Bird's eye view of difference-in-differences models with differential timing

taught by Scott Cunningham
Baylor University, Department of Economics

Roadmap

- Basic DiD
  - Welcome to DiD
  - Definitions
  - Identification
- Differential timing
  - Bacon decomposition
  - Static specification
  - Event studies
  - Imputation
- Conclusion

Welcome!

- This is a short two hour talk on just one of the new issues in difference-in-differences (DiD) – differential timing and twoway fixed effects (TWFE)
- This is a brief tour of new material, only focusing on a couple of things
- My curation of this extensive new literature is just a strategy I came up with for guiding us through differential timing

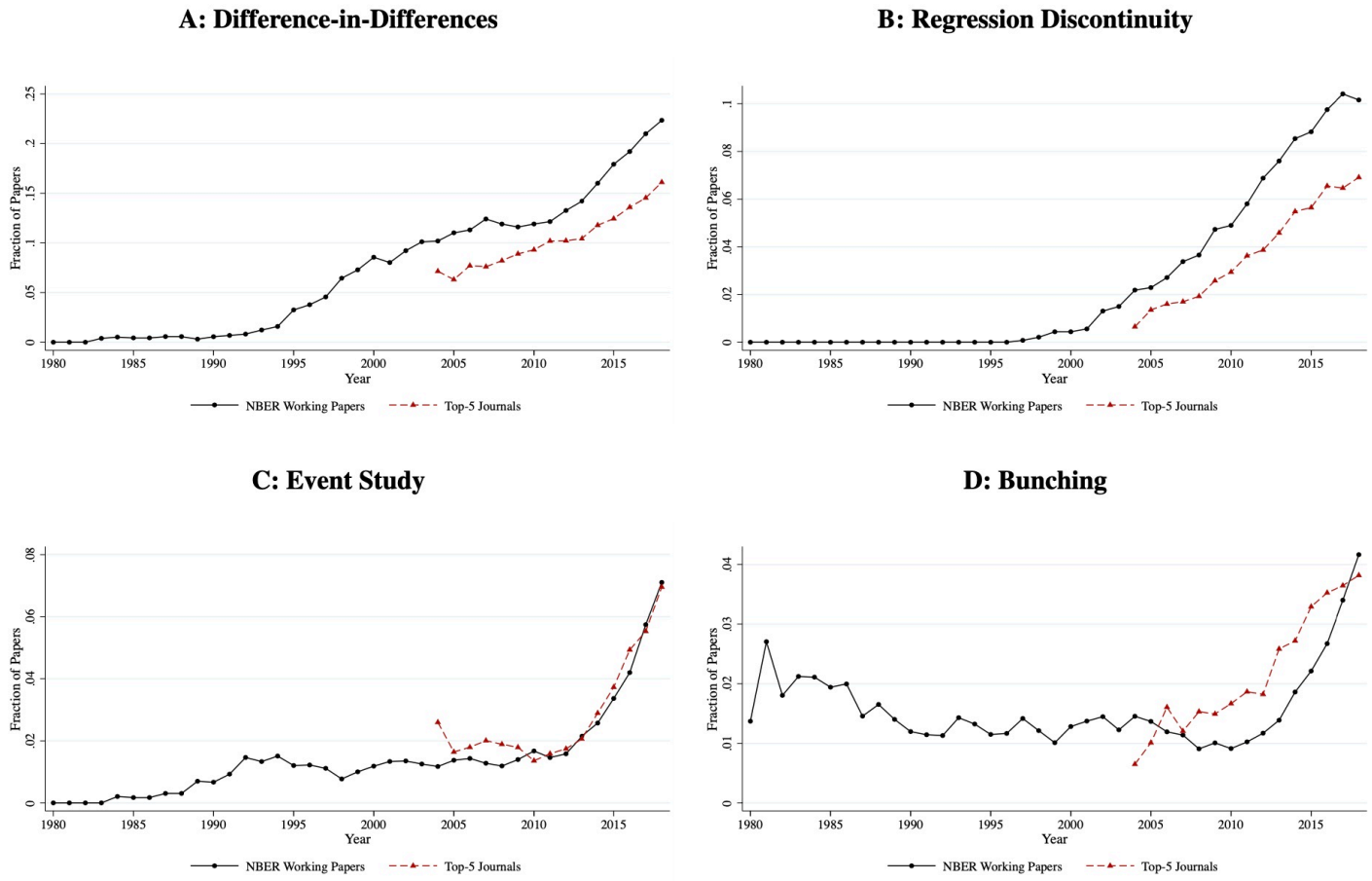
Why is learning about DiD worth our time?

