Causal Inference II

Mixtape Session

Roadmap

Background material
- Introduction
- Twoway fixed effects

TWFE Pathologies
- Bacon decomposition
- Simulation

Two solutions and a new decomposition
- CS
- SA

Example: Facebook and Mental Health

Overview

- Brief review of TWFE and idea of strict exogeneity
- Brief review of potential outcomes, ATT, and the parallel trends assumption
- Discussion of standard “constant treatment effect” TWFE pathologies using Goodman-Bacon
- Discussion of two solutions and a new study using these new methods

Beaver dam and diff-in-diff credibility crisis

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

Difference-in-differences credibility crisis

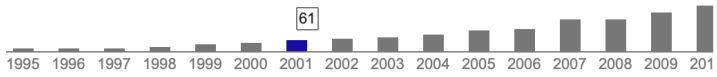
- Series of important papers starting in 2016, born independent of one another, by grad students and assistant professors found critical pathologies with TWFE and developed solutions
- Can't cover all of them, but they have a lot in common, so I'm going to cover the basic problem and some basic solutions
- Extreme meteoric rise, unusual for econometrics

Compare with LATE paper

- Compare what happened with Imbens and Angrist 1995 LATE in *Econometrica*
- 61 annual cites the year Imbens is denied tenure at Harvard for what would later win him a Nobel Prize
- Gradual rise; many famous econometrics papers like this because of the extreme firewall between econometrics and practices
- Diff-in-diff is unusual for some reason (following is only mid-2022 cites)

Identification and estimation of local average treatment effects

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

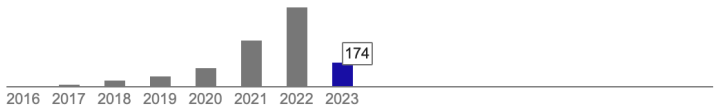


Borusyak et al

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

Revisiting event study designs: Robust and efficient estimation

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



“dCdH”

- First major hit in AER
- Very thorough decomposition of the TWFE pathology, very general solution, included code
- Very active and talented young team (assistant profs when this was done)

Two-way fixed effects estimators with heterogeneous treatment effects

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

Goodman-Bacon

- Arguably the most influential in terms of bringing attention to the problem
- Begun while grad student at Michigan, published last of the crop
- Probably Twitter network had a role as he was very active, also not an econometrician

Difference-in-differences with variation in treatment timing

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

“CS”

- Second broad solution to the problem, written while assistant professors at Vanderbilt and Ole Miss
- Colleague with Andrew Goodman-Bacon
- Introduced new terms like group-time ATT
- Seems to be in the lead

Difference-in-differences with multiple time periods

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

“SA”

- Third broad solution to the problem, very similar to CS
- Focus was on decomposing the event study
- Written while grad students at MIT

Estimating dynamic treatment effects in event studies with heterogeneous treatment effects

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

Just a drop in the bucket

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

Comments

- Because difference-in-differences is the simplest of all designs, and could be estimated with OLS, differential timing estimated with TWFE was thought to be a trivial extension
- Most people in fact though DiD and TWFE were the same thing
- Strict exogeneity, as it turned out, had functional form assumptions and constant treatment effects assumptions buried in it
- There are ways around it, but the dominant model used for decades (which probably all of us have published using) was more assumption laden than we knew

Twoway fixed effects

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

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 effects

When to use TWFE

- Traditionally, this was used for estimating constant treatment effects with unobserved time-invariant heterogeneity
- And this also made it appealing for diff-in-diff – allowed you to "control for" many unobservables, like differences in questionnaires, differences in sites, etc.
- It's a linear model, so you'll be estimating conditional mean treatment effects – if you want the median, you can't use this
- Once you enter into a world with dynamic treatment effects and differential timing, standard specifications became perverse

When not to use it

- Simultaneous equations: cannot estimate demand curves with fixed effects
- Reverse causality: Becker predicted police reduce crime, but when you regress crime onto police, it's usually positive
- Time-varying unobserved heterogeneity

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

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

- We stack individual observations into matrices shown next slide

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

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

- 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?

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 estimates coefficients that minimizes:

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

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

What does fixed effects do to the data?

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.

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

The term $(c_i - \bar{c}_i) = 0$ because the mean of a constant is the constant itself

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

- Estimation using fixed effects
  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)`
- Estimator 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

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. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Roadmap

Background material
- Introduction
- Twoway fixed effects

TWFE Pathologies
- Bacon decomposition
- Simulation

Two solutions and a new decomposition
- CS
- SA

Example: Facebook and Mental Health

Strict exogeneity

- Fixed effects estimates the parameter of interest under strict exogeneity, and since it eliminates time invariant heterogeneity in panel settings, the appeal was widespread as more panel data came online
- But does this mean fixed effects estimates the ATT under differential timing – the answer is not necessariliy

Strict exogeneity

- Strict exogeneity assumed parallel trends *and* constant treatment effects – something not clearly understood until recently
- Treatment effect heterogeneity created a new composite error term that was correlated with both group and time fixed effects and the treatment
- Traditional panel econometrics notation can show this, but it's easier to see with potential outcomes notation

2x2 versus differential timing

- For this next part, we will decompose TWFE to understand what it needs for unbiasedness under differential timing
- All of this is from Goodman-Bacon (2022) though the expression of the weights is from 2018 for personal preference
- Goodman-Bacon (2022) shows that parallel trends is **not enough** for TWFE to be unbiased when treatment adoption is described by differential timing
- TWFE with differential timing uses treated groups as controls – not all estimators do – and this can introduce bias

Decomposition Preview

- TWFE estimates a parameter that is a weighted average over all 2x2 in your sample
- TWFE assigns weights that are a function of sample sizes of each "group" and the variance of the treatment dummies for those groups
- TWFE needs two assumptions: that the variance weighted parallel trends are zero (far more parallel trends iow) and 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

K^2 distinct DDs

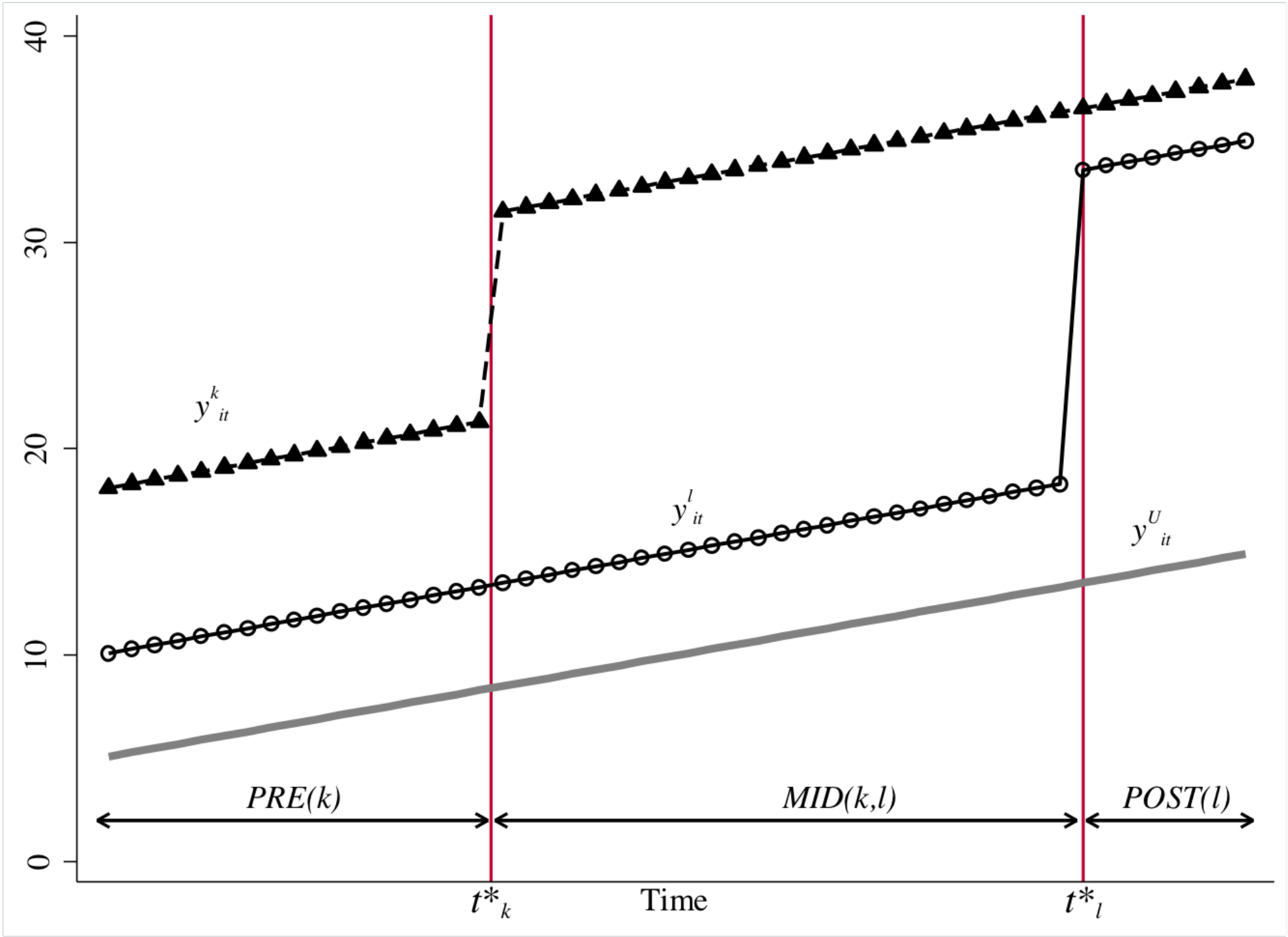
Let's look at 3 timing groups (a, b and c) and one untreated group (U). With 3 timing groups, there are 9 2x2 DDs. Here they are:

