Causal Inference II

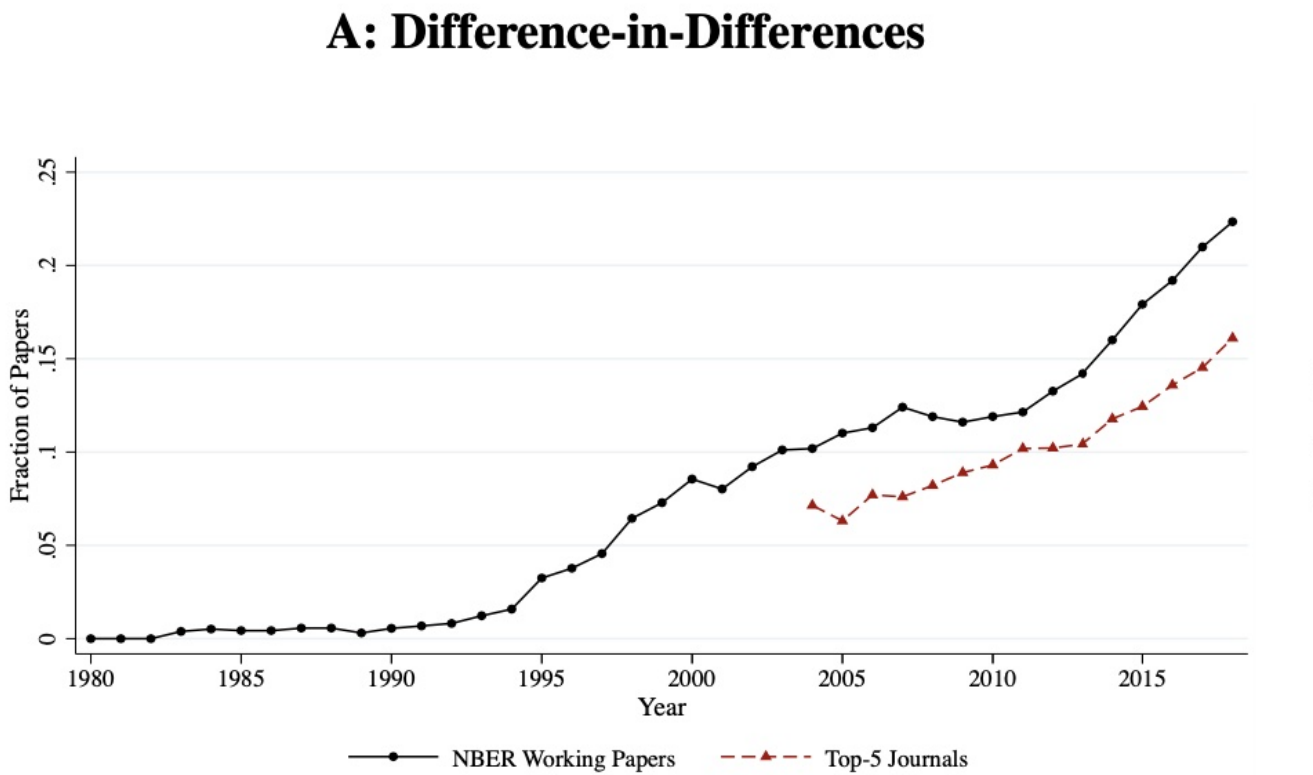
Mixtape Session

Roadmap

- Welcome to Differential Timing
  - Diff-in-diff credibility crisis
  - TWFE Pathologies
  - Simulation
- Robust Diff-in-Diff Estimators
  - CS
  - SA
  - dCH
  - Imputation based robust estimator
  - 2SDiD
- Examples
  - Facebook and Mental Health
  - Castle doctrine
- Basic suggestions going forward

Diff-in-diff had belonged to the empiricists

Figure: Currie, et al. (2020)



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

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

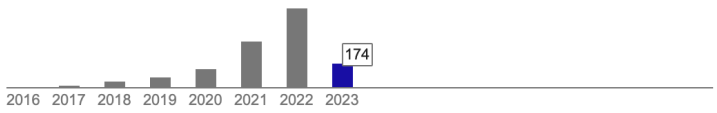
- Beginning in 2016-2017, several grad students and assistant professors found critical pathologies with standard TWFE specifications and developed solutions
- Many simultaneous discoveries, some redundancies, and **sudden** awareness of the issues started happening around 2017, eventually became a massive thing
- Extreme meteoric rise, unusual for econometrics

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, Jahn Spiess, just accepted this week 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), may have been in working paper in 2017 (at least 2018)
- Very thorough decomposition of the TWFE pathology, very general solution, included Stata 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 (but no solution)
- 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 published solution to the problem, written while assistant professors at Vanderbilt and Ole Miss,
- Pedro is a UC3M alum (2015 grad) and Brantly is a Vanderbilt grad
- Both are now coauthors with Andrew Goodman-Bacon
- Introduced new terms like group-time ATT, released very tight R code (“did”)

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 published solution to the problem, very similar to CS
- Focus was on decomposing the event study
- Written while grad students at MIT but Sophie Sun is now an assistant professor at CEMFI!

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

There’s even more and more coming

- 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

Diff-in-diff and OLS

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

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

- Orley claims that the TWFE estimator of δ and the "four averages and three subtractions" are the same thing numerically
- Without covariates, they are numerically *identical*
- But this model requires only one treated unit (i.e., Post is undefined for more than one treatment group)

Discussion of estimate

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

- The above is the workhorse model with differential timing – twoway fixed effects (year and panel unit fixed effects additive in the model)
- If you estimate this model under differential timing of the treatment, what parameter then does $\widehat{\delta}$ correspond to?
- Answer: It also corresponds to the previous "four averages and three subtractions" – but it's numerous of them, not just one

Decomposition Preview

- Andrew Goodman-Bacon decomposed $\widehat{\delta}$ and showed it is numerically identical to a weighted average of all “four averages and three subtractions”
- But, even before we get to causality there are unusual features
- TWFE model assigns its own weights which are a function of the size of a “group” and the variance of group treatment dummies

K^2 distinct DDs

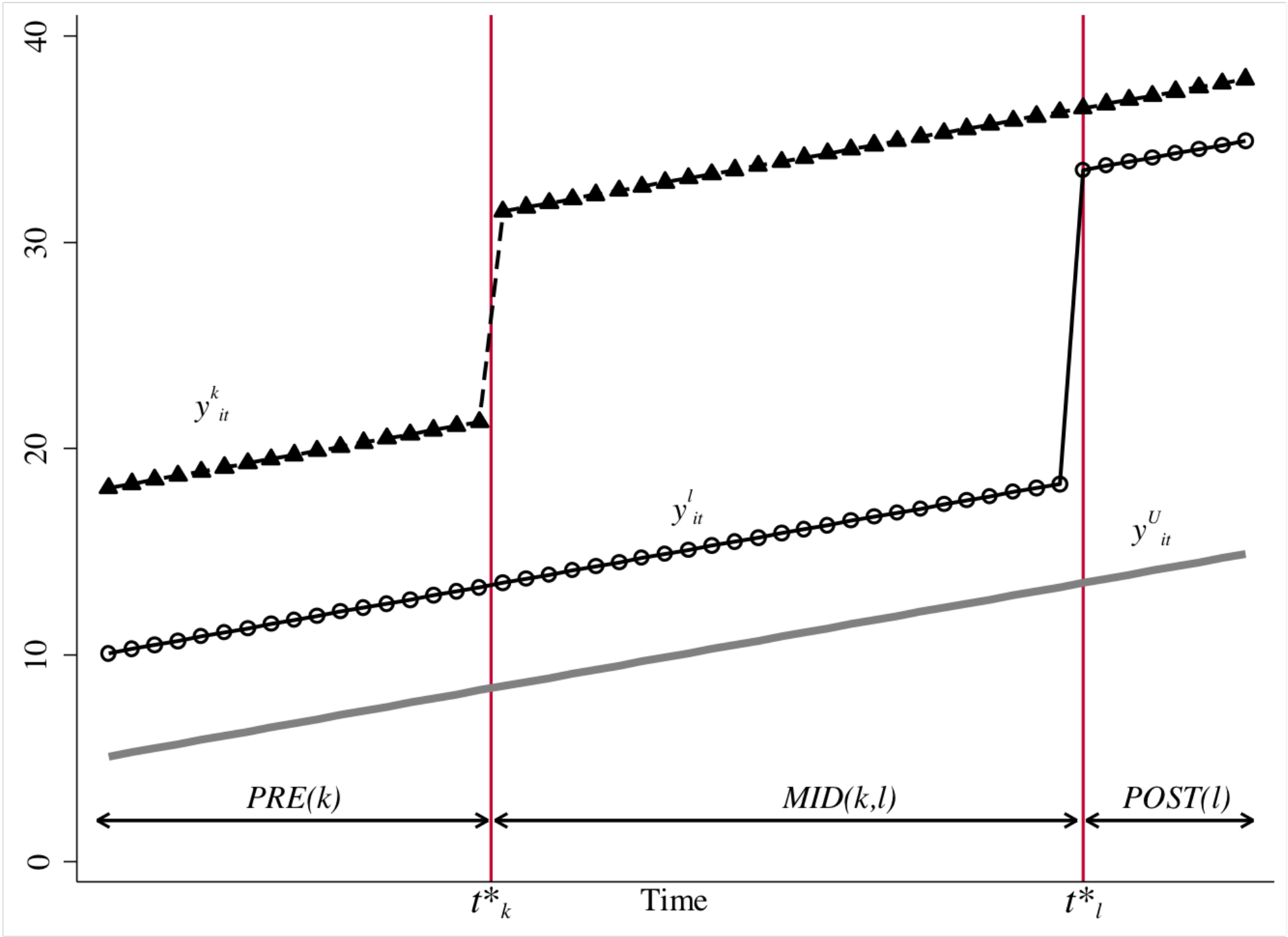
Let's look at 3 timing groups (a, b and c) and one untreated group (U). With 3 timing groups, there are 9 2x2 DDs. Here they are:

a to b	b to a	c to a
a to c	b to c	c to b
a to U	b to U	c to U

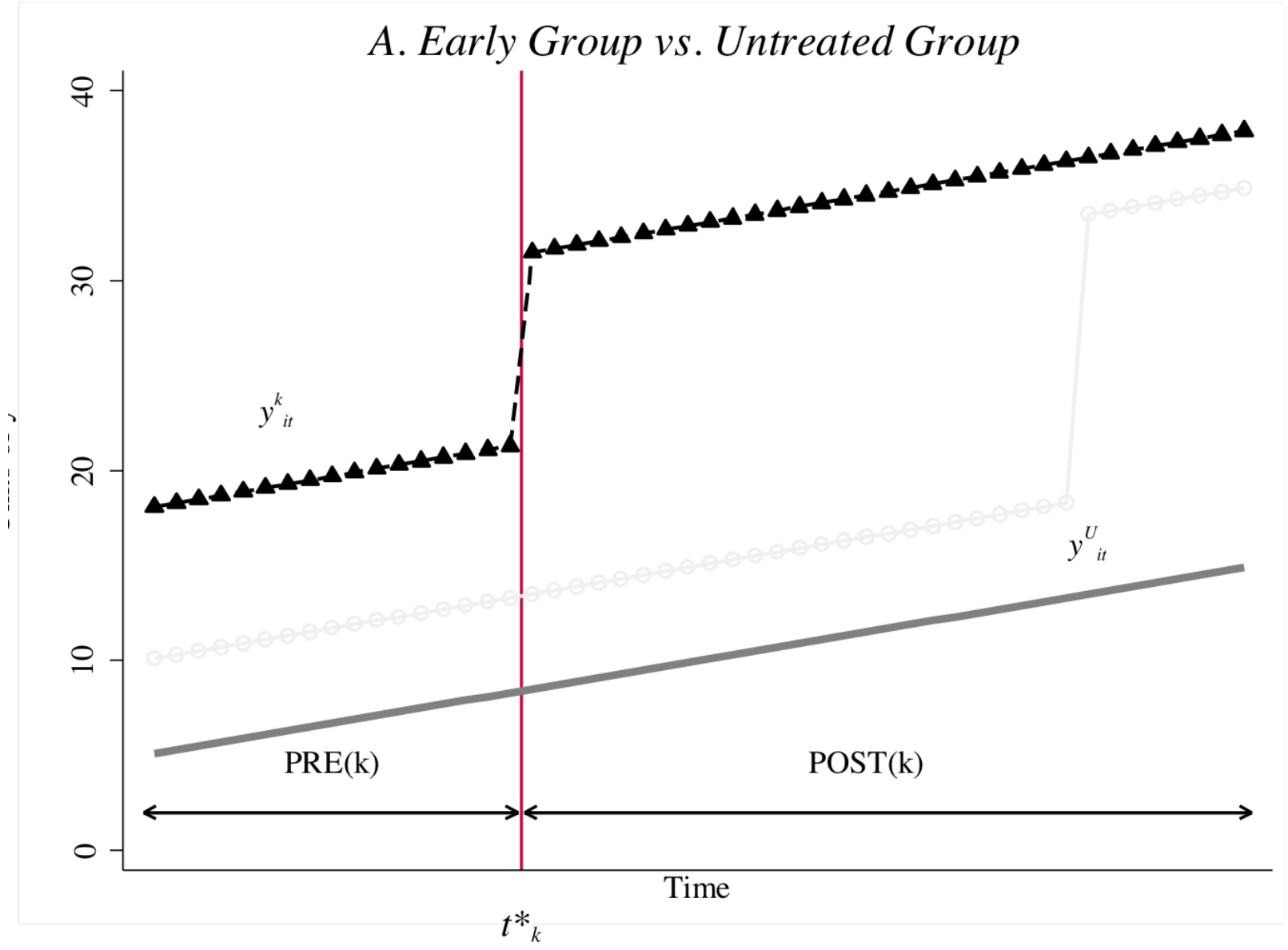
Let's return to a simpler example with only two groups – a k group treated at t_k^* and an l treated at t_l^* plus an never-treated group called the U untreated group

