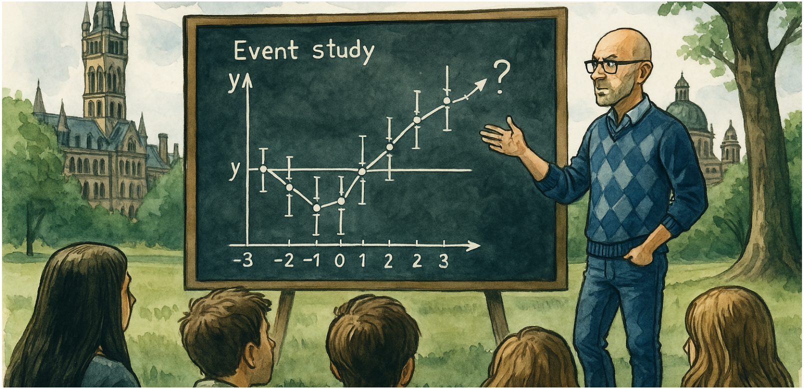
University of Glasgow

2025



Roadmap

- Differential timing or $G \times T$
  - Bacon Decomposition
  - Callaway and Sant'Anna (CS)
  - Calculating event study coefficients with CS
- Checklists and My Online Dating Project
- Alternative Estimators and Sensitivity Analysis
  - Sun and Abraham (SA)
  - de Chaisemartin and D'Haultfoeulle (dCDH)
  - Honest DID
- DDDiD

Estimator Selection for Differential Timing

- What if there had not been just one treatment cohort, but several?
- Think of there being three options with differential timing
  1. Traditional twoway fixed effects (TWFE)
  2. Aggregating ATT(g,t) using Callaway and Sant'Anna or Sun and Abraham
  3. Imputation using Borusyak, Jaravel and Speiss (BJS), Gardner or Wooldridge (Kyle will discuss these)
- Let's review now the differential timing literature with an aim to making a decision among them

Five approaches we will study

1. Twoway fixed effects (Goodman-Bacon 2021)
2. Callaway and Sant'Anna (2021), or CS
3. Sun and Abraham (2021), or SA
4. de Chaisemartin and D'Haultfouielle (2020), or dCDH
5. Borusywak, Jaravel and Spiess (2024), or BJS (Kyle)

Choosing between them

- First goal is to understand them, their assumptions, their mechanics and their code
- Second goal is to present a simple criteria for selecting between them so that you aren't ”cherry picking your diff-in-diff”

Twoway fixed effects

- When working with panel data, the so-called TWFE estimator is the workhorse estimator
- It's easy to implement, handles time-varying treatments, has a relatively straightforward interpretation under constant treatment effects, standard errors are easy to calculate and understand
- Interpretation is more complicated with heterogenous treatment effects

Discussion of estimate

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

- If you estimate with OLS with differential timing, what does $\widehat{\delta}$ correspond to?
- It is a weighted average of all possible "four averages and three subtractions"
- So similar to the 2x2 regression, except the coefficient is a weighted average over several – including one that we should have avoided all along

K^2 distinct DDs

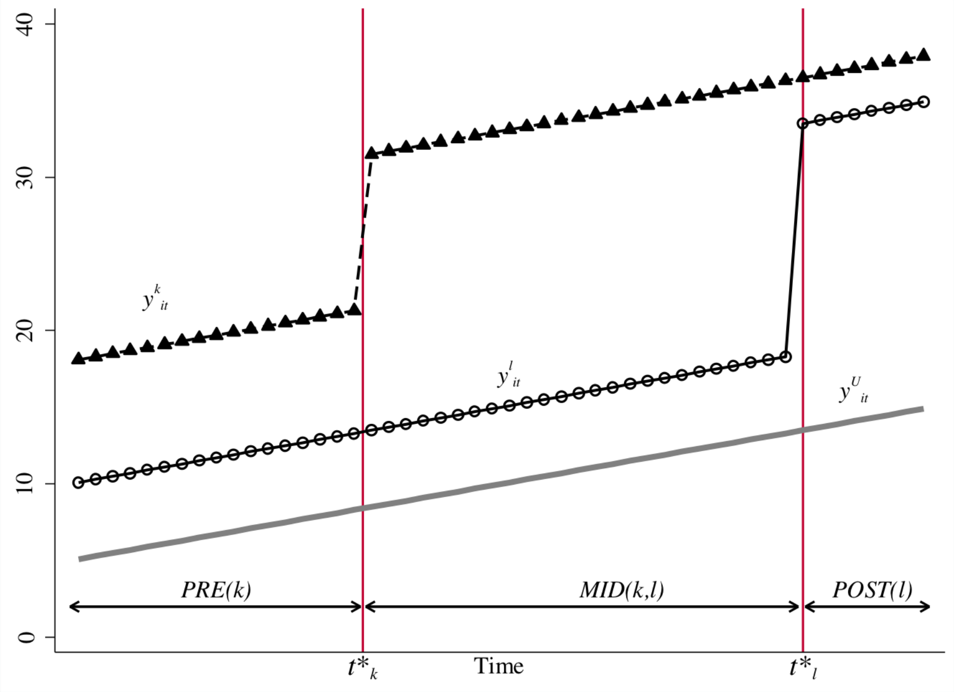
Let's look at 3 timing groups (a, b and c) and one untreated group (U). With 3 timing groups, there are 9 2x2 DDs. Here they are:

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

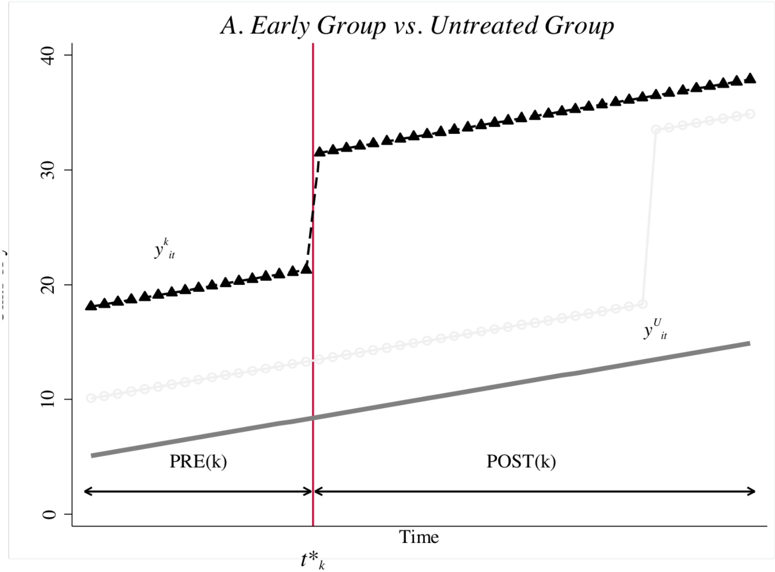
Terms and notation

- 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

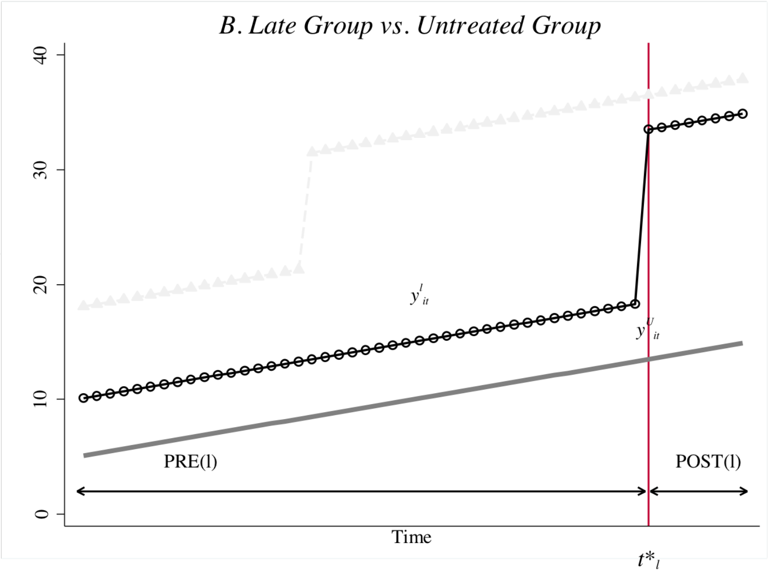


40
30
20
10
0
y^k_{it}
y^l_{it}
y^U_{it}
PRE(k)
MID(k,l)
POST(l)
t^*_k
Time
t^*_l

