Causal Inference II

Mixtape Session

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

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

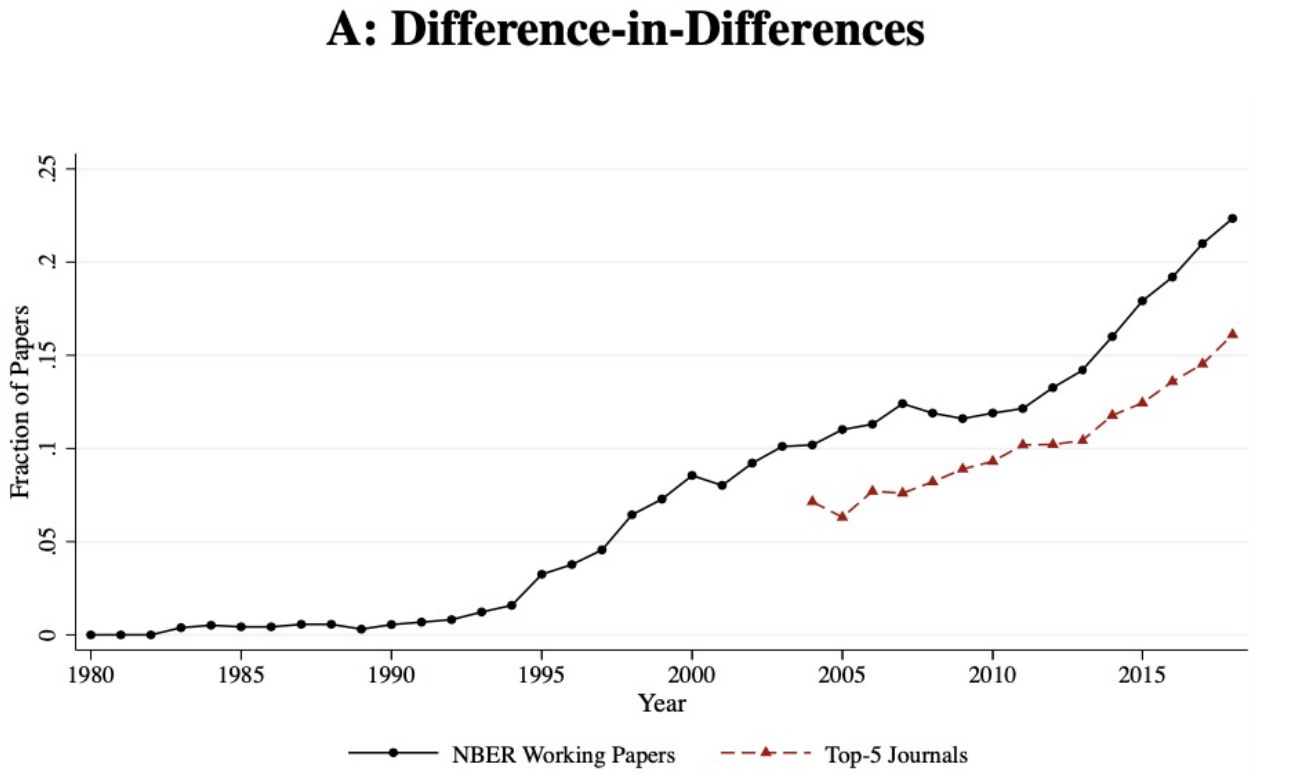
Day 3 outline

- 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

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)