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

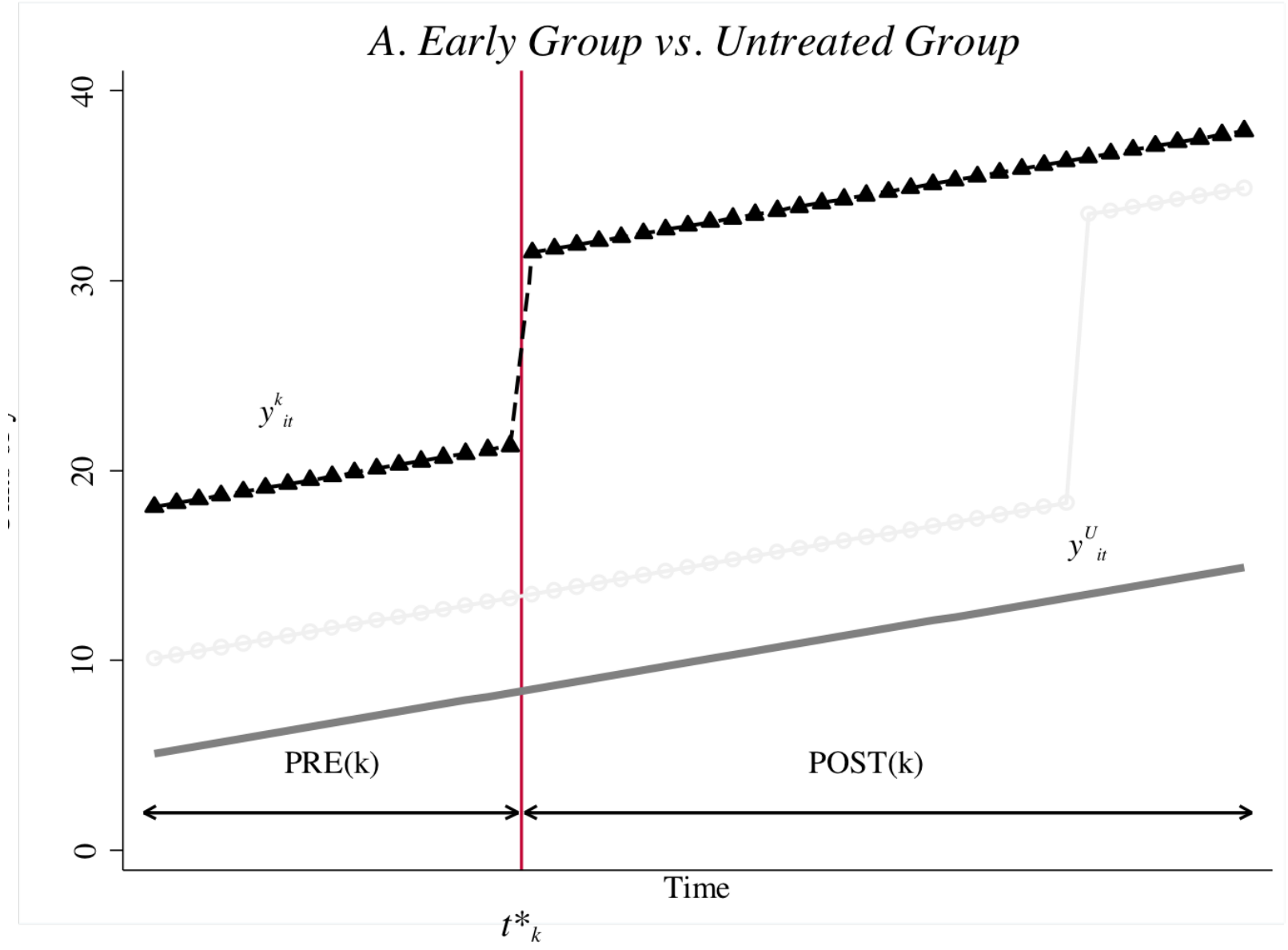
Let's return to a simpler example with only two groups – a k group treated at t_k^* and an l treated at t_l^* plus an never-treated group called the U untreated group

Terms and notation

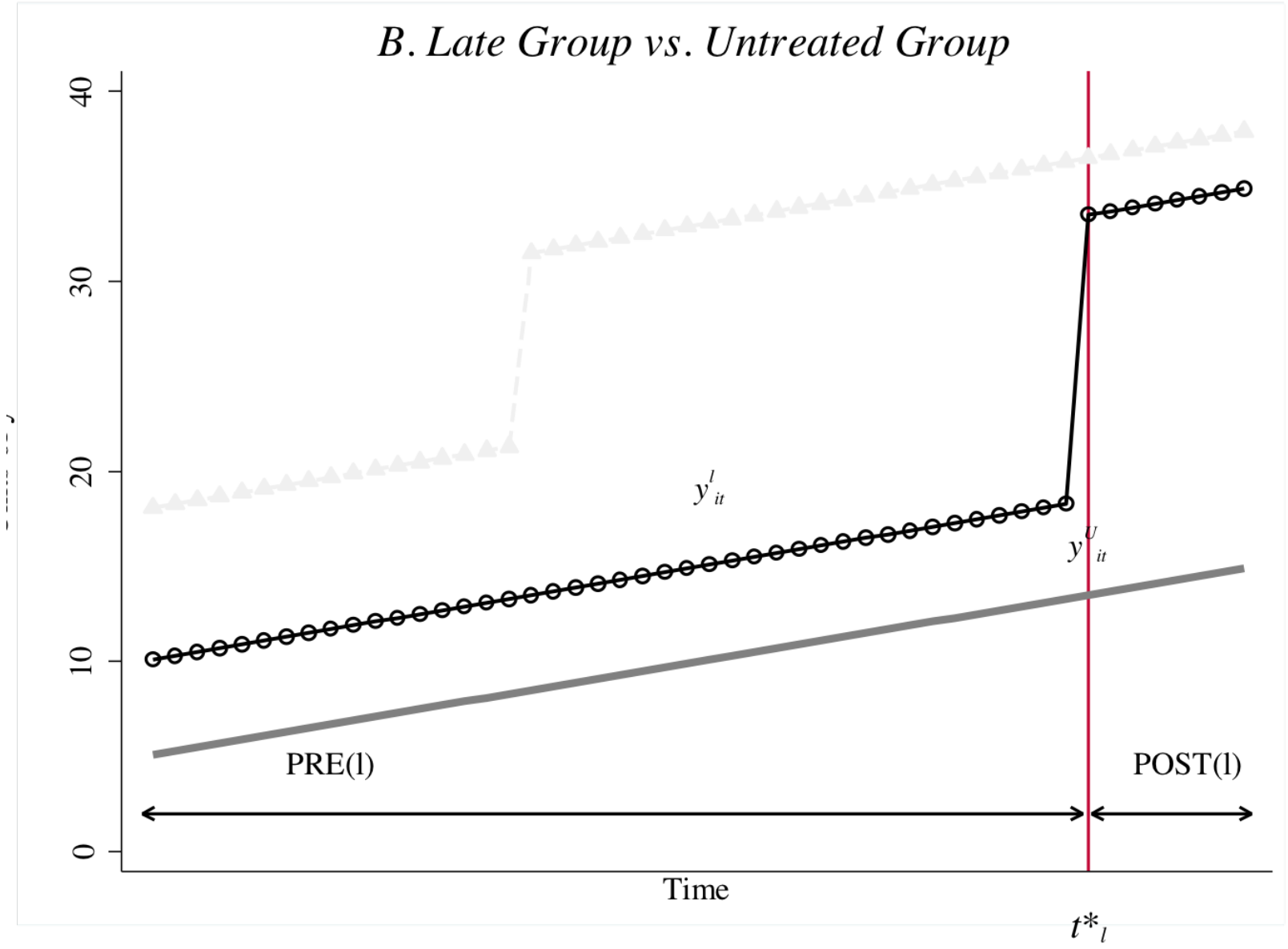
- Let there be two treatment groups (k,l) and one untreated group (U)
- k,l define the groups based on when they receive treatment (differently in time) with k receiving it earlier than l
- Denote $\overline{D}_k$ as the share of time each group spends in treatment status
- Denote $\widehat{\delta}_{jb}^{2x2}$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