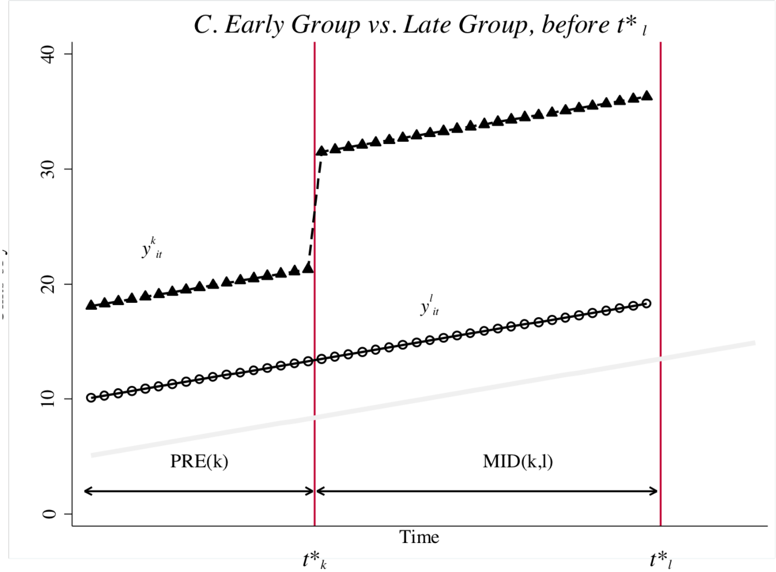
$$\widehat{\delta}_{kU}^{2x2} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)}\right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)}\right)$$



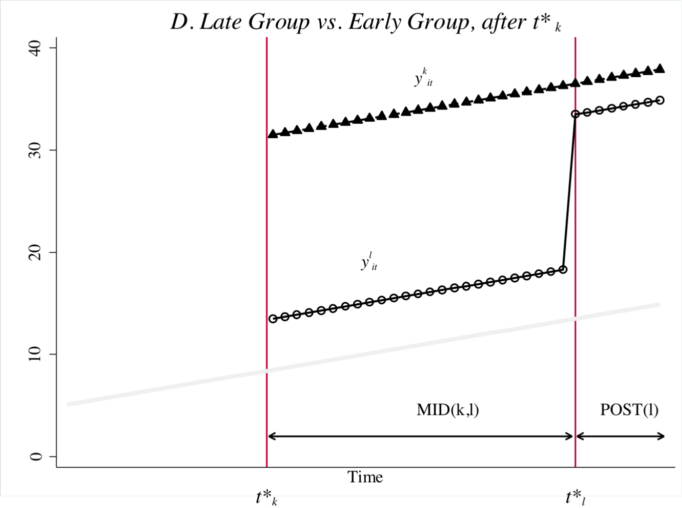
$$\widehat{\delta}_{lU}^{2x2} = \left(\overline{y}_l^{post(l)} - \overline{y}_l^{pre(l)} \right) - \left(\overline{y}_U^{post(l)} - \overline{y}_U^{pre(l)} \right)$$



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



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



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)

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

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

Third, the Weights

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

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

Weights discussion

- Two things to note:
  - $\rightarrow$ More units in a group, the bigger its 2x2 weight is
  - $\rightarrow$ 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.
  - $\rightarrow$ $\overline{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
  - $\rightarrow$ $\overline{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
  - $\rightarrow$ $\overline{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$
  - $\rightarrow$ $\overline{D} = 0.6$. Then $0.6 \times 0.4 = 0.24$
- This means the weight on treatment variance is maximized for *groups treated in middle of the panel*

More weights discussion

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

Summarizing TWFE centralities

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

Back to TWFE

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

- So we know that the estimate is a weighted average over all "four averages and three subtractions" but is that good or bad?
- It's good if it's unbiased; it's bad if it isn't, and the decomposition doesn't tell us which unless we replace realized outcomes with potential outcomes
- Bacon shows that TWFE estimate of δ needs two assumptions for unbiasedness:
  1. variance weighted parallel trends are zero and
  2. no dynamic treatment effects (not the case with 2x2)
- Under those assumptions, TWFE estimator estimates the variance weighted ATT as a weighted average of all possible ATTs (not just weighted average of DiDs)

Moving from 2x2s to causal effects and bias terms

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

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

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:

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

Substitute all this stuff into the decomposition formula

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

where we will make these substitutions

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

Notice all those potential sources of biases!

Potential Outcome Notation

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

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

Simulated data

- 1000 firms, 40 states, 25 firms per states, 1980 to 2009 or 30 years, 30,000 observations, four groups
- I'll impose "unit level parallel trends", which is much stronger than we need (we only need average parallel trends)
- Also no anticipation of treatment effects until treatment occurs but does *not* guarantee homogenous treatment effects
- Two types of situations: constant versus dynamic treatment effects

Constant vs Dynamic Treatment Effects

Calendar Time	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1981	0	0	0	0
1982	0	0	0	0
1983	0	0	0	0
1984	0	0	0	0
1985	0	0	0	0
1986	10	0	0	0
1987	10	0	0	0
1988	10	0	0	0
1989	10	0	0	0
1990	10	0	0	0
1991	10	0	0	0
1992	10	8	0	0
1993	10	8	0	0
1994	10	8	0	0
1995	10	8	0	0
1996	10	8	0	0
1997	10	8	0	0
1998	10	8	6	0
1999	10	8	6	0
2000	10	8	6	0
2001	10	8	6	0
2002	10	8	6	0

Calendar Time	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1981	0	0	0	0
1982	0	0	0	0
1983	0	0	0	0
1984	0	0	0	0
1985	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0

Group-time ATT

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

- Heterogenous treatment effects across time and across groups
- Cells are called "group-time ATT" (Callaway and Sant'Anna 2021) or "cohort ATT" (Sun and Abraham 2021)
- 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

Reverse Engineering

- Heckman (1990) showed that with noncompliance and heterogenous treatment effects, IV identified the ATE only in the most extreme case (when the instrument pushed everyone into treatment called ”identification at infinity”
- Imbens and Angrist (1994) showed that instrumental variables identified the ”local average treatment effect” (LATE)
- Note: they did not propose an IV estimator that would identify the ATE which was Heckman’s point – rather they ”reverse engineered” what IV did

