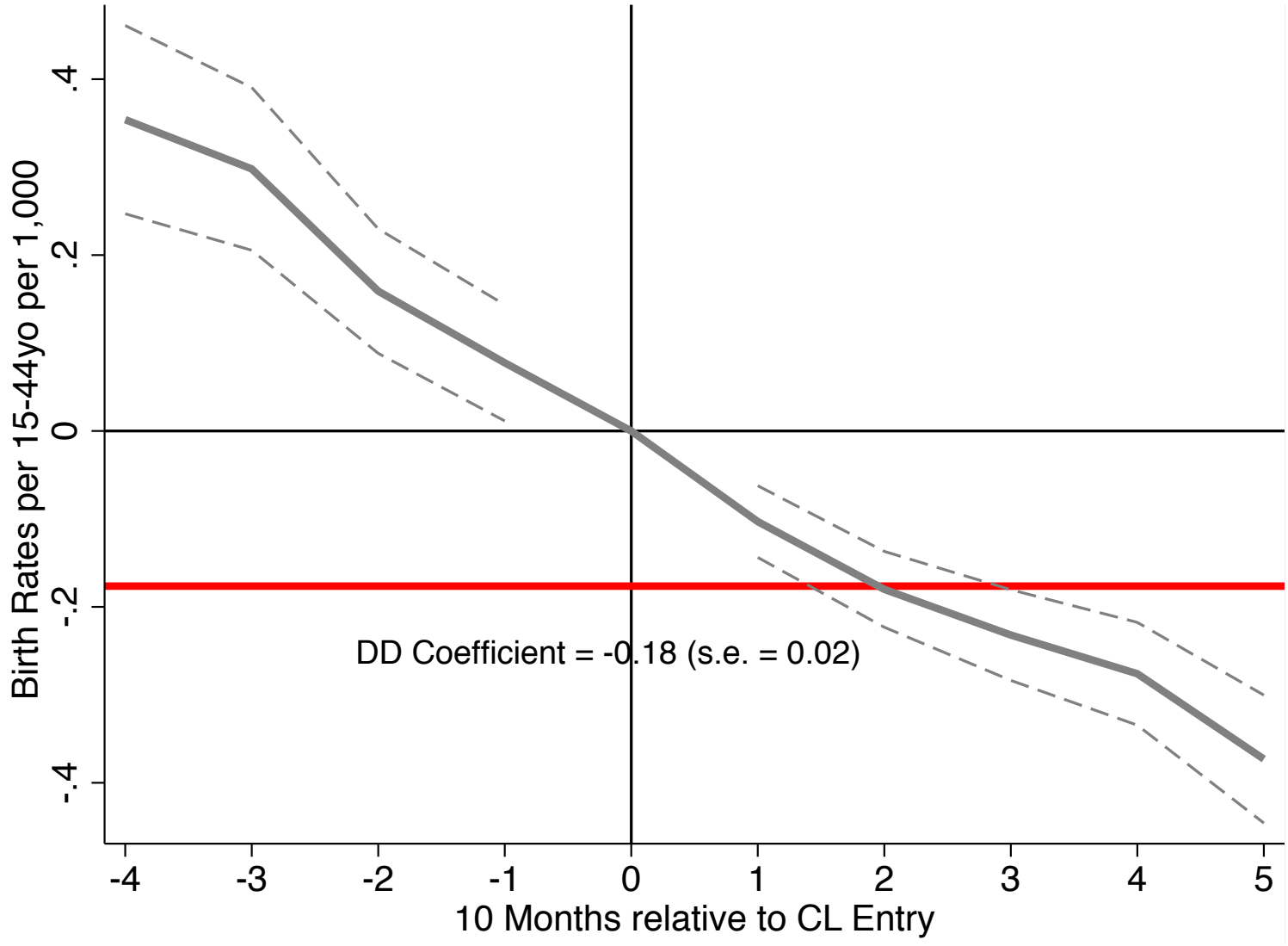


Figure: Roll out of Craigslist "personal ads" for casual intimate encounters and birth rates using the "randomized treatment assignment" approach for visualization

Event study specification with TWFE

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{g \in G} \mu_g 1\{t - E_i \in g\} + \varepsilon_{i,t}$$

Coefficient μ_g on a dummy measuring the number of years prior to or after that unit was treated.



