

Causal Inference II

MIXTAPE SESSION



Roadmap

Background material

Introduction

Three waves of Difference-in-Differences

TWFE Pathologies

Historical links

Potential outcomes

Bacon decomposition

Simulation

Two solutions and a new decomposition

CS

SA

Interesting Applications using staggered adoption: Facebook and Mental Health

Why diff-in-diff and differential timing?

- We will be discussing some of the newer material on difference-in-differences with differential timing
- Fairly rapid advances over last six years in methodologies
- Arguably the most recent wave of updates to our understanding of a technique that we thought was probably pretty straightforward
- Difference-in-differences is as we'll see very straightforward, but regression is less straightforward

Differential timing outline

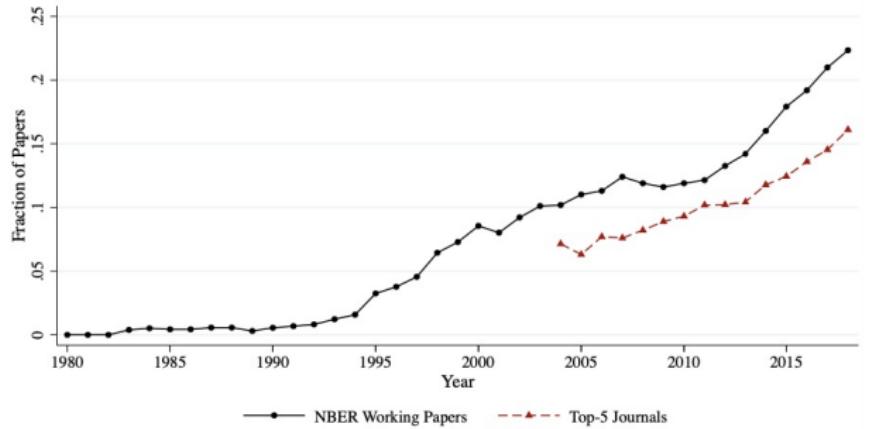
We will cover some of the properties of two way fixed effects (TWFE), some solutions and my personal opinions

1. Brief review of DiD history, potential outcomes and the ATT
2. Difference-in-differences equation ("four averages and three differences") and the parallel trends assumption
3. TWFE Pathologies in static specification
 - Goodman-Bacon decomposition as diagnosis of the problem
 - Callaway and Sant'Anna estimator as a cure
4. TWFE Pathologies in event study specification
 - Sun and Abraham as both a diagnosis and a cure
5. Application, practical advice and code

Diff-in-diff had belonged to the empiricists

Figure: Currie, et al. (2020)

A: Difference-in-Differences



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

First Wave of Diff-in-Diff

- Difference-in-differences has evolved in three waves from 1983 to present in economics
- First wave lasts from 1983 to 2011; second wave from 2011 to 2018; third wave from 2018 to present
- Initially, the pick up was slow, mostly used by labor economists throughout the 1990s in the “program evaluation” area
- No potential outcomes notation, no mention of parallel trends, no event studies – only Ashenfelter’s Dip, evidence of time varying selection bias

Ashenfelter and Card coin the phrase difference-in-differences

Figure: Ashenfelter and Card (1983)

In this case, a simple estimate of the training effect is obtained from a comparison of the change in earnings for the trainees between some pre-training period ($\tau-j$) and the post-training period ($\tau+1$) with the change in earnings for the control group over the same period. This "difference-in-differences" estimator provides an unbiased estimate of the training effect because

$$(2) \quad E(y_{i\tau+1} - y_{i\tau-j} \mid D_{i\tau+1} = 1)$$

$$- E(y_{i\tau+1} - y_{i\tau-j}) = \beta(1-p)$$

Card on Dips and Parallel Trends

Scott: "This diff-in-diff thing. ... I had the impression from stuff you'd written with Orley that you were a little lukewarm about the method a little bit. That it wasn't always something that you'd found credible. Is that accurate?"

Dave: "Well in the original [job training program evaluation] case, you didn't have parallel trends. ... [Diff-in-diff] wasn't something you could just count on to use out of the box."

Scott: "So you were thinking about parallel trends from the very start?"

Dave: "Oh no, it's right there in that 1985 paper with Orley that introduces the term difference-in-differences. Unfortunately we didn't have a graph. When I teach the paper now I show the graph. It has a very large dip for the treatment group. And that's the thing everyone nowadays calls the Ashenfelter Dip." (my emphasis)

But parallel trends does not yet appear

Figure: No mention of parallel trends in Ashenfelter and Card (1985)

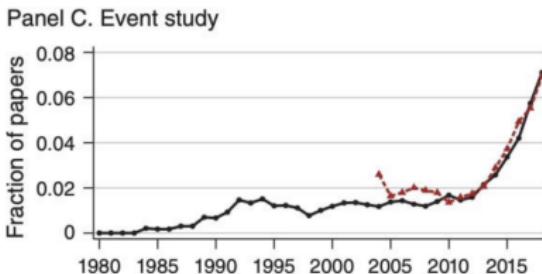
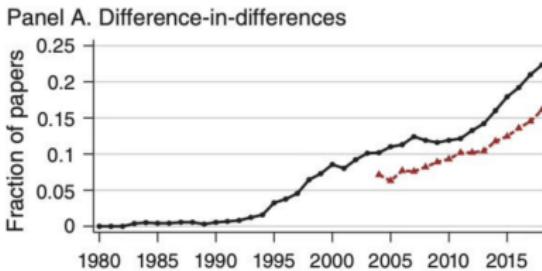
Not found Begins with

Second wave grows faster than first

- From 2011 to 2018, the share of NBER working papers that used diff-in-diff went from 11-12% to 23% in 2018, a little over 11 percentage points in only 7 years.
- It had taken 24 years to reach 11% the first time, but only 7 years the second time.
- The second wave had a very fast growth rate.
- The second wave's energy was stronger than the first.

Second wave of DiD: event study and parallel trends

Figure: Currie, et al. (2020)

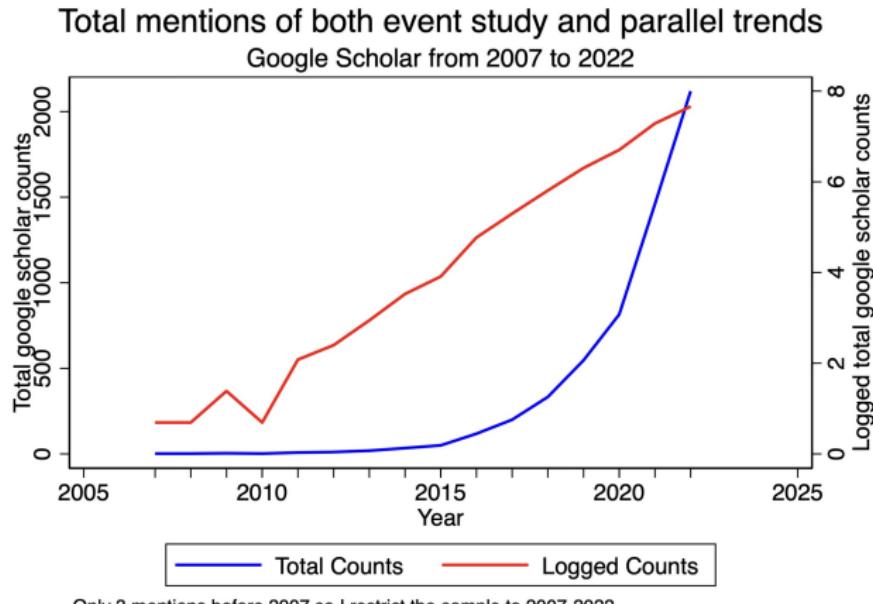


Second wave incorporates parallel trends and event studies

- Prior to 2011, parallel trends was not a phrase used in combination with diff-in-diff
- Event studies were used in finance to study abnormal stock market returns, not difference-in-differences
- From 2011 to 2018, the share of NBER working papers that used diff-in-diff went from 11-12% to 23% and event study grows from 2% to 7%.
- It's more than just correlation – papers that say both "event study" and "parallel trends" suddenly start appearing early to mid teens with exponential growth of 66% a year

Event study and parallel trends

Figure: Exponential growth in both event study and parallel trends

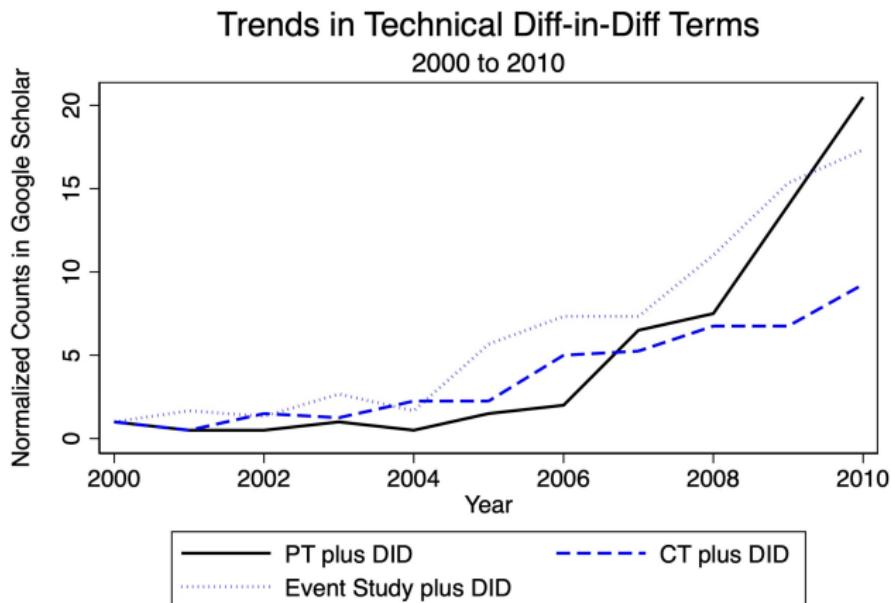


Second wave captures event studies

- Second wave DiD brings in parallel trends and event studies to DiD designs, perhaps riding on the long tradition of Ashenfelter's Dip
- But the explicit notation and identifying assumption is worked out gradually by different authors before being broadly adopted by practitioners
- Plotting of the words "parallel trends", "event study" and "common trends" in combination with papers with phrase "difference-in-differences"