- **DiD is useful tool**: Helps us study large social policies (e.g., Medicaid, minimum wages), plus it identifies the ATT which is a useful parameter
- **Turbulent last 3-4 years**: Many econometricians analyzed canonical DiD model specifications, found problems, proposed solutions
- **Died down**: New solutions yield similar answers, code is stable, widely available and in both R and Stata (but not python)

What is difference-in-differences (DiD)

- A group of units (treatment) are assigned some treatment and then compared to a group of units (control, or comparison) that weren't
- Brought into labor economics with Orley Ashenfelter in the 1970s and 1980s
- Now the most widely used quasi-experimental method

Figure IV: Quasi-Experimental Methods



Notes: This figure shows the fraction of papers referring to each type of quasi-experimental approach. See Table A.I for a list of terms. The series show 5-year moving averages.

Potential outcomes review

- DiD really can't be understood without committing to some common causality language
- Contemporary causal language is expressed using the potential outcomes model which frames causality in terms of counterfactuals
- Potential outcomes are pre-existing hypothetical worlds and may sometimes feel like science fiction (I like them partly for that reason)

Potential outcomes notation

I want to know the effect, δ_i, of my PhD, D, on my happiness, Y

Define treatment as 0 or 1:

$$D_{i,t} = \begin{cases} 1 \text{ if I finished my PhD} \\ 0 \text{ if I hadn't finished my PhD} \end{cases}$$

where i indexes an individual observation, such as a person

Treatments can be continuous too (e.g., minimum wages, vaccinations, prices), but we're sticking to binary treatments

Potential outcomes notation

Potential outcomes are hypothetical outcomes under different states of the world:

$$Y_{i,t}^{j} = \begin{cases} 1 \text{ happiness at time } t \text{ if I had finished my PhD} \\ 0 \text{ happines at time } t \text{ if I had not finished my PhD} \end{cases}$$

where j indexes a potential state of the world

I'll drop t subscript, but just remember - this is for the same person and at the same moment in time

Important definitions

Definition 1: Individual treatment effect

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

The causal effect of my PhD on my happiness is $Y_i^1 - Y_i^0$ and since I do have a PhD, I don't observe the second term and so *can't be sure*

Important definitions

Definition 2: Average treatment effect (ATE)

The average treatment effect is the population average of all i individual treatment effects

$$\begin{aligned} E[\delta_i] &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

This is average treatment effect is based on everyone, but since we cannot calculate $E[Y_i^0]$ for PhDs, and cannot calculate $[Y_i^1]$ for non-PhDs, we can't calculate the ATE

Conditional Average Treatment Effects

Definition 3: 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 the average treatment effect for our treatment group (PhDs), but we since we cannot calculate $E[Y_i^0]$ for PhDs, we cannot calculate the ATT

Important definitions

Definition 4: Switching equation

An individual's observed happiness outcomes, Y, is determined by PhD assignment, D_i, and corresponding potential outcomes:

$$
\begin{aligned}
Y_i &= D_i Y_i^1 + (1 - D_i) Y_i^0 \\
Y_i &= \begin{cases} Y_i^1 \text{ if } D_i = 1 \\ Y_i^0 \text{ if } D_i = 0 \end{cases}
\end{aligned}
$$

We don't observe Y^1 (hypotheticals). We observe Y (data) by the switching equation. Big difference.

Our challenge

Definition 5: Fundamental problem of causal inference

Since you need both potential outcomes to know causal effects, then since it is impossible to observe both Y_i^1 and Y_i^0 for the same individual, δ_i, is *unknowable*.

Chinese proverb: A farmer and his son had a beloved stallion who helped the family earn a living. One day, the horse ran away and their neighbors exclaimed, “Your horse ran away, what terrible luck!” The farmer replied, “Maybe so, maybe not. Who knows.”

Individual treatment effects are **unknowable**. Our aims are more modest than that. We *estimate* average causal effects using *groups of data*, assumptions and appropriate statistical models

DiD equation

I call this the DiD equation, but Goodman-Bacon calls it the “2x2”

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

k index people with PhDs, U index people without PhDs, $Post$ is after k individuals got their PhD, Pre before k group had gotten their PhDs (baseline), and $E[y]$ mean happiness.

“Pre” and “Post” refer to when our treatment group, k, was treated and thus is the same for both k and U groups

Potential outcomes and the switching equation

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

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

Parallel trends bias

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

Identification

Parallel trends

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

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

"The evolution of happiness for PhDs *had they not gotten their PhDs* is the same as the evolution of happiness for those who never got their PhDs". Nontrivial assumption.

OLS Specification

OLS specification gives *the same answer* as the DiD equation, or 2x2, from earlier with some advantages (like including multiple time periods and easy calculation of standard errors)

$$Y_{it} = \alpha + \gamma D_k + \lambda Post_t + \delta(D_k \times Post_t) + \varepsilon_{it}$$

If parallel trends holds, then $\widehat{\delta}_{OLS} = \delta$, which is the ATT.