Terms and notation

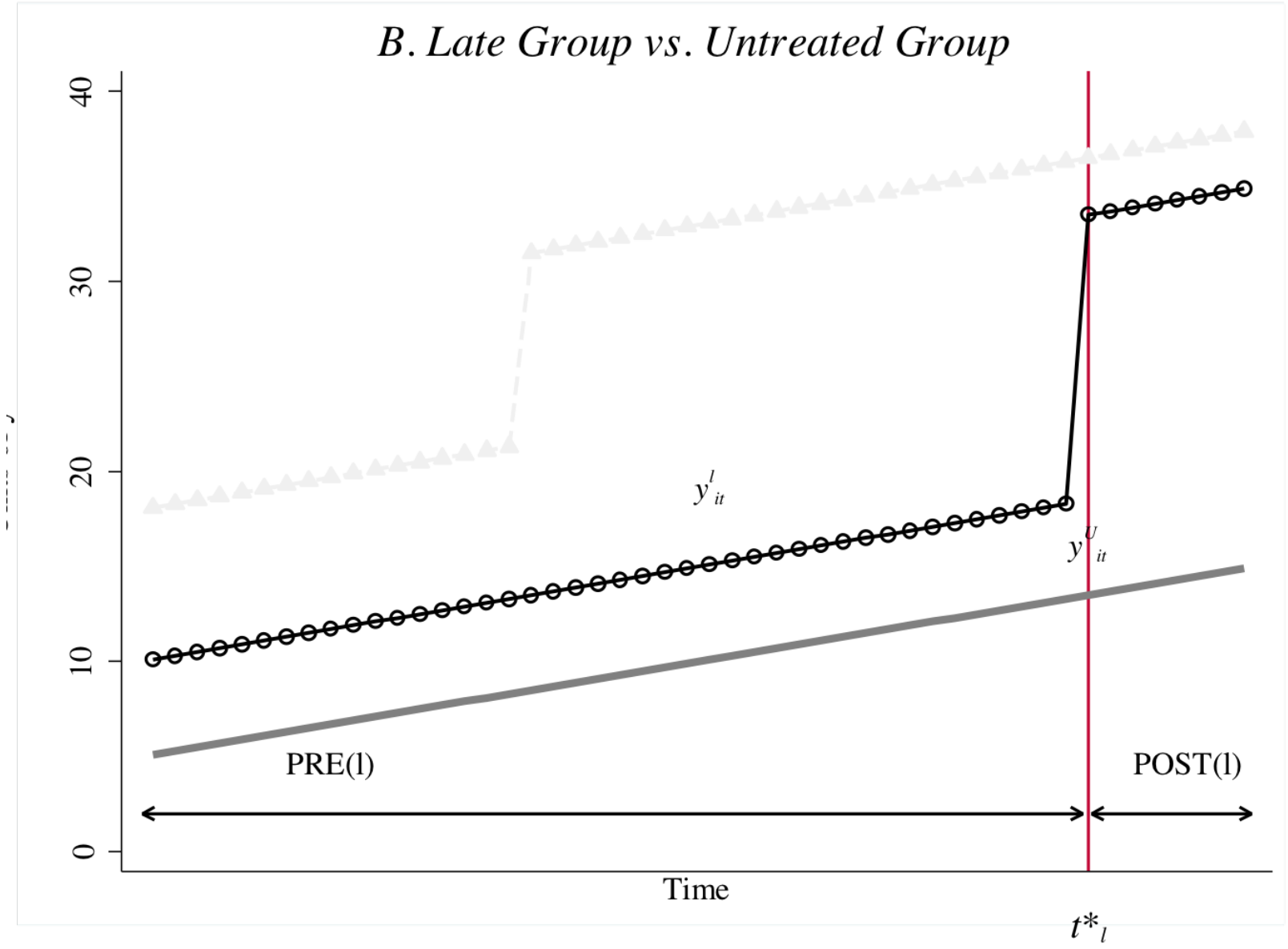
- Let there be two treatment groups (k,l) and one untreated group (U)
- k,l define the groups based on when they receive treatment (differently in time) with k receiving it earlier than l
- Denote $\overline{D}_k$ as the share of time each group spends in treatment status
- Denote $\widehat{\delta}_{jb}^{2x2}$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


40
30
20
10
0
y^{k}_{it}
y^{l}_{it}
y^{U}_{it}
PRE(k)
MID(k,l)
POST(l)
t^{*}_{k}
Time
t^{*}_{l}

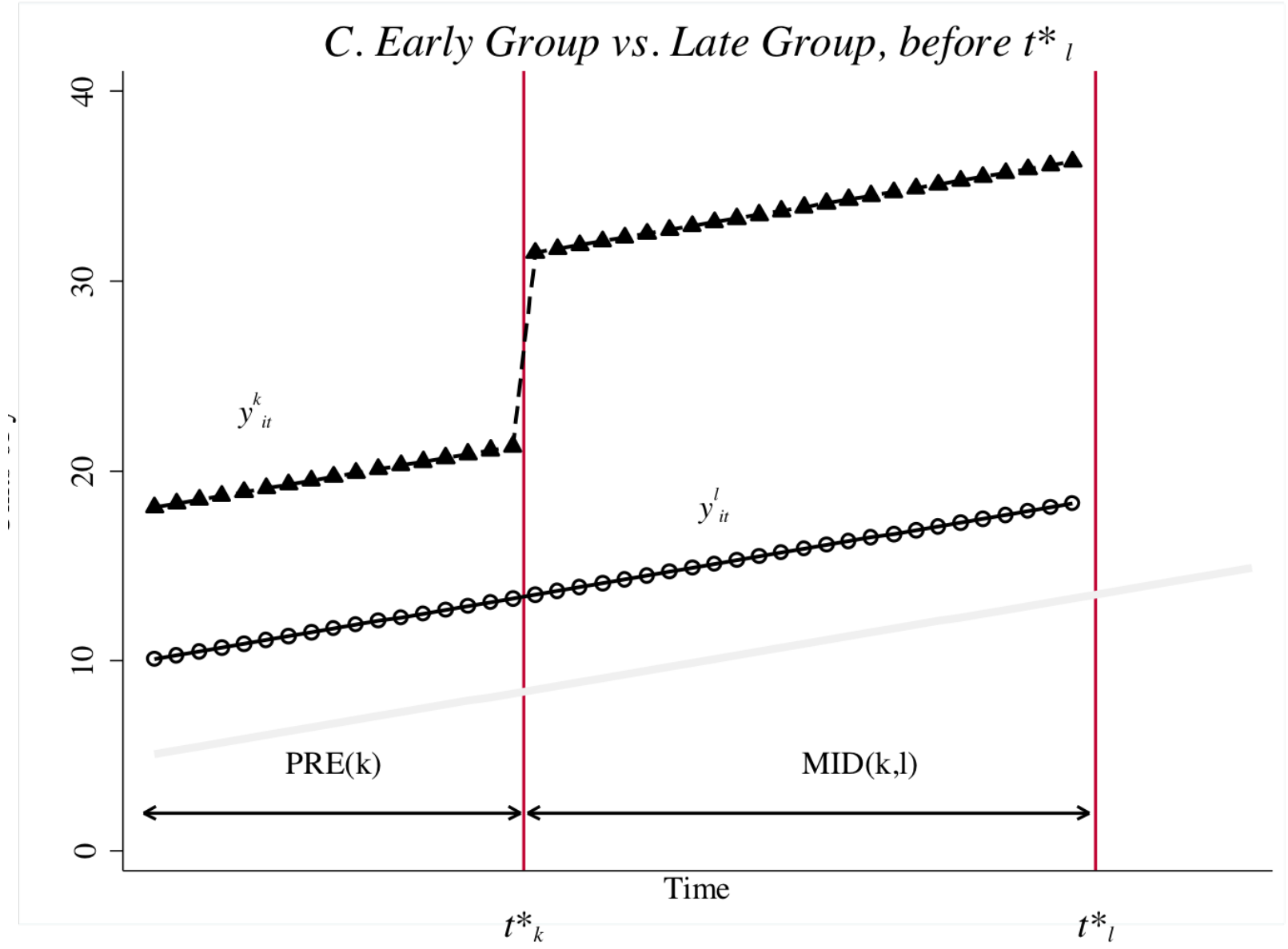
$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)} \right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)} \right)$$



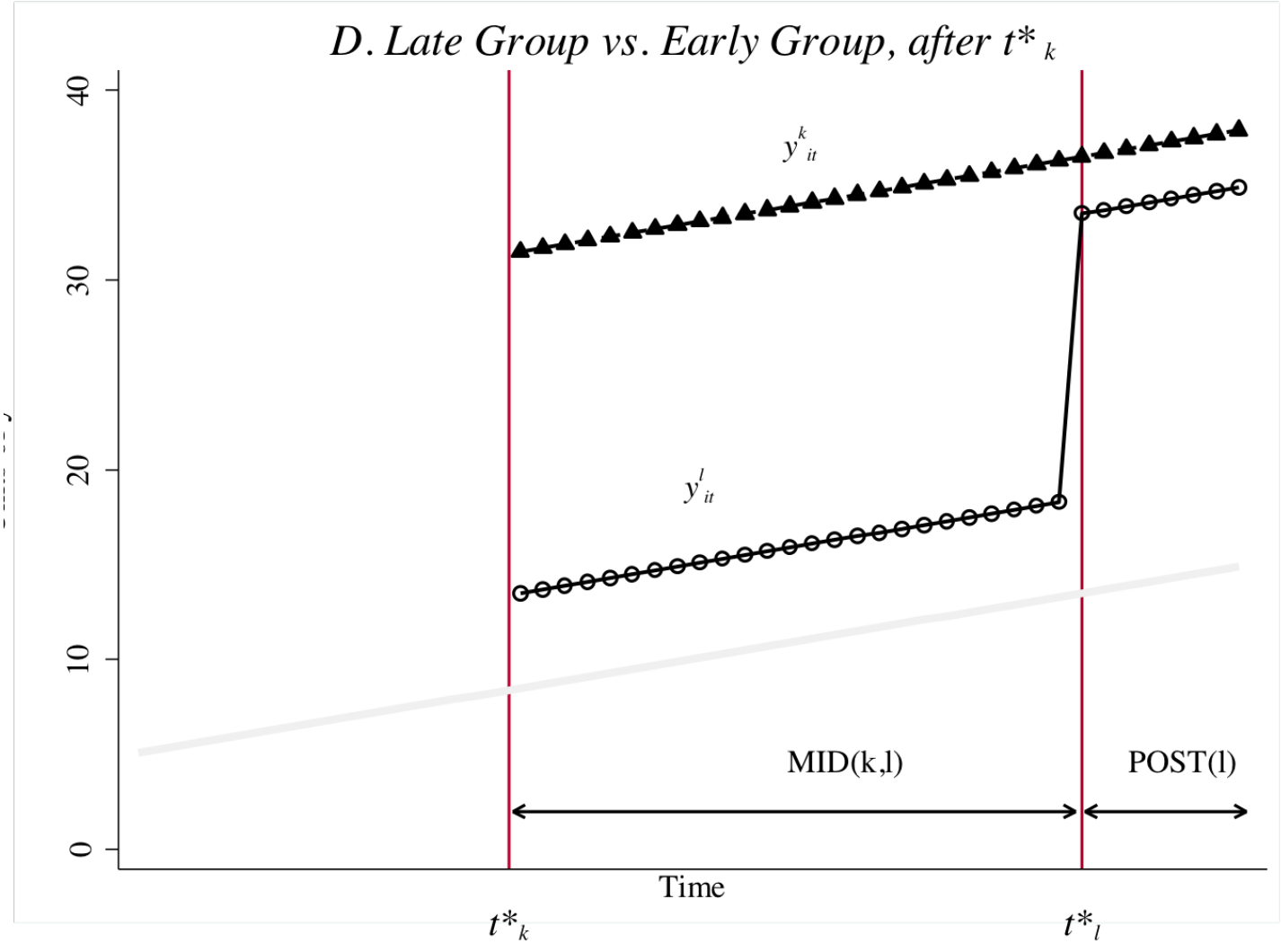
$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_{l}^{post(l)} - \overline{y}_{l}^{pre(l)}\right) - \left(\overline{y}_{U}^{post(l)} - \overline{y}_{U}^{pre(l)}\right)$$