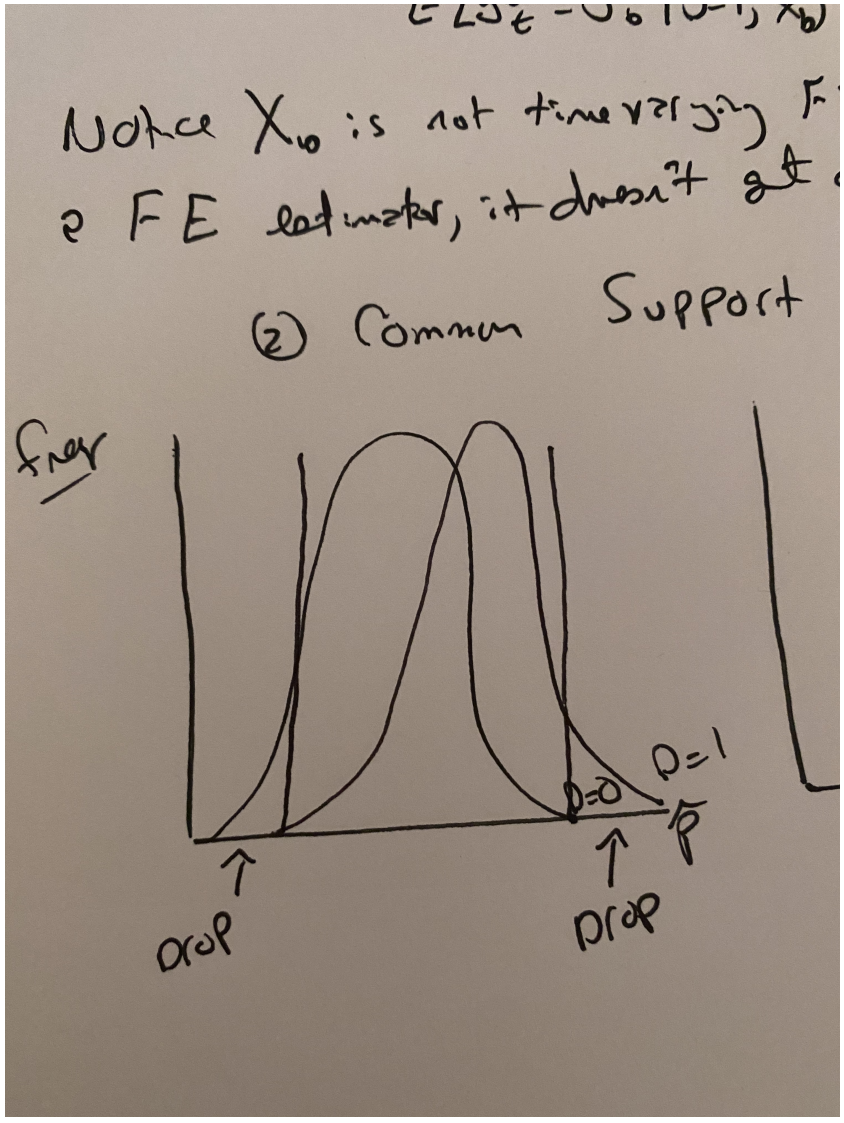
How do these work together?

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

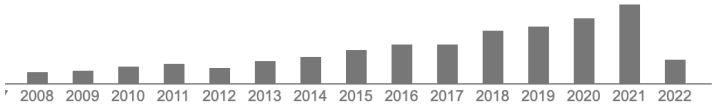
- It's common to hear people say that we don't know the propensity score; we can only estimate it. Same here – we approximate it with regressions
- Paper is titled "Semi-parametric DiD" because Abadie imposes structure on the polynomials used to construct the propensity score ("series logit")

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



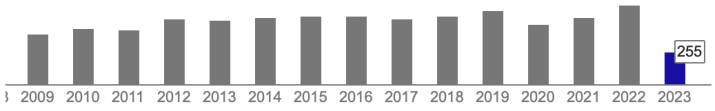
Scholar articles
Semiparametric difference-in-differences estimators
A Abadie - The Review of Economic Studies, 2005
Cited by 2330 Related articles All 12 versions