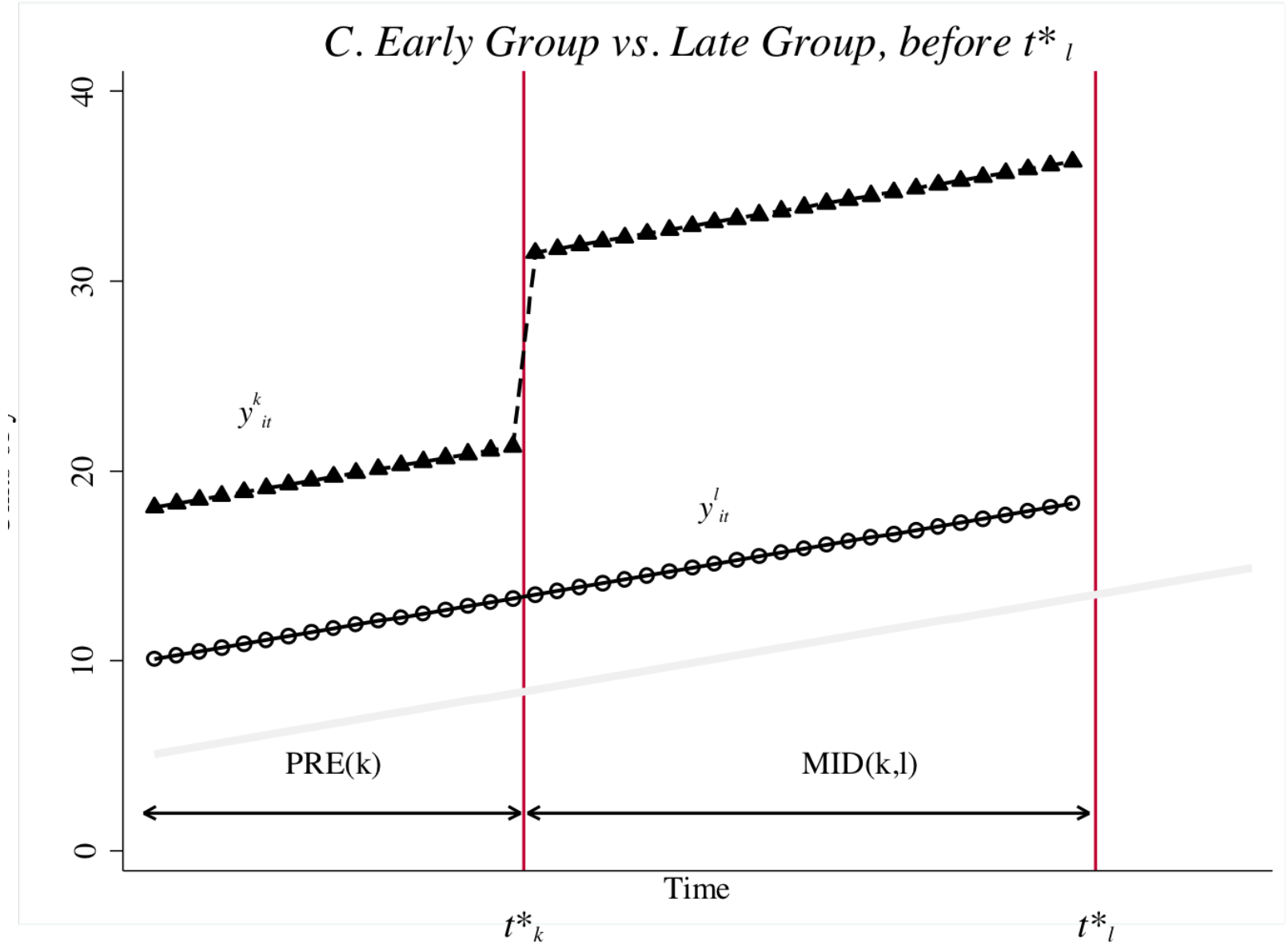
$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)} \right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)} \right)$$



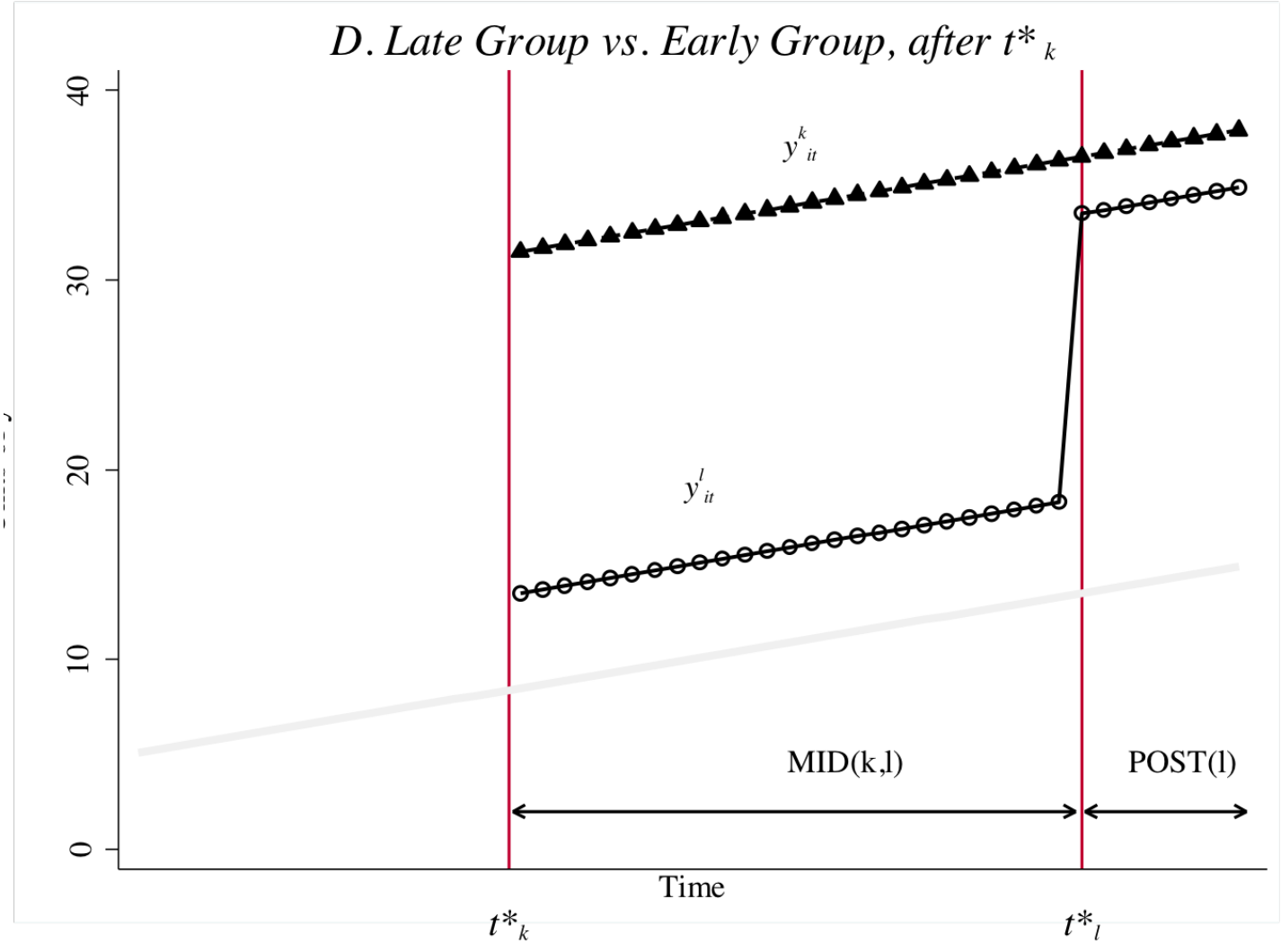
$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_{l}^{post(l)} - \overline{y}_{l}^{pre(l)}\right) - \left(\overline{y}_{U}^{post(l)} - \overline{y}_{U}^{pre(l)}\right)$$



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



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



Bacon decomposition

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

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

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

Third, the Weights

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

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

Weights discussion

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

More weights discussion

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

Summarizing TWFE centralities

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

Moving from 2x2s to causal effects and bias terms

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

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

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

The dangerous 2x2

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

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

Substitute all this stuff into the decomposition formula

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

where we will make these substitutions

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

Notice all those potential sources of biases!

Potential Outcome Notation

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

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

Simulated data

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

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

- Also no anticipation of treatment effects until treatment occurs but does *not* guarantee homogenous treatment effects

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
- Twoway fixed effects

TWFE Pathologies
- Bacon decomposition
- Simulation

Two solutions and a new decomposition
- CS
- SA

Example: Facebook and Mental Health

Causal inference is imputation

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

Causal inference involves imputation

- Causal inference is a missing data problem – we are missing counterfactuals
- And recall that estimating the ATT necessarily involved correctly imputing the counterfactual using parallel trends
- OLS, therefore, is *implicitly* imputing counterfactuals for estimating the ATT

Callaway and Sant'Anna 2020

CS is a DiD model used for estimating ATT parameters under differential timing and conditional parallel trends

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

- 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

- New Jersey treated in late 1992, New York in late 1993, Pennsylvania never treated
- What years are each state's post-treatment?
  - → New Jersey: post-1992
  - → New York: post-1993
  - → Pennsylvania: ?
- How did people go about event studies then?

