Causal Inference II

Mixtape Session

Roadmap

- Differential timing
  - Introduction
  - TWFE Estimator
  - Applications
- TWFE Pathologies
  - Potential outcomes
  - Bacon decomposition
  - Simulation
- Aggregating building blocks
  - CS
  - SA
  - dCH
- Short-gap vs Long-Difference Calculations in Event Studies

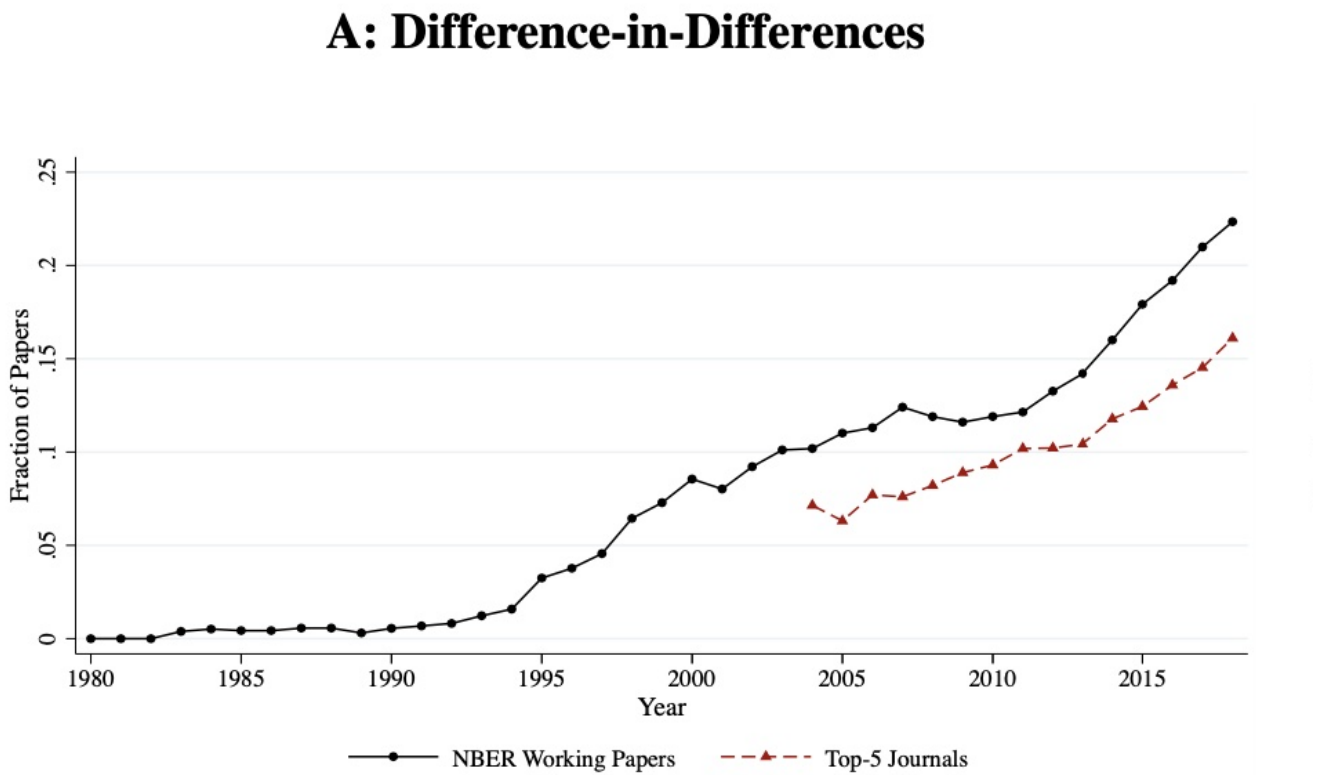
Differential timing outline

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

Diff-in-diff is a dominant design

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

Difference-in-differences credibility crisis

- Many simultaneous discoveries, some redundancies, and **sudden** awareness of the issues started happening around 2017
- Extreme meteoric rise, unusual for econometrics, has become very influential and expectations have changed
- Important to learn this material given the popularity of diff-in-diff (at minimum for your readings)
- Many survey articles out there, so this has to be selective

Differential Timing

- Differential timing refers to when groups of panel units (e.g., states, countries) receive the same treatment at different calendar dates
- Ordinarily, we are interested in knowing something about its effect, but what if there is not single one effect?
- What if the effects of the intervention differ for the earlier groups treated than the late ones, or if if the effects change over time?
- Then the standard twoway fixed effects estimator may be biased
- But first we learn what I mean by the ”standard twoway fixed effects estimator”

Two Regression 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

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

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

- Regression estimates of δ and the "four averages and three subtractions" are the same thing numerically (show code)
- And under a particular assumption, they are also unbiased estimates of an aggregate causal parameter
- But this is in the case of a 2x2 – one group treated in time period t, and another group that isn't.
- What would people do when it wasn't a 2x2? They'd use a different specification of TWFE.

Twoway fixed effects

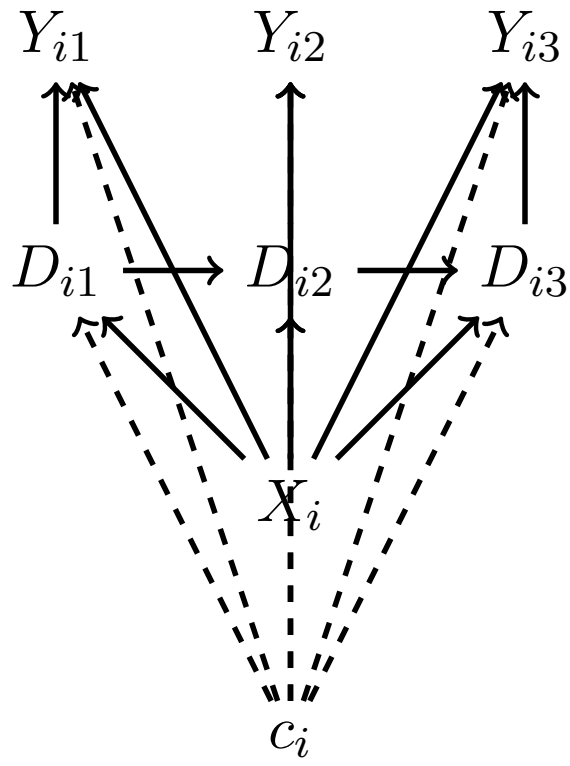
- When working with panel data, the so-called TWFE estimator is the workhorse estimator
- It's easy to implement, handles time-varying treatments, has a relatively straightforward interpretation under constant treatment effects, standard errors are easy to calculate and understand
- Interpretation is 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 not to use it