Same data as a couple slides ago, leads don't look good, so I abandoned the project.

Bias of TWFE Event Study Specificaiton

- Bacon only focused on the static specification, and that's where the biases due to dynamics revealed itself
- He was unable to get into the leads and lags using the FWL method he was using ("it's hard!" - Bacon)
- Sophie Sun and Sarah Abraham did though – prompted by a stray comment by their professor
- But they also unlike Bacon present a solution (which is like CS, but discovered independently)

Sun and Abraham 2020

1. SA shows a decomposition of the population regression coefficient on event study leads and lags with differential timing estimated with TWFE
2. They show that the population regression coefficient is "contaminated" by information from other leads and lags (which is then later generalized by Goldsmith-Pinkham, Hull and Kolsar 2022)
3. SA presents an alternative estimator that is a version of CS only using the "last cohort" as the treatment group (not the not-yet-treated)
4. Derives the variance of the estimator instead of bootstrapping, handles covariates differently than CS, but otherwise identical

Summarizing (cont.)

- Under homogenous treatment profiles, weights sum to zero and“cancel out” the treatment effects from other periods
- Under treatment effect heterogeneity, they do not cancel out and leads and lags are biased
- They present a 3-step TWFE based alternative estimator which addresses the problems that they find

Some notation and terms

- As people often **bin** the data, we allow a lead or lag l to appear in bin g so sometimes they use g instead of l or $l \in g$
- Building block is the "cohort-specific ATT" or $CATT_{e,l}$ – same as ATT(g,t)
- Our goal is to estimate $CATT_{e,l}$ with population regression coefficient μ_l
- They focus on irreversible treatment where treatment status is non-decreasing sequence of zeroes and ones

Difficult notation (cont.)

- The ∞ symbol is used to either describe the group ($E_i = \infty$) or the potential outcome (Y^∞)
- $Y_{i,t}^\infty$ is is the potential outcome for unit i if it had never received treatment (versus received it later), also called the baseline outcome
- Other counterfactuals are possible – maybe unit i isn't "never treated" but treated later in counterfactual

More difficult notation (cont.)

- Treatment effects are the difference between the observed outcome relative to the never-treated counterfactual outcome: $Y_{i,t} - Y_{i,t}^{\infty}$
- We can take the average of treatment effects at a given relative time period across units first treated at time $E_i = e$ (same cohort) which is what we mean by $CATT_{e,l}$
- Doesn't use t index time ("calendar time"), rather uses l which is time until or time after treatment date e ("relative time")
- Think of it as l =year - treatment date

Relative vs calendar event time

```
. list state-treat time_til in 1/10

     +-------------------------------------------------------------------------------------+
     | state      firms   year    n   id   group   treat_~e   treat   time_til |
     |-------------------------------------------------------------------------------------|
  1. |     1   .3257218   1980    1    1       1       1986       0         -6 |
  2. |     1   .3257218   1981    2    1       1       1986       0         -5 |
  3. |     1   .3257218   1982    3    1       1       1986       0         -4 |
  4. |     1   .3257218   1983    4    1       1       1986       0         -3 |
  5. |     1   .3257218   1984    5    1       1       1986       0         -2 |
     |-------------------------------------------------------------------------------------|
  6. |     1   .3257218   1985    6    1       1       1986       0         -1 |
  7. |     1   .3257218   1986    7    1       1       1986       1          0 |
  8. |     1   .3257218   1987    8    1       1       1986       1          1 |
  9. |     1   .3257218   1988    9    1       1       1986       1          2 |
 10. |     1   .3257218   1989   10    1       1       1986       1          3 |
     +-------------------------------------------------------------------------------------+
```

Definition 1

Definition 1: The cohort-specific ATT l periods from initial treatment date e is:

$$CATT_{e,l} = E[Y_{i,e+l} - Y^{\infty}_{i,e+l} | E_i = e]$$

Fill out the second part of the Group-time ATT exercise together.