See Heckman, et al. 1997; Abadie 2005; Sant'Anna and Zhao 2020 for handling covariates and why TWFE *may* be biased

Roadmap

Basic DiD
- Welcome to DiD
- Definitions
- Identification

Differential timing
- Bacon decomposition
- Static specification
- Event studies
- Imputation

Conclusion

Differential timing

- Previous setup had been relatively simple design with only one treatment date (so one treated group)
- More common design uses multiple treatment groups treated at different points in time
- Most common way researchers estimated the ATT was with TWFE

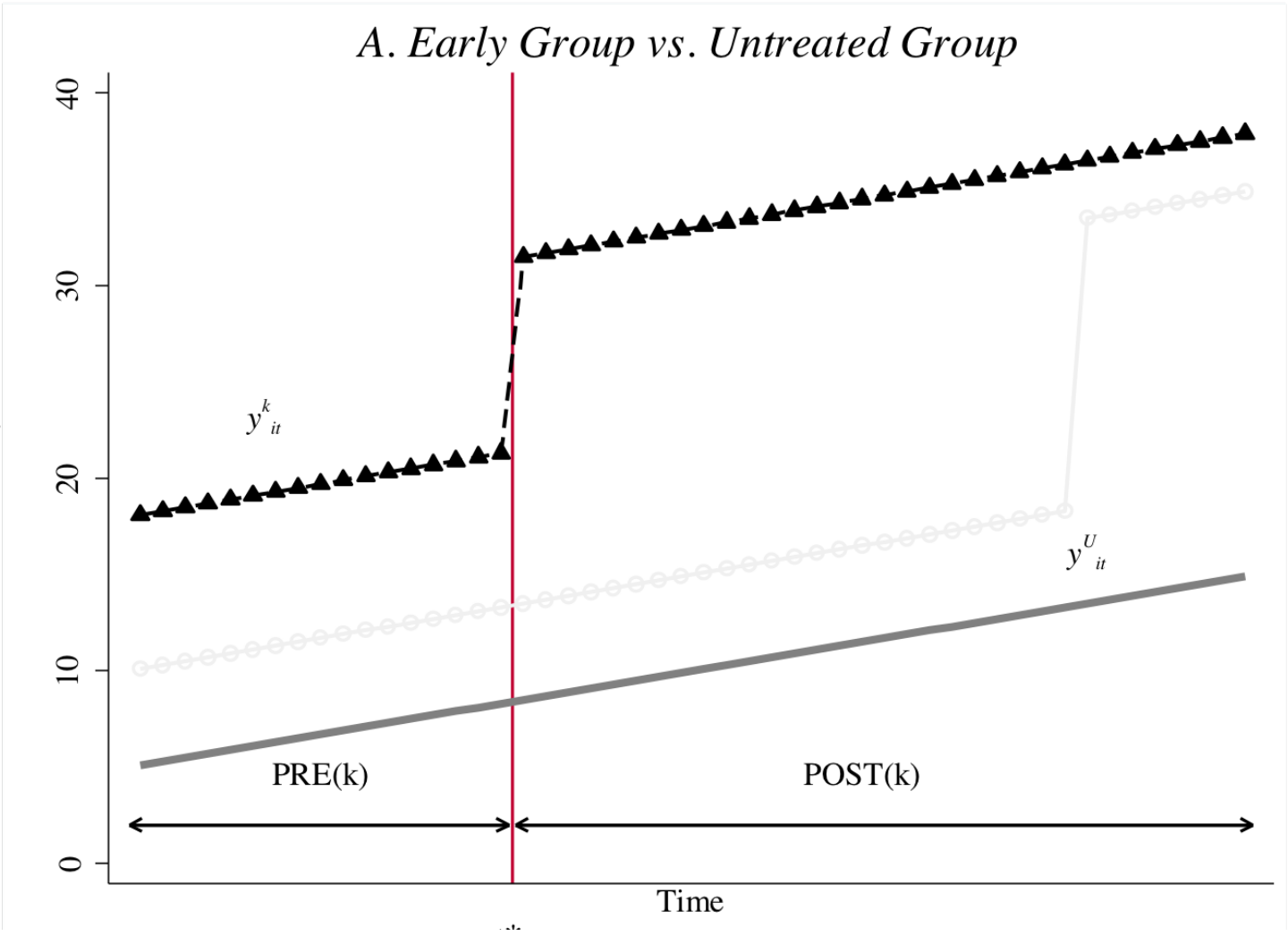
TWFE decomposition

- Two types of TWFE decompositions
  1. **Numerical decomposition of TWFE coefficient**: what adds up to TWFE coefficient?
  2. **Theoretical decomposition**: what does TWFE coefficient "mean"?
- Theoretical decompositions show negative weights on treatment effects, but numerical decompositions show positive weights
- Many authors theoretically decompose TWFE (de Chaisemartin and d'Haultfoeille 2020 for instance) but Goodman-Bacon does both