Early efforts at event studies

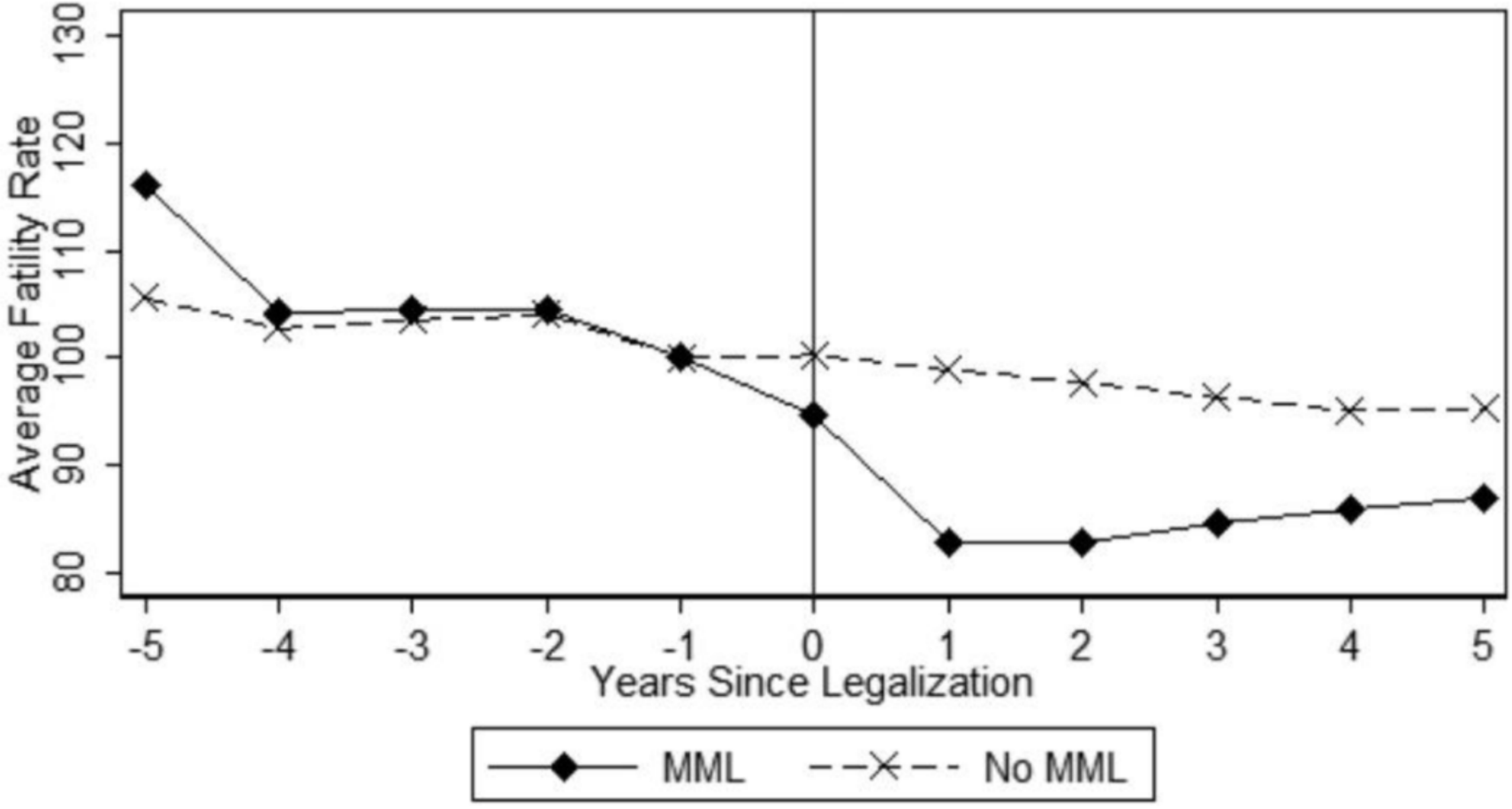


Figure: Anderson, et al. (2013) display of raw traffic fatality rates for re-centered treatment states and control states with randomized treatment dates

Replicated from a project of mine

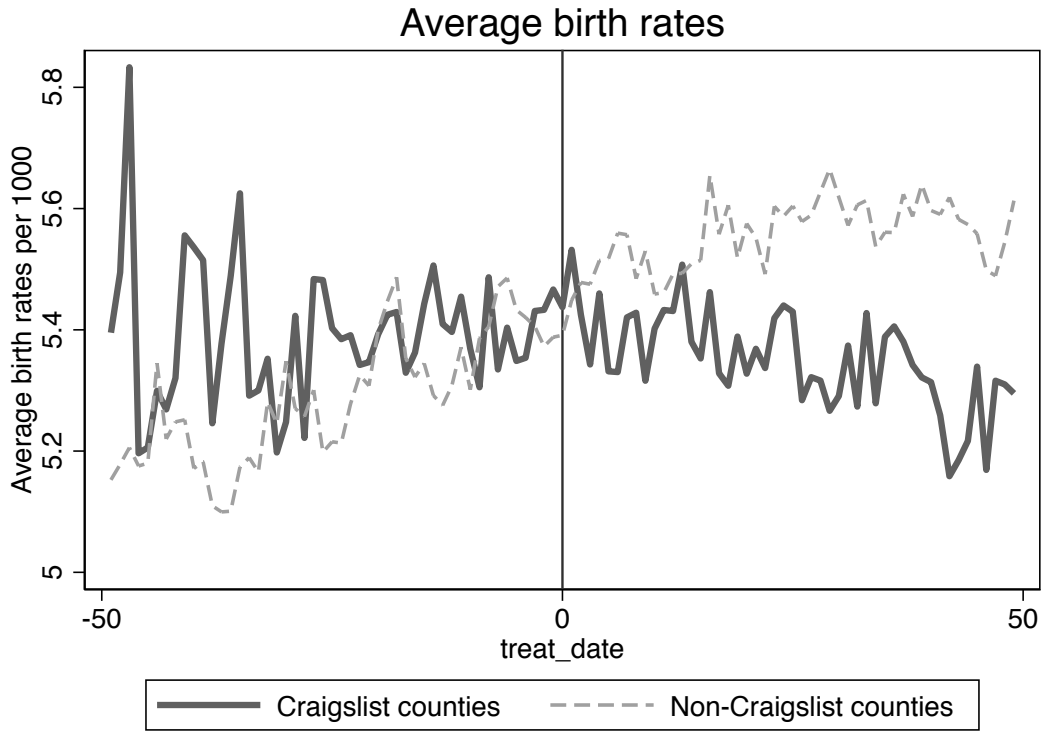
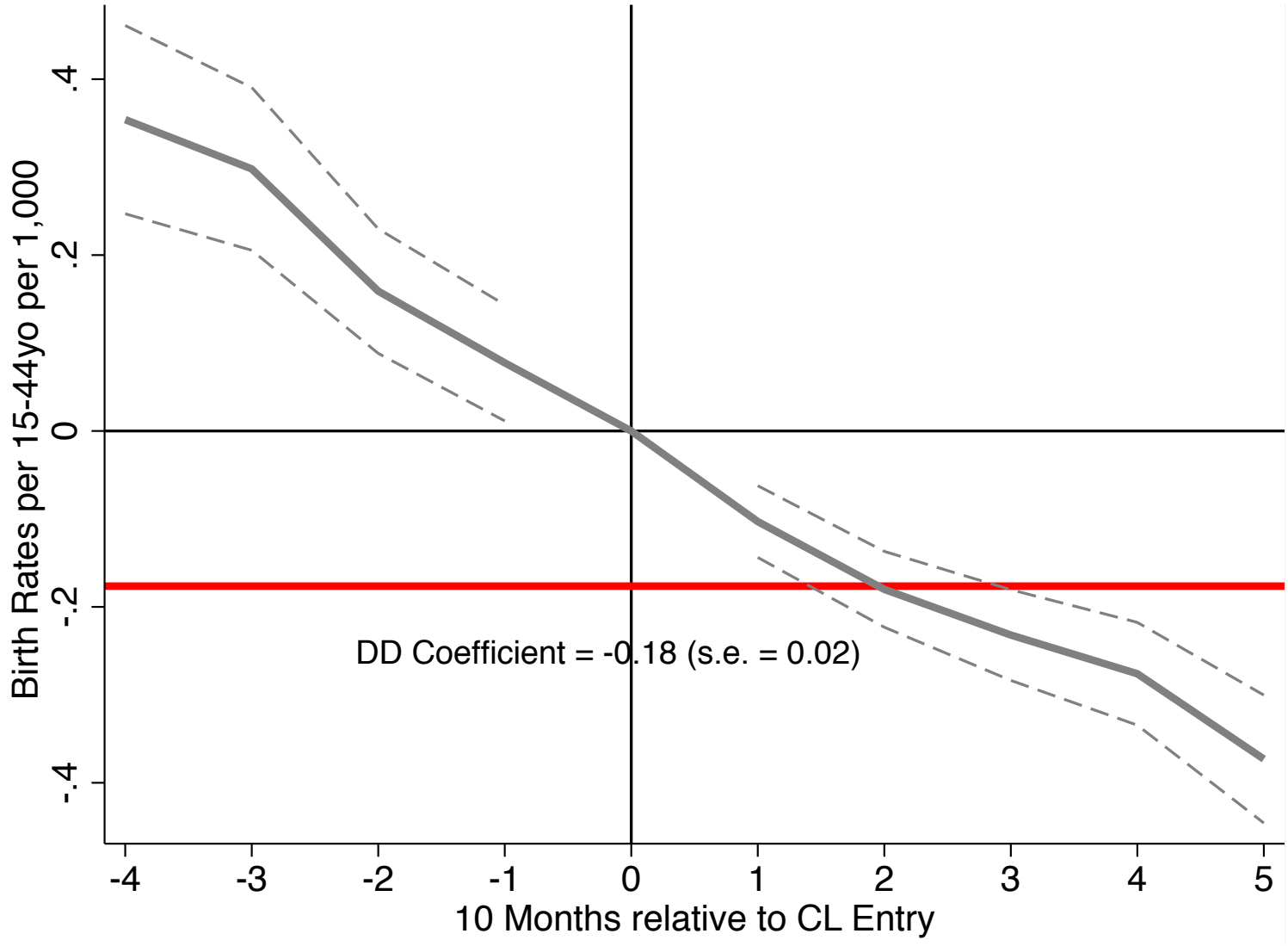


Figure: From one of my studies. Looks decent right?