Reverse Engineering

- Reverse (or backwards) engineering is when someone takes a model, and simply shows what it means
- The new diff-in-diff literature has many papers that did that – the most commonly known being Goodman-Bacon (2021) and his ”Bacon decomposition”
- All Bacon did was take a particular twoway fixed effects regression specification and show it was equal to a variance weighted average of ”all 2×2” calculations (some bad)
- Bacon **did not** propose a solution just like Imbens and Angrist (1994) did not propose a solution – reverse engineering is not ”solutions oriented”, per se

Forward Engineering

- When people design new estimators designed to overcome various problems, they are not reverse engineering – they are **forward engineering**
- Our next model is by Callaway and Sant'Anna (2021) and is an example of this
- They do not decompose what TWFE does (reverse engineering) but build a model that does not depend on as many assumptions

Callaway and Sant'Anna 2021

CS is a diff-in-diff estimator used with differential timing and a binary treatment that turns on (and stays on) to estimate smaller ”building block” causal parameters called group-time $ATT(g,t)$ and is used in situations like these:

1. Treatment effects differ in the shortrun than the longrun
2. Treatment effects differ by time period
3. Treatment effects have different dynamics for different groups
4. In other words, *unrestricted heterogenous 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**			

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 or analytical standard errors based on the influence function

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: Irreversible treatment

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: Common support (propensity score)

Assumption 4: No Anticipation

CS Estimator (the IPW version)

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

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 or influence function

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.

What did regressions do?

- In regressions, when you estimate leads and lags, you would drop one year dummy variable
- By dropping one year dummy variable, all coefficients were ”long difference” calculations relative to that baseline – both the $ATT(g,t)$ but also the event study pre-treatment coefficients
- Which means every event study plot you’ve ever seen with TWFE always was interpreted relative to a universal baseline (usually $t-1$)
- Some of the new estimators will allow for a different calculation, and this has created inconsistent estimators across new papers, which are relevant for checking for pre-trends

Event study lead calculation

- In CS and dCDH (covered later), there are two ways to calculate the pre-treatment coefficients once you have obtained the ATT(g,t) parameter estimates:
  1. **Long difference**. Uses a "universal baseline" with a fixed baseline at $t-1$.

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

  2. **Short gap**. Uses a "rolling" method in which a new baseline is used for each 2x2 comparison.

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

- Both are diff-in-diff 2×2, but pre-trends can be zero in one, but not

What is the event study for?

- The purpose of the event study is as a *falsification*, not as a proof of the parallel trends
- Falsifications should be the *same* model used for estimation but applied to something that it cannot cause – ”same group, different outcome”
- In this case, future treatments under no anticipation cannot affect the past
- Since the estimation of post-treatment ATT use $t-1$ for ”pre” period, the falsifications in the event studies should too

Detecting it in papers

- Very easy to spot which method someone used
- Long differences *never* have a coefficient at baseline, because long differences uses baseline as the comparison *always*
- Short gaps will *always* have a coefficient at baseline

Detecting it in papers

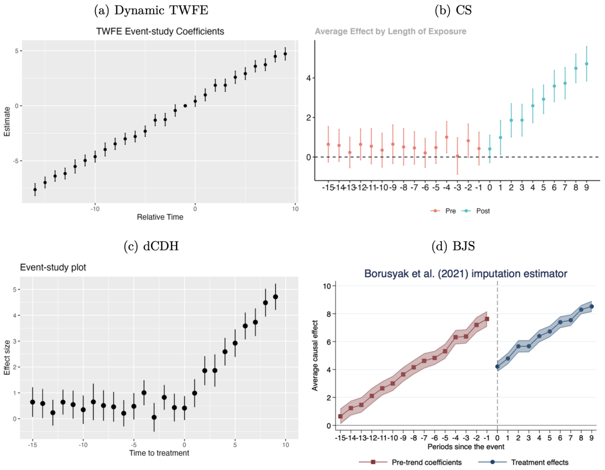
- Interestingly, long differences and short gaps have the same information – you can easily convert one into the other – but one will show trends and the other won't.
- So which one is right?
- Remember – parallel trends is a long difference, and all treatment effect calculations i are long differences, even in CS, so in my opinion so should the event study coefficients
- Plus remember – because TWFE uses long differences, which therefore is what we have come to expect
- Subtle switching out of visualizing pre-trends from our expectation

Roth simulation

- Let's look at a simulation by Jon Roth (2024)
- He generated the data with rising pre-trends
- He then estimated with TWFE (long difference), CS (short gaps), dCDH (short gaps) and BJS (Kyle can discuss this tomorrow)

Short Gaps and Long Differences

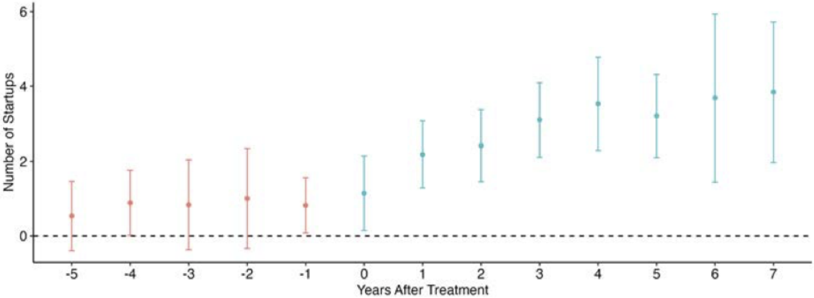
Figure 1: Comparison of event-study plots in a non-staggered setting



Software Syntax

- Stata's `csdid` has default syntax where if you don't indicate which way to do it, it only does it using the short gap method
- And in the Stata user command from Stata 18, you actually **cannot** do short differences
- You have to select `long2` in `csdid` and the universal baseline in R's `did` but if you want short gaps, you do not specify anything
- So what is going on is that since it's not documented well, `csdid` is very population, and default is short gaps, people are using short gaps and probably don't know, and don't explain it in the papers
- Here is an NBER WP and AER 2022 that are very clearly *not* using long differences

C. *All First Starbucks*



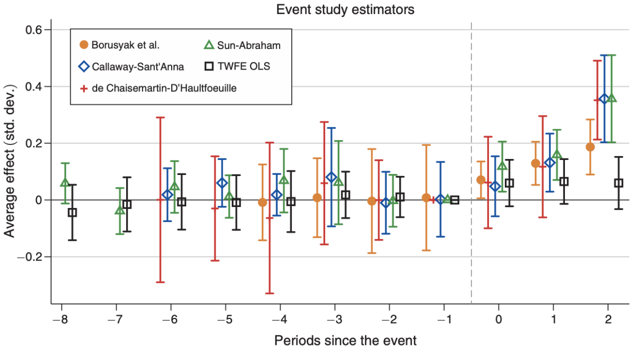


FIGURE 2. EFFECTS OF FACEBOOK ON THE INDEX OF POOR MENTAL HEALTH BASED ON DISTANCE TO/FROM FACEBOOK INTRODUCTION

Roadmap

Differential timing or $G \times T$

- Bacon Decomposition
- Callaway and Sant'Anna (CS)
- Calculating event study coefficients with CS

Checklists and My Online Dating Project

Alternative Estimators and Sensitivity Analysis

- Sun and Abraham (SA)
- de Chaisemartin and D'Haultfoeulle (dCDH)
- Honest DID

DDDiD

How Checklists Saved Lives in Medicine

Key Idea: Checklists reduce errors and improve patient safety by standardizing procedures and preventing overlooked steps.