TWFE assumptions

- For consistent estimates of the coefficient leads and lags using TWFE model, we need three assumptions
- For SA and CS, we only need two
- Let’s look then at the three

Assumption 1: Parallel trends

Assumption 1: Parallel trends in baseline outcomes:
$E[Y_{i,t}^{\infty} - Y_{i,s}^{\infty} | E_i = e]$ is the same for all $e \in supp(E_i)$ and for all s, t and is equal to $E[Y_{i,t}^{\infty} - Y_{i,s}^{\infty}]$

Lead and lag coefficients are DiD equations but once we invoke parallel trends they can become causal parameters. This reminds us again how crucial it is to have appropriate controls

Assumption 2: No anticipation

Assumption 2: No anticipator behavior in pre-treatment periods: There is a set of pre-treatment periods such that $E[Y^{e}_{i,e+l} - Y^{\infty}_{i,e+l} | E_i = e] = 0$ for all possible leads.

Essentially means that pre-treatment, the causal effect is zero. Most plausible if no one sees the treatment coming, but even if they see it coming, they may not be able to make adjustments that affect outcomes

Assumption 3: Homogeneity

Assumption 3: Treatment effect profile homogeneity: For each relative time period l, the $CATT_{e,l}$ doesn't depend on the cohort and is equal to $CATT_l$.

Treatment effect heterogeneity

- Assumption 3 is violated when different cohorts experience different paths of treatment effects
- Cohorts may differ in their covariates which affect how they respond to treatment (e.g., if treatment effects vary with age, and there is variation in age across units first treated at different times, then there will be heterogeneous treatment effects)
- Doesn’t rule out parallel trends

Event study model

Dynamic TWFE model

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{g \in G} \mu_g 1\{t - E_i \in g\} + \varepsilon_{i,t}$$

We are interested in the properties of μ_g under differential timing as well as whether there are any never-treated units

Interpreting $\widehat{\mu_g}$ under no to all assumptions

Proposition 1 (no assumptions): The population regression coefficient on relative period bin g is a linear combination of differences in trends from its own relative period $l \in g$, from relative periods $l \in g'$ of other bins $g' \neq g$, and from relative periods excluded from the specification (e.g., trimming).

$$
\begin{aligned}
\mu_g \quad = \quad & \underbrace{\sum_{l \in g} \sum_{e} w^g_{e,l} (E[Y_{i,e+l} - Y^{\infty}_{i,0} | E_i = e] - E[Y^{\infty}_{i,e+l} - Y^{\infty}_{i,0}])}_{\text{Targets}} \\
+ \quad & \underbrace{\sum_{g' \neq g} \sum_{l \in g'} \sum_{e} w^g_{e,l} (E[Y_{i,e+l} - Y^{\infty}_{i,0} | E_i = e] - E[Y^{\infty}_{i,e+l} - Y^{\infty}_{i,0}])}_{\text{Contamination from other leads and lags}} \\
+ \quad & \underbrace{\sum_{l \in g^{excl}} \sum_{e} w^g_{e,l} (E[Y_{i,e+l} - Y^{\infty}_{i,0} | E_i = e] - E[Y^{\infty}_{i,e+l} - Y^{\infty}_{i,0}])}_{\text{Contamination from dropped periods}}
\end{aligned}
$$

Weight ($w^g_{e,l}$) summation cheat sheet

1. For relative periods of μ_g own $l \in g$, $\sum_{l \in g} \sum_e w^g_{e,l} = 1$
2. For relative periods belonging to some other bin $l \in g'$ and $g' \neq g$, t $\sum_{l \in g'} \sum_e w^g_{e,l} = 0$
3. For relative periods not included in G, $\sum_{l \in g^{excl}} \sum_e w^g_{e,l} = -1$

Estimating the weights

Regress $D_{i,t}^{l} \times 1\{E_i = e\}$ on:

1. all bin indicators included in the main TWFE regression,
2. $\{1\{t - E_i \in g\}\}_{g \in G}$(i.e., leads and lags) and
3. the unit and time fixed effects