Canonical 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. This model, it turned out, suffered from model misspecification.



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

Sun and Abraham 2020

- Now that we know about the biases of the constant treatment effect model estimated with TWFE, let's revisit event studies under differential timing
- Goodman-Bacon (2021, forthcoming) focused on decomposition of TWFE to show bias under differential timing
- Callaway and Sant'anna (2020) presents alternative estimator that yields unbiased estimates of group-time ATTs which can be aggregated or put into event study plots
- Sun and Abraham (SA) is like a combination of the two papers

Summarizing (cont.)

1. SA is 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
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)

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

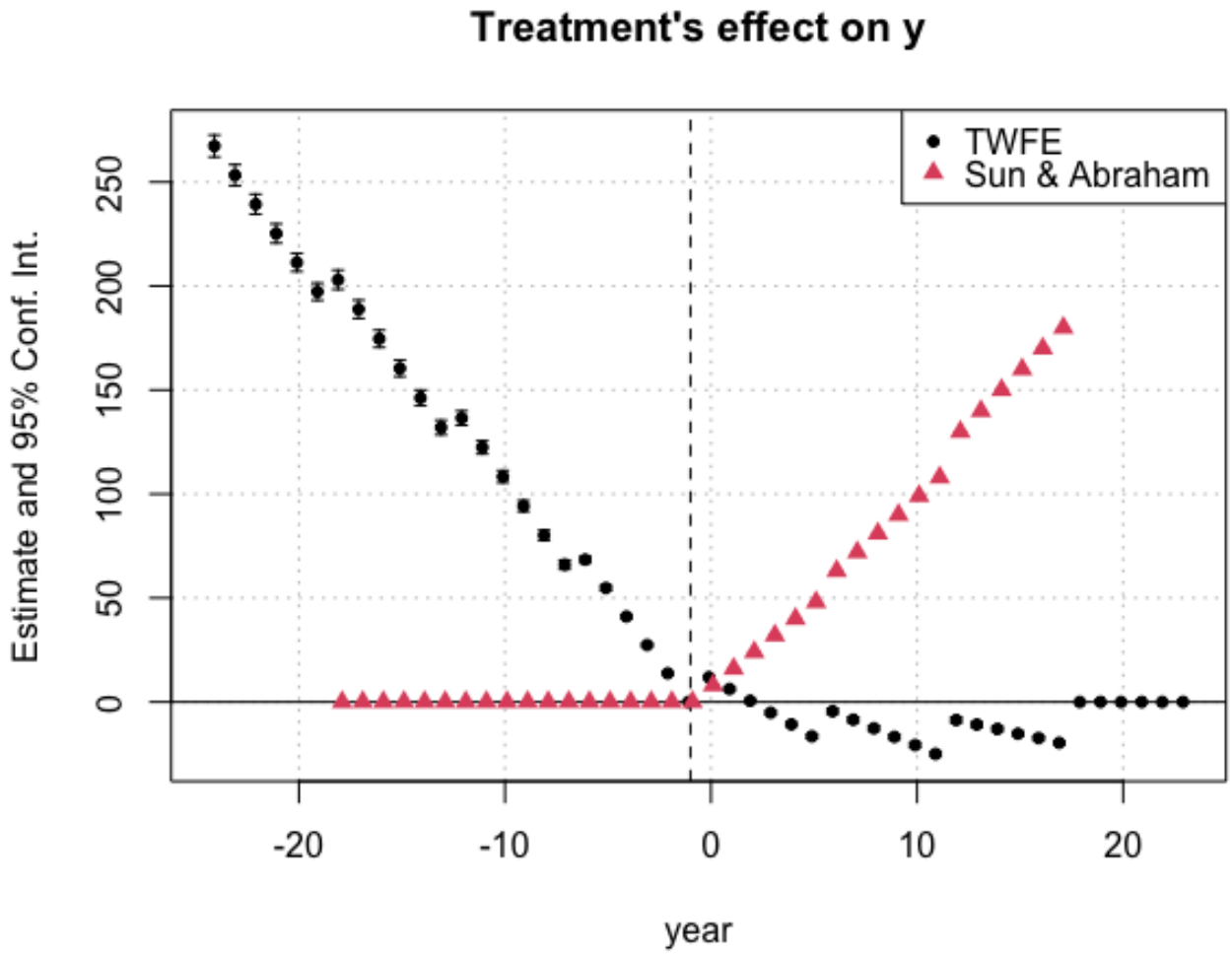
	(Truth)	**(TWFE)**	**(CS)**	**(SA)**	**(BJS)**
$\widehat{Feasible\ ATT}$	68.33	26.81***	68.34***	68.33***	

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?

Roadmap

Background material
- Introduction
- Twoway fixed effects

TWFE Pathologies
- Bacon decomposition
- Simulation

Two solutions and a new decomposition
- CS
- SA

Example: Facebook and Mental Health

Bringing them together

- Examine a paper that is contemporary with respect to all that we've reviewed
- Staggered rollout, important question, high quality data
- Braghieri, Levy and Makarin (2022), "Social Media and Mental Health", *American Economic Review*, 112(11): 3660-3693

Big picture

- Widely cited that social media causes mental health problems in youth
- Anecdotal, documentaries, but no causal evidence (“slim to none”)
- Study will use staggered rollout of Facebook platform to college campuses from 2004 to 2006 to estimate the effect on aggregate mental health scores from a survey
- You be the judge, but they present what in most cases would be strong evidence that Facebook harmed college students mental health

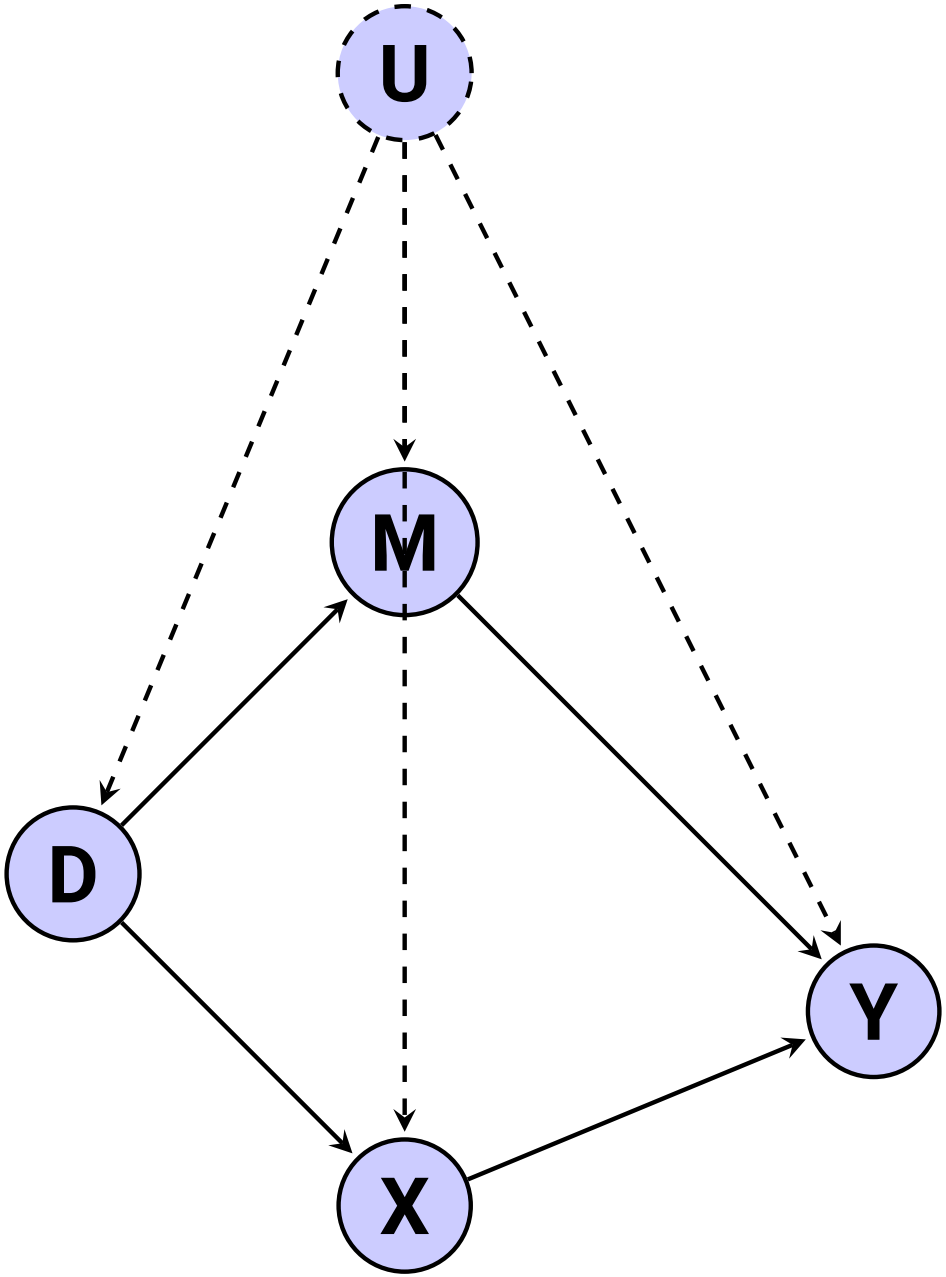
Many things to like

- Important question: mental health, suicide, review descriptive stats together
- Strong design: staggered rollout
- Event study is eye popping
- Mechanism and main results
- Very interesting dataset