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



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



Bacon decomposition

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

TWFE estimate of $\widehat{\delta}$ is equal to a weighted average over all group 2x2 (of which there are 4 in this example)

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

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

Third, the Weights

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

where n refer to 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

Back to TWFE

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

- So we saw that with differential timing, $\widehat{\delta}$ is a weighted average over all “four averages and three subtractions” but is that good or bad?
- Bacon’s decomposition doesn’t say anything about bias (yet)
- We will need to replace realized outcomes with potential outcomes in order to understand the nature of bias and causal identification

Back to TWFE

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

- Bacon shows that TWFE estimate of δ needs two assumptions for unbiasedness:
  1. variance weighted parallel trends are zero and
  2. no dynamic treatment effects (not the case with 2x2)
- Under those assumptions, TWFE estimator estimates the variance weighted ATT as a weighted average of all possible ATTs (not just weighted average of DiDs)

Moving from 2x2s to causal effects and bias terms

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

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

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

The dangerous 2x2

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

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

Remember earlier when I made the distinction between true and counterfeit diff-in-diff?

Substitute all this stuff into the decomposition formula

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

where we will make these substitutions

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

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
- 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

- Welcome to Differential Timing
  - Diff-in-diff credibility crisis
  - TWFE Pathologies
  - Simulation
- Robust Diff-in-Diff Estimators
  - CS
  - SA
  - dCH
  - Imputation based robust estimator
  - 2SDiD
- Examples
  - Facebook and Mental Health
  - Castle doctrine
- Basic suggestions going forward

Callaway and Sant'Anna 2020

CS is a DiD estimator used for estimating and then summarizing smaller ATT parameters under differential timing and conditional parallel trends into more policy relevant ATT parameters (either dynamic or static)

Difference-in-differences with multiple time periods

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

When is CS used

Just some examples of when you'd want to consider it:

1. When treatment effects differ depending on when it was adopted
2. When treatment effects change over time
3. When shortrun treatment effects are different than longrun effects
4. When treatment effect dynamics differ if people are first treated in a recession relative to expansion years

CS estimates the ATT by identifying smaller causal effects and aggregating them using non-negative weights

Group-time ATT

Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			

Each cell contains that group's ATT(g,t)

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

CS identifies all feasible ATT(g,t)

Group-time ATT

Group-time ATT is the ATT for a specific group and time

- Groups are basically cohorts of units treated at the same time
- Group-time ATT estimates are simple (weighted) differences in means
- Does not directly restrict heterogeneity with respect to observed covariates, timing or the evolution of treatment effects over time
- Allows us ways to choose our aggregations
- Inference is the bootstrap

Notation

- T periods going from $t = 1, \ldots, T$
- Units are either treated ($D_t = 1$) or untreated ($D_t = 0$) but once treated cannot revert to untreated state
- G_g signifies a group and is binary. Equals one if individual units are treated at time period t.
- C is also binary and indicates a control group unit equalling one if "never treated" (can be relaxed though to "not yet treated")
  - $\rightarrow$ Recall the problem with TWFE on using treatment units as controls
- Generalized propensity score enters into the estimator as a weight:

$$\widehat{p(X)} = Pr(G_g = 1 | X, G_g + C = 1)$$

Assumptions

Assumption 1: Sampling is iid (panel data, but repeated cross-sections are possible)

Assumption 2: Conditional parallel trends (for either never treated or not yet treated)

$$E[Y_t^0 - Y_{t-1}^0 | X, G_g = 1] = [Y_t^0 - Y_{t-1}^0 | X, C = 1]$$

Assumption 3: Irreversible treatment

Assumption 4: Common support (propensity score)

Assumption 5: Limited treatment anticipation (i.e., treatment effects are zero pre-treatment)

CS Estimator (the IPW version)

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

This is the inverse probability weighting estimator. Alternatively, there is an outcome regression approach and a doubly robust. Sant'Anna recommends DR. CS uses the never-treated or the not-yet-treated as controls but never the already-treated

Aggregated vs single year/group ATT

- The method they propose is really just identifying very narrow ATT per group time.
- But we are often interested in more aggregate parameters, like the ATT across all groups and all times
- They present two alternative methods for building “interesting parameters”
- Inference from a bootstrap

Group-time ATT

Truth				
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			
Feasible ATT	68.3333333			

CS estimates				
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1981	-0.0548	0.0191	0.0578	0
1986	10.0258	-0.0128	-0.0382	0
1987	20.0439	0.0349	-0.0105	0
1988	30.0028	-0.0516	-0.0055	0
1989	40.0201	0.0257	0.0313	0
1990	50.0249	0.0285	-0.0284	0
1991	60.0172	-0.0395	0.0335	0
1992	69.9961	8.013	0	0
1993	80.0155	16.0117	0.0105	0
1994	89.9912	24.0149	0.0185	0
1995	99.9757	32.0219	-0.0505	0
1996	110.0465	40.0186	0.0344	0
1997	120.0222	48.0338	-0.0101	0
1998	129.9164	56.0051	6.027	0
1999	139.9235	63.9884	11.969	0
2000	150.0087	71.9924	18.0152	0
2001	159.9702	80.0152	23.9656	0
2002	169.9857	88.0745	29.9757	0
2003	179.981	96.0161	36.013	0
2004				
2005				
2006				
2007				
2008				
2009				
Total ATT	n/a			
Feasible ATT	68.33718056			

Question: Why didn't CS estimate all ATT(g,t)? What is "feasible ATT"?

Reporting results

Table: Estimating ATT using only pre-2004 data

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

TWFE is no longer negative, interestingly, once we eliminate the last group (giving us a never-treated group), but is still suffering from attenuation bias.

Event study and differential timing

- Sometimes we care about a simple summary, and sometimes we care about separating it out in time and sometimes in even more interesting ways
- Event studies with one treatment group and one untreated group were relatively straightforward
- Interact treatment group with calendar date to get a series of leads and lags
- But when there are more than one treatment group, specification challenges emerge

Replicated from a project of mine

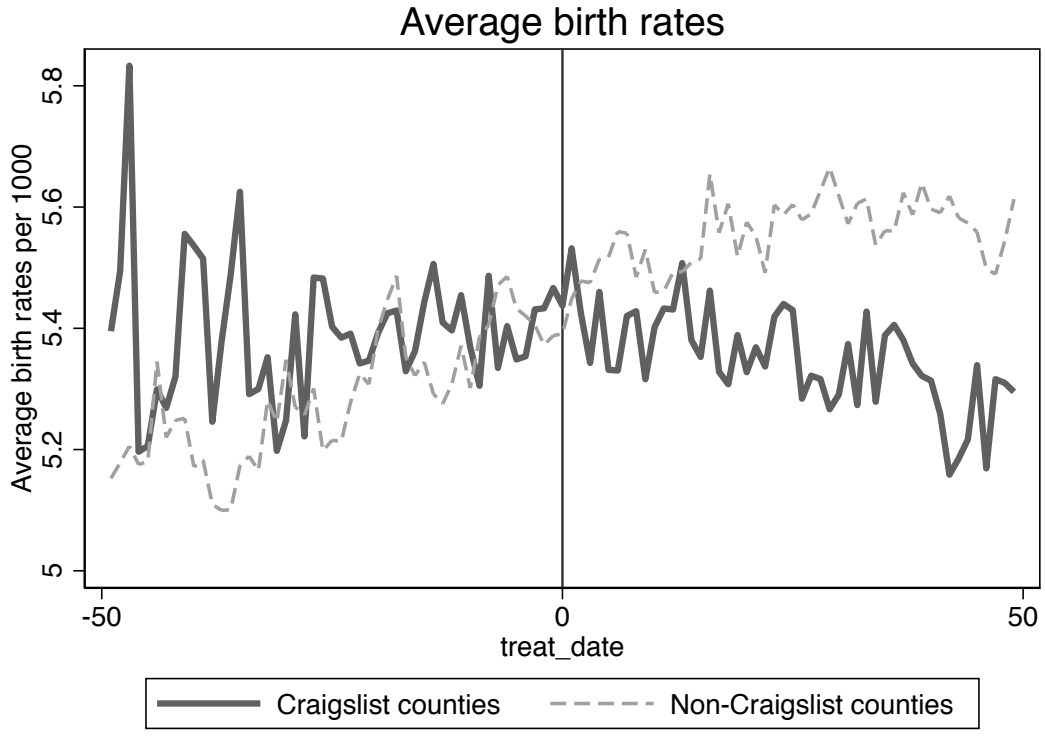
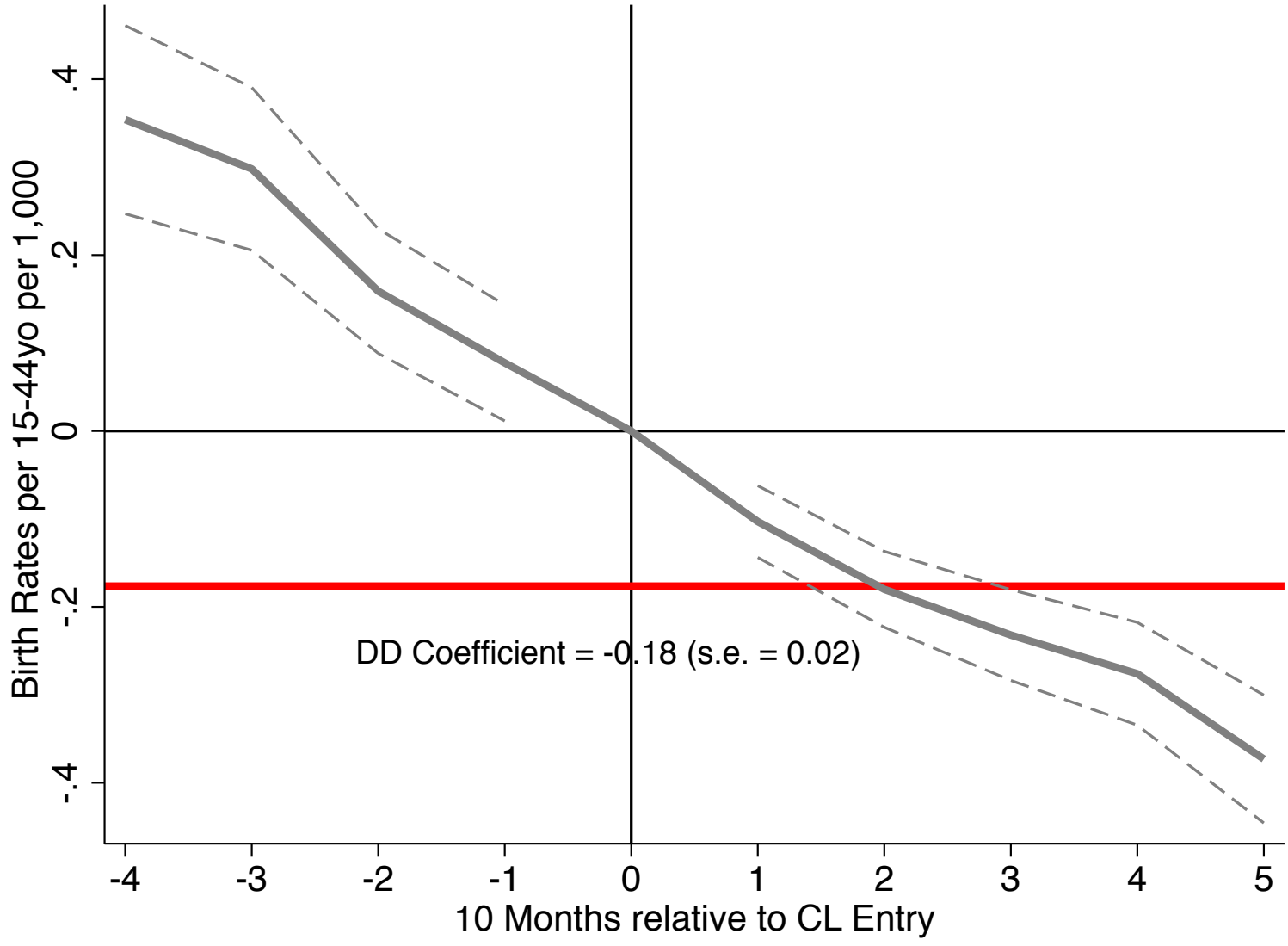