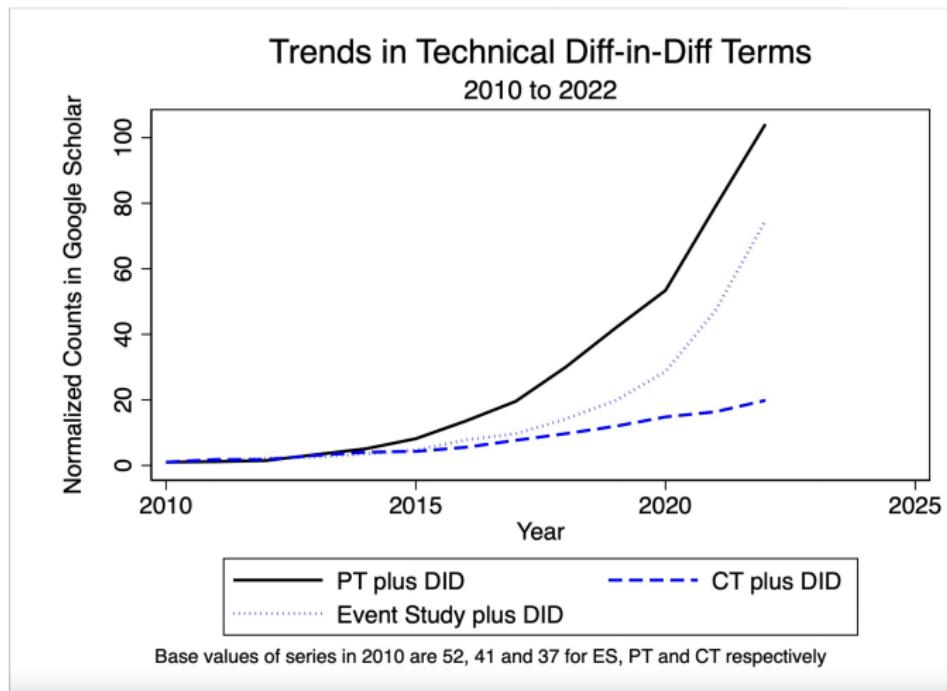
Incorporating PT and event study into diff-in-diff

Figure: Growth in both event study and parallel trends



Incorporating PT and event study into diff-in-diff

Figure: Growth in both event study and parallel trends



Parallel trends is introduced late

- Early focus had tended to emphasize the year fixed effects, but authors did not typically mean “parallel trends”
“Moreover, given the increasing dispersion of incomes and wages among all groups during that period, the common time effects (common trends) assumption among the unobservable components across the two groups may not be satisfied.” (Blundell and Dias 1999)
- Early identification was an independent assumption by Blundell, but this was viewed as too strong by Athey and Imbens (2006)
“Note that the full independence assumption $\varepsilon_i \perp\!\!\!\perp (G_i, T_i)$ (e.g., Blundell and MaCurdy (2000)) is stronger than necessary for [DiD] to give the average treatment effect.”

Heckman, et al. (1997) introduce potential outcomes

- First time we see parallel trends expressed (without name) using potential outcomes, though, is a conditional parallel trends assumption by Heckman, Ichimura and Todd (1997)

Term $D_{t,r}(X)$ identifies $E(\Delta | X, D=1)$ if the following assumption holds

$$E(Y_{0t} - Y_{0r'} | X, D=1) = E(Y_{0t} - Y_{0r'} | X, D=0). \quad (\text{A-5})$$

- Angrist and Pischke (2009) chapter 5 is entitled "Parallel Worlds: Fixed Effects, Difference-in-Differences and Panel Data"
- But phrase "parallel trends" does not appear (actually unclear when it starts getting used) – but we do see this diagram

Mostly Harmless and Parallel Trends

172

CHAPTER 5. FIXED EFFECTS, DD, AND PANEL DATA

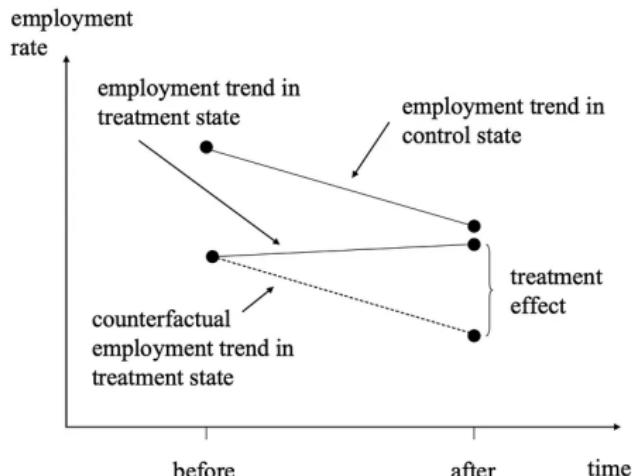


Figure 5.2.1: Causal effects in the differences-in-differences model

Introducing Event Studies to Diff-in-Diff

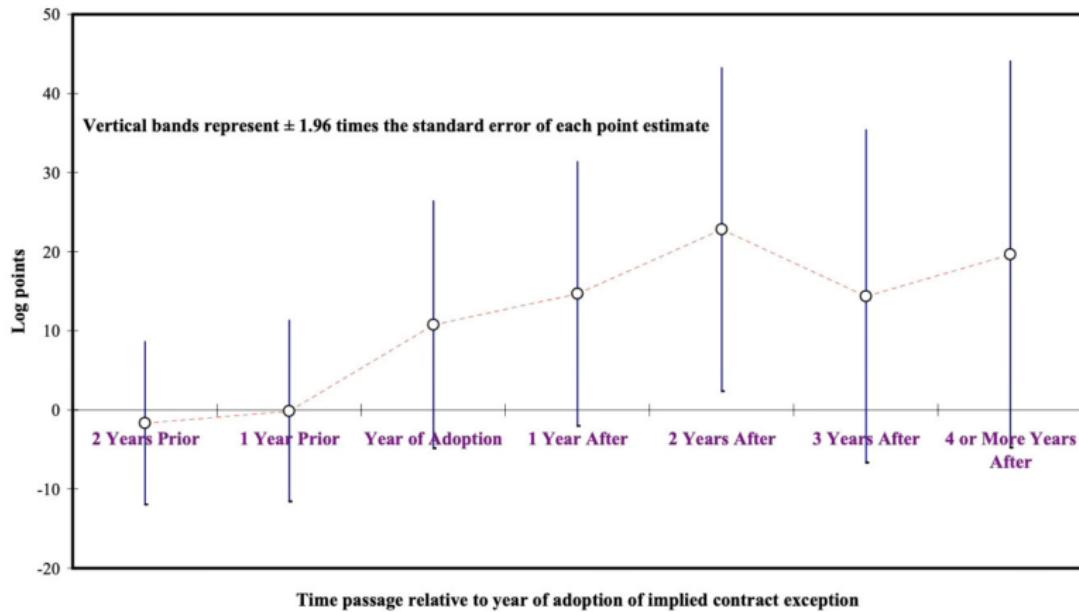


Figure: David Autor (2003) "unjust dismissals" paper

Constant effects model

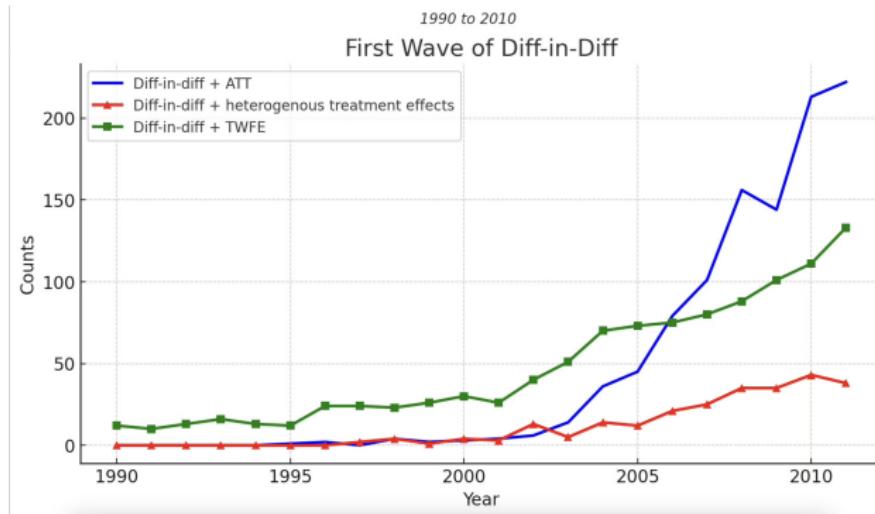
In applying the difference-in-difference framework to the data, it is important to consider carefully the ‘experiment’ created by these court decisions. In the ideal case, the court decisions would be independent, random events that varied in timing and had no spillover effects to non-adopting states. Assuming further that the ‘treatment effect’ of the common law exceptions was constant, equation (8) if correctly specified will provide an unbiased estimate of the average ‘treatment’ effect, δ .