- *The Checklist Manifesto* by Atul Gawande explores how checklists improve outcomes in medicine, aviation, and other fields.
- The WHO Surgical Safety Checklist reduced complications and mortality in surgeries worldwide. A study found a 36% reduction in major surgical complications.
- Checklists are thought to work because they:
  - → Ensure that critical steps are not skipped.
  - → Encourage teamwork and communication.
  - → Create a structured, repeatable process for complex tasks.

Power to the Platform

- I'm going to now walk you through a simple checklist I have in my new book
- Got this from Roth (2022), Pedro Sant'Anna informal conversations, principles from Rubin (2008) and many of Guido Imbens' survey articles
- We ran into a lot of problems by following this checklist, and a lot of heartbreak followed by excitement
- It's a study of online dating's effect on the American family, but I'll just focus on birth rates
- We have a great title – Power to the Platform – as a homage to "Power to the Pill" by Goldin and Katz and "More Power to the Pill" by Bailey

Online Dating and Birth Rates

- Me, Christine Durrance, and Melanie Guldi have been working on a project for years looking at online dating's effect on birth rates
  - → Thickening of relationship markets
  - → Reduced search costs
  - → Formation of better relationships meant for forming families
- But online dating companies have an incentive to perpetuate dating despite claims to the contrary, which may reduce the formation of families

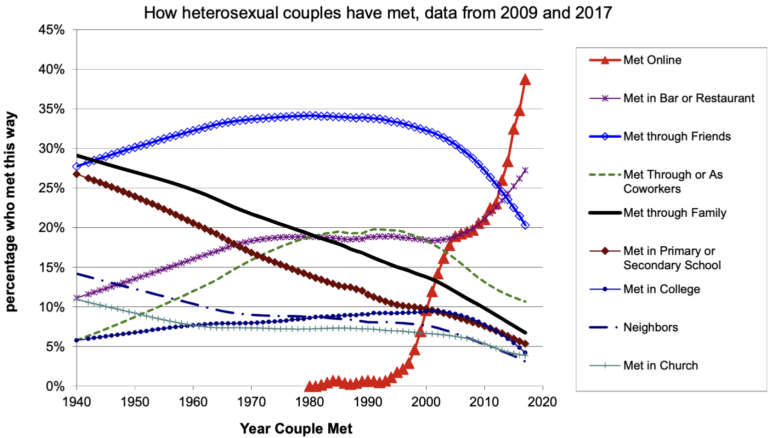


Fig. 1. Source: HCMST 2009 and HCMST 2017 waves. Consistent with Rosenfeld and Thomas (3), all trends are from unweighted Lowess regression with bandwidth 0.8 (39), except for meeting online, which is a 5-y moving average because meeting online takes place in the more recent and data-rich part of the data (N = 2,473 for HCMST 2009 and N = 2,997 for HCMST 2017). Friends, family, and coworkers can belong to either respondent or partner. Percentages do not add to 100% because the categories are not mutually exclusive; more than one category can apply.

Confounders

- Two problems:
  1. Online dating either hits the US with an open website anyone can get on (early period) or a ”swiping app” that anyone can get on (late period)
  2. After 2008, the American economy birth rates plummeted and never recovered (demographic transition)
- Both are massive hurricane like winds and it’s going to be challenging to deal with them with diff-in-diff
- But, amazingly, we have a solution that gets at both and is probably much closer to our target parameter – Craigslist Personals

What was Craigslist Personals?

- Craigslist is one of the most visited websites in the United States
- Two sided matching website that devastated classified advertising revenue in newspapers
- Primarily made money from housing and jobs, but you can get *anything* on it

What was Craigslist Personals?

- But then in 2000, in the Bay Area (i.e., San Francisco), they introduce ”People matching technology” (their words)
- Casual sex, serious relationships, men seeking men, men seeking women, women seeking men, women seeking men
- And great for us – cities got this on different dates giving us ”staggered rollout”

Personals


INTERNET ARCHIVE
WayBackMachine
http://www.craigslist.org/
79,156 captures
2 Dec 1998 - 4 May 2025
craigslist
San Francisco
Bay Area
online community
need help?
post housing/stuff/resumes
post a job
subscribe or unsubscribe
new category info
Search
jobs
resumes
housing
other
our giving program
stuff about us
2/3 party pictures
7 April 2000 (updated)
housing
rooms/shared available
rooms/shared wanted
housing/offices available
housing/offices wanted
stuff
community/etc
small/personal biz ads
general for sale
general wanted
car/motorcycle/etc
system/pc stuff
events/entertainment
tech events
volunteers
internet jobs
web/info design
engineering
business
jobs
software/other engineering
marketing/sales/PR
accounting/finance
administrative/office/HR
et cetera
writing/editing
arts/print/other media
healthcare/medical
retail/hospitality/food
nonprofit
resumes
Webby Awards: it's an honor just to be nomiated
summer community builder internships in DC/Virginia
Sneak-preview : our new people-matching technology

Personals

INTERNET ARCHIVE WayBackMachine

http://www.craigslist.org/need.your.help.html

2 captures

20 Jun 2000 - 20 Jun 2000

Folks, we've been listening to your requests for new categories and new functionality, and have worked out some possibilities. We need your feedback and your ok on this.

You'll see, in particular, a temporary discussion board we've put up, which is available to give us feedback on:

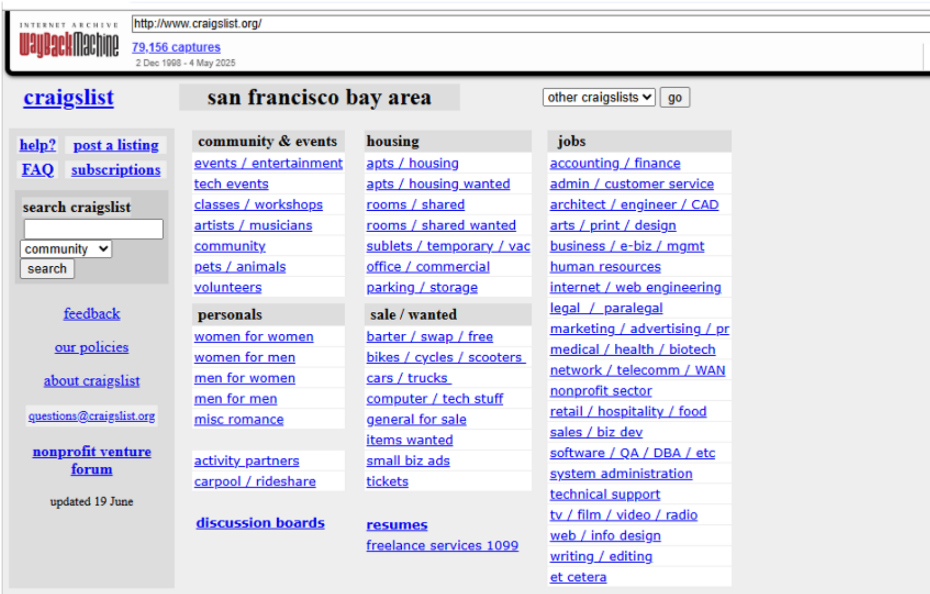
- **new categories**
- **new overall look and feel**
- **discussion board functionality**
- **new craigslist area sites & navigation between them**
- **search functionality**

you'll be telling us what to do. We'll look at your feedback and produce more based on what we think the consensus is.