Figure: Roll out of Craigslist "personal ads" for casual intimate encounters and birth rates using the "randomized treatment assignment" approach (Anderson, et al. 2013) for visualization

Event study specification with TWFE

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

Coefficient μ_g on a dummy measuring the number of years prior to or after that unit was treated.



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

Bias of TWFE Event Study 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)

Sun and Abraham 2020

1. SA shows a decomposition of the population regression coefficient on event study leads and lags with differential timing estimated with TWFE
2. They show that the population regression coefficient is "contaminated" by information from other leads and lags (which is then later generalized by Goldsmith-Pinkham, Hull and Kolsar 2022)
3. SA presents an alternative estimator that is a version of CS only using the "last cohort" as the treatment group (not the not-yet-treated)
4. Derives the variance of the estimator instead of bootstrapping, handles covariates differently than CS, but otherwise identical

Summarizing (cont.)

- Under homogenous treatment profiles, weights sum to zero and“cancel out” the treatment effects from other periods
- Under treatment effect heterogeneity, they do not cancel out and leads and lags are biased
- They present a 3-step TWFE based alternative estimator which addresses the problems that they find

Some notation and terms

- As people often **bin** the data, we allow a lead or lag l to appear in bin g so sometimes they use g instead of l or $l \in g$
- Building block is the "cohort-specific ATT" or $CATT_{e,l}$ – same as ATT(g,t)
- Our goal is to estimate $CATT_{e,l}$ with population regression coefficient μ_l
- They focus on irreversible treatment where treatment status is non-decreasing sequence of zeroes and ones

Difficult notation (cont.)

- The ∞ symbol is used to either describe the group ($E_i = \infty$) or the potential outcome (Y^∞)
- $Y_{i,t}^\infty$ is is the potential outcome for unit i if it had never received treatment (versus received it later), also called the baseline outcome
- Other counterfactuals are possible – maybe unit i isn't "never treated" but treated later in counterfactual

More difficult notation (cont.)

- Treatment effects are the difference between the observed outcome relative to the never-treated counterfactual outcome: $Y_{i,t} - Y_{i,t}^{\infty}$
- We can take the average of treatment effects at a given relative time period across units first treated at time $E_i = e$ (same cohort) which is what we mean by $CATT_{e,l}$
- Doesn't use t index time ("calendar time"), rather uses l which is time until or time after treatment date e ("relative time")
- Think of it as l =year - treatment date

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: Homogeneous dynamics

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 $CATT_l$.

Homogeneous dynamics

- Assumption 3 allows for dynamic treatment effects; it just requires that different cohorts have the *same* dynamics
- 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)
- Original SA does not include covariates, though code has often been developed to include it – just remember what we saw with the double robust discussion
- Treatment effect doesn't rule out parallel trends – parallel trends turns simple DiD into causal parameters

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

Interacted weighted estimator

- Sun and Abraham (2020) propose a 3-step interacted weighted estimator (IW) using last treated group as control group
- Contrast with 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
- They are numerical identical with balanced panels, no covariates and using the never treated as controls

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; just can't accommodate the not-yet-treated. 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:

$$\widehat{\delta_{e,l}} = \frac{E_N[(Y_{i,e+l} - Y_{i,s}) \times 1\{E_i = e\}]}{E_N[1\{E_i = e\}]} - \frac{E_N[(Y_{i,e+l} - Y_{i,s}) \times 1\{E_i \in C\}]}{E_N[1\{E_i \in C\}]}$$

Proposition 5: If parallel trends and no anticipation both hold for all pre-periods, then the DD estimator using any pre-period and non-empty control cohorts (never-treated or not-yet-treated) is an unbiased estimate for $CATT_{e,l}$

Software

- **Stata**: eventstudyinteract (can be installed from ssc)
- **R**: fixest with subab() option (see `https://lrberge.github.io/fixest/reference/sunab.html/`)

Reporting results

Table: Estimating ATT

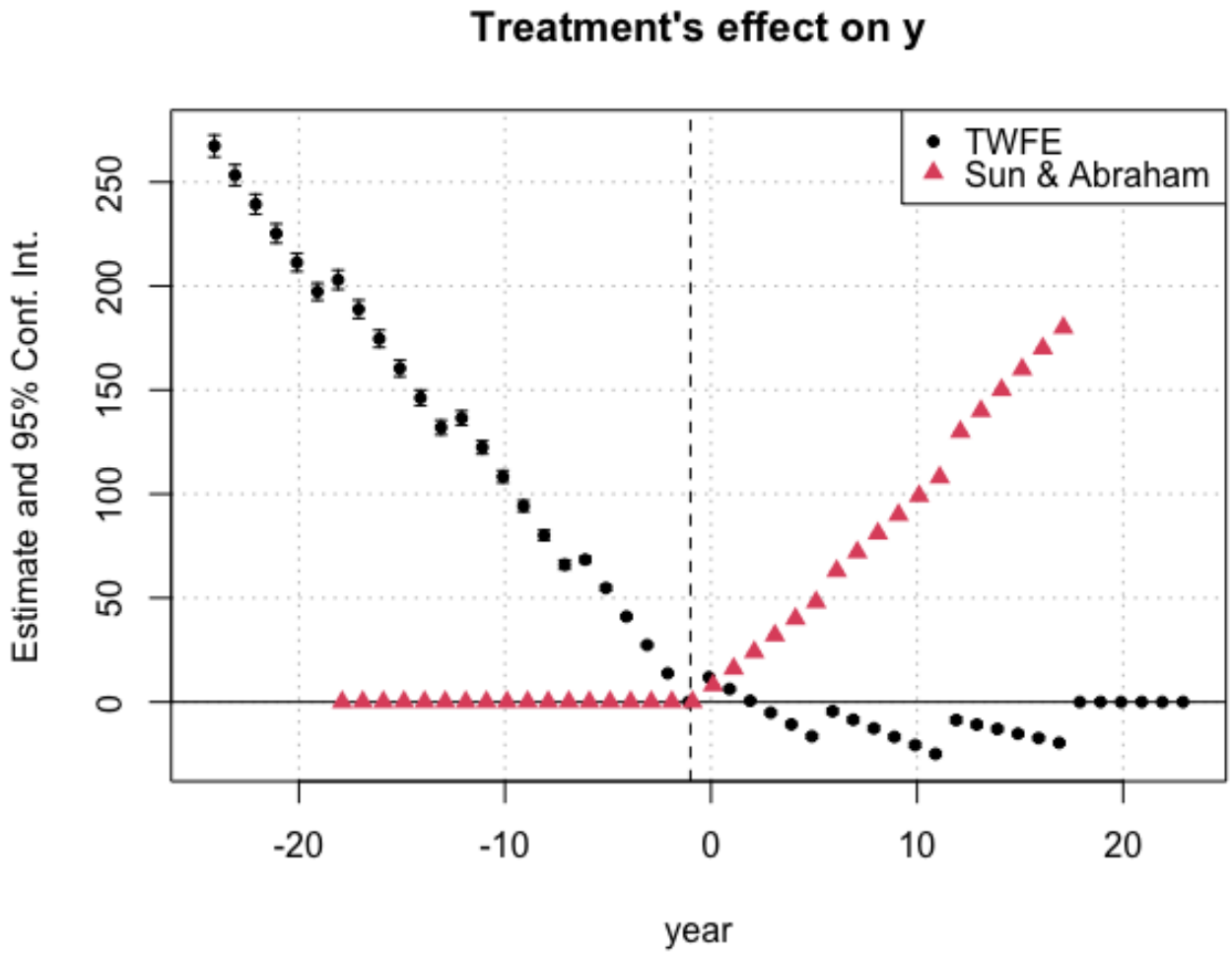
	(Truth)	**(TWFE)**	**(CS)**	**(SA)**	**(BJS)**
$\widehat{Feasible\ ATT}$	68.33	26.81***	68.34***	68.33***	

Computing relative event time leads and lags

Truth						Relative time coefficients		
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)		Leads	Truth	SA
1980	0	0	0	0		t-2	0	0.02
1986	10	0	0	0	(10+8+6)/3 = 8	t	8	8.01
1987	20	0	0	0	(20+16+12)/3 = 16	t+1	16	16.00
1988	30	0	0	0		t+2	24	24.00
1989	40	0	0	0		t+3	32	31.99
1990	50	0	0	0		t+4	40	40.00
1991	60	0	0	0		t+5	48	48.01
1992	70	8	0	0		t+6	63	62.99
1993	80	16	0	0		t+7	72	72.00
1994	90	24	0	0		t+8	81	80.99
1995	100	32	0	0		t+9	90	89.98
1996	110	40	0	0		t+10	99	99.06
1997	120	48	0	0		t+11	108	108.01
1998	130	56	6	0		t+12	130	129.92
1999	140	64	12	0		t+13	140	139.92
2000	150	72	18	0		t+14	150	150.01
2001	160	80	24	0		t+15	160	159.97
2002	170	88	30	0		t+16	170	169.99
2003	180	96	36	0		t+17	180	179.98
2004	190	104	42	4				
2005	200	112	48	8				
2006	210	120	54	12				
2007	220	128	60	16				
2008	230	136	66	20				
2009	240	144	72	24				

Two things to notice: (1) there only 17 lags with robust models but will be 24 with TWFE; (2) changing colors mean what?

Comparing TWFE and SA



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

de Chaisemartin and D'Haultfoeulle 2020

de Chaisemartin and D'Haultfouelle 2020 (dCdH) is different from the other papers in several ways

- Like SA, it's a diagnosis and a cure
- TWFE decomposition shows coefficient a weighted average of underlying treatment effects, but weights can be negative negating causal interpretation
- Propose a solution for both static and dynamic specification which does not use already treated as controls
- Treatment can turn on and off

Comment on Bacon

- Recall the Bacon decomposition – TWFE coefficients are decomposed into weighted average of all underlying 2x2s. Weights were non-negative and summed to one.
- But this decomposition was more a numerical decomposition – what exactly adds up to equal the TWFE coefficient using the data we observe?
- Bacon's decomposition is not "theoretical" – not in the way that other decompositions are. He is just explaining what OLS "does" when it calculates $\widehat{\delta}$
- Just explains what comparisons OLS is using to calculate the TWFE coefficient – just peels back the curtain.