Still biased under parallel trends

Proposition 2: Under the parallel trends only, the population regression coefficient on the indicator for relative period bing g is a linear combination of $CATT_{e,l \in g}$ as well as $CATT_{d,l'}$ from other relative periods $l' \notin g$ with the same weights stated in Proposition 1:

$$\mu_g = \underbrace{\sum_{l \in g} \sum_{e} w^g_{e,l} CATT_{e,l}}_{\text{Desirable}} + \underbrace{\sum_{g' \neq g, g' \in G} \sum_{l' \in g'} \sum_{e} w^g_{e,l'} CATT_{e,l'}}_{\text{Bias from other specified bins}} + \underbrace{\sum_{l' \in g^{excl}} \sum_{e} w^g_{e,l'} CATT_{e,l'}}_{\text{Bias from dropped relative time indicators}}$$

Still biased under parallel trends and no anticipation

Proposition 3: If parallel trends holds and no anticipation holds for all $l < 0$ (i.e., no anticipatory behavior pre-treatment), then the population regression coefficient μ_g for g is a linear combination of post-treatment $CATT_{e,l'}$ for all $l' \geq 0$.

$$
\begin{aligned}
\mu_g \quad = \quad & \sum_{l' \in g, l' \geq 0} \sum_{e} w^g_{e,l'} CATT_{e,l'} \\
& + \sum_{g' \neq g, g' \in G} \sum_{l' \in g', l' \geq 0} \sum_{e} w^g_{e,l'} CATT_{e,l'} \\
& + \sum_{l' \in g^{excl}, l' \geq 0} \sum_{e} w^g_{w,l'} CATT_{e,l'}
\end{aligned}
$$

Proposition 3 comment

Notice how once we impose zero pre-treatment treatment effects, those terms are gone (i.e., no $l \in g, l < 0$). But the second term remains unless we impose treatment effect homogeneity (homogeneity causes terms due to weights summing to zero to cancel out). Thus μ_g may be non-zero for pre-treatment periods *even though parallel trends hold in the pre period.*

Proposition 4

Proposition 4: If parallel trends and treatment effect homogeneity, then $CATT_{e,l} = ATT_l$ is constant across e for a given l, and the population regression coefficient μ_g is equal to a linear combination of $ATT_{l \in g}$, as well as $ATT_{l' \notin g}$ from other relative periods

$$\begin{aligned} \mu_g &= \sum_{l \in g} w_l^g ATT_l \\ &\quad + \sum_{g' \neq g} \sum_{l' \in g'} w_{l'}^g ATT_{l'} \\ &\quad + \sum_{l' \in g^{excl}} w_{l'}^g ATT_{l'} \end{aligned}$$

Simple example

Balanced panel $T = 2$ with cohorts $E_i \in \{1, 2\}$. For illustrative purposes, we will include bins $\{-2, 0\}$ in our calculations but drop $\{-1, 1\}$.

Simple example

$$\mu_{-2} = \underbrace{CATT_{2,-2}}_{\text{own period}} + \underbrace{\frac{1}{2}CATT_{1,0} - \frac{1}{2}CATT_{2,0}}_{\text{other included bins}}$$
$$+ \underbrace{\frac{1}{2}CATT_{1,1} - CATT_{1,-1} - \frac{1}{2}CATT_{2,-1}}_{\text{Excluded bins}}$$

- Parallel trends gets us to all of the $CATT$
- No anticipation makes $CATT = 0$ for all $l < 0$ (all $l < 0$ cancel out)
- Homogeneity cancels second and third terms
- Still leaves $\frac{1}{2}CATT_{1,1}$ – you chose to exclude a group with a treatment effect

Lesson: drop the relative time indicators on the left, not things on the right, bc lagged effects will contaminate through the excluded bins

Robust event study estimation

- All the robust estimators under differential timing have solutions and they all skip over forbidden contrasts.
- Sun and Abraham (2020) propose a 3-step interacted weighted estimator (IW) using last treated group as control group
- Callaway and Sant'anna (2020) estimate group-time ATT which can be a weighted average over relative time periods too but uses "not-yet-treated" as control

Interaction-weighted estimator

- **Step one**: Do this DD regression and hold on to $\widehat{\delta}_{e,l}$

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e \notin C} \sum_{l \neq -1} \delta_{e,l} (1\{E_i = e\} \cdot D_{i,t}^l) + \varepsilon_{i,t}$$

Can use never-treated or last-treated cohort. Drop always treated. The $\delta_{e,l}$ is a DD estimator for $CATT_{e,l}$ with particular choices for pre-period and cohort controls