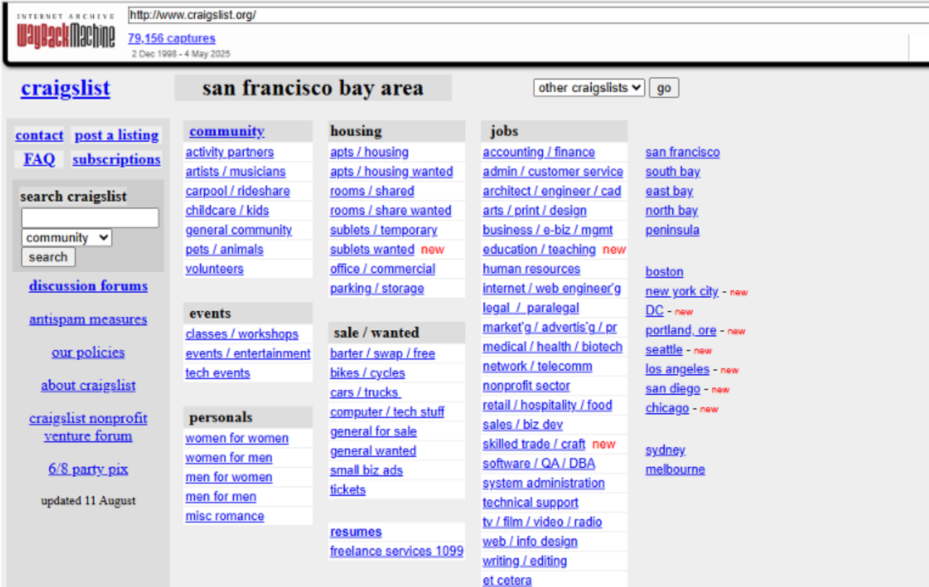
here's the new stuff - enjoy!

note: if you want to critique the new stuff via email, you can do so at feedback@craigslist.org

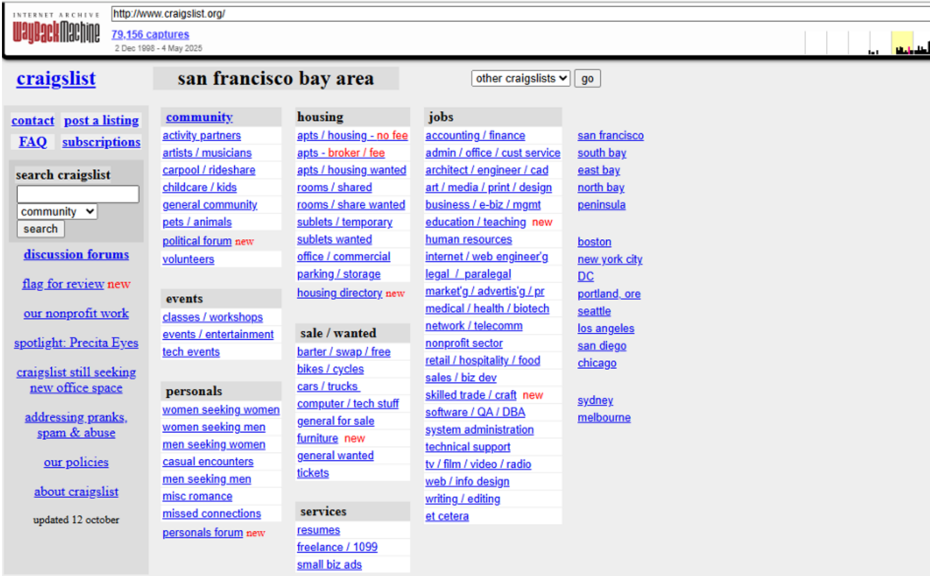
Personals starts in San Francisco



Personals begins spreading



Personals gets larger



Personals is popular!

Personals is popular!

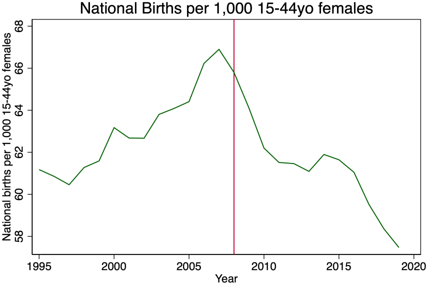
Personals is popular!

Our project

- So we have geographic rollout from 2000 to 2010
  1. But after 2008, Great Recession leads to plummeting birth rates
  2. And after 2008, we have social media, smart phones, all of which maybe had their own independent effects on matching (”sex recessions”)
- So we will use 1995 (pre-treatment) to 2007 as our sample period
- But we will include the 2008-2010 ”later treated counties” as our control group
- Used the wayback machine to get every county’s craigslist start date (three of us and an RA!)

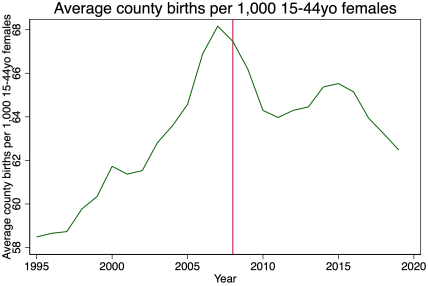
Recovery of Birth Rates (National vs. County)

Average National Birth Rates

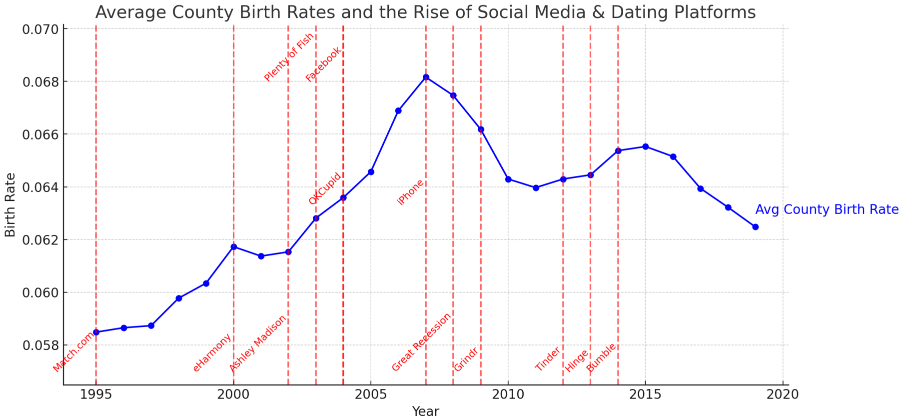


Total births divided by total female population. Vertical line marks the Great Recession.

Average County Birth Rates



Data averages over county level births per 15-44 yo female. Vertical line marks the Great Recession.