Negative weights

- dCdH impose causal assumptions and try a different decomposition strategy
- Uses as its building block the unit-specific treatment effects
- Their decomposition will reveal negative weights on the underlying treatment effects (similar to negative weight on dynamics with Bacon)
- Remember though: the Bacon decomposition weights were *always* positive, because they were numerical weights (not theoretical weights) on the underlying 2x2s (not the treatment effects)

Turning on and off

- CS and SA both require interventions to turn on and stay on
- dCdH allows for “switching” on and off
- Before we move quickly into that, please note that the researcher bears the burden of knowing whether in fact you want to impose symmetry on turning on and off
- Roe v Wade “turned on” legalized abortion and 2022 it was “turned off” – do we want to treat these as simply a single policy flipping of the switch or two separate policies?

dCdH notation

- Individual treatment effects (iow, not the group-time ATT):

$$\Delta_{i,t}^{g} = Y_{i,t}^{1} - Y_{i,t}^{\infty}$$

  but where the treatment is in time period g. Notice –it's not the ATT (it's i individual treatment effect)
- with defined error term as $\varepsilon_{i,t}$:

$$D_{i,t} = \alpha_i + \alpha_t + \varepsilon_{i,t}$$

- Weights:

$$w_{i,t} = \frac{\varepsilon_{i,t}}{\frac{1}{N^T}\sum_{i,t:D_{i,t}=1}\varepsilon_{i,t}}$$

Parallel trend assumption

Strong unconditional PT

Assume that for every time period t and every group g, g',

$$E[Y_t^\infty - Y_{t-1}^\infty | G = g] = E[Y_t^\infty - Y_{t-1}^\infty | G = g']$$

Assume parallel trends for every unit in every cohort in every time period.

What then does TWFE estimate with differential timing?

dCdH Theorem

Theorem – dCdH decomposition

Assuming SUTVA, no anticipation and the strong PT, then let δ be the TWFE estimand associated with

$$Y_{i,t} = \alpha_i + \alpha_t + \delta D_{i,t} + \varepsilon_{i,t}$$

Then it follows that

$$\delta = E\left[\sum_{i,t:D_{i,t}=1} \frac{1}{N^T} w_{i,t} \cdot \Delta^g_{i,t} \right]$$

where $\sum_{i,t:D_{i,t}=1} \frac{w_{i,t}}{N^T} = 1$ but $w_{i,t}$ can be negative

Origins

- So once you run that specification, $\widehat{\delta}$ is going to recover a "non-convex average" over all unit level treatment effects (weights can be negative, more on this).
- Not sure who came first, because there were working papers before publications, but my understanding is dCdH was the first to prove this
- Very important theorem – established the "no sign flip property" for OLS with differential timing in the canonical static specification

Negative weights

- Very common now to hear about negative weights, and furthermore, that negative weights wipe out any causal interpretation, but why?
- Thought experiment: imagine every unit gained from the treatment, but their treatment effect when estimated was multiplied by a negative number
- It's possible it could flip the sign, but it would definitely at least pull the estimate away from the true effect
- This is dangerous – and it's caused by the forbidden contrasts (comparing treated to already treated) which is what the canonical TWFE static specification is doing (for many of us unknowingly)

Negative weights

- Doesn’t always pose a problem, but no proofs for this intuition known yet
- A large number of never-treated seems to make this less an issue
- Shrinking the spacing between treatment dates also can drive it down
- But does that mean that TWFE works, and what does it mean to work?
- TWFE still even when all the weights are positive the weighted average may not aggregate to what we think it does

Weighting

- The weights in OLS all come out of the model itself, *not the economic question*
- The economic question is “what parameter do you want? What does it look like? Who is in it?”
- And when you define the parameter up front, you’ve more or less defined the economic question you’re asking
- But OLS sort of ignores your question and just gives you what it wants

Weighting

- What makes something a good vs a bad weight?
- Not being negative is the absolute minimal requirement
- But it’s also not a good sign if you can’t really explain the weights

dCdH Solution

- dCdH propose an alternative that doesn't have the problems of TWFE – both avoiding negative weights and improving interpretability
- Recall, their model can handle reversible treatments

Imputation method by BJS (2023)

- The origin of the robust diff-in-diff papers identifying pathologies in TWFE was Borusyak and Jaravel (2016) working paper
- Both problems with static and dynamic specifications were discussed, and the identification of the “already treated” as controls
- Paper remained in working paper until 2021 when Jan Speiss was brought on; the team developed a new estimator (now published at Restud)
- We will focus primarily on the estimator, to avoid redundancies

ATT parameter

Estimation target will be unit level treatment effects (as opposed to group-time ATT) aggregated to a higher level like the ATT:

$$\tau_w = \sum_{it \in \Omega_1} w_{it} \tau_{it} = w_1' \tau$$

Note the weights – they need not add up to one. Weights could be $\frac{1}{N}$ for all $it \in \Omega_1$. We have a number of options.

Standard TWFE Assumptions

1. Parallel trends – expressed as a TWFE model with unit-level parallel trends (stronger than needed)
2. No anticipation
3. Homogenous treatment effects
4. SUTVA (as always)

A1: Parallel trends

Assumption 1: Parallel trends. There exist non-stochastic α_i and β_t such that:

$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}$$

with

$$E[\varepsilon_{it}] = 0$$

for all $it \in \Omega$. Can be extended (e.g., unit-specific trends). Only imposes restrictions on $Y(0)$, not treatment effects themselves. Notice how it is a TWFE assumption – it's actually the same data generating process as in baker.do.

A2: No anticipation

- No anticipation rules out anticipatory behavior that would cause treatment effects to materialize even before the treatment occurred:

$$Y_{it} = Y_{it}(0)$$

  for all $it \in \Omega_0$.
- Notice how as an assumption, it literally imposes $\tau = 0$ for all pre-treatment periods.
- It's crucial for the pre-trends to be zero, far more than parallel trends, as parallel trends is *only* about post-treatment from baseline, whereas event studies are about pre-trends and that's where no anticipation reigns

A3: Restricted causal effects

This is the one that places restrictions on what treatment effects can and cannot be (i.e., homogenous treatment effects). Notice the very detailed expression:

Assumption 3 (Restricted causal effects): $B\tau 0$ for a known $M \times N_1$ matrix B of full row rank.

If we can assume something like homogenous treatment effects, then TWFE actually is best because its ability to *correctly* extrapolate will increase efficiency. But it's when A3 is not tenable or not really ex ante justified by theory that we should be worried. There's an A3' that is a slight modification.

Critique of Common Practice

1. Under-identification in event studies
2. Negative weighting
3. Spurious identification of long-run casual effects

Critique: Underidentification problem

Lemma 1: If there are no never-treated units, the path of [pre-treatment lead population regression coefficients] is not point identified in the fully dynamic OLS specification. In particular. adding a linear trend to this path $\{\tau_h + k(h+1)\}$ for any $k \in R$ fits the data equally well with the fixed effects coefficients appropriately modified.

In english, it means you're going to have a multicollinearity problem even worse than you thought when estimating the fully dynamic event study model (i.e., dropping only one lead for all base comparisons)

Underidentification of lead coefficients

Under-identification problem

Formally the problem arises because a linear time trend t and a linear term in the cohort E_i (subsumed by the unit FEs) can perfectly reproduce a linear term in relative time $K_{it} = t - E_i$. Therefore a complete set of treatment leads and lags, which is equivalent to the FE of relative time, is collinear with the unit and period FEs.

Just one additional normalization is needed – drop $\tau_{-a} = 0$ and $\tau_{-1} = 0$. This will break the multicollinearity. We saw this in SA also. So multiple people saw this at the same time.

Under-identification and theoretical justifications

- Imposing any $-a$ lead and -1 lead to equal zero is somewhat ad hoc. Why those two and not some other two?
- Recall with SA – it mattered which ones you dropped because otherwise leads were contaminated
- This is again about NA – if you chose $-a$ and -1, then you had some theoretical reason to assume NA held for them and not some other periods

Negative weighting and violations of A3

Assume some simple static model with a single dummy for treatment. Then they lay out a second lemma

Lemma 2: If A1 and A2 hold, then the estimand of the static OLS specification satisfies $\tau^{static} = \sum_{it \in \Omega_1} w_{it}^{OLS} \tau_{it}$ for some weights w_{it}^{OLS} that do not depend on the outcome realizations and add up to one $\sum_{it \in \Omega_1} = 1$.

The static OLS estimand cannot be interpreted as a “proper” weighted average, as some weights can be negative.

Simple illustration

Table: TWFE dynamics

$E(y_{it})$	$i = A$	$i = B$
t=1	α_A	α_B
t=2	$\alpha_A + \beta_2 + \delta_{A2}$	$\alpha_B + \beta_2$
t=3	$\alpha_A + \beta_3 + \delta_{A3}$	$\alpha_B + \beta_3 + \delta_{B3}$
Event date	$E_i = 2$	$E_i = 3$

Static: $\delta = \delta_{A2} + \frac{1}{2}\delta_{B3} - \frac{1}{2}\delta_{A3}$.

Notice the negative weight on the furthest lag. This is what you get when A3 is not satisfied..

Short-run bias of TWFE

- TWFE OLS has a severe short-run bias
- the long-run causal effect, corresponding to the early treated unit A and the late period 3, enters with a negative weight (-1/2)
- The larger the effects in the long-run, the smaller the coefficient will be
- It's caused by "forbidden comparisons" (late to early treated) – we saw this with Goodman-Bacon (2021)
- Forbidden comparisons create downward bias on long-run effects with treatment effect heterogeneity, *but not with treatment effect homogeneity* – so it really is an A3 violation

Spurious Long-Run Causal Effects

More A3 problems, this time finding long-run effects where there are none. Basically, you need to impose a lot of pre-trend restrictions to get estimates of long-run population regression coefficients. Even then you can't get them all.

OLS estimates are fully driven by unwarranted extrapolations of treatment effects across observations and may not be trusted unless strong ex ante justifications for A3 exist

Lemma 4: Suppose there are no never-treated units and let $H = max_i E_i - min_i E_i$. Then for any non-negative weights w_{it} defined over the set of observations with $K_{it} \geq \overline{H}$ (that are not identically zero), the weighted sum of causal effects $\sum_{it:K_{it} \geq \overline{H}} w_{it} \tau_{it}$ is not identified by A1 and A2.

Modifications of general model

Modification of A1 to A1':

$$Y_{it}(0) = A'_{it}\lambda_i + X'_{it}\delta + \varepsilon_{it}$$

Assumption 4 is introduced (homoskedastic residuals). This is key, because they will be building an "efficient estimator" with BLUE like OLS properties.

Using A1' to A4, we get the "efficient estimator" which is for all linear unbiased estimates of δ_W, the unique efficient estimator $\widehat{\delta^*_W}$ can be obtained with 3 steps

Role of the untreated observations

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

"The idea is to estimate the model of Y_{it}^0 using the untreated observations and extrapolate it to impute Y_{it}^0 for treated observations."

Steps

1. Estimate expected potential outcomes using OLS fitted values of year and panel unit fixed effects but only the untreated observations. Similar to Heckman, Ichimura and Todd 1997; See Ichimura tell the origins of this (4 min):
   https://www.youtube.com/watch?v=qb8-cP998yM
2. Then calculate $\widehat{\delta}_{it} = Y_{it}^1 - \widehat{Y}_{it}^0$
3. Then estimate target parameters as weighted sums

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

Why is this working?

- Think back to that original statement of the PT assumption – you're modeling $Y(0)_{it}$.
- That is, without treatment – so the potential outcomes do not depend on any treatment effect
- Hence where we get treatment heterogeneity
- We obtain consistent estimates of the fixed effects which are then used to extrapolate to the counterfactual units for all missing unit-level $Y(0)_{it \in \Omega_1}$
- Computationally fast and flexible to unit-trends, triple diff, covariates etc. (though remember what we said about covariates)

Comparisons to other estimators

Table 3: Efficiency and Bias of Alternative Estimators

Horizon	Estimator	Baseline simulation		More pre-periods	Heterosk. residuals	AR(1) residuals	Anticipation effects
		Variance (1)	Coverage (2)	Variance (3)	Variance (4)	Variance (5)	Bias (6)
$h = 0$	Imputation	0.0099	0.942	0.0080	0.0347	0.0072	-0.0569
	DCDH	0.0140	0.938	0.0140	0.0526	0.0070	-0.0915
	SA	0.0115	0.938	0.0115	0.0404	0.0066	-0.0753
$h = 1$	Imputation	0.0145	0.936	0.0111	0.0532	0.0143	-0.0719
	DCDH	0.0185	0.948	0.0185	0.0703	0.0151	-0.0972
	SA	0.0177	0.948	0.0177	0.0643	0.0165	-0.0812
$h = 2$	Imputation	0.0222	0.956	0.0161	0.0813	0.0240	-0.0886
	DCDH	0.0262	0.958	0.0262	0.0952	0.0257	-0.1020
	SA	0.0317	0.950	0.0317	0.1108	0.0341	-0.0850
$h = 3$	Imputation	0.0366	0.928	0.0255	0.1379	0.0394	-0.1101
	DCDH	0.0422	0.930	0.0422	0.1488	0.0446	-0.1087
	SA	0.0479	0.952	0.0479	0.1659	0.0543	-0.0932
$h = 4$	Imputation	0.0800	0.942	0.0546	0.3197	0.0773	-0.1487
	DCDH	0.0932	0.950	0.0932	0.3263	0.0903	-0.1265
	SA	0.0932	0.954	0.0932	0.3263	0.0903	-0.1265

Notes: See Section 4.6 for a detailed description of the data-generating processes and reported statistics.