Figure: David Autor clues us into the problems of TWFE

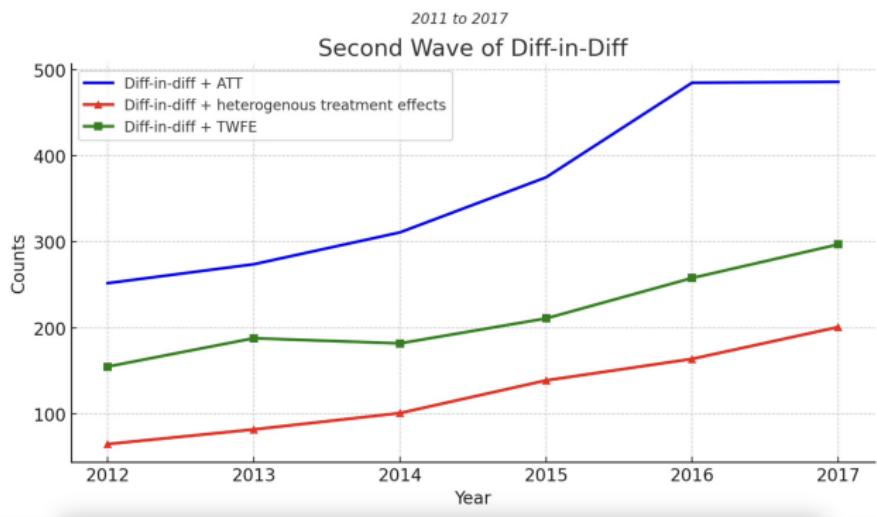
Third wave of diff-in-diff

- Starts with Kirill Borusyak and Xavier Jaravel (2016) working paper posted to SSRN (later published with Jan Speiss 2023)
- Key elements: explicit mention of the average treatment effect on the treated (ATT), heterogenous treatment effects and the pathologies of twoway fixed effects (TWFE)

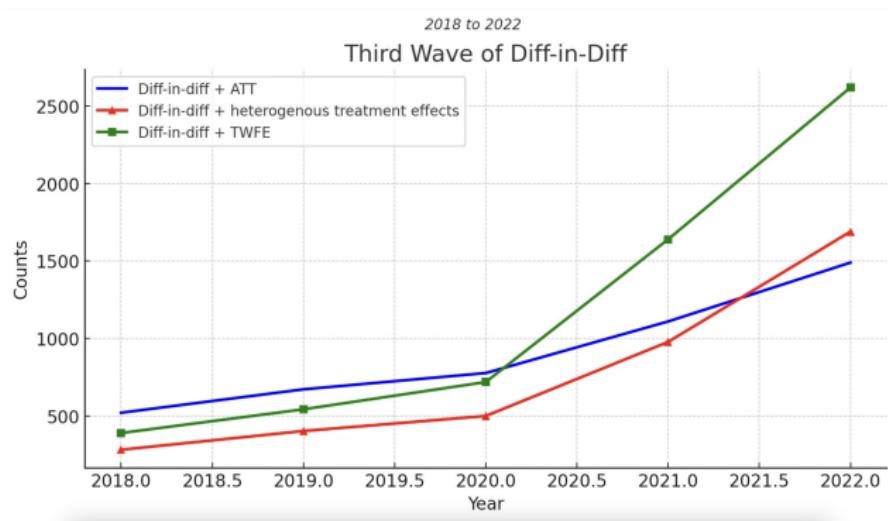
First Wave



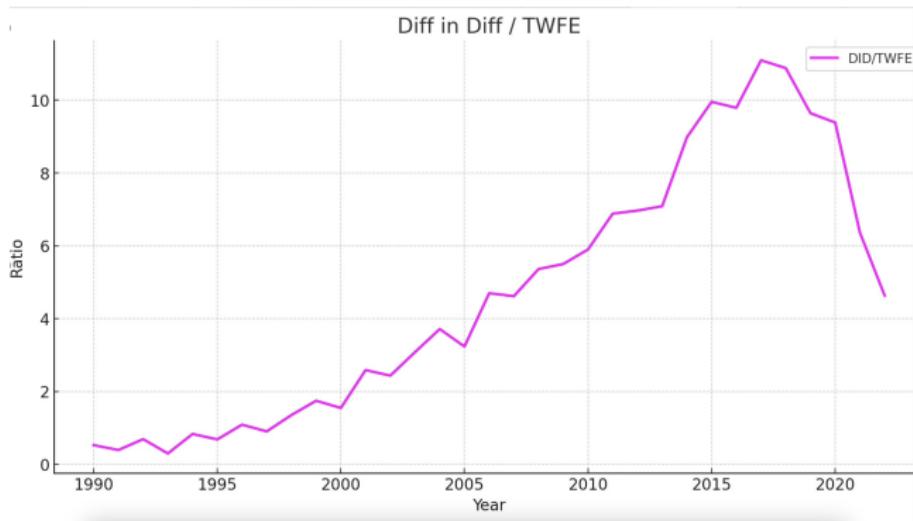
Second Wave



Third Wave



TWFE relative to DID



Roadmap

Background material

Introduction

Three waves of Difference-in-Differences

TWFE Pathologies

Historical links

Potential outcomes

Bacon decomposition

Simulation

Two solutions and a new decomposition

CS

SA

Interesting Applications using staggered adoption: Facebook and Mental Health

Two-way fixed effects

- When working with panel data, the so-called “two-way fixed effects” (TWFE) estimator was the workhorse estimator
- And from the start, regression was used with diff-in-diff, and as more complex designs began studied, panel fixed effects were a natural choice
- But DiD is not regression, nor is it TWFE – it is four averages and three subtractions as we saw earlier
- Under differential timing, interpreting the TWFE coefficient became challenging given there weren’t just two groups

Two OLS Models

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

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

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

The coefficient on the interaction is the DiD equation of "four averages and three subtractions" but the coefficient on equation 2 delta under heterogenous treatment effects – what was it?

Equivalence

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

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

- The δ in the regression is numerically identical to “four averages and three subtractions”
- And under a particular assumption, they are also unbiased estimates of an aggregate causal parameter
- But to see this we need new notation – potential outcomes

Potential outcomes notation

- Let the treatment be a binary variable:

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

where i indexes an individual observation, such as a person

Potential outcomes notation

- Potential outcomes:

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

where j indexes a counterfactual state of the world

Treatment effect definitions

Individual treatment effect

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

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

Conditional Average Treatment Effects

Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

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

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

Potential outcomes vs data

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

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

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

DiD equation

Orley's "four averages and three subtractions", or what Bacon will call the 2x2

$$\hat{\delta} = \left(E[Y_k|Post] - E[Y_k|Pre] \right) - \left(E[Y_U|Post] - E[Y_U|Pre] \right)$$

k are the people in the job training program, U are the untreated people not in the program, $Post$ is after the trainees took the class, Pre is the period just before they took the class, and $E[y]$ is mean earnings.

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

Potential outcomes and the switching equation

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

Parallel trends bias

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

Identification through parallel trends

Parallel trends

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

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

In words: “The evolution of earnings for our trainees *had they not trained* is the same as the evolution of mean earnings for non-trainees”.

It's in red because parallel trends is untestable and critically important to estimation of the ATT using any method, OLS or “four averages and three subtractions”

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 $\hat{\delta}$ correspond to?
- It relates to the previous "four averages and three subtractions" – but it's numerous of them, not just one
- In differential timing, the coefficient, $\hat{\delta}^{TWFE}$, is a weighted average of *all* possible DiD calculations – including some we should not be calculating

K^2 distinct DDs

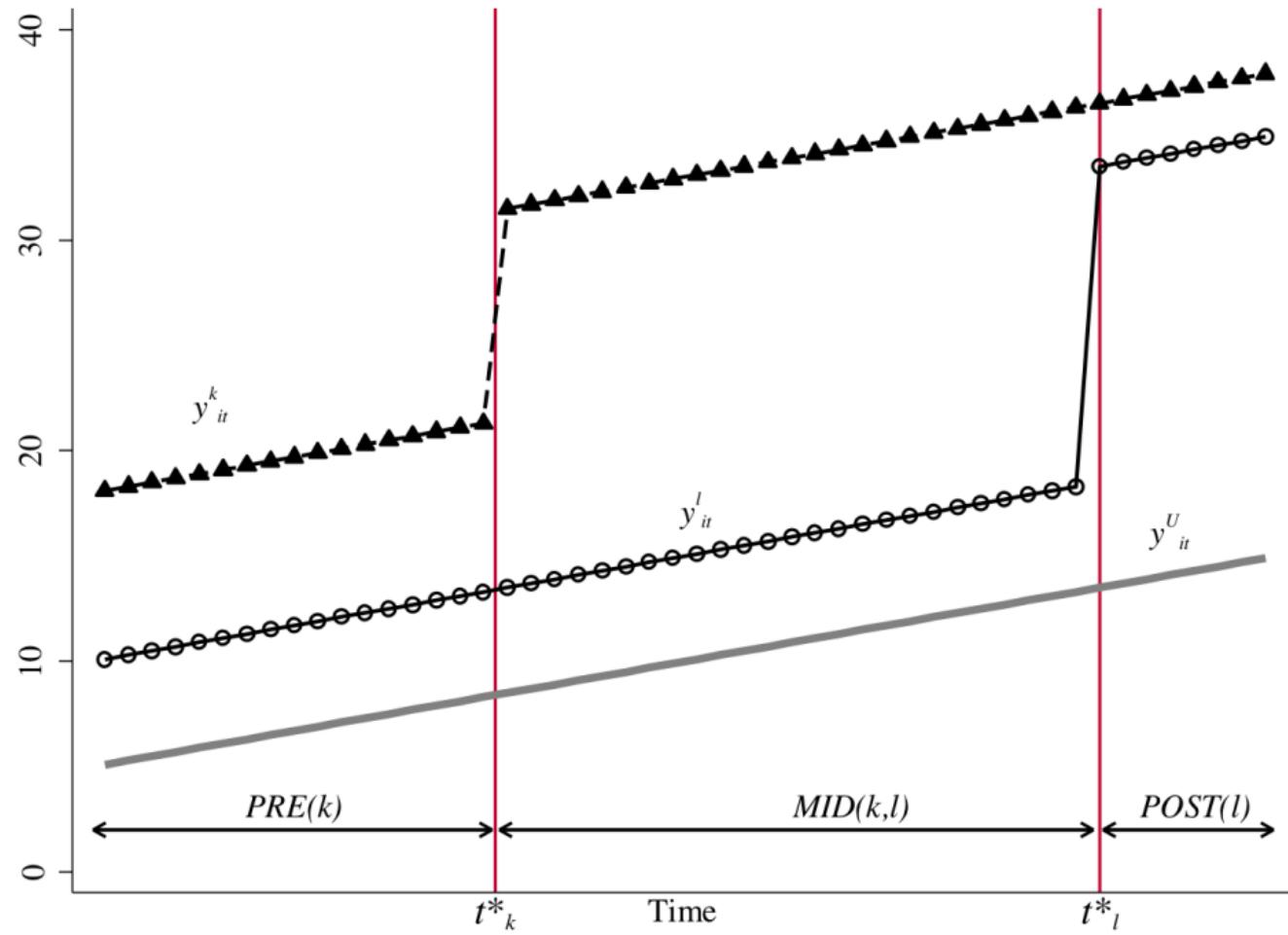
Let's look at 3 timing groups (a, b and c) and one untreated group (U).
With 3 timing groups, there are 9 2x2 DDs. Here they are:

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

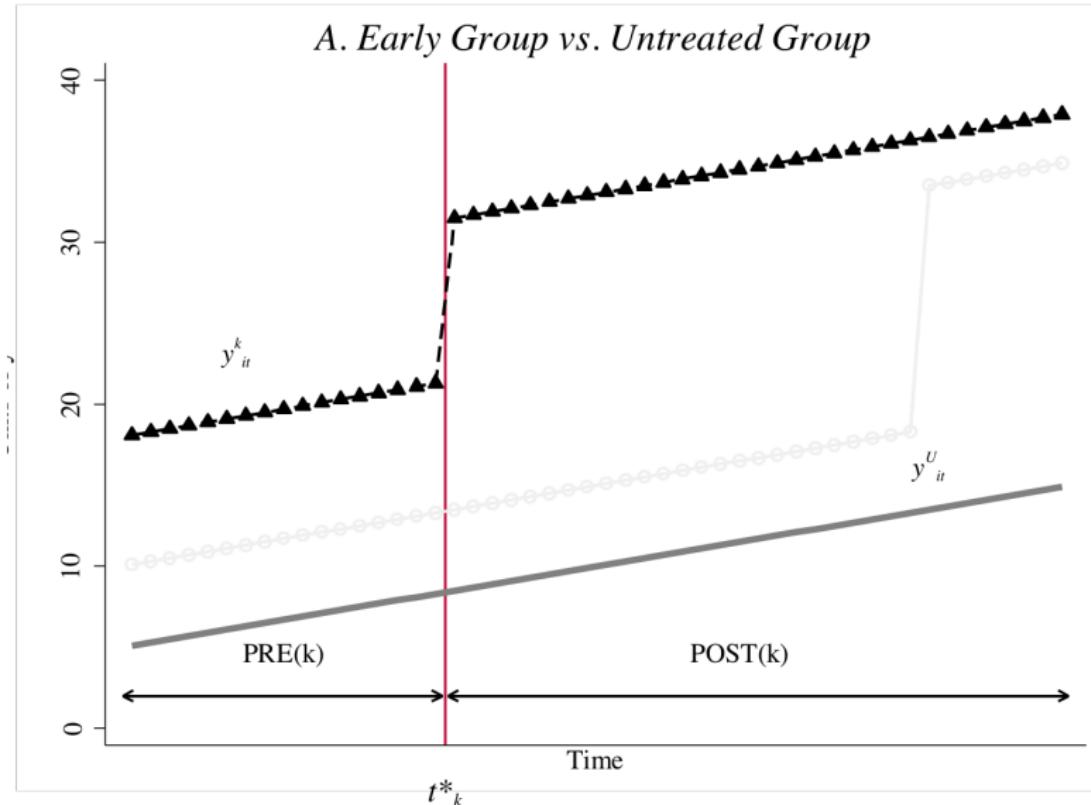
Let's return to a simpler example with only two groups – a k group treated at t_k^* and an l treated at t_l^* plus an never-treated group called the U untreated group

Terms and notation

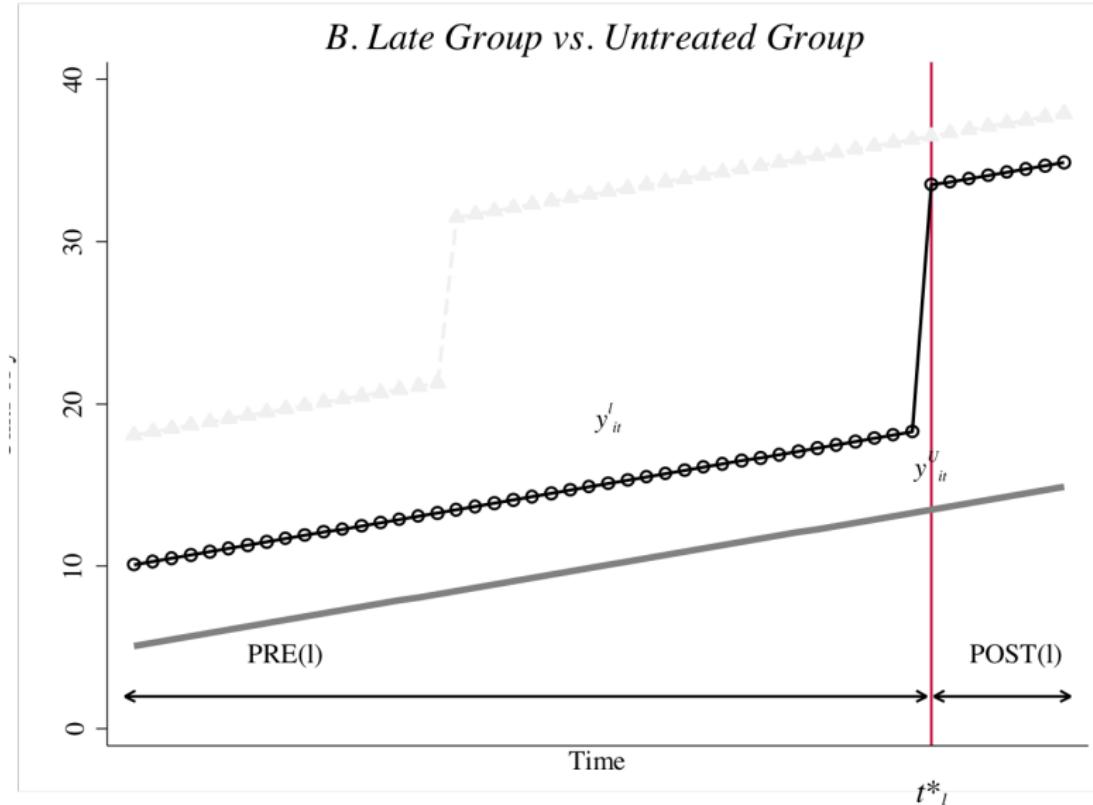
- Let there be two treatment groups (k, l) and one untreated group (U)
- k, l define the groups based on when they receive treatment (differently in time) with k receiving it earlier than l
- Denote \bar{D}_k as the share of time each group spends in treatment status
- Denote $\hat{\delta}_{jb}^{2x2}$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



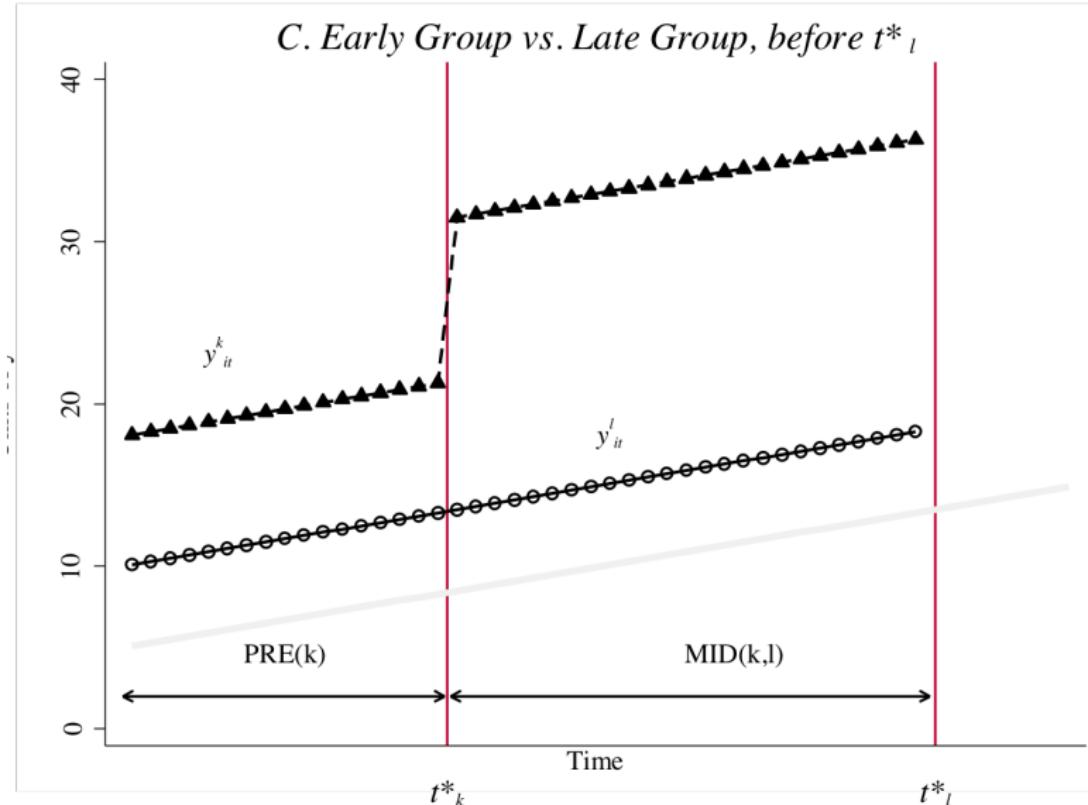
$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)} \right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)} \right)$$