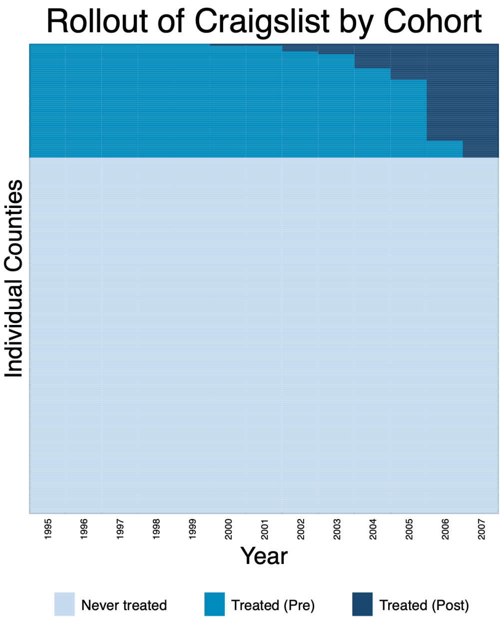
Average County Birth Rates and the Rise of Social Media & Dating Platforms
Birth Rate
Year
0.070
0.068
0.066
0.064
0.062
0.060
0.058
1995
2000
2005
2010
2015
2020
Match.com
eHarmony
Ashley Madison
Plenty of Fish
Facebook
OKCupid
iPhone
Great Recession
Grindr
Tinder
Hinge
Bumble
Avg County Birth Rate

Number of counties by treatment cohorts

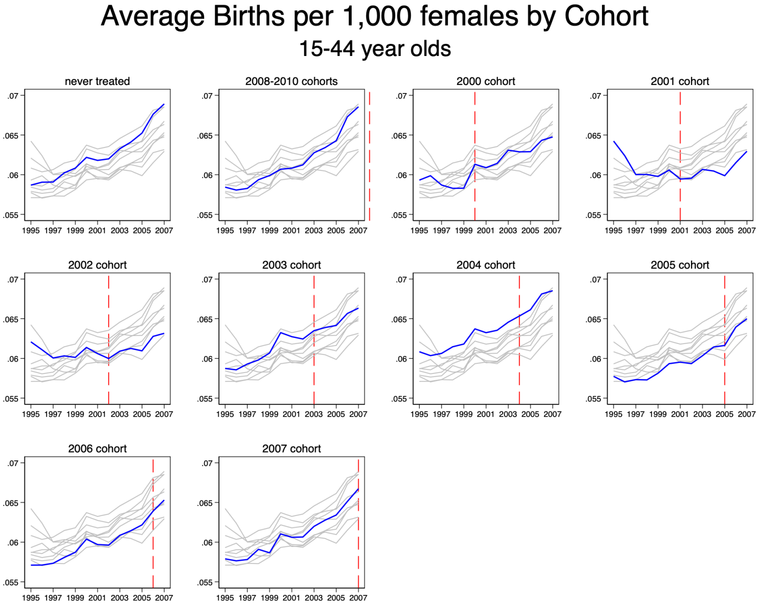
Table: Number of counties by year Craigslist personals appeared

Treatment Cohort	Number of Counties Treated
Never treated	1,779
2000 cohort	9
2001 cohort	5
2002 cohort	12
2003 cohort	36
2004 cohort	58
2005 cohort	69
2006 cohort	341
2007 cohort	65
2008 cohort	66
2009 cohort	215
2010 cohort	1
Total counties	2,656

Rollout of Craigslist's Personals



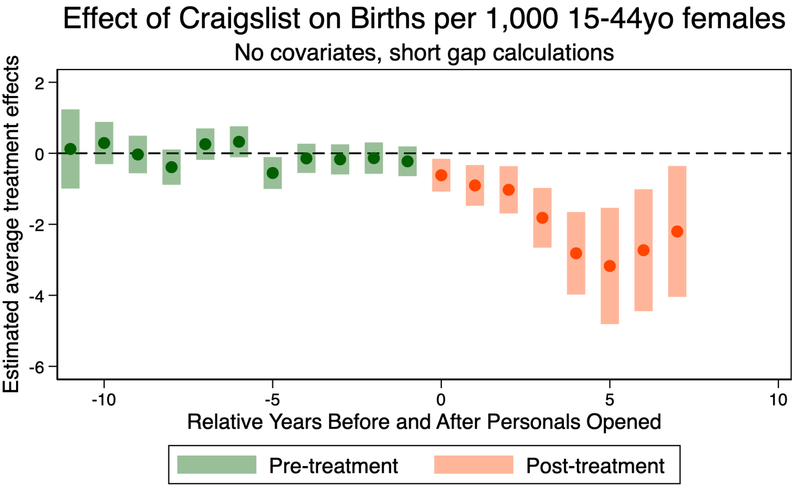
Births by Treatment Cohort



Estimator selection

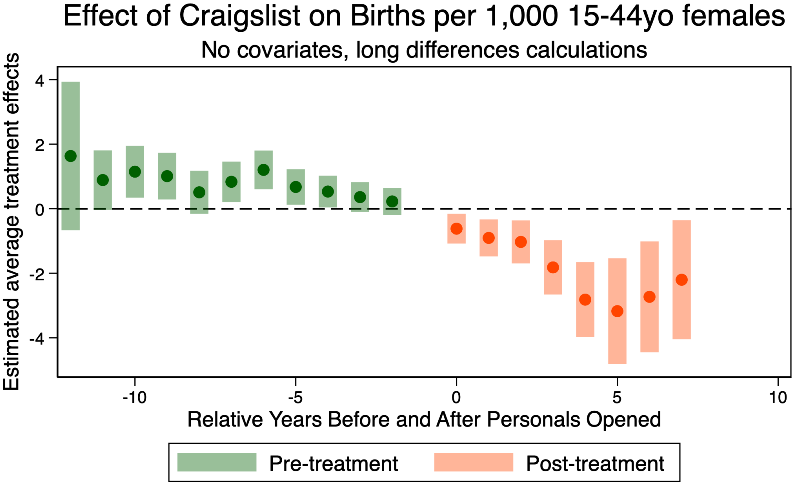
- We used CS to estimate the average effect of the rollout on birth rates
- It’s unweighted by population so ”average effect on the average county for now
- We decided to make our control group the ”2008-2010” as our control group because at least they were eventually treated, but doesn’t really matter as never-treated gets same things
- First, I want to just illustrate for you the ”short vs long difference” issue

Short gap event study, no covariates



Note: Uses the 2008–2010 eventually treated and the not-yet-treated counties as controls, but no covariates. Circles are ATT(g,t) estimates by relative event time. All groups and bands are 95% uniform confidence intervals. Mean birth rate was approximately 0.062 in 2000.

Long differences event study, no covariates



Note: Uses the 2008–2010 eventually treated and the not-yet-treated counties as controls, but no covariates. Circles are ATT(g,t) estimates by relative event time. All groups and bands are 95% uniform confidence intervals. Mean birth rate was approximately 0.062 in 2000.

Trends or No Trends?!

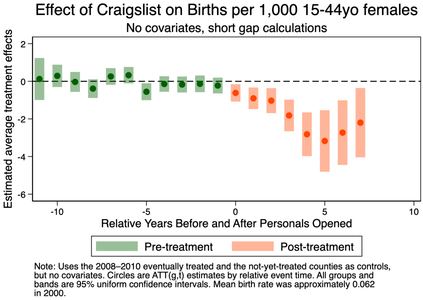


Figure: Short Gap

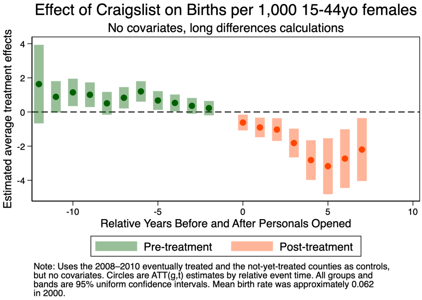


Figure: Long Difference

Scrutinizing that event study

- I was so excited when I saw those flat pre-trends until I learned that the default syntax in Stata's `csdid` calculated short gaps!
- We had a whole story that online dating caused ”permanent dating” until I realized that!! lol
- But more than that – was unconditional parallel trends really plausible?
- We are assuming the mean change in ”untreated birth rates” is the same in our timing cohorts as the comparison cohort, but that assumes a particular ”similar control group” and we have no covariates