Two Stage DiD

> *“It seems natural that TWFE should identify the ATT” – Gardner (2021)*

It just seems like TWFE with a DiD will estimate the ATT with weights that we’ll find intuitive. Was this just a conjecture and was never true? Why isn’t this working?

Model misspecification

- Why does TWFE fail under differential timing? Violates strict exogeneity under heterogeneity
- The logic of the failure suggests an obvious, but previously unknown, solution which is the 2SDiD
- “Misspecified DiD regression models project heterogenous treatment effects onto group and period fixed effects rather than the treatment status itself”
- If you can get consistent and unbiased estimates of group and relative time fixed effects, then you can delete them through residualizing the outcome and run normal analysis

2SDiD

- First stage – estimate the group and relative time fixed effects using only the $D = 0$ observations (like BJS)
- Second stage – using predicted values based off those fixed effect coefficients, run your model off the residualized outcome
- Get the standard errors right by taking the first stage into account (uses GMM)

More high level

- The second step recovers the average difference in outcomes between treated and untreated units after removing group and period fixed effects
- Strong parallel trends assumption compared to CS and SA, but unclear if this is a big deal in general

Notation

i: panel units
t: calendar time – think of real dates
$g \in \{0, 1, \ldots, G\}$ – groups
$p \in \{0, 1, \ldots, P\}$ – relative time or “periods”

Periods are successive. Group 0 – never treated. Group 1 – treated in period 1, 2, and on. Group 2 – treated in period 2, etc.

Parameters

$$\beta_{gp} = E\left[Y^1_{gpit} - Y^0_{gpit} | g, p\right]$$

It's a group-time ATT but expressed in a more traditional econometric notation that you could easily find in Wooldridge or some such

Modeling basics

Under parallel trends, mean outcomes will satisfy the following equation

$$E\left[Y_{gpit}|g,p,D_{gp}\right] = \lambda_g + \gamma_p + \beta_{gp}D_{gp}$$

In two-group, group and period effects are eliminated with dummies because TWFE uses dummies to demean across multiple dimensions. Then TWFE identifies ATT. But this does not hold when average effects vary across group and period. There are many ways to express a treatment effect's across group and time, but Gardner presented it as a weighted average of the coefficients for only that group-period situation:

$$E\left(\beta_{gp}|D_{gp}=1\right) = E\left(Y^1_{gpit} - Y^0_{gpit}|D_{gp}=1\right)$$

Strict exogeneity violation

Rewriting the above we get:

$$
\begin{aligned}
E\left[Y_{gpit}|g,p,D_{gp}\right] &= \lambda_g + \gamma_p + E\left[\beta_{gp}|D_{gp}=1\right]D_{gp} \\
&\quad \left[\beta_{gp} - E(\beta_{gp}|D_{gp}=1)\right]D_{gp}
\end{aligned}
$$

The problem is there's this weird new error term and it isn't mean zero under heterogenous treatment effects spread across group and period. Unlike the two group case, the coefficient on D_{gp} from TWFE doesn't identify the average $E(\beta_{gp}|D_{gp}=1)$

DiD regression estimand

- So if TWFE isn't recovering $E(\beta_{gp}|D_{gp}=1)$, then what is it recovering?
- He shows that under PT, the coefficient on D_{gp} is:

$$\beta^* = \sum_{g=1}^{G}\sum_{p=g}^{P} w_{gp}\beta_{gp}$$

- So then – what are the weights w_{gp}? They are variance weights

Estimation

$$Y_{gpit} = \lambda_g + \gamma_p + \beta D_{gp} + \varepsilon_{gpit}$$

This specification assumes a conditional expectation function that is linear in group, period and treatment status. But when the model is misspecified, it will attribute some of the heterogeneity impacts of the treatment to group and period fixed effects. The longer the treatment, the greater $\overline{D}$ is, the more that group's treatment effects will be absorbed by group fixed effects. When misspecified, TWFE doesn't recover $E[\beta|D = 1]$.

Statistical issues

- Common support: “as long as there are untreated and treated observations for each group and period, λ_g and γ_p are identified from the subpopulation of untreated groups and periods.”
- Identification: “the overall group $\times$ period ATT is identified from a comparison of mean outcomes between treated and untreated groups after removing group and period effects.”

Estimation: First stage

First stage:

$$Y_{gpit} = \lambda_g + \gamma_p + \varepsilon_{gpit}$$

using only $D_{gp} = 0$, retaining the fixed effects. Collect the $\widehat{\lambda_g}$ and $\widehat{\gamma_p}$.

Estimation: Second stage

Second stage:

$$\begin{aligned} \widehat{y}_{gpit} &= y_{gpit} - \widehat{\lambda_g} - \widehat{\gamma_p} \\ \widehat{y}_{gpit} &= \alpha + \beta D_{gp} + \psi_{gpit} \end{aligned}$$

Why does this work? Parallel trends assumption implies:

$$E(y_{gpit}|g,p,D_{gp}) - \lambda_g - \gamma_p = E\Big[\beta_{gp}|D_{gp}=1\Big]D_{gp} + \Big[\beta_{gp} - E(\beta_{gp}|D_{gp}=1)\Big]D_{gp}$$

But because

$$E\Big\{[\beta_{gp} - E(\beta_{gp}|D_{gp}=1)]D_{gp}|D_{gp}\Big\} = 0$$

Estimand

Then this procedure will identify $E(\beta_{gp}|D_{gp}=1)$. Consistency and unbiasedness proofs.

This is $E(\beta_{gp}|D_{gp}=1) = \sum^{G}\sum^{P}\beta_{gp}P(g,p|D_{gp}=1)$. It will tend to put more weight, by definition, on groups earlier into their treatment. But this isn't the same as the negative weighting that BJS say occurs oof the long lags. It just means there are more of them.

Event studies are:

$$y_{gpit} = \lambda_g + \gamma_p + \sum_{r=-R}^{P}\beta_r D_{rgp} + \varepsilon_{gpit}$$

Just change the second stage with the transformed outcome.

Inference

- Standard errors are wrong on the second stage because the dependent variable uses estimates obtained from the first stage.
- The asymptotic distribution of the second stage can be obtained by interpreting the two-stage procedure as a joint GMM

Roadmap

- Welcome to Differential Timing
  - Diff-in-diff credibility crisis
  - TWFE Pathologies
  - Simulation
- Robust Diff-in-Diff Estimators
  - CS
  - SA
  - dCH
  - Imputation based robust estimator
  - 2SDiD
- Examples
  - Facebook and Mental Health
  - Castle doctrine
- Basic suggestions going forward

Motivation

- Widely cited that social media causes mental health problems in youth but no causal evidence ("slim to none")
- Braghieri, Levy and Makarin (2022), "Social Media and Mental Health", *American Economic Review*, 112(11): 3660-3693
- 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

Fourth part of a strong DiD

1. **Bite**: They cannot really show much here. No data on Facebook usage. More an ITT
2. **Falsifications**: Interestingly, they dod none
3. **Event studies**: Now canonical kind of presentation of main results
4. Mechanism: Weakly suggestive (you decide)

Outcome Data

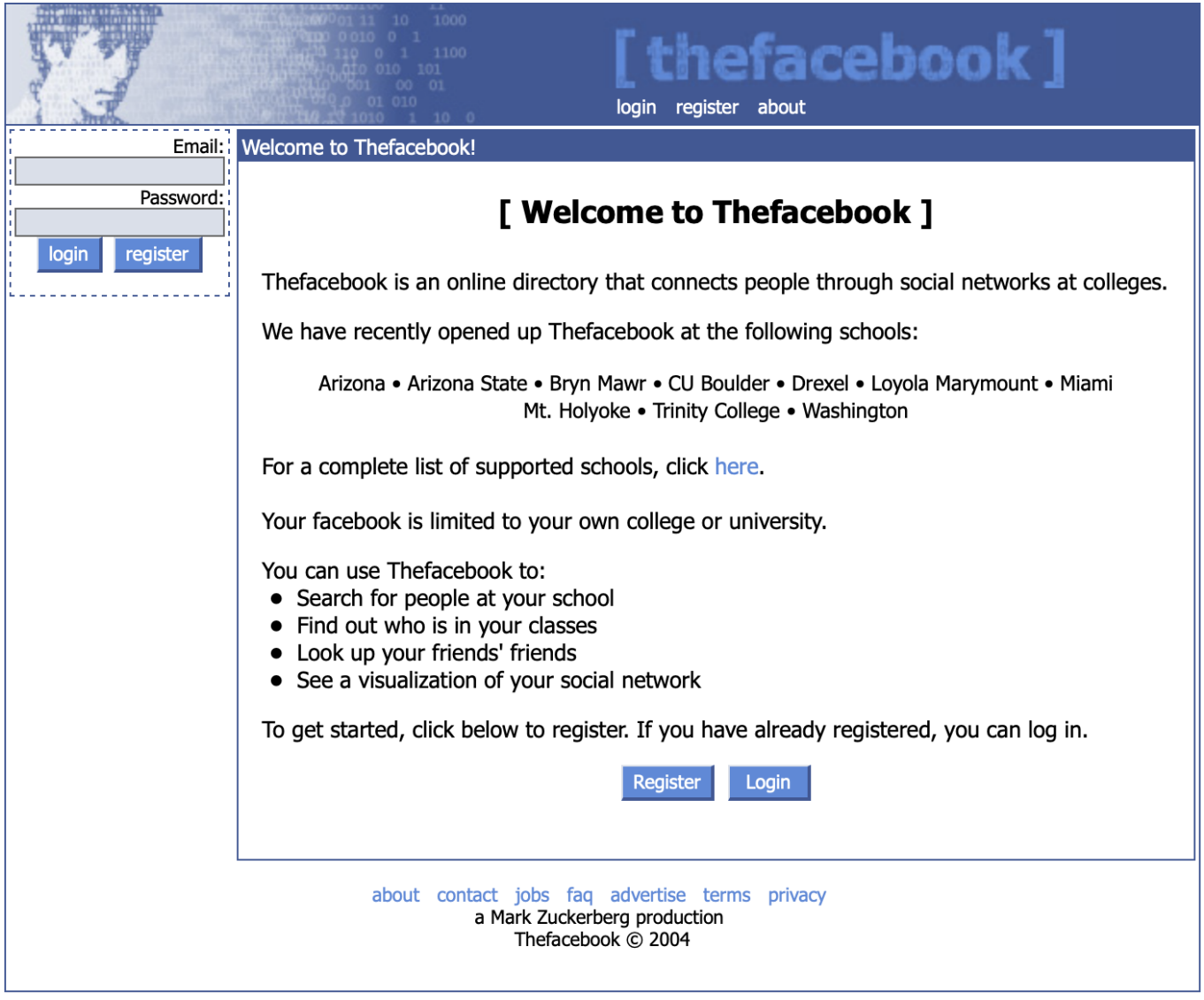
- Long-running health survey conducted at colleges in the US
- Requested that the survey owners match their facebook data with the survey data
- They said yes and did it but then never returned any calls (though the DUA had been signed)
- Outcomes are normalized so that coefficients on treatment is in standard deviations (e.g., z-score)

Treatment Data and Wayback machine

- Used Wayback machine from when it was `thefacebook.com`
- Identified all schools in the data by looking at each each front page to identify precise dates of rollout

Treatment Data and Wayback machine

Treatment Data and Wayback machine



TWFE

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

Their primary model is TWFE – done largely to appease referees

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.