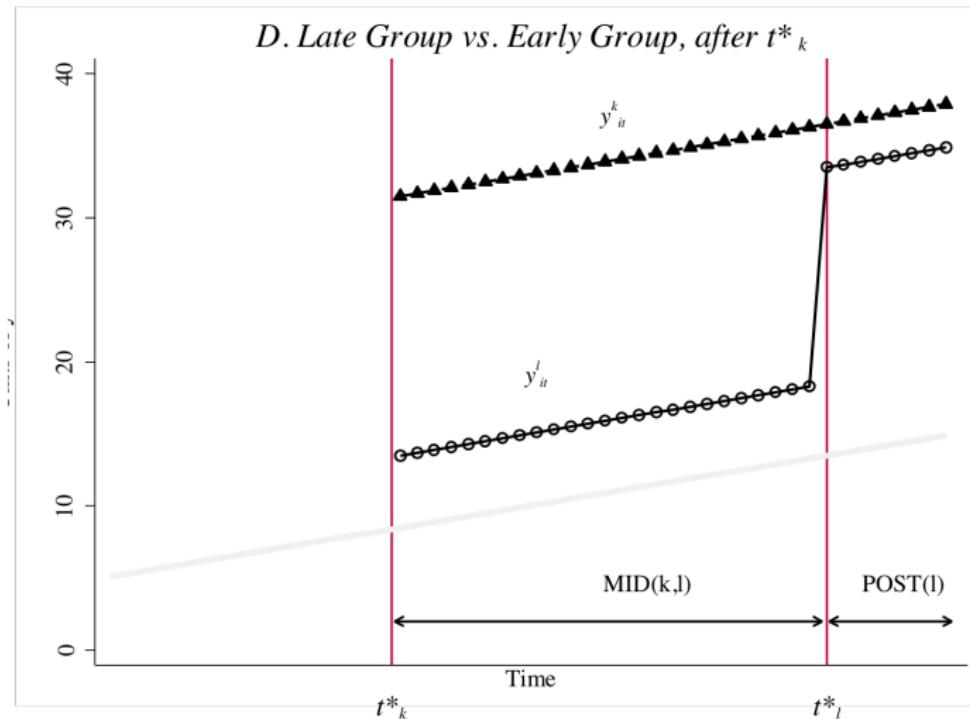
$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_l^{post(l)} - \overline{y}_l^{pre(l)} \right) - \left(\overline{y}_U^{post(l)} - \overline{y}_U^{pre(l)} \right)$$



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



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



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)

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

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

Third, the Weights

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

where n refer to sample sizes, $\bar{D}_k(1 - \bar{D}_k)$ ($\bar{D}_k - \bar{D}_l$) $(1 - (\bar{D}_k - \bar{D}_l))$ 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.
 - $\bar{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
 - $\bar{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
 - $\bar{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$
 - $\bar{D} = 0.6$. 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., $\bar{D}_k - \bar{D}_l$)
- 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 $\bar{D}_k - \bar{D}_l = 0.52$. 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

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

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

The dangerous 2x2

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

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

Substitute all this stuff into the decomposition formula

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

where we will make these substitutions

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

Notice all those potential sources of biases!

Potential Outcome Notation

$$p \lim_{n \rightarrow \infty} \hat{\delta}_{n \rightarrow \infty}^{TWFE} = VWATT + VWPT - \Delta ATT$$

- 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

Background material

Introduction

Three waves of Difference-in-Differences

TWFE Pathologies

Historical links

Potential outcomes

Bacon decomposition

Simulation

Two solutions and a new decomposition

CS

SA

Interesting Applications using staggered adoption: Facebook and Mental Health

Robust DiD

- There are probably about 6-8 possible robust DiD methods that can be used now
- All of them have certain things in common – don't use already treated units as controls
- To review all of them is impossible in a short class so I will focus on two
- Callaway and Sant'Anna (CS) and Sun and Abraham (SA), but I will point you to more

Callaway and Sant'Anna

CS is a DiD estimator used for estimating and then summarizing specific ATT parameters for each treatment group for each period ("group-time ATT") under differential timing assuming SUTVA, no anticipation, common support and conditional parallel trends

CS is a straightforward estimator – identify smaller causal effects per group and time ("building blocks") and then aggregate them into whatever parameter you're interested in using non-negative weights while avoiding "bad controls"

Examples of when you'd use 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

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 $t = 1, \dots, T$
- 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”) → 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					CS estimates				
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)	Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0	1981	-0.0548	0.0191	0.0578	0
1986	10	0	0	0	1986	10.0258	-0.0128	-0.0382	0
1987	20	0	0	0	1987	20.0439	0.0349	-0.0105	0
1988	30	0	0	0	1988	30.0028	-0.0516	-0.0055	0
1989	40	0	0	0	1989	40.0201	0.0257	0.0313	0
1990	50	0	0	0	1990	50.0249	0.0285	-0.0284	0
1991	60	0	0	0	1991	60.0172	-0.0395	0.0335	0
1992	70	8	0	0	1992	69.9961	8.013	0	0
1993	80	16	0	0	1993	80.0155	16.0117	0.0105	0
1994	90	24	0	0	1994	89.9912	24.0149	0.0185	0
1995	100	32	0	0	1995	99.9757	32.0219	-0.0505	0
1996	110	40	0	0	1996	110.0465	40.0186	0.0344	0
1997	120	48	0	0	1997	120.0222	48.0338	-0.0101	0
1998	130	56	6	0	1998	129.9164	56.0051	6.027	0
1999	140	64	12	0	1999	139.9235	63.9884	11.969	0
2000	150	72	18	0	2000	150.0087	71.9924	18.0152	0
2001	160	80	24	0	2001	159.9702	80.0152	23.9656	0
2002	170	88	30	0	2002	169.9857	88.0745	29.9757	0
2003	180	96	36	0	2003	179.981	96.0161	36.013	0
2004	190	104	42	4	2004				
2005	200	112	48	8	2005				
2006	210	120	54	12	2006				
2007	220	128	60	16	2007				
2008	230	136	66	20	2008				
2009	240	144	72	24	2009				
ATT	82				Total ATT	n/a			
Feasible ATT	68.3333333				Feasible ATT	68.33718056			

Question: Why didn't CS estimate all $\text{ATT}(g,t)$? What is "feasible ATT"?

Reporting results

Table: Estimating ATT using only pre-2004 data

	(Truth)	(TWFE)	(CS)	(SA)	(BJS)
<i>Feasible ATT</i>	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

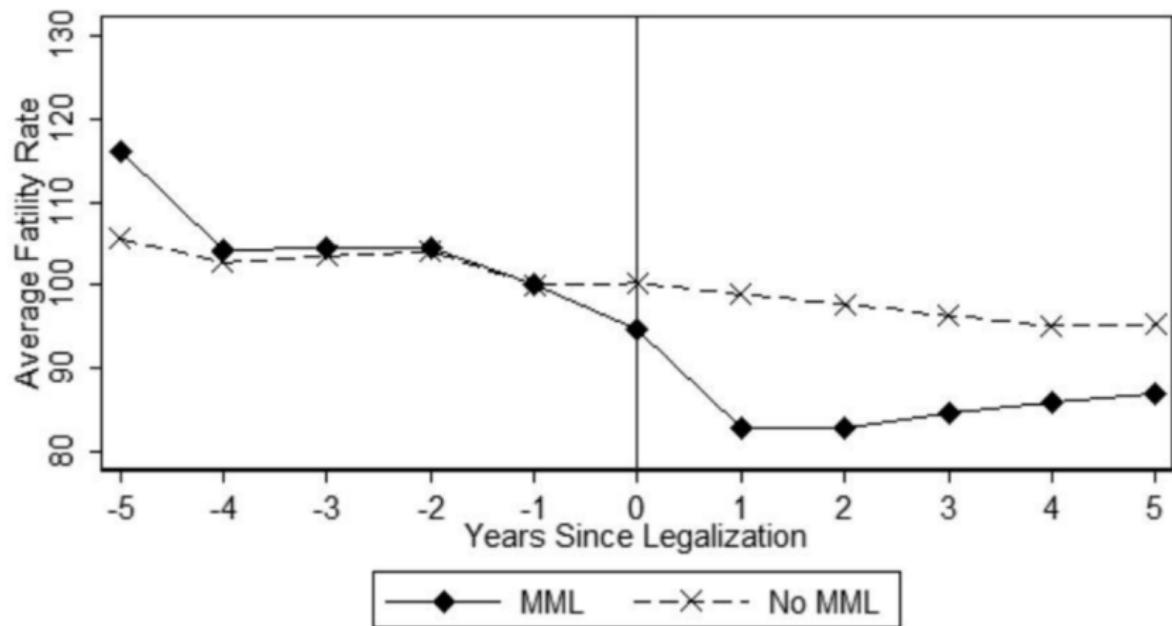


Figure: Anderson, et al. (2013) display of raw traffic fatality rates for re-centered treatment states and control states with randomized treatment dates