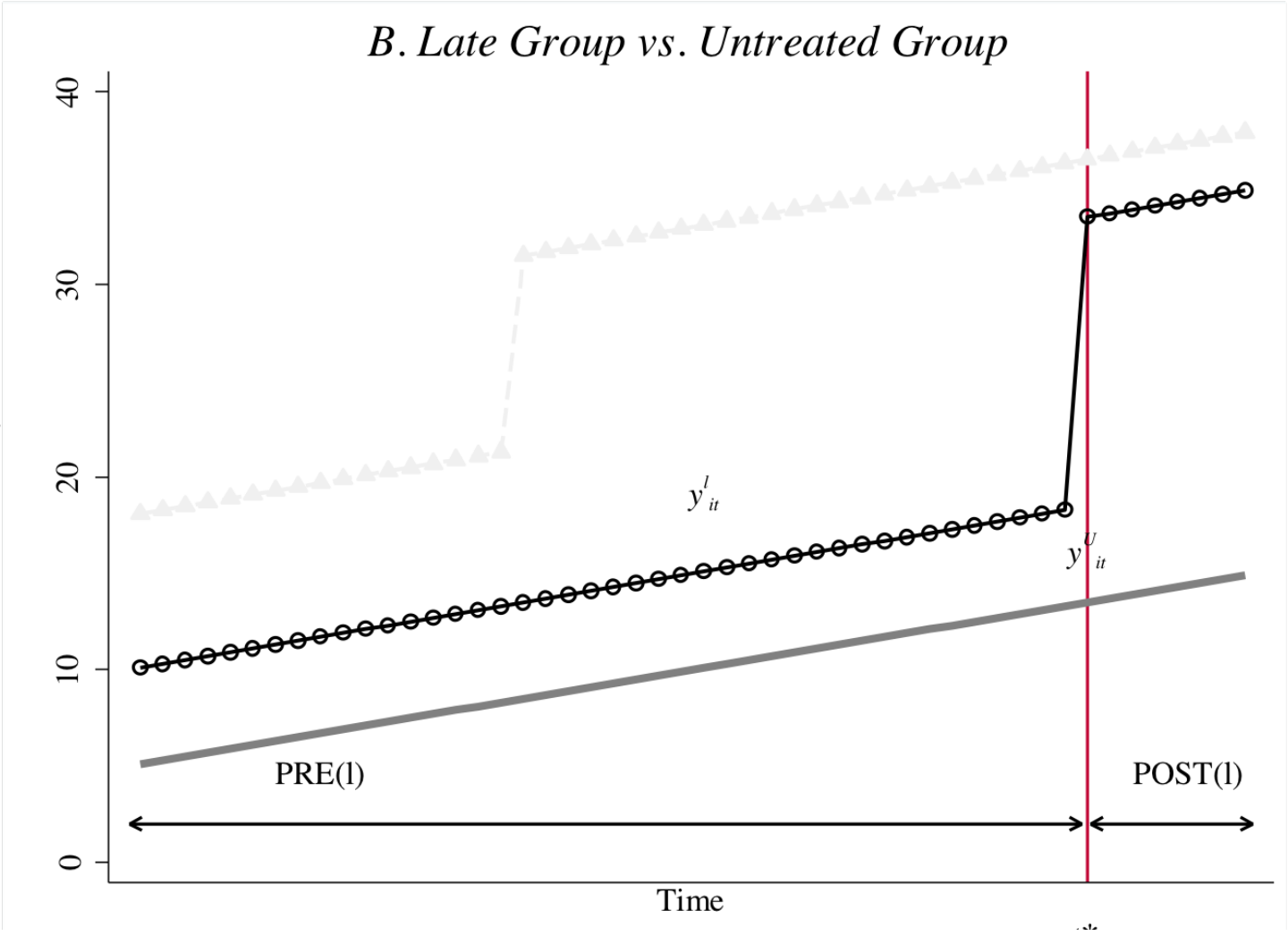
Terms and notation

- Now there will be two treatment groups (k,l) and one untreated group (U)
- k,l are defined by their treatment date with k receiving their treatment earlier than l
- Weights, s_{jb}, are based on variance of treatment and group size
- 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