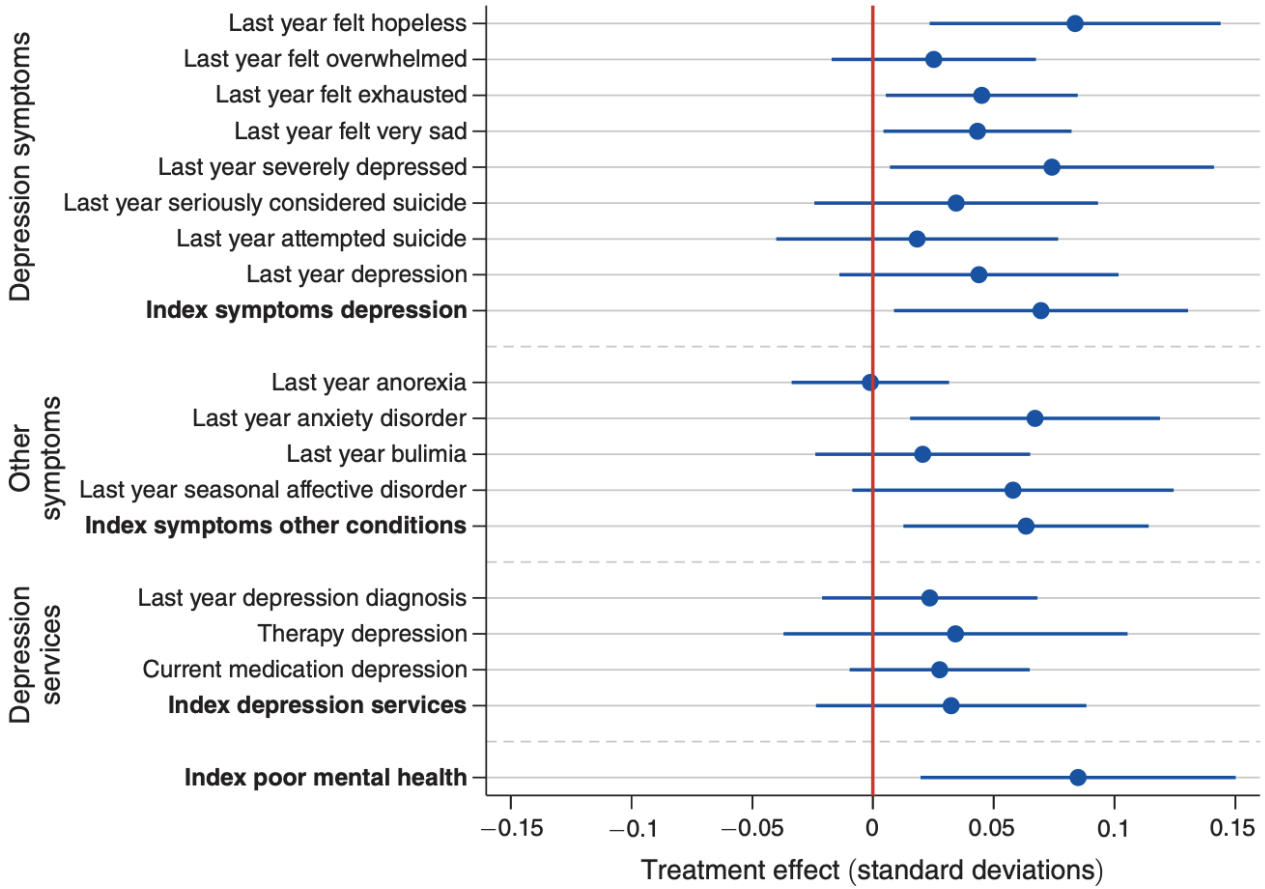


FIGURE 1. EFFECTS OF THE INTRODUCTION OF FACEBOOK ON STUDENT MENTAL HEALTH

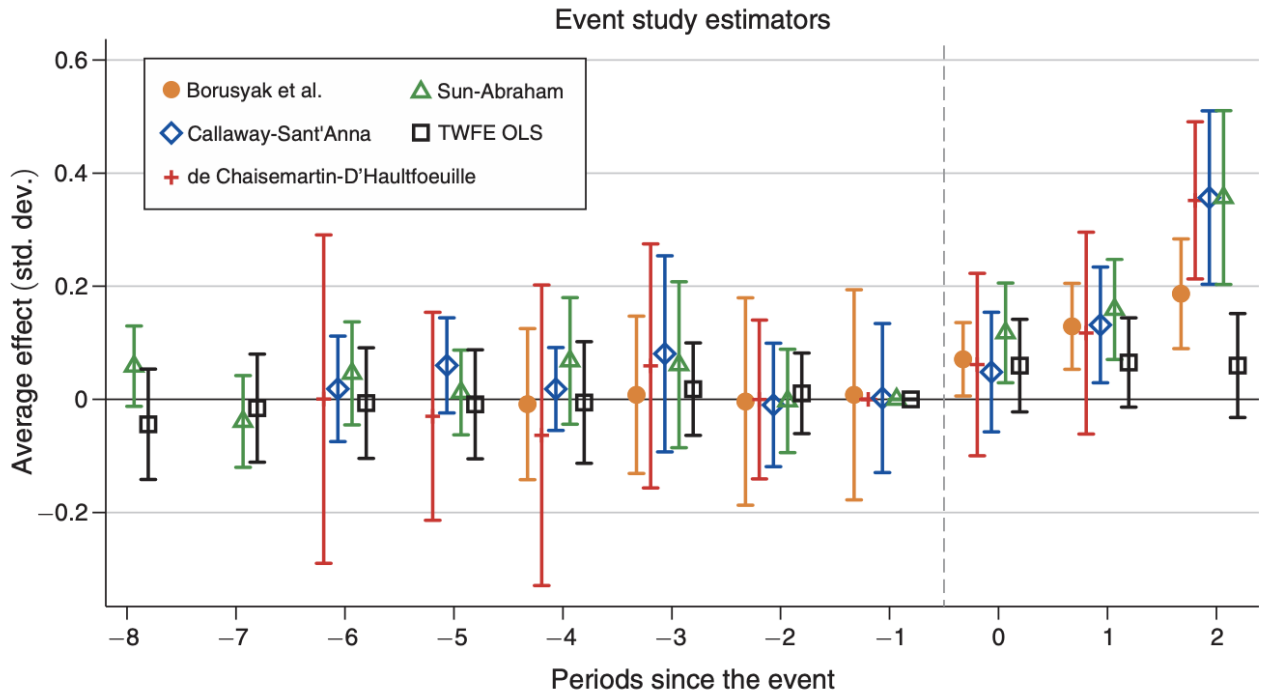


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

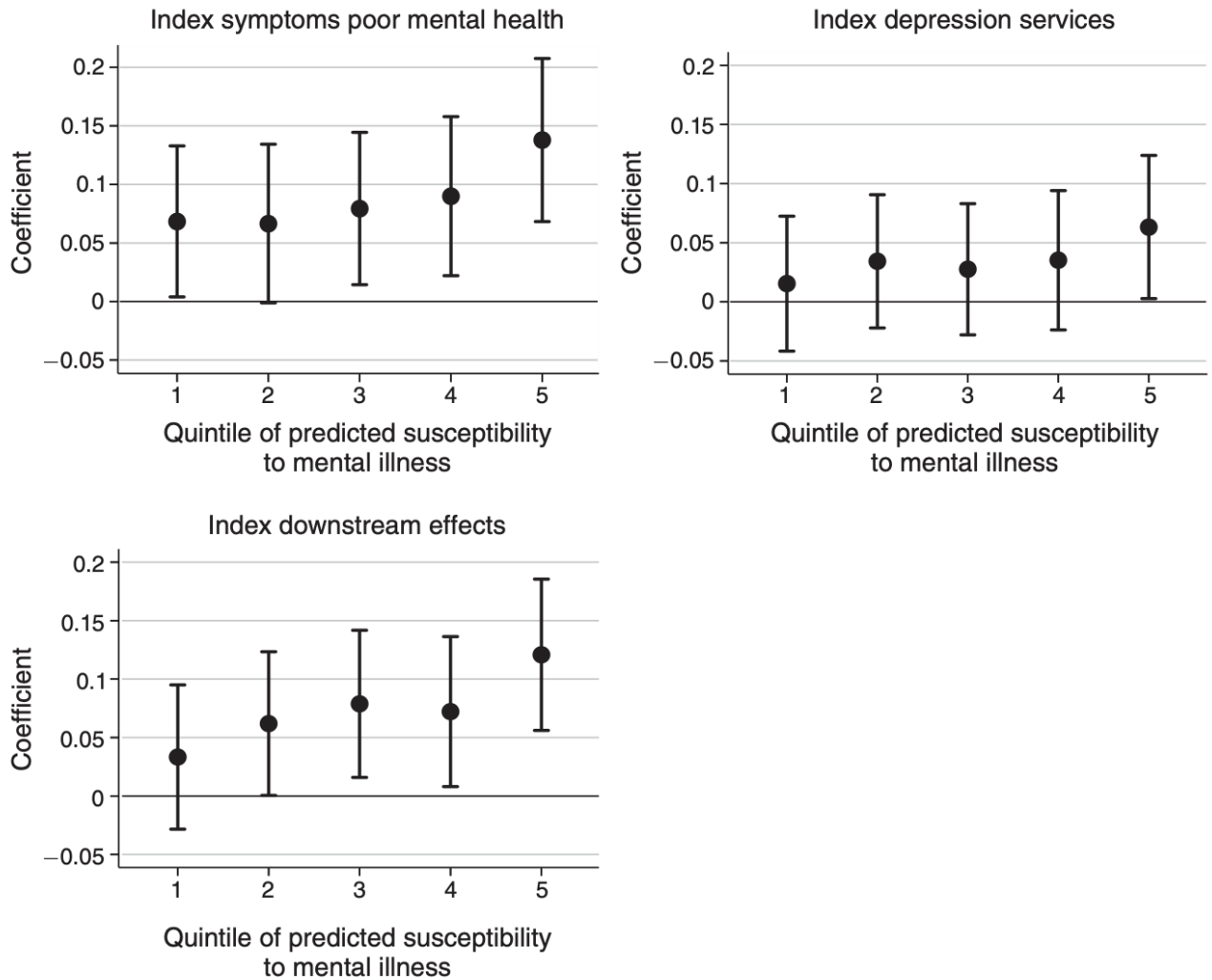


FIGURE 3. HETEROGENEOUS EFFECTS BY PREDICTED SUSCEPTIBILITY TO MENTAL ILLNESS

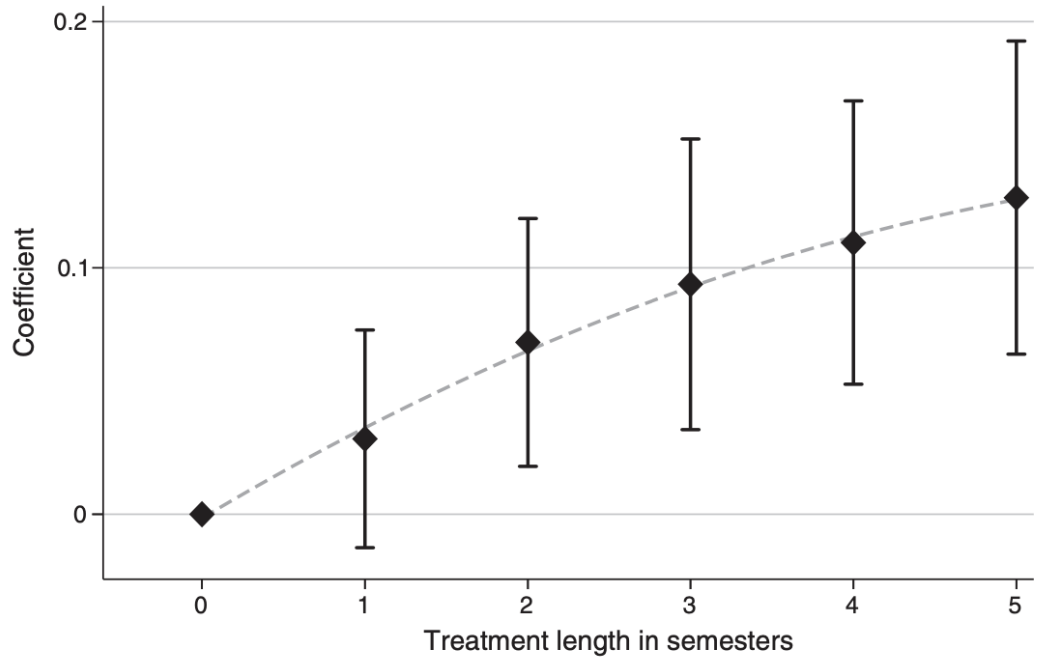


FIGURE 4. EFFECT ON POOR MENTAL HEALTH BY LENGTH OF EXPOSURE TO FACEBOOK

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.