Interaction-weighted estimator

- **Step two**: Estimate weights using sample shares of each cohort in the relevant periods:

$$Pr(E_i = e | E_i \in [-l, T - l])$$

Interaction-weighted estimator

- **Step three**: Take a weighted average of estimates for $CATT_{e,l}$ from Step 1 with weight estimates from step 2

$$\widehat{v}_g = \frac{1}{|g|} \sum_{l \in g} \sum_{e} \widehat{\delta}_{e,l} \widehat{Pr}\{E_i = e | E_i \in [-l, T-l]\}$$

Consistency and Inference

- Under parallel trends and no anticipation, $\widehat{\delta}_{e,l}$ is consistent, and sample shares are also consistent estimators for population shares.
- Thus IW estimator is consistent for a weighted average of $CATT_{e,l}$ with weights equal to the share of each cohort in the relevant period(s).
- They show that each IW estimator is asymptotically normal and derive its asymptotic variance. Doesn't rely on bootstrap like CS.

DD Estimator of CATT

Definition 2: DD estimator with pre-period s and control cohorts C estimates $CATT_{e,l}$ as:

$$\widehat{\delta_{e,l}} = \frac{E_N[(Y_{i,e+l} - Y_{i,s}) \times 1\{E_i = e\}]}{E_N[1\{E_i = e\}]} - \frac{E_N[(Y_{i,e+l} \times 1\{E_i \in C\}]}{E_N[1\{E_i \in C\}]}$$

Proposition 5: If parallel trends and no anticipation both hold for all pre-periods, then the DD estimator using any pre-period and non-empty control cohorts (never-treated or not-yet-treated) is an unbiased estimate for $CATT_{e,l}$

Software

- **Stata**: eventstudyinteract (can be installed from ssc)
- **R**: fixest with subab() option (see `https://lrberge.github.io/fixest/reference/sunab.html/`)

Reporting results

Table: Estimating ATT

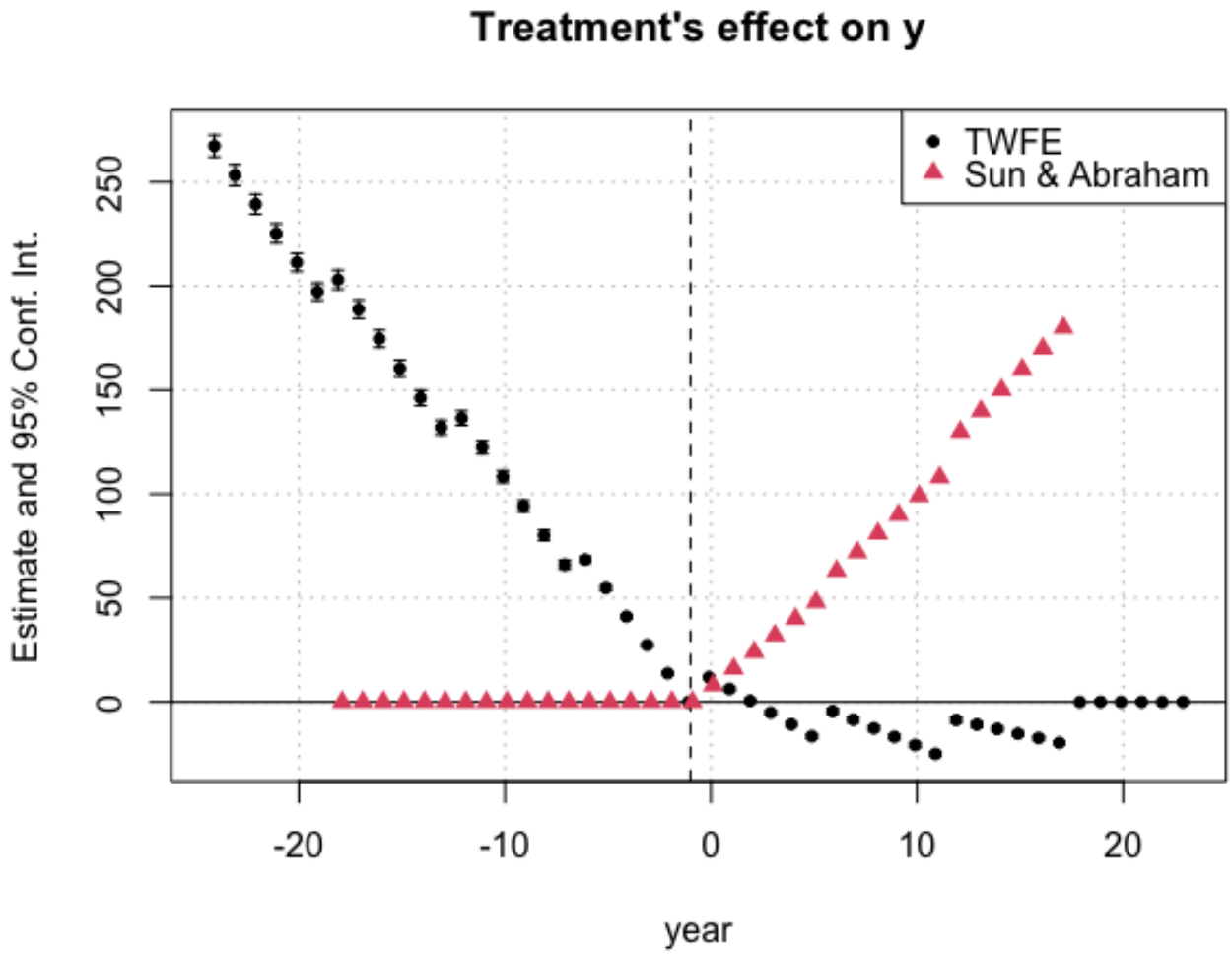
	(Truth)	**(TWFE)**	**(CS)**	**(SA)**	**(BJS)**
$\widehat{Feasible\ ATT}$	68.33	26.81***	68.34***	68.33***	