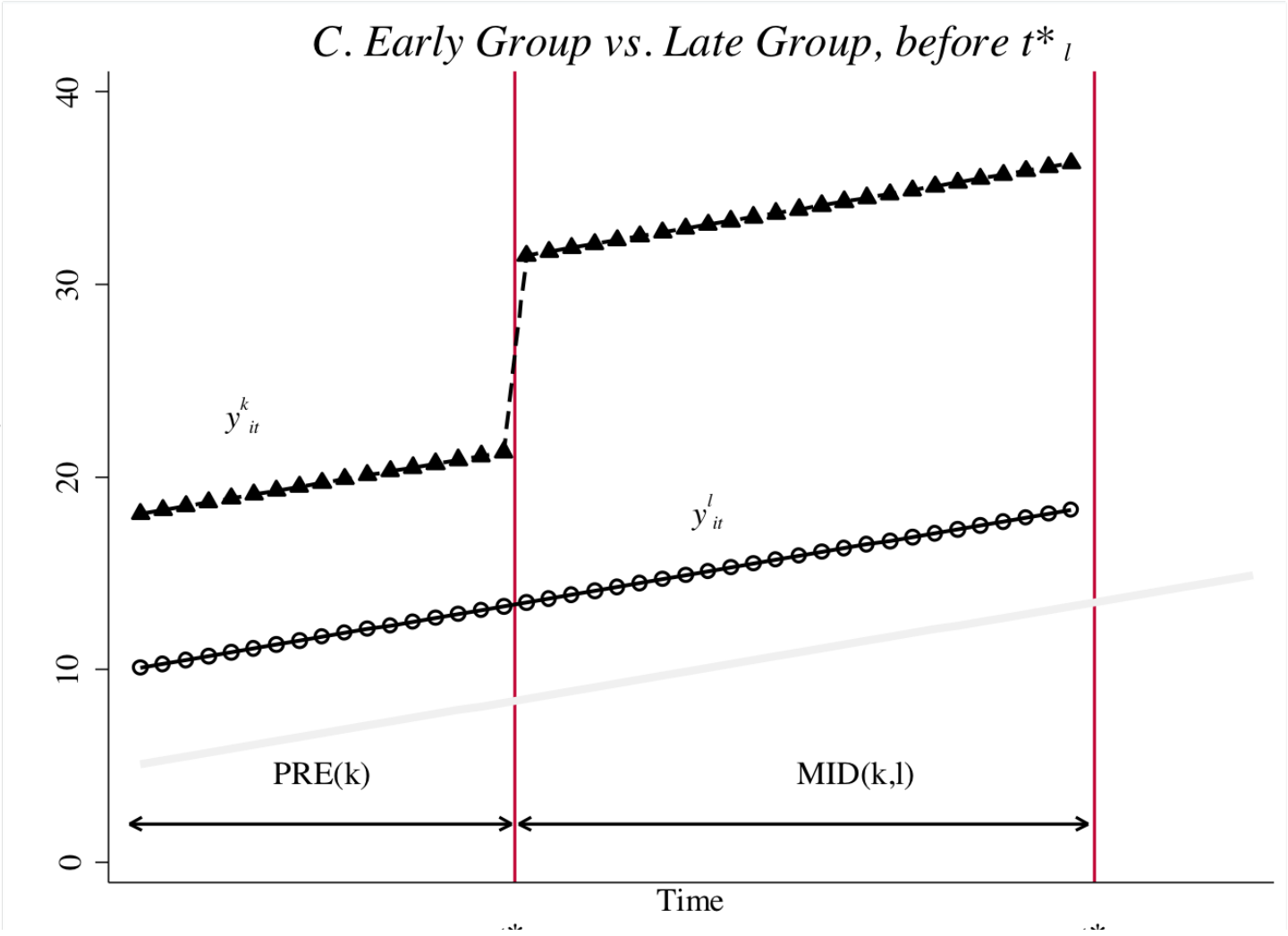
$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)} \right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)} \right)$$