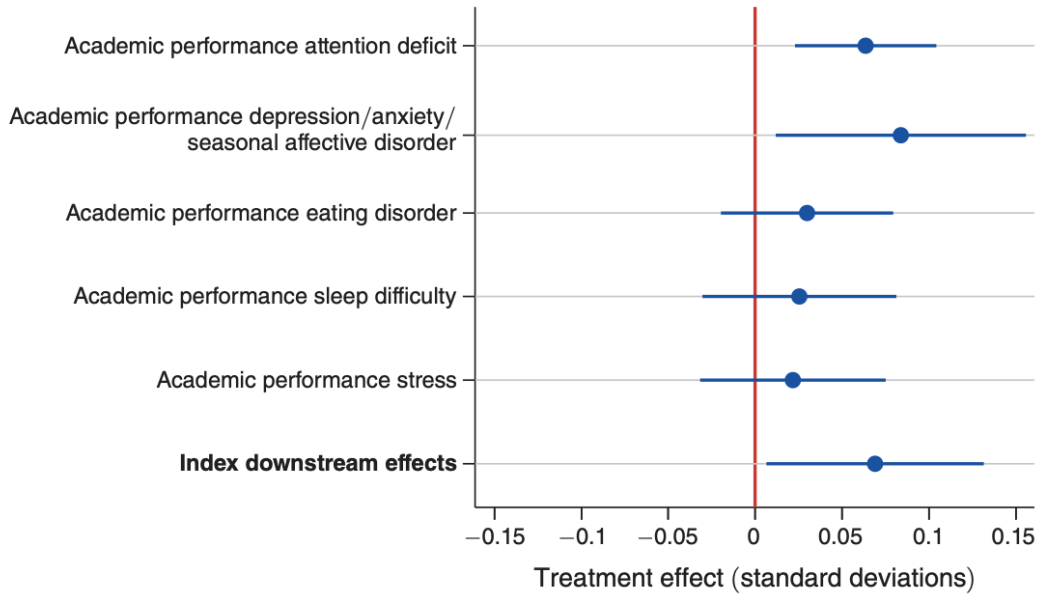


FIGURE 5. DOWNSTREAM EFFECTS ON ACADEMIC PERFORMANCE

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.

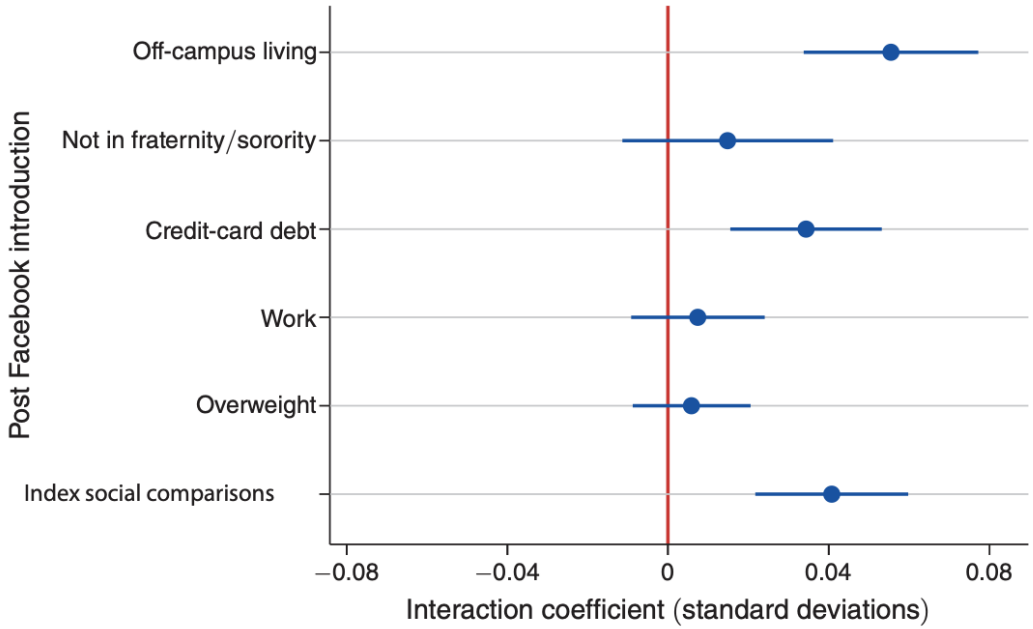


FIGURE 6. HETEROGENEOUS EFFECTS AS EVIDENCE OF UNFAVORABLE SOCIAL COMPARISONS

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.

Case study: Castle doctrine reforms

- We will replicate this together
- Cheng and Hoekstra (2013) is a good, clean example of a differential timing for us to practice on
- In 2005, Florida passed a law called Stand Your Ground that expanded self-defense protections beyond the house
- More “castle doctrine” reforms followed from 2006 to 2009

Description

Details of castle doctrine reforms

- “Duty to retreat” is removed versus castle doctrine reforms; expanded where you can use lethal force
- Presumption of reasonable fear is added
- Civil liability for those acting under the law is removed

Ambiguous predictions

Castle reforms $\rightarrow$ homicides: Increase by removing homicide penalties and increasing opportunities

- Castle doctrine expansions lowered the (expected) cost of killing someone in self-defense
- Lowering the price of lethal self-defense should increase lethal homicides

Castle reforms $\rightarrow$ homicides: decrease through deterrence

Cheng and Hoekstra's estimation model

- TWFE model

$$Y_{it} = \beta_1 D_i + \beta_2 T_t + \beta_3 (CDL_{it}) + \alpha_1 X_{it} + c_i + u_t + \varepsilon_{it}$$

- CDL is a fraction between 0 and 1 depending on the percent of the year the state has a castle doctrine law
- Preferred specifications includes "region-by-year fixed effects" (see next slide)
- Estimation with TWFE and Poisson with and without population weights
- Models will include covariates (e.g., police, imprisonment, race shares, state spending on public assistance)

Publicly available crime data

Main data: FBI Uniform Crime Reports Part 1 Offenses (2000-2010)

- Main outcomes: log homicides
- Falsification outcomes: motor vehicle theft and larceny (skipping this)
- Deterrence outcomes: burglary, robbery, assault

Region-by-year fixed effects

- **Parallel trends assumption**: imposed structurally with region-by-year dummies
- **Argument**: unobserved changes in crime are running “parallel” to the treatment states within region over time
- **SUTVA** and **No Anticipation**: No spillovers, no hidden variation in treatment, no behavioral change today in response to tomorrow’s law

Results – Deterrence

	OLS - Weighted by State Population						OLS - Unweighted					
	1	2	3	4	5	6	7	8	9	10	11	12
Panel A: Burglary	Log (Burglary Rate)						Log (Burglary Rate)					
Castle Doctrine Law	0.0780***	0.0290	0.0223	0.0164	0.0327*	0.0237	0.0572**	0.00961	0.00663	0.00277	0.00683	0.0207
	(0.0255)	(0.0236)	(0.0223)	(0.0247)	(0.0165)	(0.0207)	(0.0272)	(0.0291)	(0.0268)	(0.0304)	(0.0222)	(0.0259)
One Year Before Adoption of Castle Doctrine Law				-0.0201						-0.0154		
				(0.0139)						(0.0214)		
Panel B: Robbery	Log (Robbery Rate)						Log (Robbery Rate)					
Castle Doctrine Law	0.0408	0.0344	0.0262	0.0216	0.0376**	0.0515*	0.0448	0.0320	0.00839	0.00552	0.00874	0.0267
	(0.0254)	(0.0224)	(0.0229)	(0.0246)	(0.0181)	(0.0274)	(0.0331)	(0.0421)	(0.0387)	(0.0437)	(0.0339)	(0.0299)
One Year Before Adoption of Castle Doctrine Law				-0.0156						-0.0115		
				(0.0167)						(0.0283)		
Panel C: Aggravated Assault	Log (Aggravated Assault Rate)						Log (Aggravated Assault Rate)					
Castle Doctrine Law	0.0434	0.0397	0.0372	0.0362	0.0424	0.0414	0.0555	0.0698	0.0343	0.0305	0.0341	0.0317
	(0.0387)	(0.0407)	(0.0319)	(0.0349)	(0.0291)	(0.0285)	(0.0604)	(0.0630)	(0.0433)	(0.0478)	(0.0405)	(0.0380)
One Year Before Adoption of Castle Doctrine Law				-0.00343						-0.0150		
				(0.0161)						(0.0251)		
Observations	550	550	550	550	550	550	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes			Yes	Yes	Yes	Yes
Contemporaneous Crime Rates					Yes						Yes	
State-Specific Linear Time Trends						Yes						Yes

Results – Homicides

	1	2	3	4	5	6
Panel C: Homicide (Negative Binomial - Unweighted)						
Castle Doctrine Law	0.0565*	0.0734**	0.0879***	0.0783**	0.0937***	0.108***
	(0.0331)	(0.0305)	(0.0313)	(0.0355)	(0.0302)	(0.0346)
One Year Before Adoption of Castle Doctrine Law				-0.0352		
				(0.0260)		
Observations	550	550	550	550	550	550
Panel D: Log Murder Rate (OLS - Weighted)						
Castle Doctrine Law	0.0906**	0.0955**	0.0916**	0.0884**	0.0981**	0.0813
	(0.0424)	(0.0389)	(0.0382)	(0.0404)	(0.0391)	(0.0520)
One Year Before Adoption of Castle Doctrine Law				-0.0110		
				(0.0230)		
Observations	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes
Contemporaneous Crime Rates					Yes	
State-Specific Linear Time Trends						Yes

Interpretation

- Series of robustness checks (falsifications on larceny and motor vehicle theft; deterrence; many different specifications)
- Castle doctrine reforms are associated with an 8% net increase in homicide rates per year across the 21 adopting states
- Interpretation is these would not have occurred without castle doctrine reforms
- But is this robust to alternative models? Today we will check

Roadmap

- Welcome to Differential Timing
  - Diff-in-diff credibility crisis
  - TWFE Pathologies
  - Simulation
- Robust Diff-in-Diff Estimators
  - CS
  - SA
  - dCH
  - Imputation based robust estimator
  - 2SDiD
- Examples
  - Facebook and Mental Health
  - Castle doctrine
- Basic suggestions going forward

Robust DiD is now standard practice

- You're probably going to write a paper using DiD at least once in your life, but probably more
- Even if you don't, you're going to read a lot of papers using DiD, referee them, or advise students using them
- It's in your best interest to make the fixed cost investment in the new econometrics of DiD because the old methods are mostly harmful
- Good news is we are at the conclusion of this wave of papers, software is now widely available, solutions tend to have common features, and overall presentations (static and dynamic) aren't all that different

Handling covariates may need some finesse

- Simple 2x2 has its own problems when estimated using TWFE *if you include covariates*
- Stronger assumptions needed to include covariates, and bias can be large
- Don’t control for covariates that could be affected by the outcome

Avoid forbidden contrasts

- Main problem in differential timing is heterogeneity and the use of already-treated units as controls
- TWFE hid this under the hood more than likely
- Under differential timing, canonical TWFE assumes constant treatment effects and without it is biased (and can even flip signs)
- Robust DiD methods do not place restrictions on treatment effect heterogeneity

Moving to synthetic control

- When you lose parallel trends, you're probably going in the direction of needing synthetic control methods
- There are plenty and they also require good form fit pre-treatment
- But when you fail that, there are new advances – augmented synth, Imbens and Doudchenko both allow for lower quality fit