Fives parts of a strong DiD

1. **Bite**: They cannot really show much here. No data on Facebook usage. More an ITT
2. **Main Results**: Estimated effect on mental health measures
3. Mechanism: Speculative
4. **Falsifications**: I can't really see very strong falsifications either.
5. **Event studies**: POW. Just wait

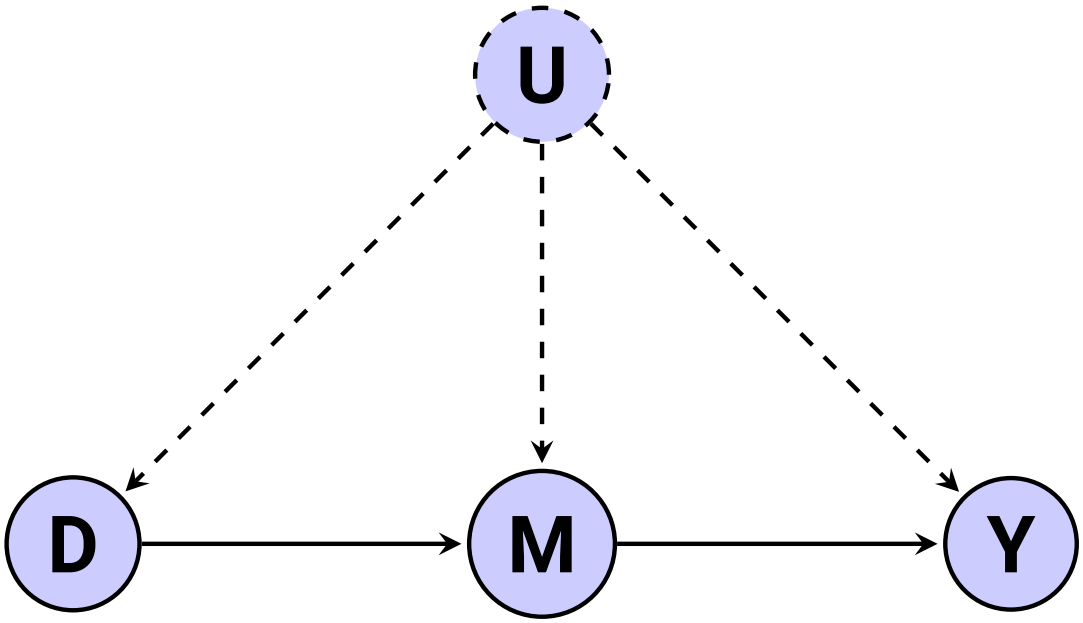
Mechanism



Mechanism

- D is the treatment variable, and the ATT is over all possible channels, but what if you want to think M is the mechanism
- When you can't rule out competing theories with falsifications, you have to try and build the case that the effect is coming through a channel
- Rule out X and provide evidence for M
- Goal here is to try and present evidence (not proof) that it's probably the story you're saying

Ruling out alternative mechanism



Mechanism

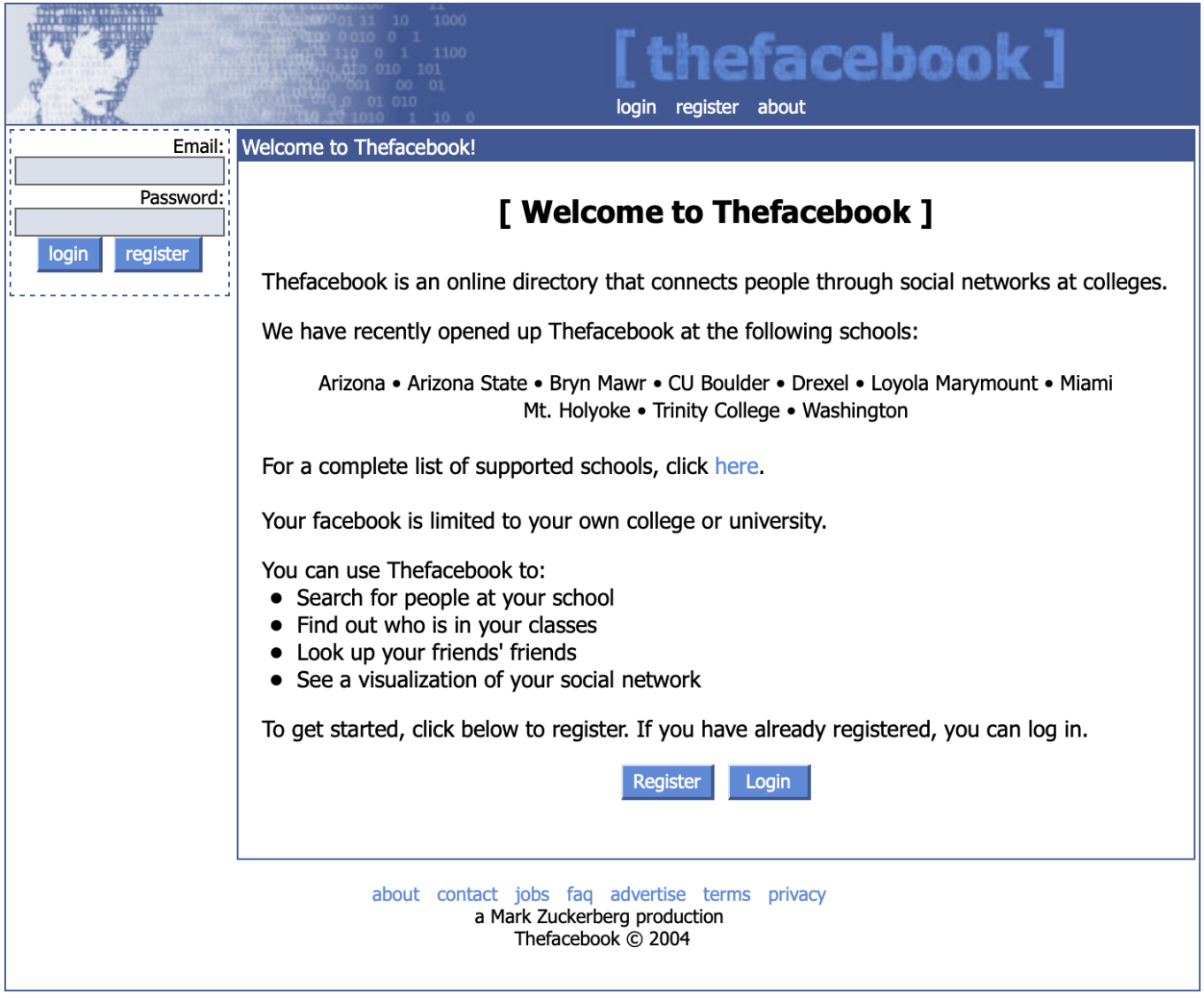
- Story is interpersonal comparisons which they try to show
- We can discuss how plausible we found it, but ask yourself at the end – did the event study help you believe it? Why/why not?

Data on Facebook

- Ingenious use of the Wayback Time Machine
- Looked at over 700 schools using Facebook screen shots
- When Facebook first mentions a school on its front page, that school is marked as having gotten Facebook

New schools being adopted

New schools being adopted



Data on college students

- NCHA Data is survey administered to college students on a semi-annual basis by American College Health Assoc
- Inquires about demographics, physical health, mental health, alcohol and drug use, sexual behaviors, and perception of these behaviors by peers
- ACHA merged a treatment indicator to each respondent based on Facebook dataset provided to them so that privacy could be maintained

Mental health

- Self-reported symptoms are standard medical practice in mental health – DSM-5 relies on self-reports such as difficulty sleeping, fatigue, feelings of guilt, suicidal ideation
- No data on Facebook or social media usage so this is ITT version of the ATT
- Respondent answers to the questions are aggregated into indices such as *poor mental health* where larger numbers are worse

Main TWFE Model

$$Y_{icgt} = \alpha_g + \delta_t + \beta \times Facebook_{gt} + X_i \times \gamma + X_c \times \psi + \varepsilon_{icgt} \quad (1)$$

Y_{icgt} is an outcome for person i in wave t attending college c in expansion group g; α_g is expansion-group or college fixed effects; δ_t are survey-wave fixed effects; $Facebook_{gt}$ indicates the respondents' campus has Facebook by time t at expansion group g; X_i and X_c are individual and college-level controls; and standard errors are clustered at college level.