Replicated from a project of mine

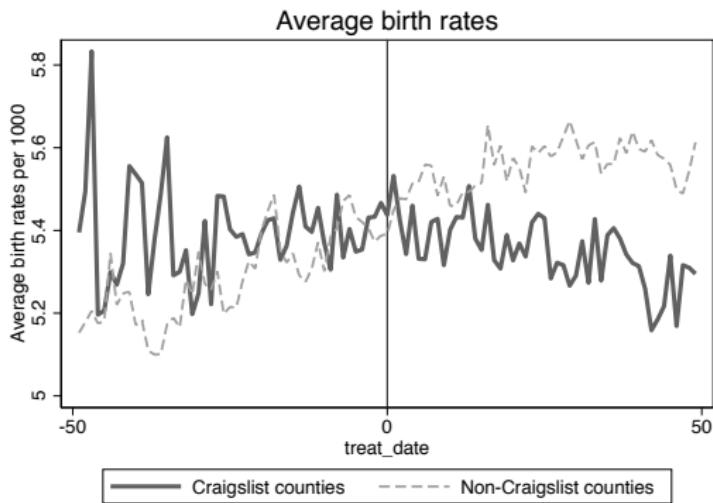
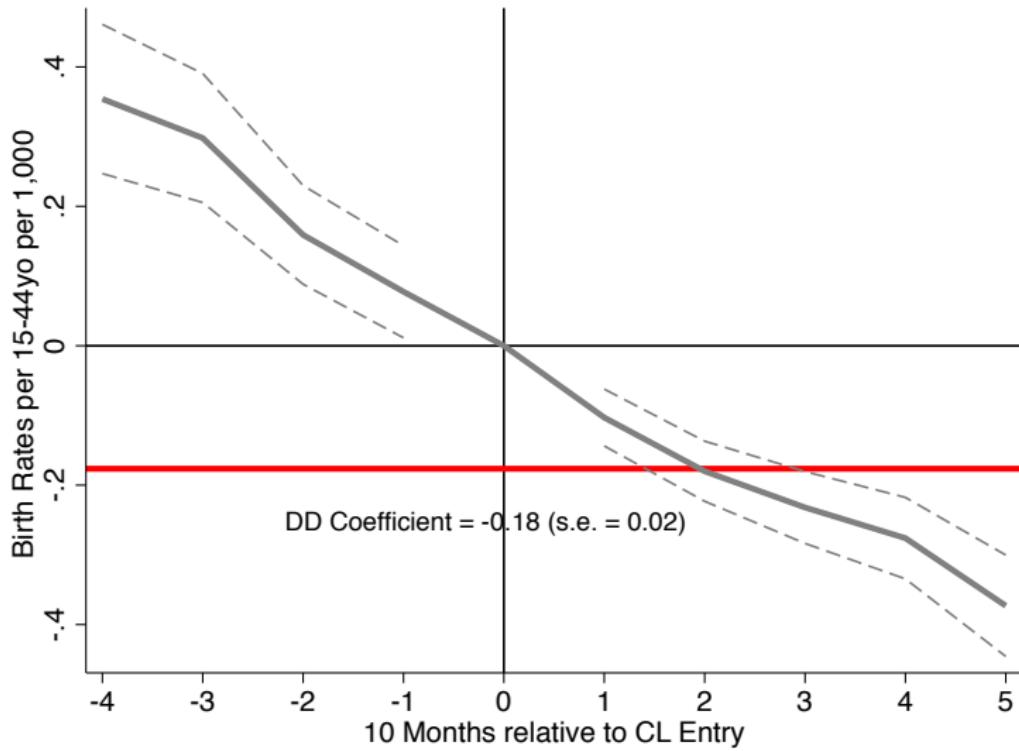


Figure: Roll out of Craigslist “personal ads” for casual intimate encounters and birth rates using the “randomized treatment assignment” approach for visualization

Event study specification with TWFE

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

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 Specification

- 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)

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 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 $l = \text{year} - \text{treatment date}$

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:

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

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

$$E[Y_{i,e+l}^e - Y_{i,e+l}^\infty | E_i = e] = 0 \text{ for all possible leads.}$$

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

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

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).

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

Weight ($w_{e,l}^g$) summation cheat sheet

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

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 bin 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:

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

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 $l' \geq 0$.

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

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

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

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

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

- 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

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

Consistency and Inference

- Under parallel trends and no anticipation, $\hat{\delta}_{e,l}$ 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

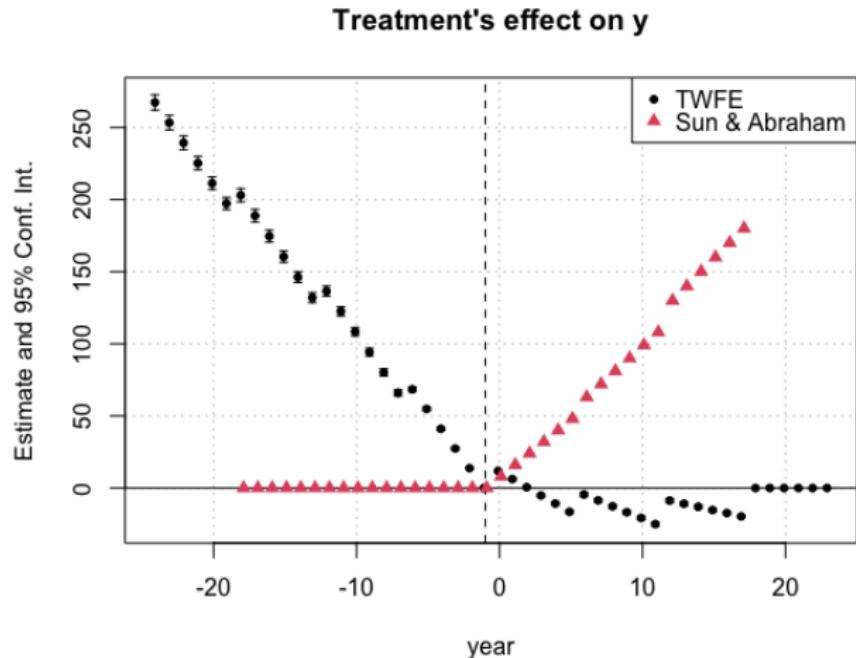
	(Truth)	(TWFE)	(CS)	(SA)	(BJS)
<i>Feasible</i> \widehat{ATT}	68.33	26.81***	68.34***	68.33***	

Computing relative event time leads and lags

Year	Truth					Relative time coefficients		
	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?

Roadmap

Background material

Introduction

Three waves of Difference-in-Differences

TWFE Pathologies

Historical links

Potential outcomes

Bacon decomposition

Simulation

Two solutions and a new decomposition

CS

SA

Interesting Applications using staggered adoption: Facebook and Mental Health

POSTED JAN 31, 2024 AT 10:32 AM EST

0 Comments (0 New)

A



ADI ROBERTSON

Mark Zuckerberg is trying to reset the conversation on social media's mental health effects. His opening testimony emphasizes that "the existing body of scientific work has not shown a causal link between using social media and young people having worse mental health outcomes." (Which isn't necessarily wrong, partly because causal links are very hard to prove — the overall body of research is complicated and seems to suggest social media has a variety of possible effects.)



All the news from Congress' Big Tech child safety hearing

ADI ROBERTSON JAN 31

Mental health and Social Media

- Unclear what he means; he may mean there is no experimental evidence
- Very difficult to imagine a convincing randomized experiment passing IRB now at this point
- Quasi-experimental evidence can step in to answer important questions like this
- Braghieri, Levy and Makarin (2022), "Social Media and Mental Health", *American Economic Review*, 112(11): 3660-3693

Overview of design and data

- Authors take advantage of a clever quirk in Facebook (then “theFacebook”) targeted different universities from 2004 to 2006
- They found an online data source that allowed them to pin point precisely when a university was “treated” with theFacebook
- They then linked that data with a longrunning health survey of college students (both before and after) in a very clever way
- Estimated the effect of a new social media platform’s presence at a university on student revealed mental health problems

Evidence I like to see

1. **Bite:** **Nothing.** They cannot really show much here. No data on Facebook usage. They had to rely on anecdote and Facebook as a "first mover", but there had been Friendster and MySpace so this does weaken the paper maybe
2. **Main Results:** Very strong evidence, mostly expressed using rich survey data and questions transformed into z-scores (standard deviations)
3. **Falsifications:** **None.** Authors do not perform falsifications. Remember Miller, Johnson and Wherry looking at Medicaid's effect on Medicare eligible population? There isn't anything like that here.
4. **Event studies:** Extremely compelling evidence and robustness across a half dozen different models
5. **Mechanism:** **Very weak in my opinion**

DiD in Court

- So in many ways the strength of the project lies in a few areas:
 1. Important question – social media and youth mental health problems is a major policy question (see Zuckerberg testifying before Congress about it)
 2. Excellent research design – difference-in-differences
 3. Meticulous data collection
 4. Data visualization is compelling
- And it publishes in the premiere journal in economics, which I think shows that the research question and high quality data combined with research design can lift a paper

TWFE

$$Y_{icgt} = \alpha_g + \delta_t + \beta \times Facebook_{gt} + X_i \times \gamma + X_c \times \psi + \varepsilon_{icgt} \quad (3)$$

Authors make the decision to lead with TWFE, most likely because the new methods have not yet fully been adopted and there is delayed "literacy" around the biases of TWFE

Data on Facebook

- When does Facebook appear at a school?
 - Facebook only publishes a fraction of that information
 - They came up with a workaround
- The Wayback Machine has been taking almost daily photographs of every website since the Internet's beginning – including the frontpage of "TheFacebook", which held some surprises

[DONATE](#)

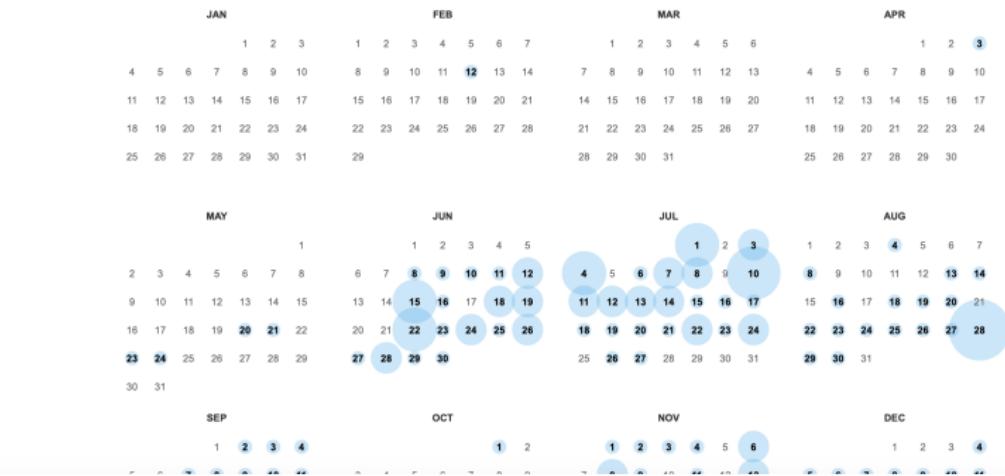
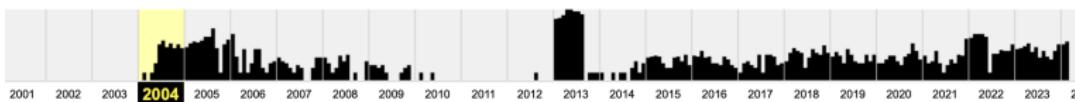
Explore more than 294 billion web pages saved over time

thefacebook.com

X

[Calendar](#)[Collections](#)[Changes](#)[Summary](#)[Site Map](#)[URLs](#)

Saved 7,057 times between February 12, 2004 and February 16, 2024.