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

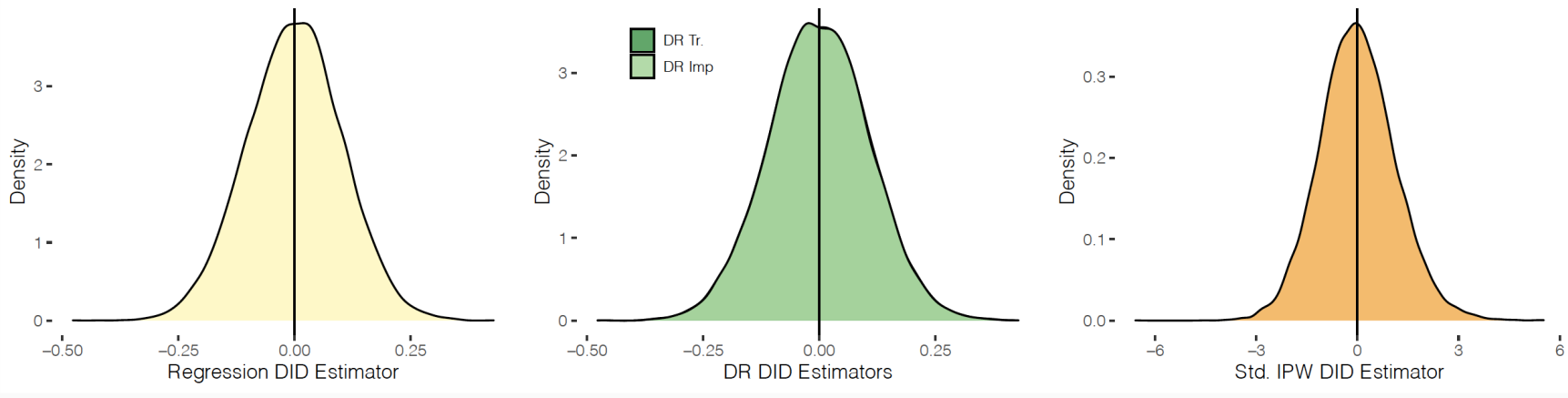
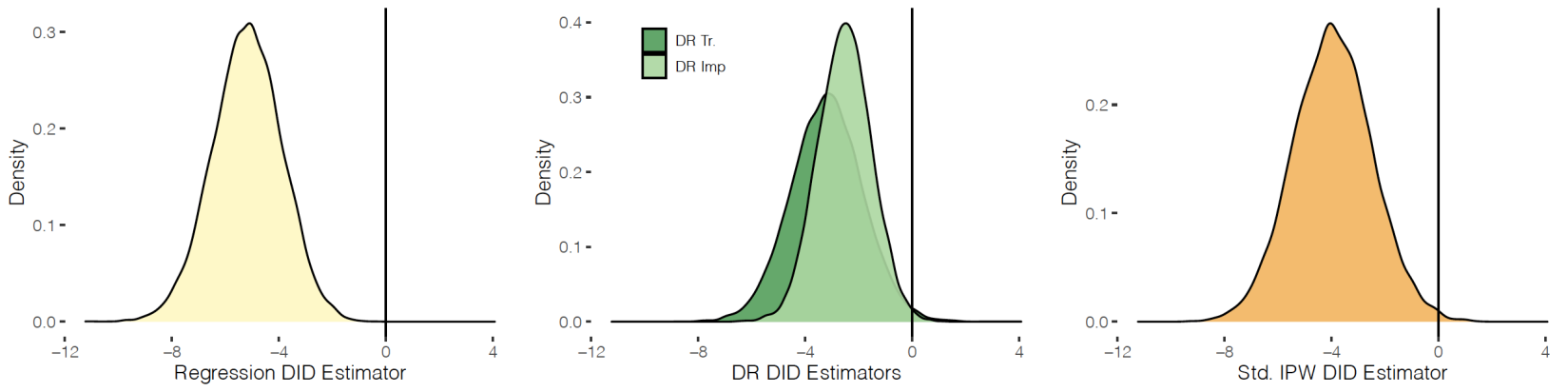


Table: Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

	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