- Reverse causality: Becker predicted police reduce crime, but when you regress crime onto police, it's usually positive
  - → TWFE requires an assumption that rules out reverse causality
  - → For those, you typically need an instrument
- Time-varying unobserved variables that are correlated with the treatment and the outcome
  - → 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 an outcome (i.e., 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

$$Y_{it} = \beta_0 + \beta(CDL_{it}) + \alpha_1 X_{it} + c_i + u_t + \varepsilon_{it}$$

- In original paper, 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 Negative Binomial 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, means conditional on covariates, untreated potential outcomes on averaged evolved similar to comparison group
- **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

Differential timing
- Introduction
- TWFE Estimator
- Applications

TWFE Pathologies
- Potential outcomes
- Bacon decomposition
- Simulation

Aggregating building blocks
- CS
- SA
- dCH

Short-gap vs Long-Difference Calculations in Event Studies

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

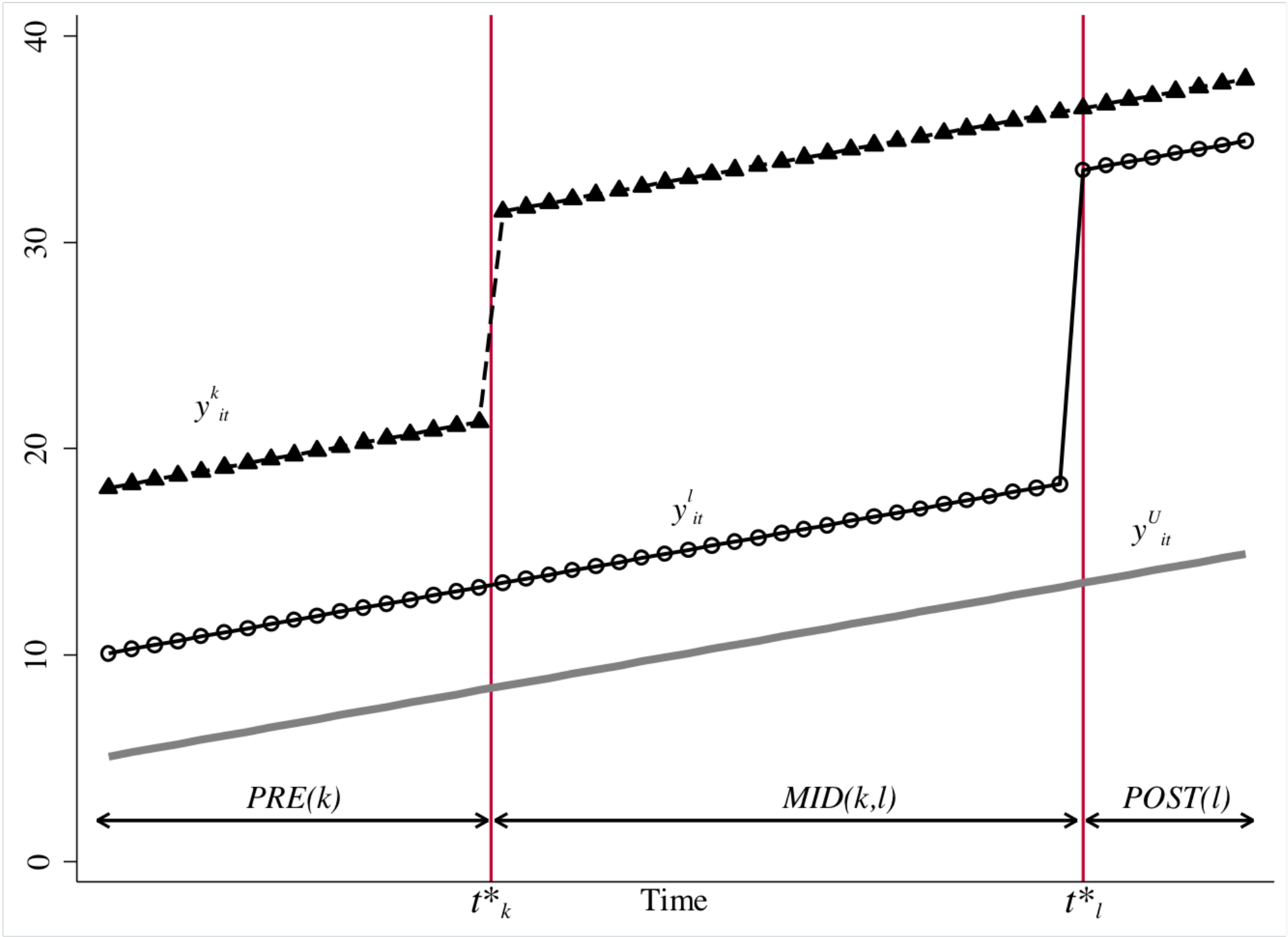
Let's look at 3 timing groups (a, b and c) and one untreated group (U). With 3 timing groups, there are 9 2x2 DDs. Here they are:

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

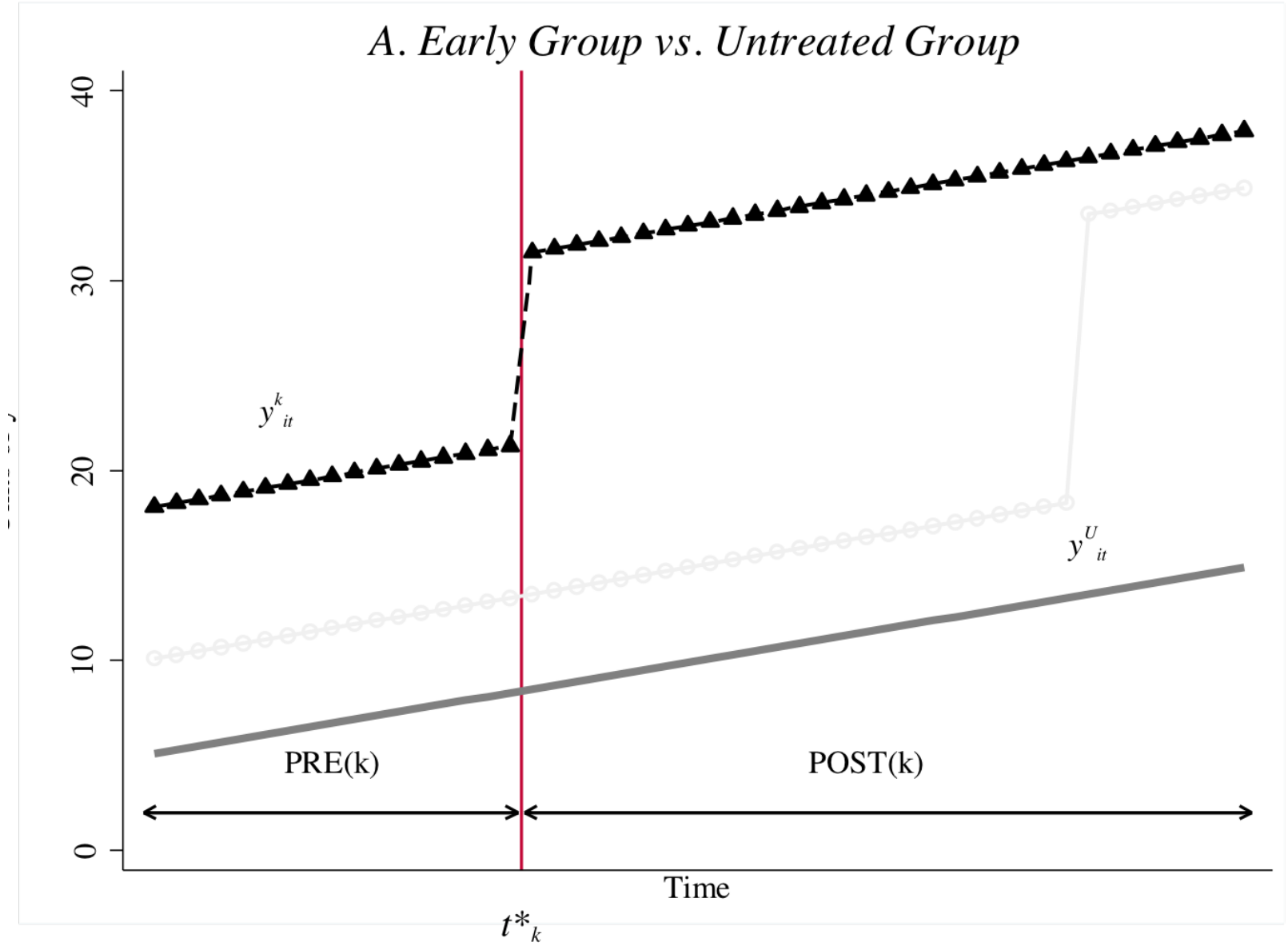
Let's return to a simpler example with only two groups – a k group treated at t_k^* and an l treated at t_l^* plus an never-treated group called the U untreated group

Terms and notation

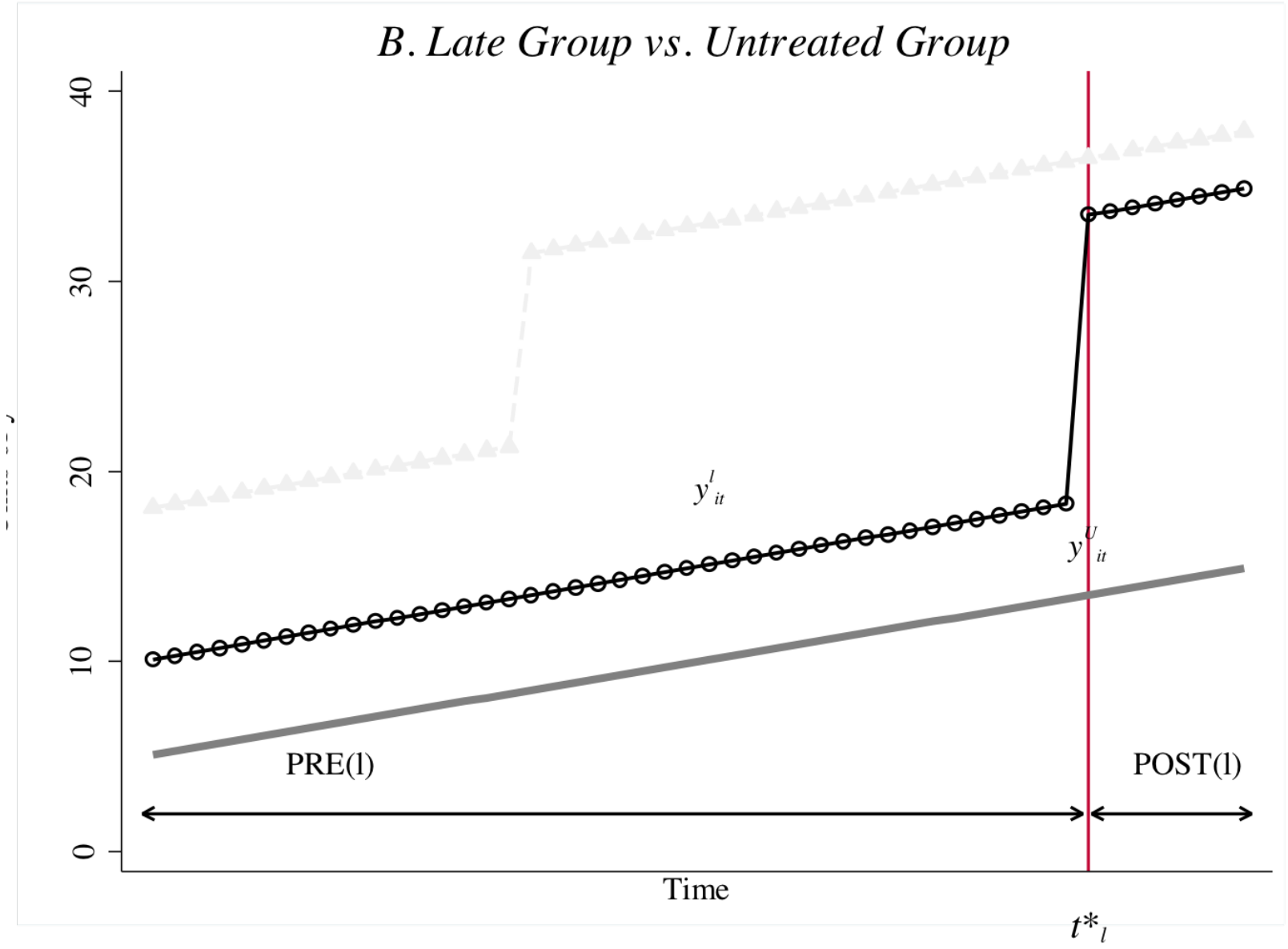
- Let there be two treatment groups (k,l) and one untreated group (U)
- k,l define the groups based on when they receive treatment (differently in time) with k receiving it earlier than l
- Denote $\overline{D}_k$ as the share of time each group spends in treatment status
- Denote $\widehat{\delta}_{jb}^{2x2}$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