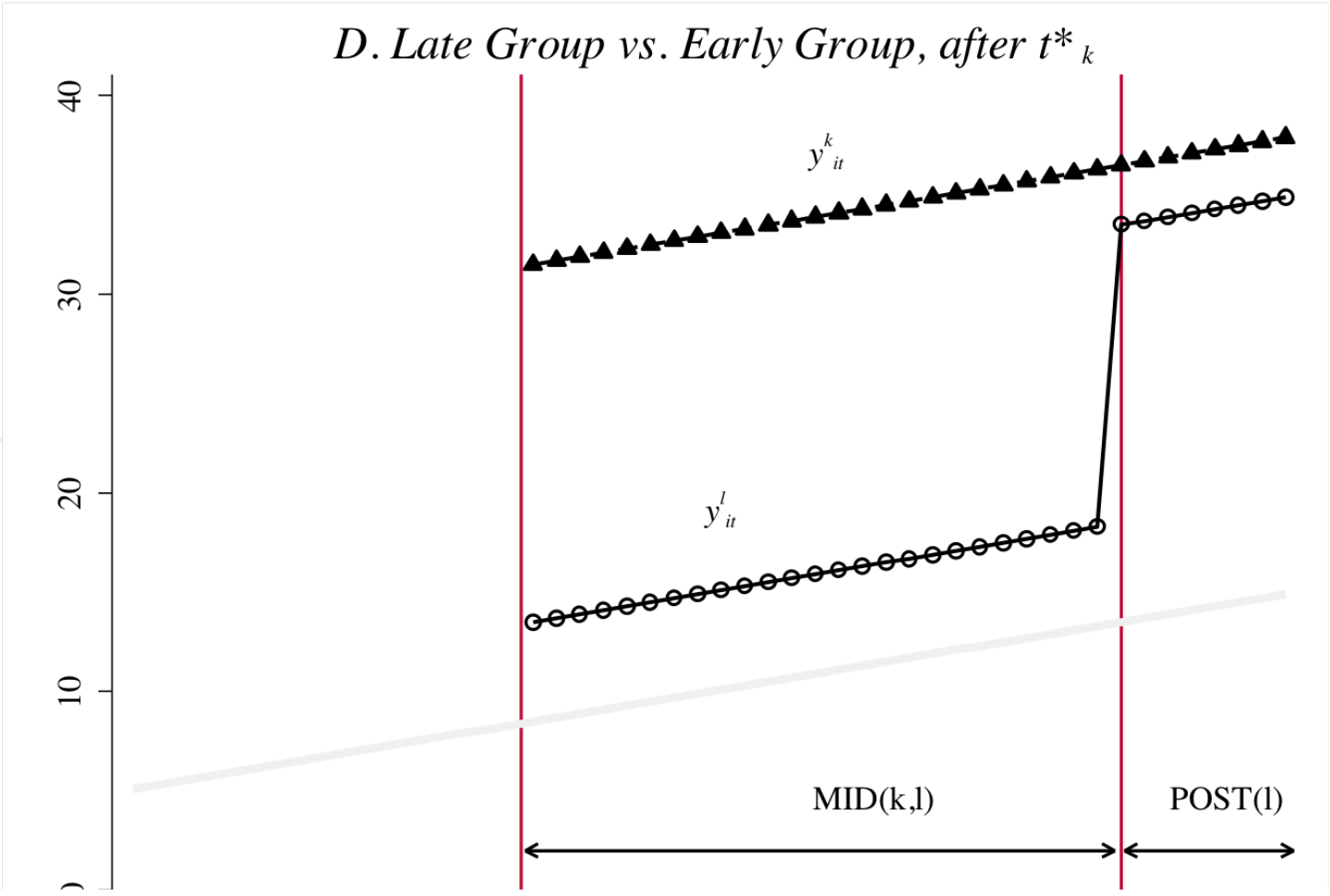
$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_{l}^{post(l)} - \overline{y}_{l}^{pre(l)}\right) - \left(\overline{y}_{U}^{post(l)} - \overline{y}_{U}^{pre(l)}\right)$$



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



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



Bacon's numerical decomposition of TWFE coefficient

Bacon decomposition

TWFE estimate yields a weighted combination of each groups' respective 2x2 (of which there are 4 in this example)

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

Variance weights, s, are **positive**, sum to one and are "strange" in that if you change your dataset's start and/or stop dates, then the weights change and so does the coefficient *regardless of treatment effects*

Never-treated and not-yet-treated 2x2s

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

The top one is comparing a treatment group k or l to a never-treated group, but the bottom one to a not yet treated.

Already-treated 2x2

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

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

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!
- Let's look at each of these three parts more closely

Simulated data

- 1000 firms, 40 states, 25 firms per states, 1980 to 2009 or 30 years, 30,000 observations, four groups
- $E[Y^0]$ satisfies “strong parallel trends” (stronger than necessary)

$$Y_{ist}^0 = \alpha_i + \gamma_t + \varepsilon_{ist}$$

- Dynamic treatment effects (next slide)

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 across groups
- Dynamic treatment effects
- Staggered rollout (all treated)
- ATT is $+82$ (equally weighted positive treatment effects)

Estimation

Recall data generating process guaranteed “parallel trends” and “no anticipation” but not homogenous treatment effects

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!

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

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

Large weight on the “late to early 2x2” is *suggestive* of an issue

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			

- Callaway and Sant'Anna (2020) is unbiased even with heterogeneity and dynamics
- Group-time ATT target

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

- Estimate each ATT(g,t), then calculate ATT through a weighted average

Question: What weight did I use?