Computing relative event time leads and lags

Truth						Relative time coefficients		
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)		Leads	Truth	SA
1980	0	0	0	0		t-2	0	0.02
1986	10	0	0	0	(10+8+6)/3 = 8	t	8	8.01
1987	20	0	0	0	(20+16+12)/3 = 16	t+1	16	16.00
1988	30	0	0	0		t+2	24	24.00
1989	40	0	0	0		t+3	32	31.99
1990	50	0	0	0		t+4	40	40.00
1991	60	0	0	0		t+5	48	48.01
1992	70	8	0	0		t+6	63	62.99
1993	80	16	0	0		t+7	72	72.00
1994	90	24	0	0		t+8	81	80.99
1995	100	32	0	0		t+9	90	89.98
1996	110	40	0	0		t+10	99	99.06
1997	120	48	0	0		t+11	108	108.01
1998	130	56	6	0		t+12	130	129.92
1999	140	64	12	0		t+13	140	139.92
2000	150	72	18	0		t+14	150	150.01
2001	160	80	24	0		t+15	160	159.97
2002	170	88	30	0		t+16	170	169.99
2003	180	96	36	0		t+17	180	179.98
2004	190	104	42	4				
2005	200	112	48	8				
2006	210	120	54	12				
2007	220	128	60	16				
2008	230	136	66	20				
2009	240	144	72	24				

Two things to notice: (1) there only 17 lags with robust models but will be 24 with TWFE; (2) changing colors mean what?

Comparing TWFE and SA



Question: why is TWFE *falling* pre-treatment? Why is SA rising, but jagged, post-treatment?

de Chaisemartin and D'Haultfoeulle 2020

de Chaisemartin and D'Haultfouelle 2020 (dCdH) is different from the other papers in several ways

- Like SA, it's a diagnosis and a cure
- TWFE decomposition shows coefficient a weighted average of underlying treatment effects, but weights can be negative negating causal interpretation
- Propose a solution for both static and dynamic specification which does not use already treated as controls
- Treatment can turn on and off

Comment on Bacon

- Recall the Bacon decomposition – TWFE coefficients are decomposed into weighted average of all underlying 2x2s. Weights were non-negative and summed to one.
- But this decomposition was more a numerical decomposition – what exactly adds up to equal the TWFE coefficient using the data we observe?
- Bacon's decomposition is not "theoretical" – not in the way that other decompositions are. He is just explaining what OLS "does" when it calculates $\widehat{\delta}$
- Just explains what comparisons OLS is using to calculate the TWFE coefficient – just peels back the curtain.