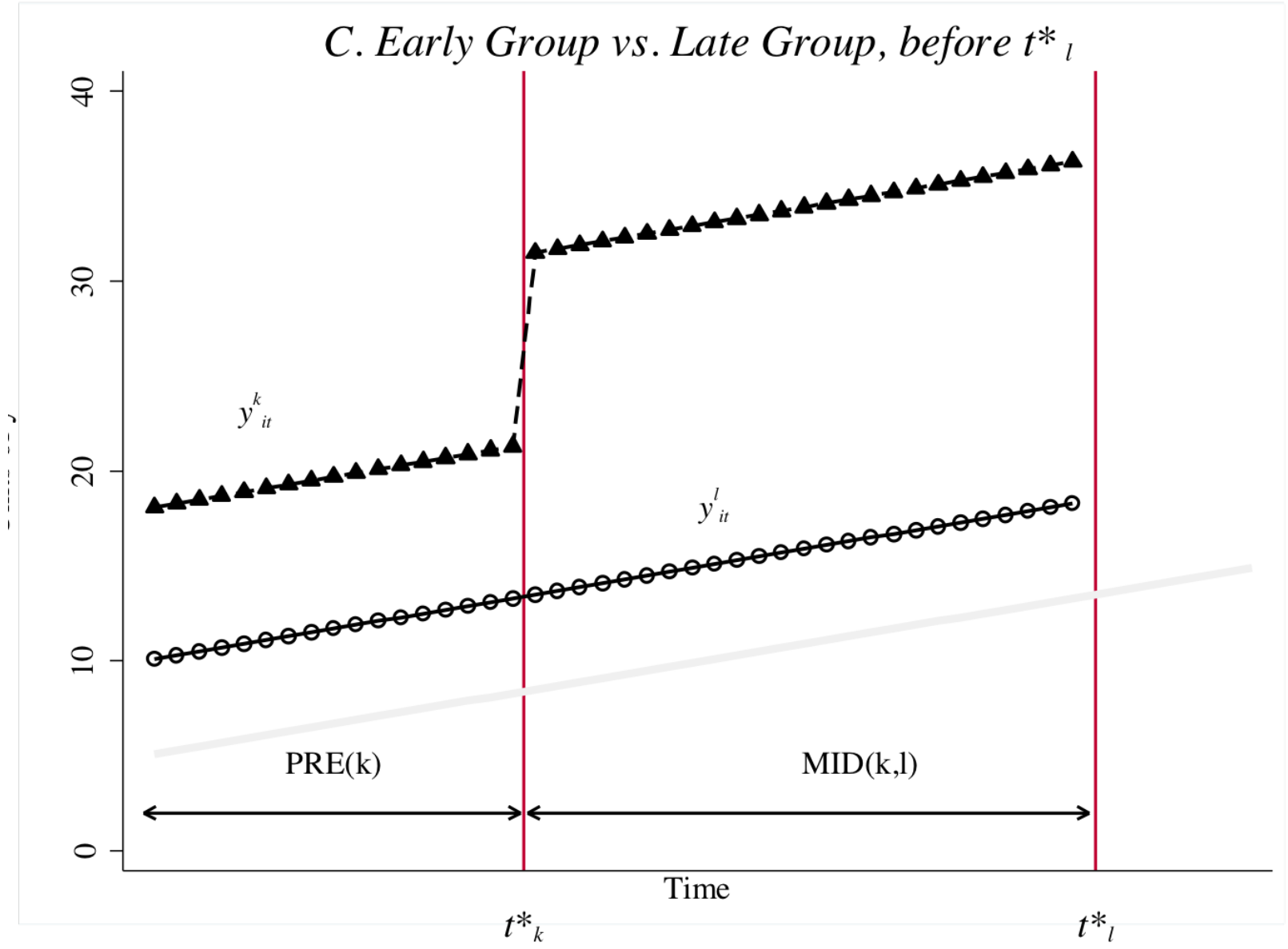
$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)} \right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)} \right)$$



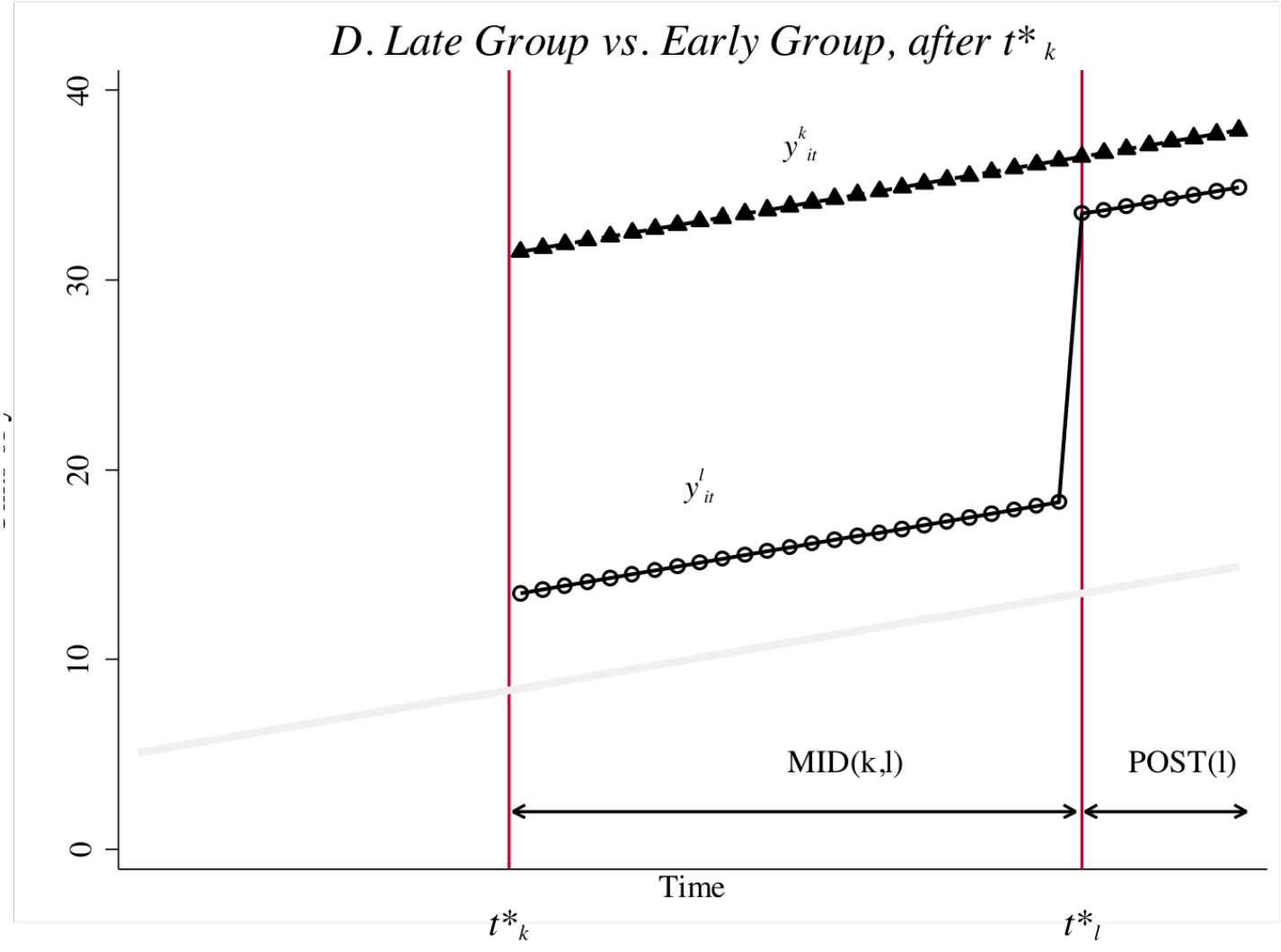
$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_{l}^{post(l)} - \overline{y}_{l}^{pre(l)}\right) - \left(\overline{y}_{U}^{post(l)} - \overline{y}_{U}^{pre(l)}\right)$$