Assumptions

1. Panel or repeated cross section data (modularity)
2. Conditional parallel trends
3. Common support
4. No anticipation (zero treatment effects before treatment)
5. Irreversible treatment

CS Estimator (the IPW version)

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

CS only uses never or not-yet treated as controls C – not the already-treated as controls (done through subsetting the data).

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

	(Truth)	(TWFE)	(CS)	(SA)	(BJS)
$\widehat{Feasible\ ATT}$	68.33	-6.69***	68.34***		

Event studies

- Randomization gives us confidence because “we know how the science works” – Don Rubin
- DiD identifies the ATT using parallel trends, but there is no science of parallel trends, so the bar is higher
- Main “test” is to examine the pre-trends and check if their changes over time are the same as the comparison group
- Historically, people estimated with TWFE but Sun and Abraham (2020) showed it was biased under differential timing and heterogenous treatment effects

Notation and terms

- When treatment occurs at the same time, we say they are part of the same cohort, e
- If we bin the data, then a lead or lag l will appear in the bin g so sometimes they use g instead of l or $l \in g$
- Building block is the "cohort-specific ATT" or $CATT_{e,l}$ which was each cell in the simulation data
- We want to estimate $CATT_{e,l}$ with a regression

Notation and terms

- 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

Assumptions

1. Parallel trends
2. No anticipation
3. Treatment effect homogeneity

TWFE will be unbiased estimate of each population regression coefficient lead and lag

TWFE Regression

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

We estimate this μ_g population regression coefficient leads and lags using TWFE and get $\widehat{\mu_g}$.

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

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, $\sum_{l \in g^{excl}} \sum_e w^g_{e,l} = -1$

Intuition for contamination

- Each population regression coefficient is the sum of three things (one good, two bad)
  1. CATT from that period
  2. CATT from the omitted period
  3. CATT from all other relative periods
- When all three assumptions hold, only the lead/lag's CATT remains (all others vanish)
- This vanishing of other period leads and lag CATT happens through bc CATT=0 (no anticipation on pre-treatment), or through the weighting scheme (like with homogenous treatment effects)

Simple example: A balanced panel $T = 2$ with cohorts $E_i \in \{1, 2\}$. We drop two relative time periods to avoid multicollinearity, so we will include bins $\{-2, 0\}$ and drop $\{-1, 1\}$.

Toy example

Second pre-treatment lead estimated with TWFE

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

- Parallel trends gets us to all of the $CATT$
- No anticipation makes $CATT = 0$ for all pre-periods
- Homogeneity cancels second and third terms (via weighting)
- Leaves $\frac{1}{2}CATT_{1,1}$ because you dropped a post-treatment period with a non-zero CATT

Interaction-weighted estimator

- They propose a 3-step interacted weighted estimator (IW) as a consistent estimator for μ_g
- It's just like CS only instead of using the "not-yet-treated" as controls, it uses the "last treated" as controls
- TWFE regression specification that interacts relative period indicators with cohort/group indicators, excluding indicators for never-treated (last treated) cohorts

IW estimator

Take a weighted average of estimates for $CATT_{e,l}$ from Step 1 with weight estimates from step 2 using last cohort as comparison (dropping already treated)

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

Reporting results

Table: Estimating ATT

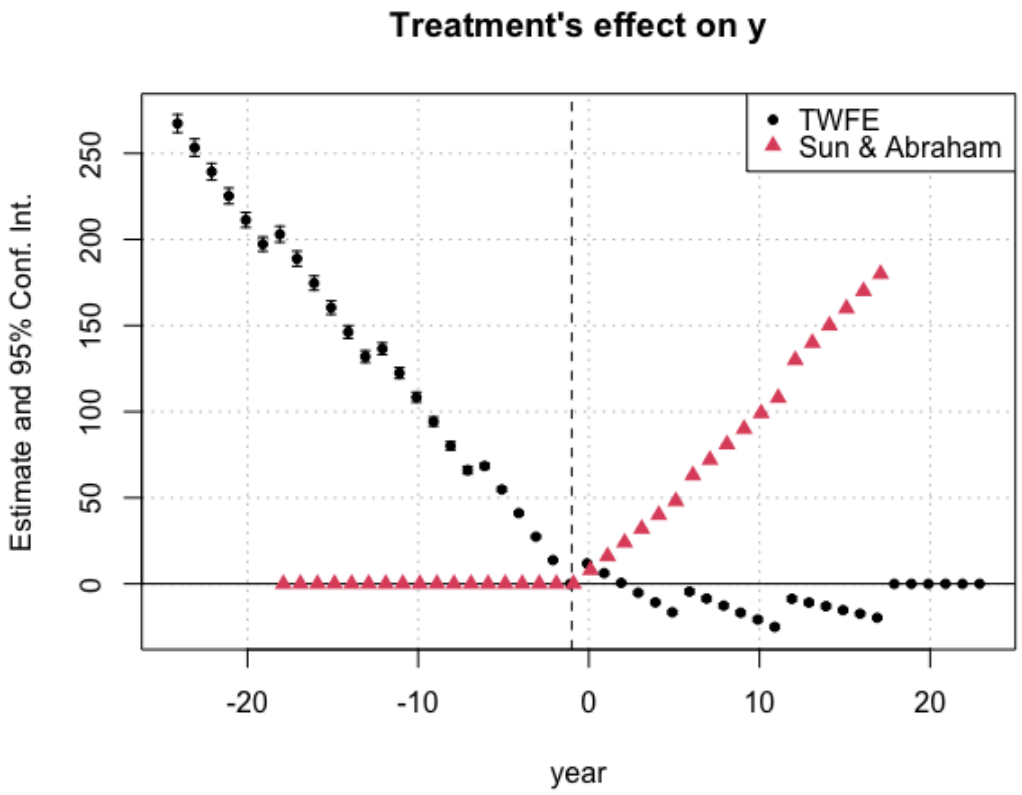
	(Truth)	(TWFE)	(CS)	(SA)	(BJS)
$\widehat{Feasible\ ATT}$	68.33	-6.69***	68.34***	68.33***	