Negative weights

- dCdH impose causal assumptions and try a different decomposition strategy
- Uses as its building block the unit-specific treatment effects
- Their decomposition will reveal negative weights on the underlying treatment effects (similar to negative weight on dynamics with Bacon)
- Remember though: the Bacon decomposition weights were *always* positive, because they were numerical weights (not theoretical weights) on the underlying 2x2s (not the treatment effects)

Turning on and off

- CS and SA both require interventions to turn on and stay on
- dCdH allows for “switching” on and off
- Before we move quickly into that, please note that the researcher bears the burden of knowing whether in fact you want to impose symmetry on turning on and off
- Roe v Wade “turned on” legalized abortion and 2022 it was “turned off” – do we want to treat these as simply a single policy flipping of the switch or two separate policies?

dCdH notation

- Individual treatment effects (iow, not the group-time ATT):

$$\Delta_{i,t}^{g} = Y_{i,t}^{1} - Y_{i,t}^{\infty}$$

  but where the treatment is in time period g. Notice –it's not the ATT (it's i individual treatment effect)
- with defined error term as $\varepsilon_{i,t}$:

$$D_{i,t} = \alpha_i + \alpha_t + \varepsilon_{i,t}$$

- Weights:

$$w_{i,t} = \frac{\varepsilon_{i,t}}{\frac{1}{N^T}\sum_{i,t:D_{i,t}=1}\varepsilon_{i,t}}$$

Parallel trend assumption

Strong unconditional PT

Assume that for every time period t and every group g, g',

$$E[Y_t^\infty - Y_{t-1}^\infty | G = g] = E[Y_t^\infty - Y_{t-1}^\infty | G = g']$$

Assume parallel trends for every unit in every cohort in every time period.

What then does TWFE estimate with differential timing?

dCdH Theorem

Theorem – dCdH decomposition

Assuming SUTVA, no anticipation and the strong PT, then let δ be the TWFE estimand associated with

$$Y_{i,t} = \alpha_i + \alpha_t + \delta D_{i,t} + \varepsilon_{i,t}$$

Then it follows that

$$\delta = E\left[\sum_{i,t:D_{i,t}=1} \frac{1}{N^T} w_{i,t} \cdot \Delta^g_{i,t} \right]$$

where $\sum_{i,t:D_{i,t}=1} \frac{w_{i,t}}{N^T} = 1$ but $w_{i,t}$ can be negative

Origins

- So once you run that specification, $\widehat{\delta}$ is going to recover a "non-convex average" over all unit level treatment effects (weights can be negative, more on this).
- Not sure who came first, because there were working papers before publications, but my understanding is dCdH was the first to prove this
- Very important theorem – established the "no sign flip property" for OLS with differential timing in the canonical static specification

Negative weights

- Very common now to hear about negative weights, and furthermore, that negative weights wipe out any causal interpretation, but why?
- Thought experiment: imagine every unit gained from the treatment, but their treatment effect when estimated was multiplied by a negative number
- It's possible it could flip the sign, but it would definitely at least pull the estimate away from the true effect
- This is dangerous – and it's caused by the forbidden contrasts (comparing treated to already treated) which is what the canonical TWFE static specification is doing (for many of us unknowingly)

Negative weights

- Doesn’t always pose a problem, but no proofs for this intuition known yet
- A large number of never-treated seems to make this less an issue
- Shrinking the spacing between treatment dates also can drive it down
- But does that mean that TWFE works, and what does it mean to work?
- TWFE still even when all the weights are positive the weighted average may not aggregate to what we think it does

Weighting

- The weights in OLS all come out of the model itself, *not the economic question*
- The economic question is “what parameter do you want? What does it look like? Who is in it?”
- And when you define the parameter up front, you’ve more or less defined the economic question you’re asking
- But OLS sort of ignores your question and just gives you what it wants

Weighting

- What makes something a good vs a bad weight?
- Not being negative is the absolute minimal requirement
- But it’s also not a good sign if you can’t really explain the weights

dCdH Solution

- dCdH propose an alternative that doesn't have the problems of TWFE – both avoiding negative weights and improving interpretability
- Recall, their model can handle reversible treatments