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



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



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 the panel group shares, $\overline{D}_k(1 - \overline{D}_k)$, as well as $(\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l))$ 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

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

where we will make these substitutions

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

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

Differential timing
- Introduction
- TWFE Estimator
- Applications

TWFE Pathologies
- Potential outcomes
- Bacon decomposition
- Simulation

Aggregating building blocks
- CS
- SA
- dCH

Short-gap vs Long-Difference Calculations in Event Studies

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

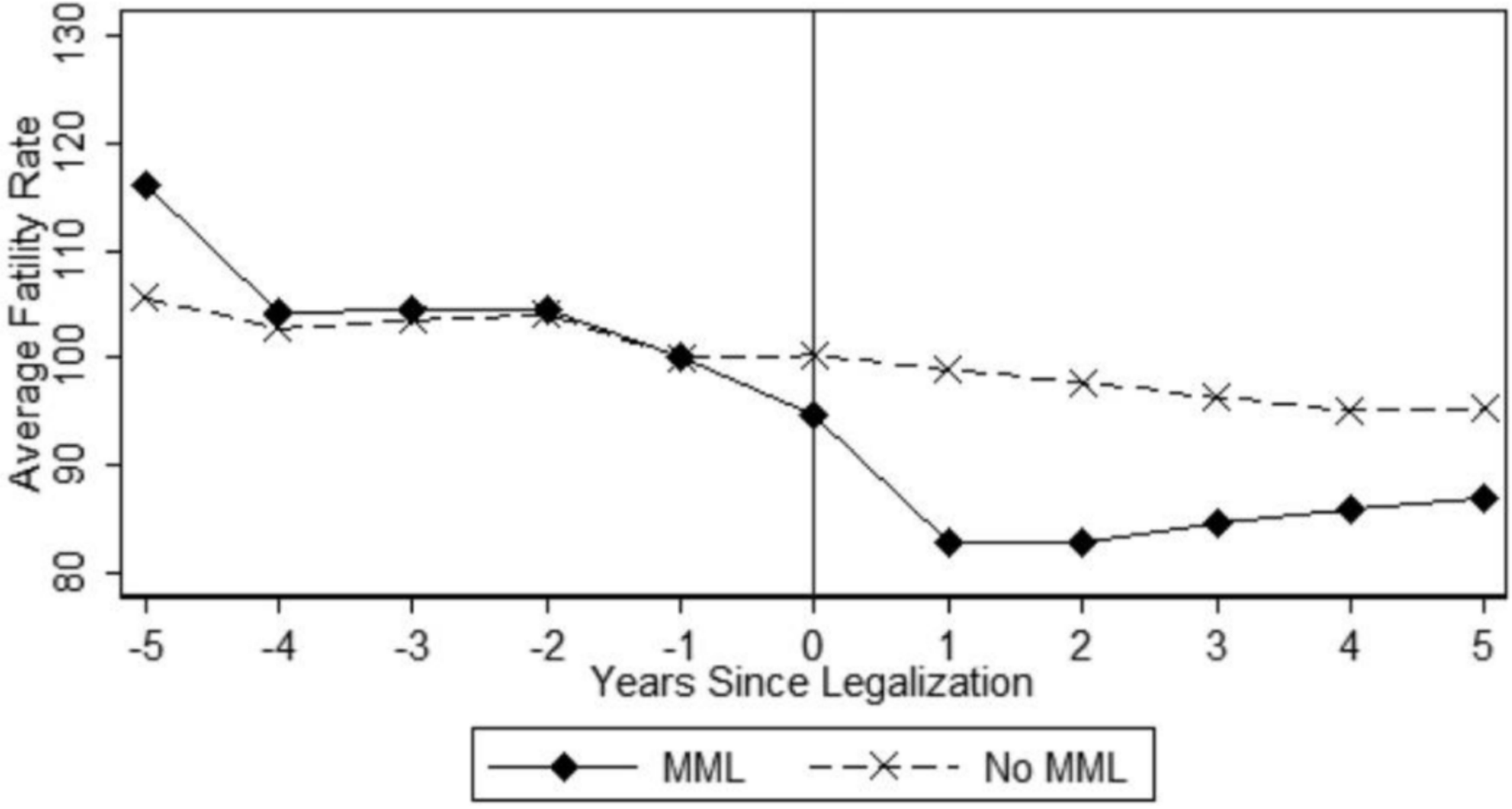


Figure: Anderson, et al. (2013) display of raw traffic fatality rates for re-centered treatment states and control states with randomized treatment dates