[thefacebook]

[login](#) [register](#) [about](#)

Email:

Password:

[login](#)

[register](#)

Welcome to Thefacebook!

[Welcome to Thefacebook]

Thefacebook is an online directory that connects people through social networks at colleges.

We have opened up Thefacebook for popular consumption at **Harvard, Columbia, Stanford, Yale, Cornell, Dartmouth, UPenn, MIT**, and now **BU** and **NYU**.

Your facebook is limited to your own college or university.

You can use Thefacebook to:

- Search for people at your school
- Find out who is in your classes
- Look up your friends' friends
- See a visualization of your social network

To get started, click below to register. If you have already registered, you can log in.

[Register](#)

[Login](#)

[about](#) [contact](#) [faq](#) [terms](#) [privacy](#)

a Mark Zuckerberg production

Thefacebook © 2004

Go JUN SEP OCT
◀ 02 ▶
2003 2004 2005 ▾ Abd



login register about

Welcome to Thefacebook!

[Welcome to Thefacebook]

Thefacebook is an online directory that connects people through social networks at colleges.

We have opened up Thefacebook for popular consumption at:

Amherst • BC • Berkeley • Bowdoin • Brown • BU • Bucknell
Caltech • Carnegie Mellon • Chicago • Columbia • Cornell • Dartmouth
Duke • Emory • Florida • Florida State • Georgetown • Georgia • GWU • Hamilton
Harvard • Illinois • Johns Hopkins • Maryland • Michigan
Michigan State • Middlebury • MIT • Northeastern • Northwestern • Notre Dame
NYU • Oberlin • Penn • Princeton • Rice • Rochester • South Florida
Stanford • Swarthmore • Syracuse • Tufts • Tulane • UCDavis
UCF • UCLA • UCSD • UNC • USC • UVA • **Vanderbilt** • WashU
Wellesley • Wesleyan • Williams • Yale

Your facebook is limited to your own college or university.

You can use Thefacebook to:

- Search for people at your school
- Find out who is in your classes
- Look up your friends' friends
- See a visualization of your social network

To get started, click below to register. If you have already registered, you can log in.

[Register](#) [Login](#)

[about](#) [contact](#) [jobs](#) [fan](#) [advertise](#) [terms](#) [privacy](#)
a Mark Zuckerberg production
Thefacebook © 2004

Go MAR APR MAY
◀ 28 ▶
2004 2005 2006



[thefacebook]

login register about faq

Email:

Password:

[login](#) [register](#)

Welcome to Thefacebook!

[Welcome to Thefacebook]

Thefacebook is an online directory that connects people through social networks at colleges.

We have recently opened up Thefacebook at the following schools:

Abilene Christian • Agnes Scott • Albright • Allegheny
Anderson • Angelo State • Arcadia • Arts • Azusa Pacific
Beloit • Boise State • C. Arkansas • C. Missouri • Cal. Lutheran
Campbell • Canisius • Capital • Carthage • Christian Brothers
Clark Atlanta • Cleveland State • Columbus State • CSCC
CUNY City • Delta State • DeSales • Edgerton • Endicott • ETSU
Evansville • Frostburg • Guilford • Gustavus • Hampden-Sydney
Hartwick • Hendrix • Illinois Tech • IPFW • Jacksonville
John Jay • Kettering • Lake Forest • Lamar • Liberty • Lock Haven
McDaniel • Messiah • Milligan • MSOE • Murray State • N. Georgia
N. Kentucky • NJIT • Nova • NYIT • Otterbein • Philadelphia
RIC • S. Alabama • S.E. Louisiana • Saginaw Valley • Salem State
Shepherd • St. Cloud • St. Rose • Sweet Briar • Tarleton
TN Chattanooga • TN Tech • UMass Lowell • Valencia • W. Florida
W. Oregon • Widener • WNEC • Wofford • Yeshiva

For a complete list of supported schools, click [here](#).

Your facebook is limited to your own college or university.

You can use Thefacebook to:

- Search for people at your school
- Find out who is in your classes
- Look up your friends' friends
- See a visualization of your social network

To get started, click below to register. If you have already registered, you can log in.

[Register](#) [Login](#)

[about](#) [contact](#) [jobs](#) [announce](#) [advertise](#) [terms](#) [privacy](#)
a Mark Zuckerberg production
Thefacebook © 2005

Timing Dates

- They went through three years of daily screenshots on Wayback machine to find when a school appeared on the front page
- The first time Penn State appears on the front page, the authors mark that as the date when the school got Facebook
- But now they need information on mental health outcomes
- They find it with an old long running repeated cross section survey of college students

NCHA survey by ACHA

*Our second main data source consists of more than 430,000 responses to the NCHA survey, a survey administered to college students on a semi-annual basis by the American College Health Association (ACHA). The NCHA survey was developed in 1998 by a team of college health professionals with the purpose of obtaining information from college students about their mental and physical health. Specifically, the NCHA survey inquires about demographics, physical health, **mental health**, alcohol and drug use, sexual behaviors, and perceptions of these behaviors among one's peers.*

No evidence of bite

The NCHA survey does not include any questions on social media use; therefore, it is not possible for us to determine whether a particular survey respondent had a Facebook account.

This is probably the problem in any study in which your treatment is more or less the first of its kind – most likely the standard surveys have not yet incorporated the questions into their surveys

Linking Facebook data with NCHA data

In order to protect the privacy of the institutions that participate in the NCHA survey while still allowing us to carry out the analysis, the ACHA kindly agreed to provide us with a customized dataset that includes a variable indicating the semester in which Facebook was rolled out at each college. Specifically, the ACHA adopted the following procedure: (i) merge our dataset containing the Facebook introduction dates to the NCHA dataset; (ii) add a variable listing the semester in which Facebook was rolled out at each college;¹⁵ (iii) strip away any information that could allow us to identify colleges (including the specific date in which Facebook was introduced at each college).

Basic facts about early and late adopters

- Colleges in earlier Facebook expansion groups are more selective in terms of test scores, larger, more likely to be on the East Coast, and have more residential undergraduate programs than colleges in later Facebook expansion groups.
- Colleges in earlier Facebook expansion groups enroll students from relatively more advantaged economic backgrounds.
- Students in colleges that received Facebook relatively earlier have worse baseline mental health outcomes than students attending colleges in later Facebook expansion groups.

Measurement

- The survey data is very rich with a lot of questions about mental health with different scales
- They create their own combinations of these questions into aggregate indices – “index of poor mental health” where higher numbers mean worse mental health
- Each outcome survey question is normalized into what is called a “z-score” which is interpreted as a fraction of a standard deviation
- Estimates are ATT parameters – average effect of Facebook on students at schools that got Facebook

TABLE 1—BASELINE RESULTS: INDEX OF POOR MENTAL HEALTH

	Index of poor mental health			
	(1)	(2)	(3)	(4)
Post-Facebook introduction	0.137 (0.040)	0.124 (0.022)	0.085 (0.033)	0.077 (0.032)
Observations	374,805	359,827	359,827	359,827
Survey-wave fixed effects	✓	✓	✓	✓
Facebook-expansion-group fixed effects	✓	✓		
Controls		✓	✓	✓
College fixed effects			✓	✓
FB-expansion-group linear time trends				✓

Notes: This table explores the effect of the introduction of Facebook at a college on student mental health. Specifically, it presents estimates of coefficient β from equation (1) with our index of poor mental health as the outcome variable. The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. Column 1 estimates equation (1) without including controls; column 2 estimates equation (1) including controls; column 3, our preferred specification, replaces Facebook-expansion-group fixed effects with college fixed effects; column 4 includes linear time trends estimated at the Facebook-expansion-group level. Our controls consist of age, age squared, gender, indicators for year in school (freshman, sophomore, junior, senior), indicators for race (White, Black, Hispanic, Asian, Indian, and other), and an indicator for international student. Column 2 also includes indicators for geographic region of college (Northeast, Midwest, West, South); such indicators are omitted in columns 3 and 4 because they are collinear with the college fixed effects. For a detailed description of the outcome, treatment, and control variables, see online Appendix Table A.31. Standard errors in parentheses are clustered at the college level.