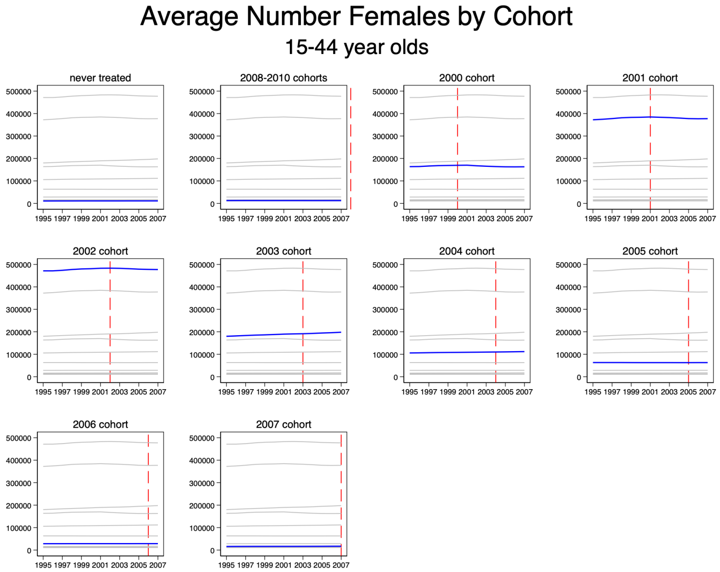
Who got a Craigslist and who didn't?

- Craigslist is this: `houston.craigslist.org` for Houston, Texas
- I grew up in a town in Mississippi called Brookhaven with a population of 11,000 and if you search for `brookhaven.craigslist.org` it's not there
- Craigslist targeted *cities* but most of the US counties are *rural*

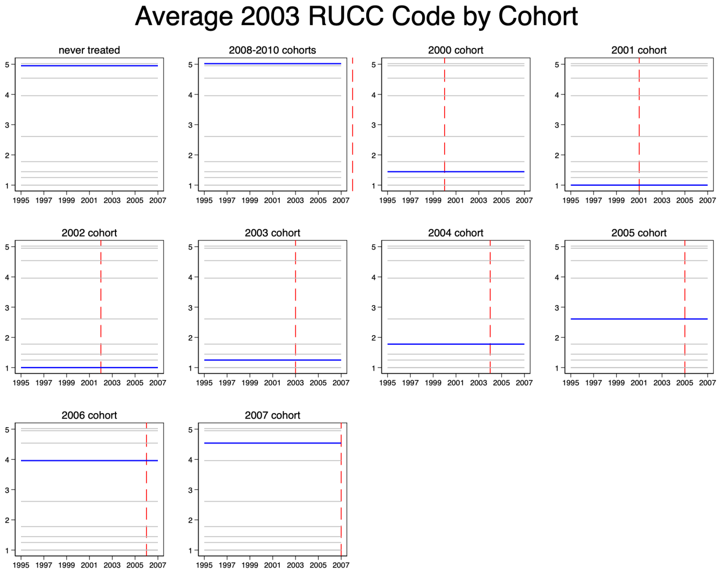
Is Unconditional parallel trends plausible?

- But my coauthors are demographers who focus a lot on maternal health and children, plus they're labor economists and they urban counties and rural counties differ:
  - → Cities have more college educated women, delayed childbearing, lower marriage rates, different racial composition, more access to female healthcare resources
  - → Rural towns have lower educated women, earlier age at first birth, higher marriage rates, more homogenous, worse healthcare access
- If Craigslist counties are more urban, then our treatment and control are imbalance on covariates that cause trends in Y^0
- Let's look at mean female population and urban measures by timing cohort

Female population by timing cohort



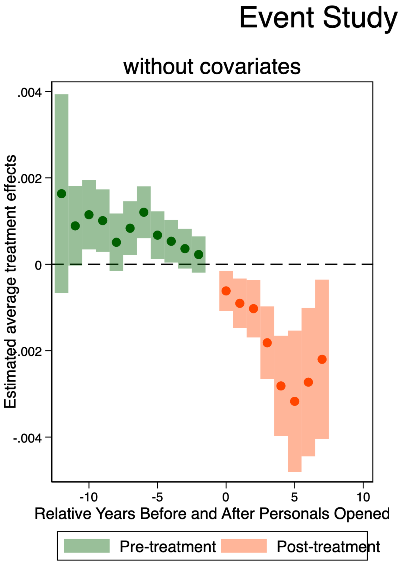
Urban and rural counties by timing cohort



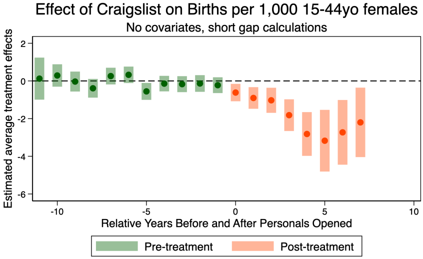
Control for 9-digit urban code

- So we decided – and I still remember this day – to just include one variable
- 9-digit RUCC code measuring ”how urban is the county?”
  - → RUCC code of 1: VERY URBAN (e.g., San Francisco)
  - → RUCC code of 9: VERY RURAL (e.g., Brookhaven)
- Effectively, what is happening when we do this, it is as though we are estimating CS for 1s, 2s, 3s, and so on
- Our control group is so large they have *every RUCC code*
- So, we estimate CS with 8 RUCC dummies (more dramatic if I show it to you)

Long Differences, with and without RUCC dummies

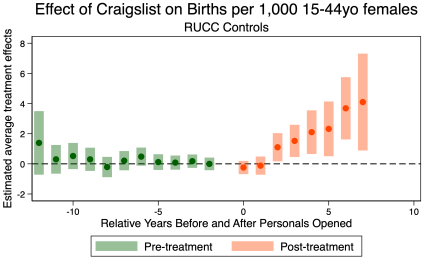


Now put yourself in my shoes!



Note: Uses the 2008–2010 eventually treated and the not-yet-treated counties as controls, but no covariates. Circles are ATT(g,t) estimates by relative event time. All groups and bands are 95% uniform confidence intervals. Mean birth rate was approximately 0.062 in 2000.

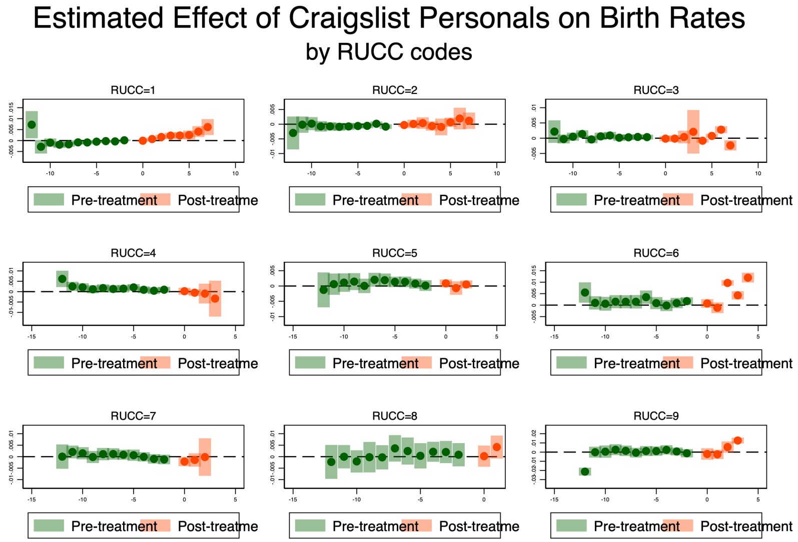
Figure: Short Gap, no covariates



Note: Uses the 2008–2010 eventually treated and the not-yet-treated counties as controls, and eight 2003 RUCC code dummies. Circles are ATT(g,t) estimates by relative time. All groups and bands are 95% uniform confidence intervals. Meanbirth rate was approximately 0.062 in 2000.