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				

Question: Who is control for CS vs SA?. Why do the leads and lags become “imbalanced” in event time?

Comparing TWFE and SA



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

Imputation methods

> *"At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others."*
> *– Imbens and Rubin (2015)*

There's three imputation models (four if you count matrix completion with nuclear norm regularization)

Imputation methods

All recent working papers

1. **2SDiD** (Gardner 2021) – imputes Y^0 using estimated fixed effects from the $D = 0$ units, residualizing into $\widehat{Y}$, regressing new $\widehat{Y}$ using GMM
2. **Robust efficient imputation** (Borusyak, et al. 2021) – very similar to 2SDID in that you impute Y^0 using $D = 0$ sample and estimated fixed effects
3. **Mundlak** (Wooldridge 2022) – TWFE with saturated interactions, is equivalent to the above two

Steps for BJS

Target parameter is individual treatment effect, δ_i

1. Estimate expected potential outcomes using OLS and only the untreated observations (this is similar to Gardner 2021)
2. Then calculate $\widehat{\delta}_{it} = Y_{it}^1 - \widehat{Y}_{it}^0$
3. Then estimate target parameters as weighted sums

$$\widehat{\delta}_W = \sum_{it} w_{it} \widehat{\delta}_{it}$$

Why is this working?

- Because we can obtain consistent estimates of the fixed effects, we can extrapolate to the counterfactual units for all $Y(0)_{it}$ that were treated
- This is the same type of trick we see with Heckman, et al. (1997) as well as Gardner (2021)
- As it is still OLS, it's computationally fast and flexible to unit-trends, triple diff, covariates and so forth (with caveats about time-varying covariates requiring more assumptions)
- Wooldridge shows the Mundak estimator maps onto BJS robust model

Reporting results

Table: Estimating ATT

	(Truth)	(TWFE)	(CS)	(SA)	(BJS)
$\widehat{Feasible\ ATT}$	68.33	-6.69***	68.34***	68.33***	68.33***

Software

1. Callaway and Sant'anna
   - → **Stata**: csdid
   - → **R**: did
2. Sun and Abraham
   - → **Stata**: eventstudyinteract
   - → **R**: fixest with subab() option
3. Borusyak, et al. (2022)
   - → **Stata**: dd_imputation
   - → **R**: didimputation

Roadmap

Basic DiD
- Welcome to DiD
- Definitions
- Identification

Differential timing
- Bacon decomposition
- Static specification
- Event studies
- Imputation

Conclusion

More models

- Continuous treatments (Callaway, Goodman-Bacon and Sant'anna 2022)
- Time varying controls (Cattaneo, et al. 2022)
- Reversible treatment (de Chaisemartin and d'Haultfoeille 2018)
- Fuzzy borders between treated and control (de Chaisemartin and d'Haultfoeille 2017)
- Geographic difference-in-differences (Butts 2022)

My two cents

Good news:

- Differential timing is very common, TWFE was historically preferred, robust solutions have surpassed it
- Results are very similar to one another when the parallel trends assumptions encompass one another
- Confidence intervals can differ
- Advice: Don’t stress because results are pretty similar across. Just know your assumptions.

My two cents

Going forward

- Define the target parameter and use the estimator that is consistent to get that parameter (as opposed to a weird variance weighting that is biased)
- Read when the benefit outweighs the cost if you're having "DiD fatigue" (for instance I think continuous treatments might fit that)
- Be glad it's not 3 years ago anymore – some of us all really stressed out

Thank you!