$\widehat{\beta}$ identifies the ATT under parallel trends in the robust models

Robustness

- Main static results will all be in TWFE, but appendix shows other methods like CS and SA
- Event studies will show all models including some we haven't reviewed
- Growing popularity to show "all the robust DiD" models so that readers can see you aren't cherry picking

TABLE 1—BASELINE RESULTS: INDEX OF POOR MENTAL HEALTH

	Index of poor mental health			
	(1)	(2)	(3)	(4)
Post-Facebook introduction	0.137	0.124	0.085	0.077
	(0.040)	(0.022)	(0.033)	(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.

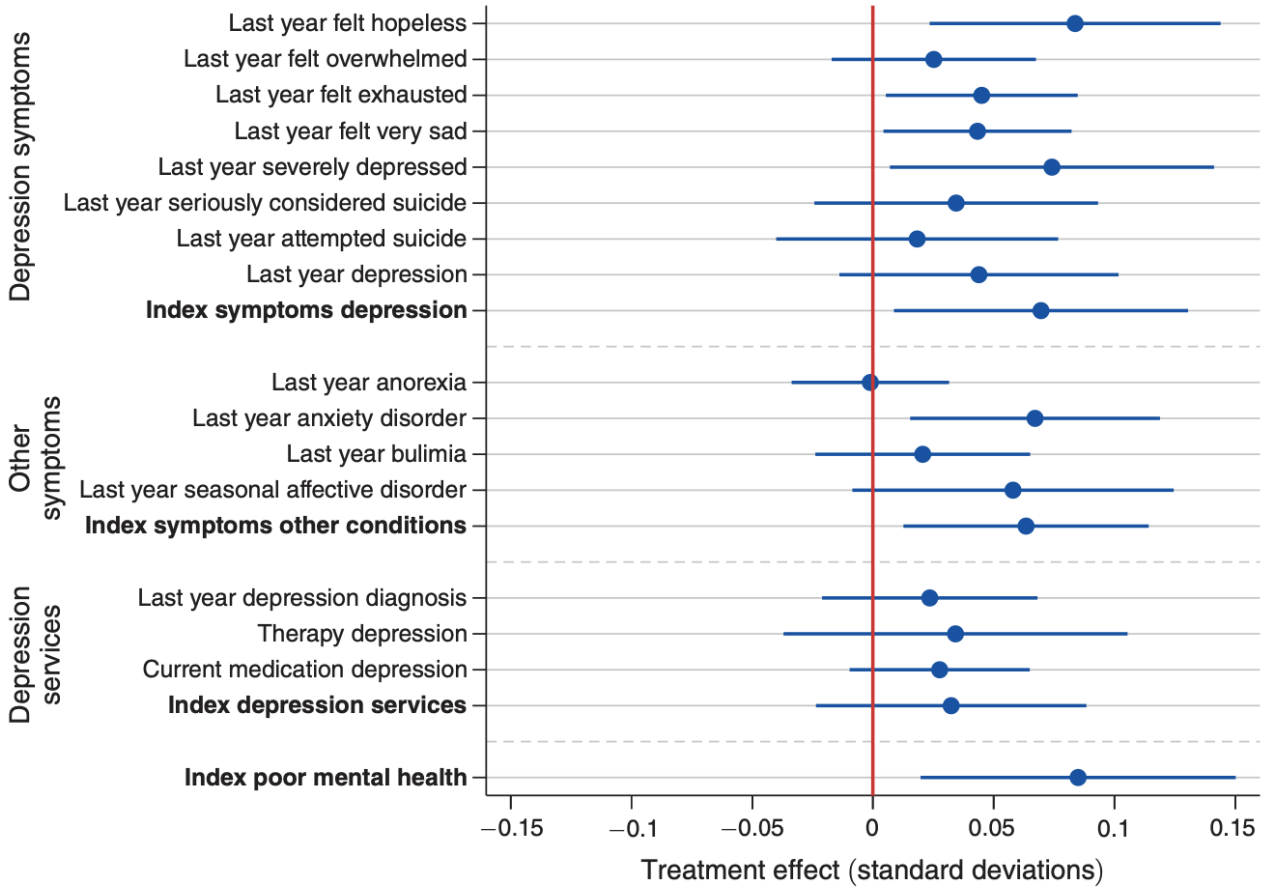


FIGURE 1. EFFECTS OF THE INTRODUCTION OF FACEBOOK ON STUDENT MENTAL HEALTH

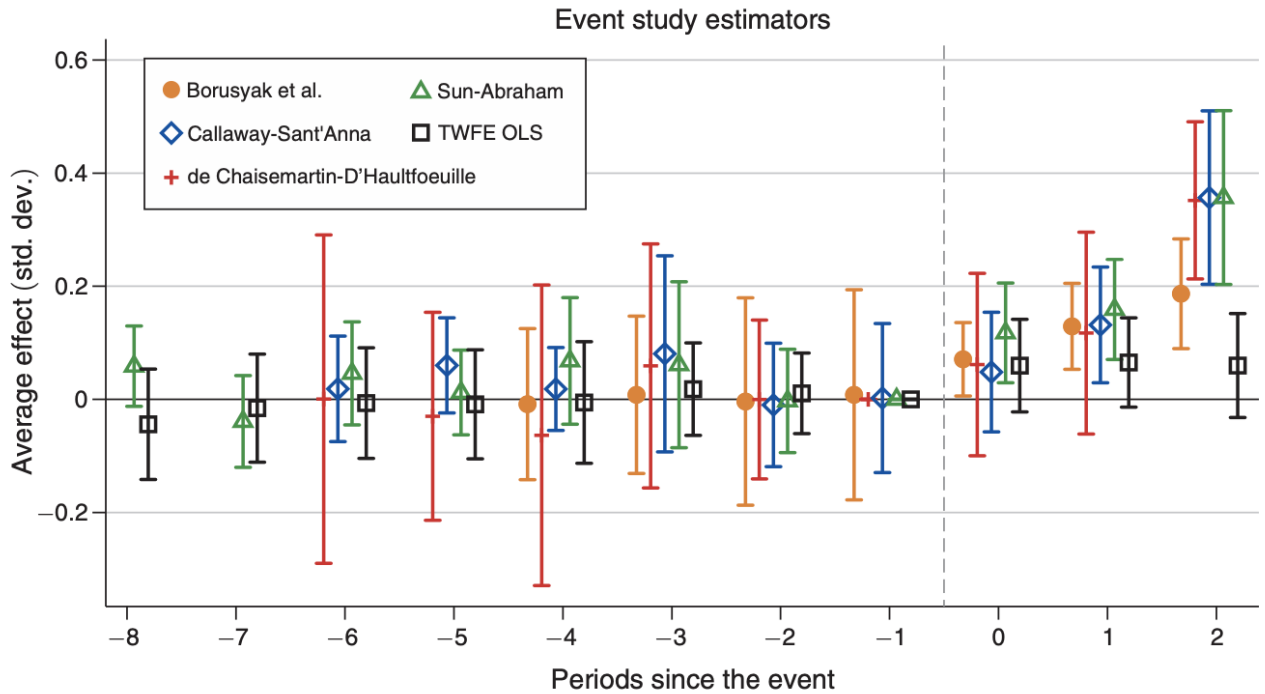


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

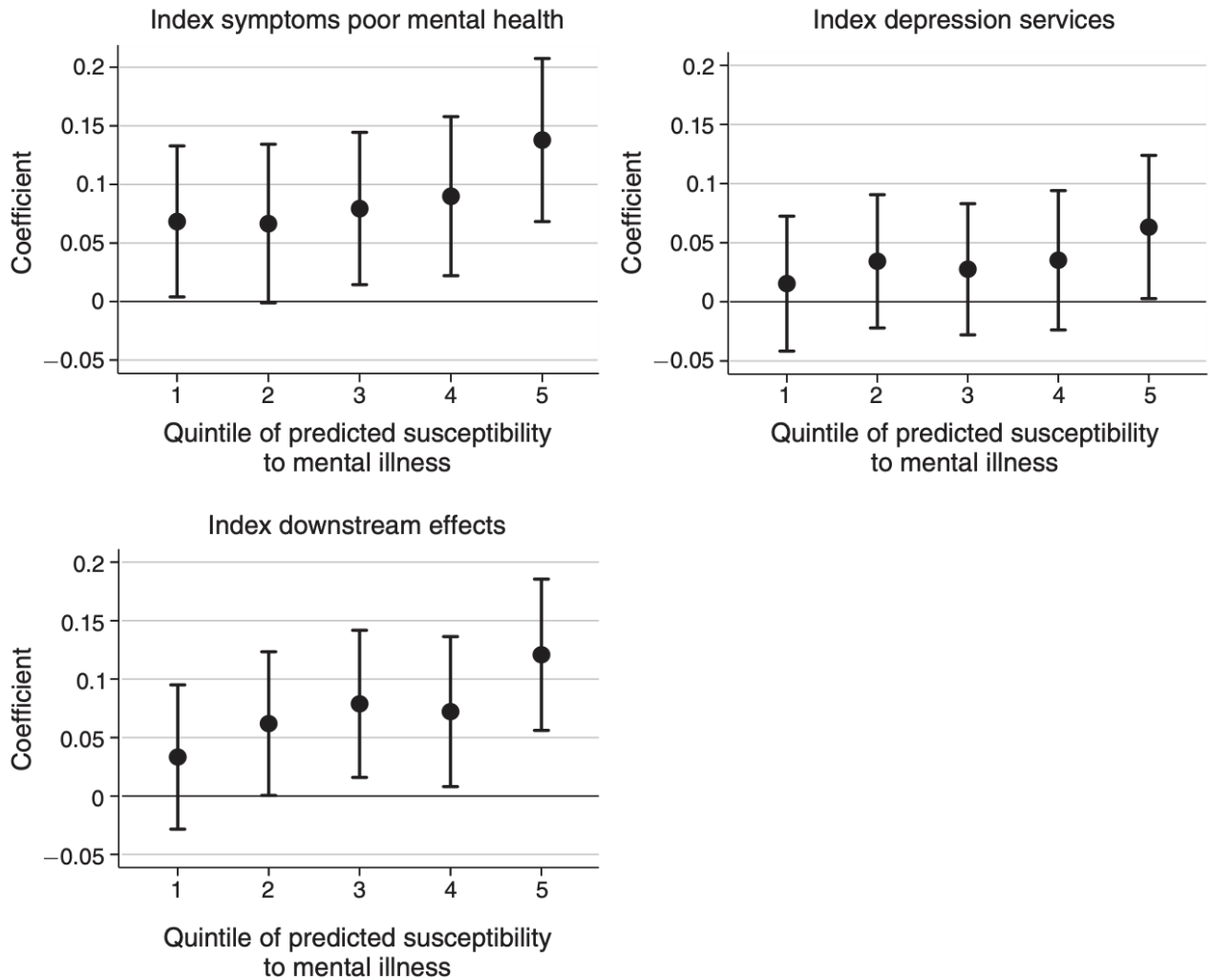


FIGURE 3. HETEROGENEOUS EFFECTS BY PREDICTED SUSCEPTIBILITY TO MENTAL ILLNESS