Replicated from a project of mine

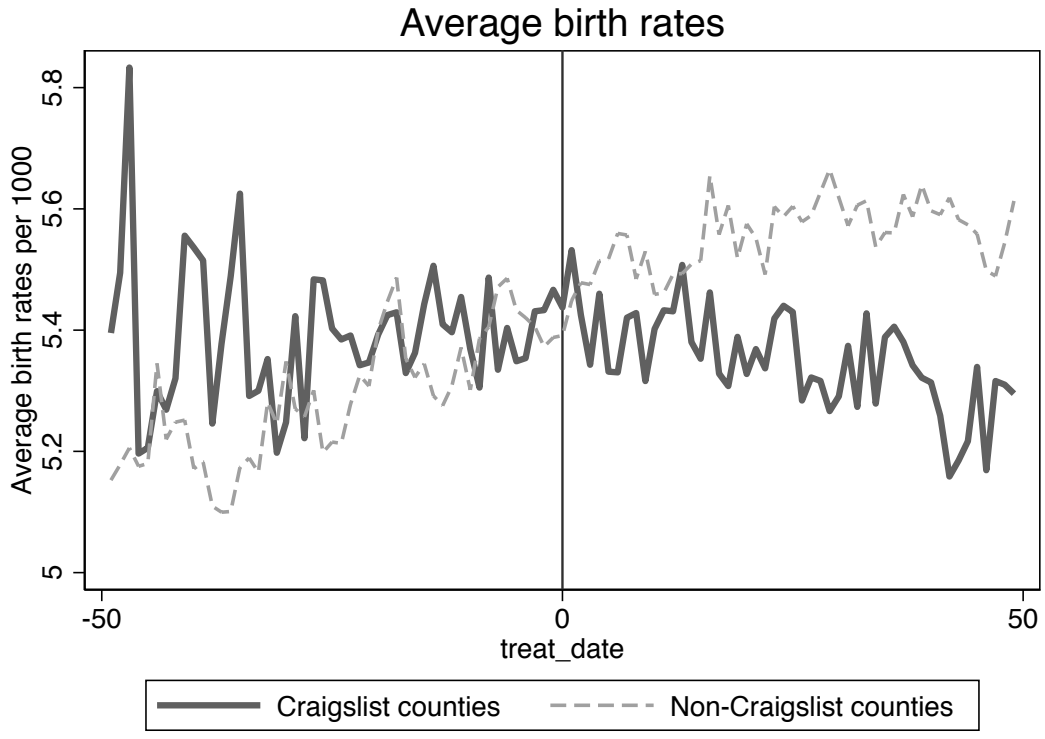
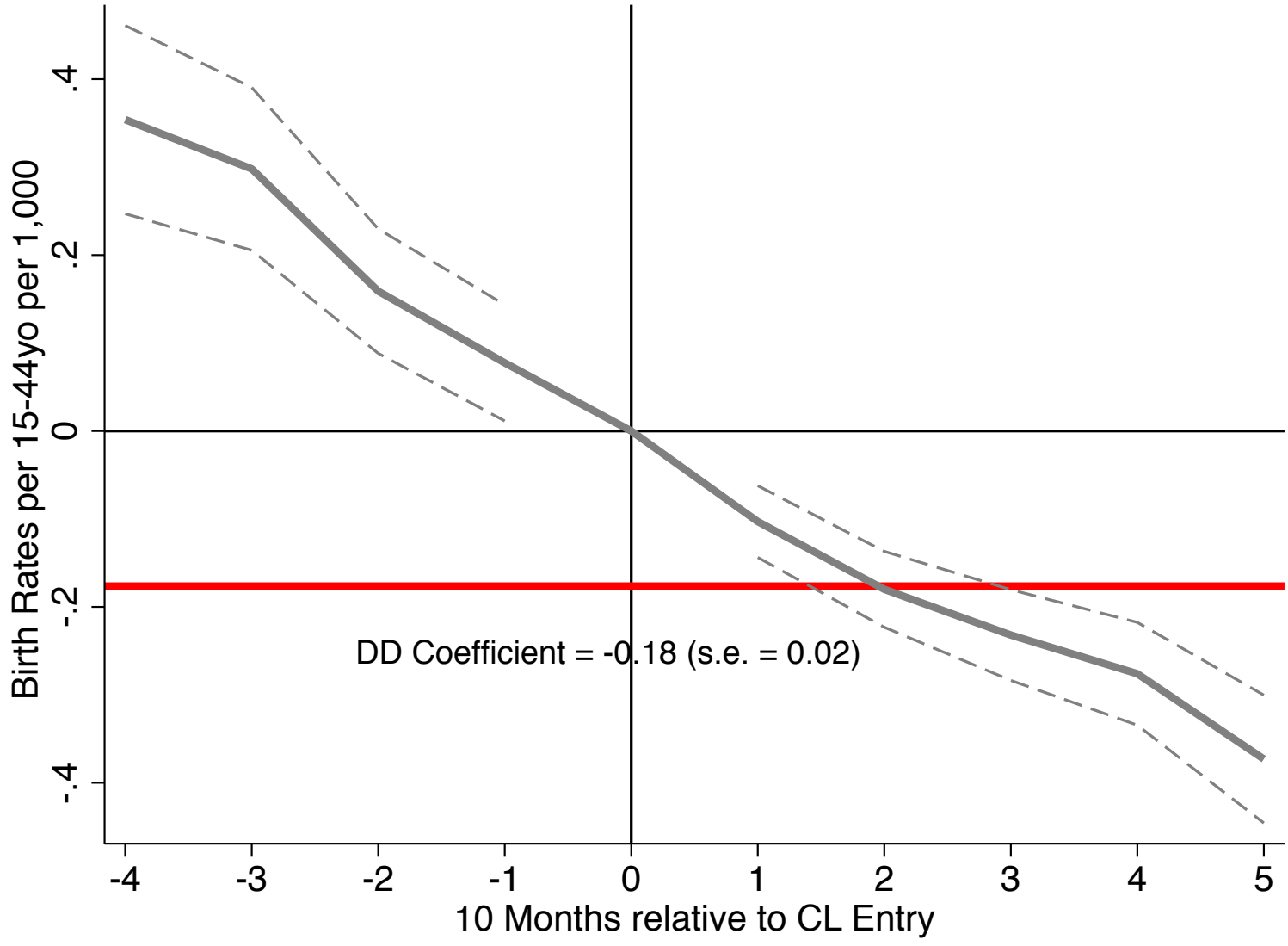


Figure: Roll out of Craigslist "personal ads" for casual intimate encounters and birth rates using the "randomized treatment assignment" approach for visualization