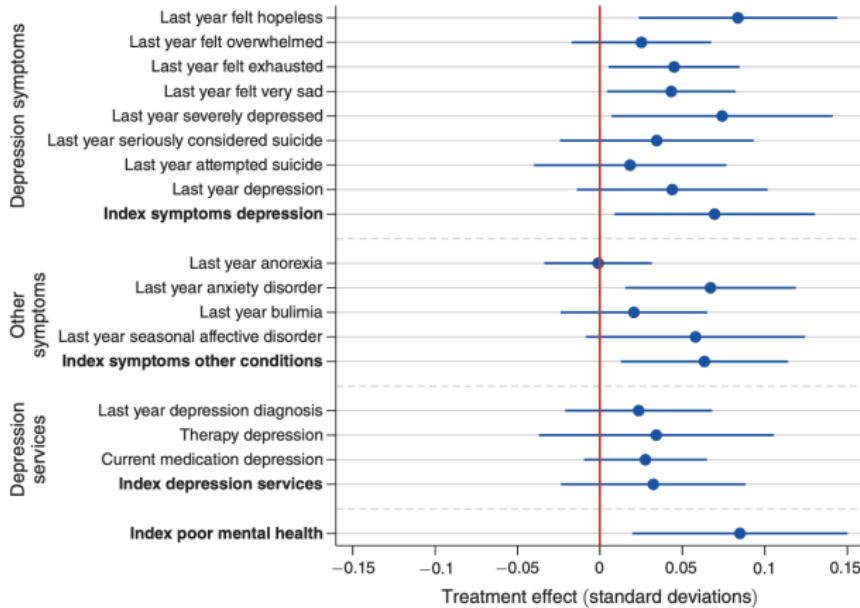


FIGURE 1. EFFECTS OF THE INTRODUCTION OF FACEBOOK ON STUDENT MENTAL HEALTH

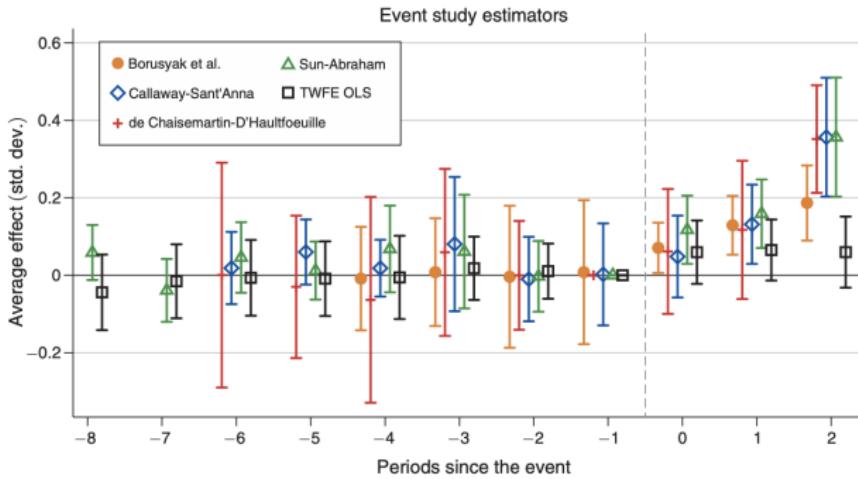


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

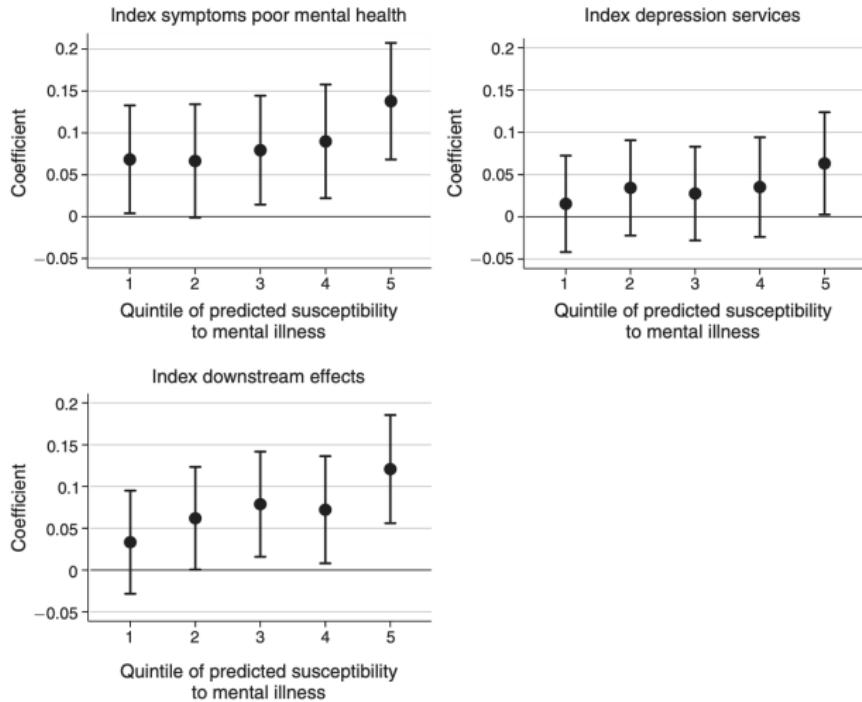


FIGURE 3. HETEROGENEOUS EFFECTS BY PREDICTED SUSCEPTIBILITY TO MENTAL ILLNESS

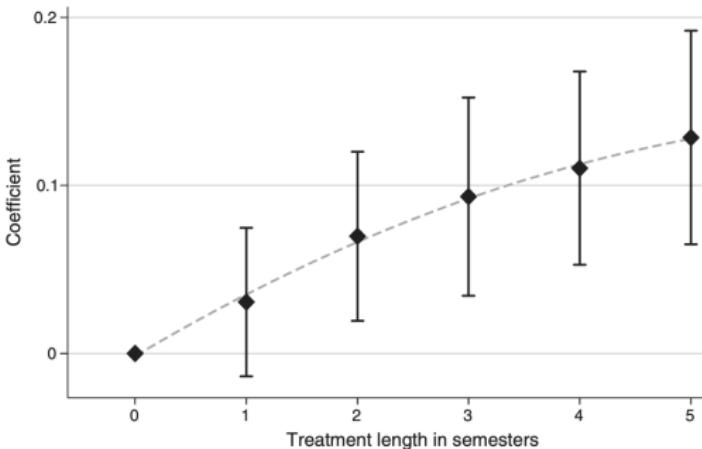


FIGURE 4. EFFECT ON POOR MENTAL HEALTH BY LENGTH OF EXPOSURE TO FACEBOOK

Notes: This figure explores the effects of length of exposure to Facebook on our index of poor mental health by presenting estimates of equation (4). The index is standardized so that, in the preperiod, it has a mean of zero and a standard deviation of one. The dashed curve is the quadratic curve of best fit. Our controls consist of age, age squared, gender, indicators for year in school (freshman, sophomore, junior, senior), indicators for race (White, Black, Hispanic, Asian, Indian, and other), and an indicator for international student. Students who entered college in 2006 might have been exposed to Facebook already in high school, because, starting in September 2005, college students with Facebook access could invite high school students to join the platform. Such students are excluded from the regression. For a detailed description of the outcome, treatment, and control variables, see online Appendix Table A.31. The bars represent 95 percent confidence intervals. Standard errors are clustered at the college level.

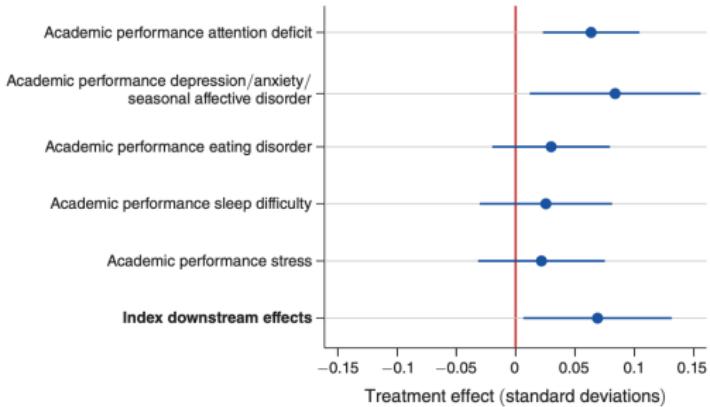


FIGURE 5. DOWNSTREAM EFFECTS ON ACADEMIC PERFORMANCE

Notes: This figure explores downstream effects of the introduction of Facebook on the students' academic performance. It presents estimates of coefficient β from equation (1) using our preferred specification, including survey-wave fixed effects, college fixed effects, and controls. The outcome variables are answers to questions inquiring as to whether various mental health conditions affected the students' academic performance and our index of downstream effects. All outcomes are standardized so that, in the preperiod, they have a mean of zero and a standard deviation of one. For a detailed description of the outcome, treatment, and control variables, see online Appendix Table A.31. The bars represent 95 percent confidence intervals. Standard errors are clustered at the college level.

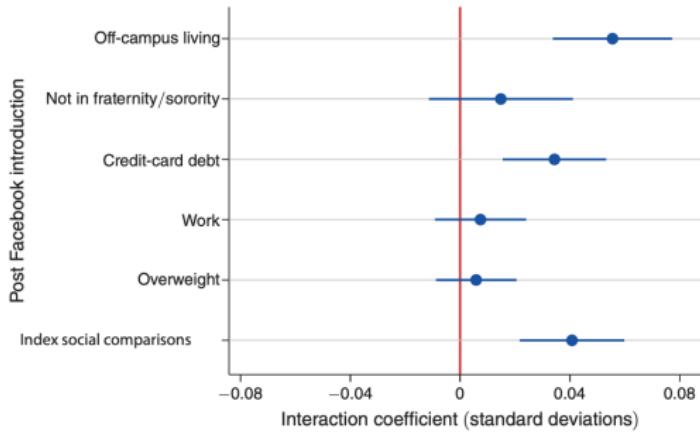


FIGURE 6. HETEROGENEOUS EFFECTS AS EVIDENCE OF UNFAVORABLE SOCIAL COMPARISONS

Notes: This figure explores the mechanisms behind the effects of Facebook on mental health. It presents estimates from a version of equation (1) in which our treatment indicator is interacted with a set of indicators for belonging to a certain subpopulation of students. The outcome variable is our overall index of poor mental health. The estimates are obtained using our preferred specification, namely the one including survey-wave fixed effects, college fixed effects, and controls. For a detailed description of the outcome, treatment, interaction, and control variables, see online Appendix Table A.31. The bars represent 95 percent confidence intervals. Standard errors are clustered at the college level.

Questions and Comments

- Had this study been done five years ago, the conclusion of the event study would've been much different, and the paper may not have been as easily published let alone in the AER
- No doubt more work is needed, but you can see here the power of the well designed DID and its acceptance at answering causal questions for major policy problems
- While the RCT would definitively answer nearly anything we can imagine, it is oftentimes not practical or even sometimes ethical raising the marginal value of the quasi-experimental design
- But these are not straightforward either and can be easily done badly even in good faith

My Preferred Ladder of Evidence

1. Main results are your claim, not your evidence
2. Show bite – first order effects are essential evidence
3. Event study graphs are one of your strongest pieces of evidence
4. Falsifications – If you can find falsifications, use them
5. Mechanisms – can you find any explanation?

Synthetic control

- But what if parallel trends really isn't realistic – what then?
- Then you may need to create your own control group that follows the same approximately trajectory as your treatment group pre-treatment
- A method by Abadie and Gardazebal (2003) and follow up papers worked out a method for this called synthetic control