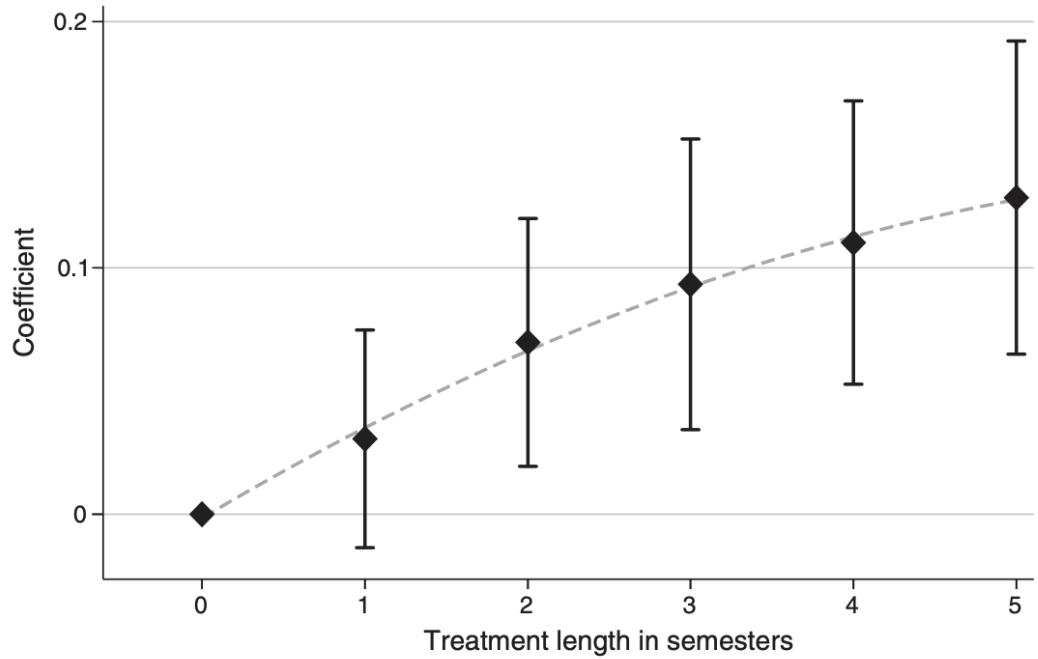


FIGURE 4. EFFECT ON POOR MENTAL HEALTH BY LENGTH OF EXPOSURE TO FACEBOOK

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.

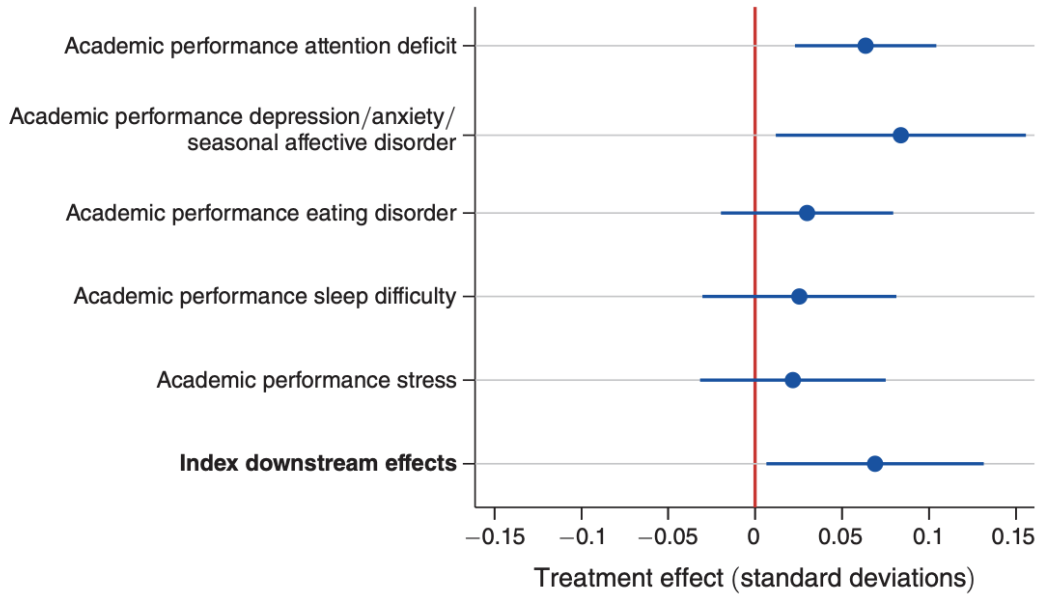


FIGURE 5. DOWNSTREAM EFFECTS ON ACADEMIC PERFORMANCE

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.

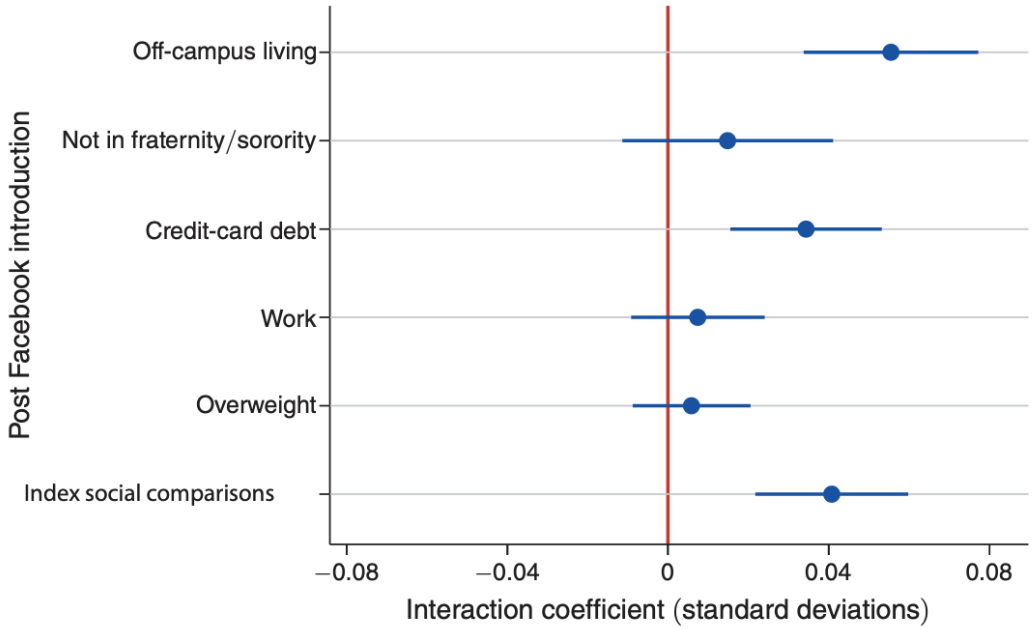


FIGURE 6. HETEROGENEOUS EFFECTS AS EVIDENCE OF UNFAVORABLE SOCIAL COMPARISONS

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.

Conclusion

- Good question, good data, and you can publish well with DiD
- First evidence that social media caused mental health to worsen
- Hardly definitive, but the staggered design is a solution to our inability to run the RCT
- Remember – many questions can be randomized in theory but not practice (e.g., smoking)
- Clearly more work needs to be done, but this is a start