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

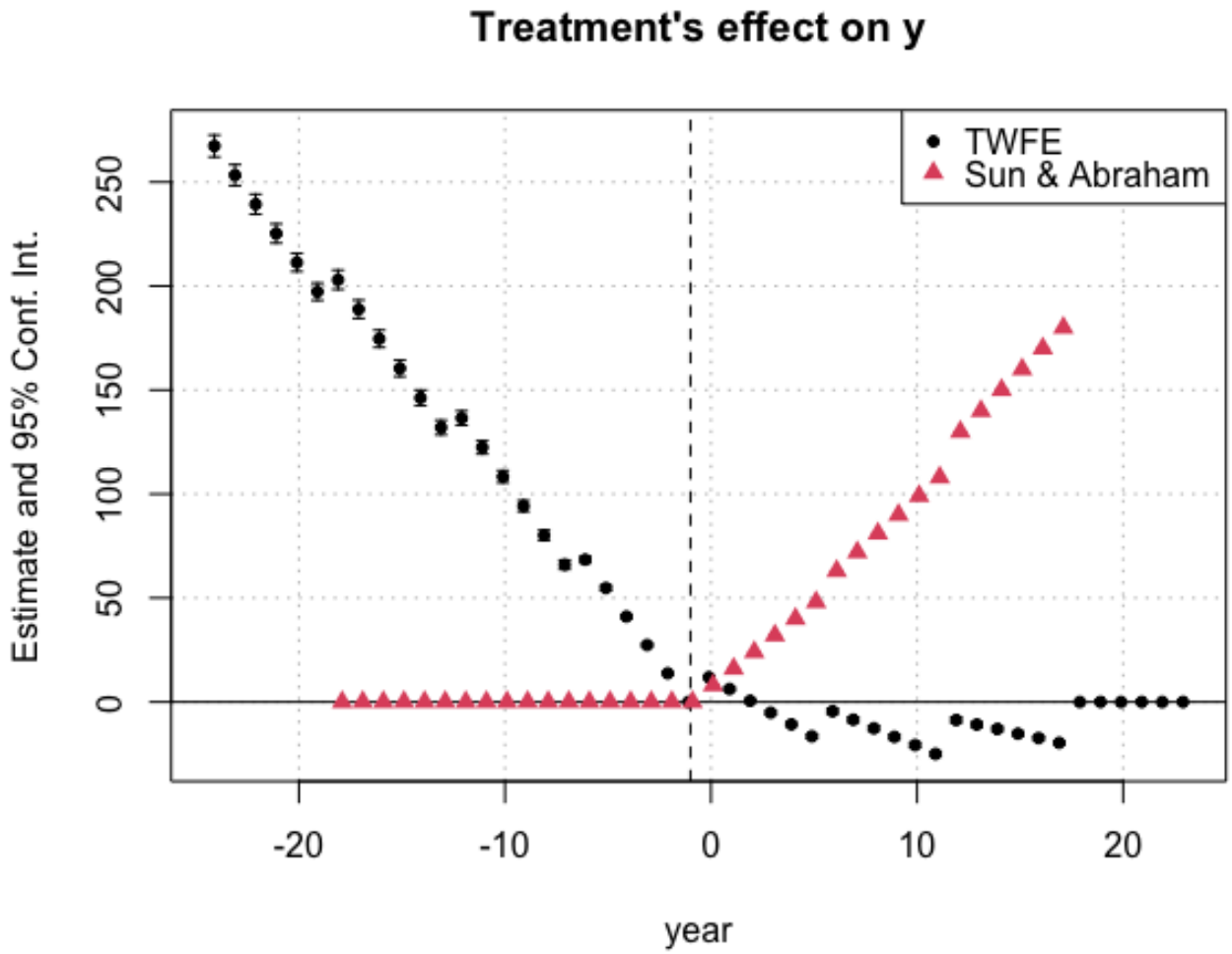
	(Truth)	**(TWFE)**	**(CS)**	**(SA)**	**(BJS)**
$\widehat{Feasible\ ATT}$	68.33	26.81***	68.34***	68.33***	

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
- Their model can handle reversible treatments, but in the context of differential timing is equivalent to CS and SA with a particular choice of weights
- For diagnostic purposes, they recommend reporting the number/fraction of group-time ATTs that receive negative weights, as well as the degree of heterogeneity in treatment effects that would be necessary for the estimated treatment effect to have the "wrong sign"

Roadmap

- Differential timing
  - Introduction
  - TWFE Estimator
  - Applications
- TWFE Pathologies
  - Potential outcomes
  - Bacon decomposition
  - Simulation
- Aggregating building blocks
  - CS
  - SA
  - dCH

Short-gap vs Long-Difference Calculations in Event Studies

Different ways

- In CS and dCDH, there are two ways to calculate the pre-treatment coefficients once you have obtained the ATT(g,t) parameter estimates:
  1. **Short gap**. Uses a "rolling" method in which a new baseline is used for each 2x2 comparison.

$$\begin{aligned}\widehat{\delta}_{t-3} = &(E[Y|D=1,t-3] - E[Y|D=1,t-2]) \\ &- (E[Y|D=0,t-3] - E[Y|D=0,t-2])\end{aligned}$$

  2. **Long difference**. Uses a "universal baseline" with a fixed baseline at $t-1$.

$$\begin{aligned}\widehat{\delta}_{t-3} = &(E[Y|D=1,t-3] - E[Y|D=1,t-1]) \\ &- (E[Y|D=0,t-3] - E[Y|D=0,t-1])\end{aligned}$$

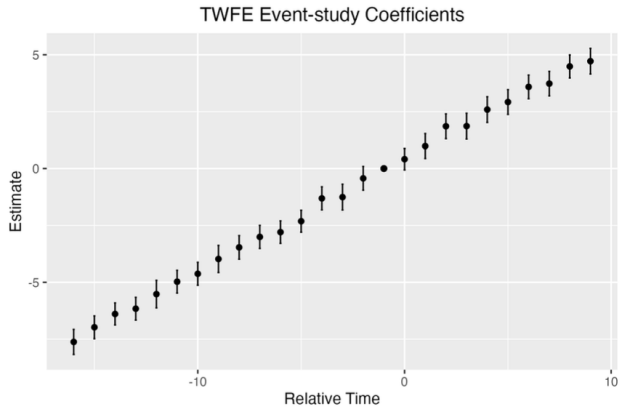
TWFE, Imputation, and CS

TWFE and imputation methods (discussed next) always use a fixed universal baseline. To ensure comparability with CS or dCDH, be sure to use the universal baseline syntax in Stata, R, or Python.

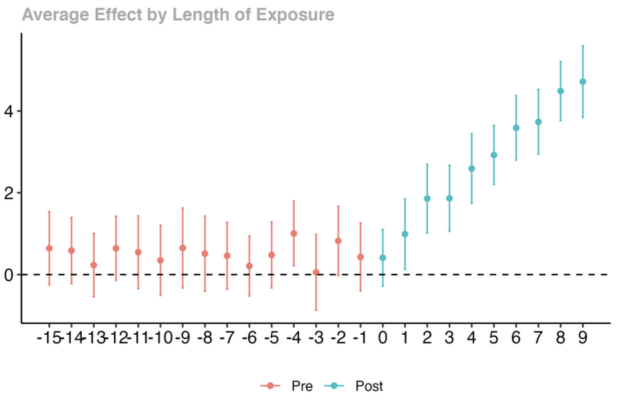
Short Gaps and Long Differences

Figure 1: Comparison of event-study plots in a non-staggered setting

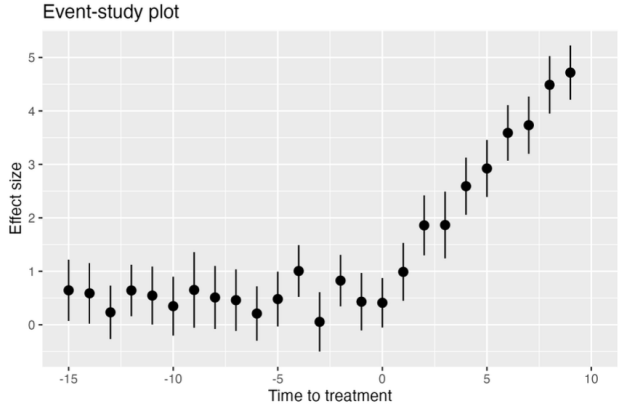
(a) Dynamic TWFE



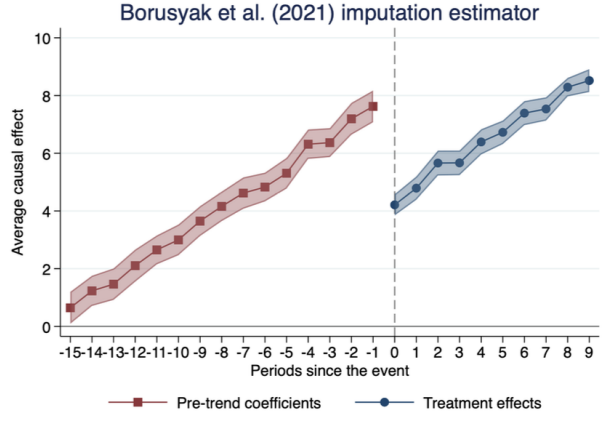
(b) CS



(c) dCDH



(d) BJS



Recommendations

"A natural follow-up question is to what extent the event-study plots for these alternative methods can be modified to be more comparable to conventional event-study plots. For CS/dCDH, there is a straightforward solution, which is to use "long-differences" for the pre-treatment coefficients as well as the post-treatment coefficients (i.e. always use the period before treatment as the baseline). This can be implemented, for example, in the `did` *R and Stata packages using the options* `base_period = \universal"` *and* `long2`*, respectively. Using the default options in the* `did_multiplegt` *Stata package also yields comparisons using "long differences", in contrast to the results for the R package shown above. Using these settings, the event-study estimates are numerically equivalent to the dynamic TWFE specification in the non-staggered setting considered here." - Roth (2024)*

Short vs long2 syntax in `csdid`

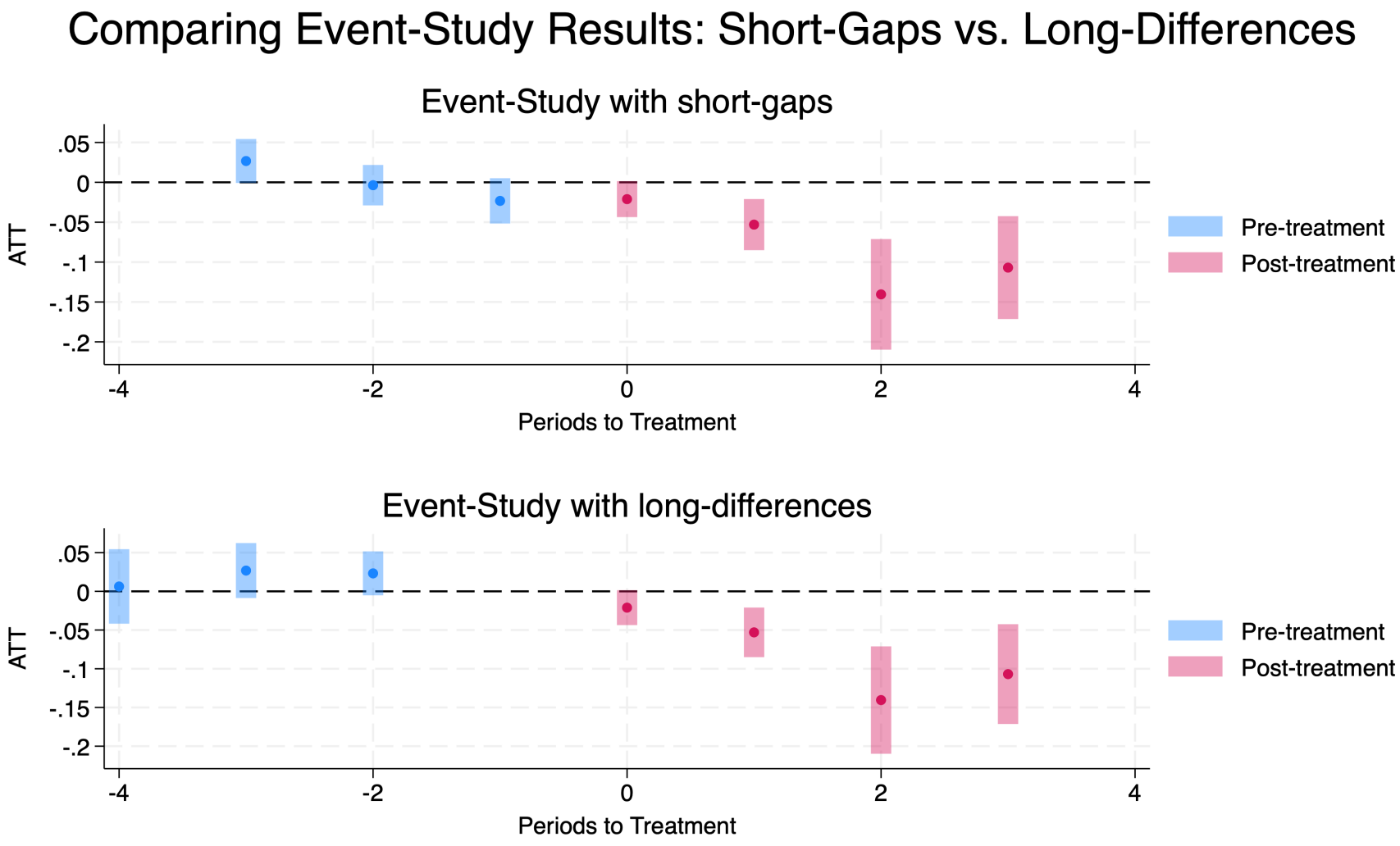
Short-gap method

```
csdid lemp lpop, ivar(countyreal) time(year) gvar(first_treat) method(dripw)
estat event
csdid_plot
```

Long differences method

```
csdid lemp lpop, ivar(countyreal) time(year) gvar(first_treat) method(dripw) long2
estat event
csdid_plot
```

Short Gaps and Long Differences



Warning

- Unfortunately the default syntax for `csdid` and `did` in R uses short-gap
- So you *must* use `long2` or `base_period = universal` to get the event study coefficient that will correctly illustrate the pre-treatment coefficients
- In addition to Roth (2024), Brantly Callaway has a note on his website about it `https://bcallaway11.github.io/posts/event-study-universal-v-varying-base-period`

Conclusion

- The previous methods are fairly comparable, but note, these models all assume parallel trends or conditional parallel trends
- Question is which *comparison group* is more sensible to use – the never-treated or the not-yet-treated?
- There is no single answer to that – the more the treatment was quasi-random, the more the never-treated is appealing
- But maybe the not-yet-treated is better as why didn't the never-treated adopt the treatment?
- Don't get lost in the decisions and forget the importance of *designing*, focus on treatment assignment mechanism, checking imbalance