Figure: Long Difference, 8 RUCC dummies

Breaking down CS by RUCC Code



Smaller RUCC codes are more urbanized counties. Models estimated separately with CS and not-yet-treat

Any more covariates?

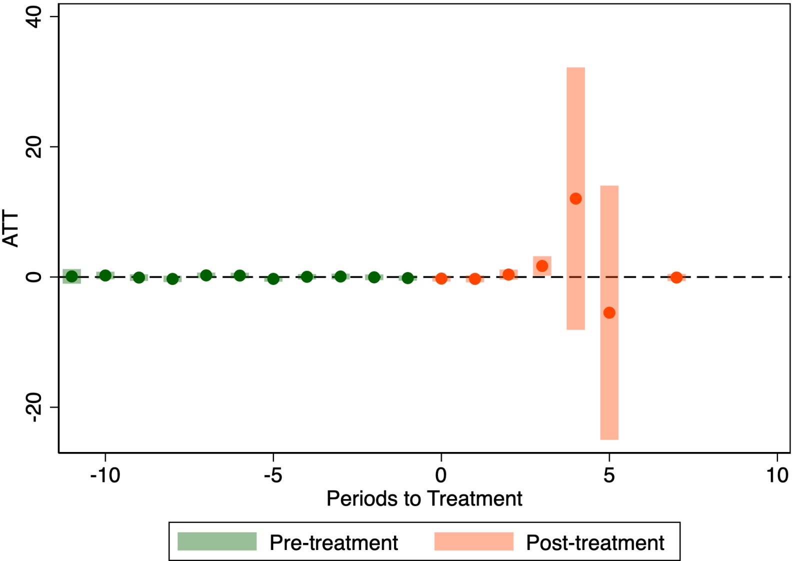
- So then we started asking ourselves – is urban enough? And how will we decide?
- We decided we urban dummies captured a lot of things, but we wanted more, but what and how will we decide so limit specification searching
- We decided to LASSO on $\Delta Birth\_rates$ using only the pretreatment periods because $\Delta Y_{t-\tau} = \Delta Y^0_{t-\tau}$ then

Specific LASSO steps

1. Regress birth rates on state-year interactions for 1995-1999 (pre-treatment)
2. Took first difference per county for residuals
3. LASSO regression of first difference residuals (county-level)
4. Cross validation

LASSO selected male to female sex ratio, per capita income, and unemployment rates (out of 11 covariates)

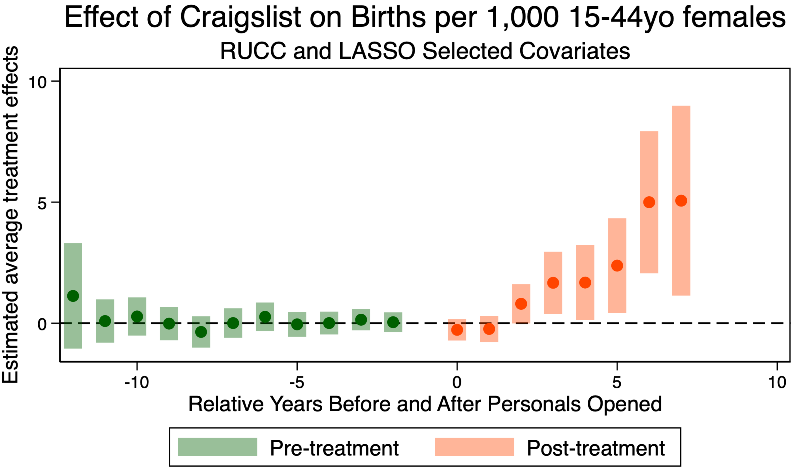
CS estimator with 8 RUCC codes, sex ratio, per capita income and unemployment rate



Yikes!

- What’s going on? Overlap problems
- We are creating ”separation” caused by curse of dimensionality
- All of the propensity scores are ending up not overlapping
- So we had to discretize the covariates into quantiles

CS estimator with 8 RUCC codes, sex ratio, per capita income and unemployment rate quantiles



Note: Uses the 2008–2010 eventually treated and the never treated counties as controls. Covariates include 2003 RUCC code dummies, above-median sex ratio, above-median per capita income, and above-median unemployment rate dummies. Circles are ATT(g,t) estimates by relative time. All groups and bands are 95% uniform confidence intervals. Mean birth rate was approximately 0.062 in 2000.

Interpreting those late lags

- We have staggered rollout – why does that matter?
- All cohorts contribute $t = 0$, but we lose one cohort at each lag
- We don't find positive effects until 3rd lag and there's two things happening
    1. On the one hand, it takes at least 9 months after meeting "the one" to have a child, so lags are reassuring
    2. On the other hand, CS is literally losing cohorts with each lag so mechanical sample selection

Number of counties by treatment cohorts

Table: Number of counties by year Craigslist personals appeared

Treatment Cohort	Number of Counties Treated
Never treated	1,779
2000 cohort	9
2001 cohort	5
2002 cohort	12
2003 cohort	36
2004 cohort	58
2005 cohort	69
2006 cohort	341
2007 cohort	65
2008 cohort	66
2009 cohort	215
2010 cohort	1
Total counties	2,656

Differentially sized cohorts

- But there’s more – those ”late adopters” are *massive* which in CS is causing that alone them to shift the effects
- Simple ATT is not significant with the 2006 cohort, but it is when it’s gone
- But it’s just size – those late adopters are *rural*, they are *much later*
- Lots of different things about them and we are working on that now

Writing this up will be horrible!

- I don’t mind showing you all this because it’s a workshop about doing the work
- Mark and Dan are going to talk about writing the papers!
- But, we have *no idea* how we are going to somehow collect all of this and make one coherent paper (we will be taking Mark and Dan’s hidden curriculum workshop!)
- Point is, though, that that’s a different part of the process – we want to understand *mechanically* what CS is doing, *numerically* the calculations (e.g., the weighting, the ”dropping out cohorts”)
- We don’t just want to jump to some gestation story in other words

Be Curious, Not Judgment – Ted Lasso

- Do everything you can to figure out the treatment assignment mechanism
- Shoeleather dominates technical knowledge (Rubin 2008 ”Design Trumps Analysis”; Freedman 1991)
- Consider interviewing key people! But don’t ask them if they randomized – they don’t know what that word means
- Ask them ”hey, why did you do that thing you did?” or ”hey, why did some people get put into your program but not others?”

2:52 5G E

New message

craig newmark

decision at all. I don't want to
ask you to share things you're not comfortable sharing, and understand if you can't, but I thought I'd ask. Hope you are well.

Scott

1/17/18, 12:43 PM ✓

Scott, thanks! but I haven't been a craigslist spokesman nor in management since 2000, have had no part in decision making. However, you're asking a question involving sophisticated marketing, economic analysis, and business development, and as a matter of public record, craigslist doesn't do that. Thanks!

1/17/18, 12:47 PM

Thanks Craig. That's a

GIF Start a message

1

Roadmap

Differential timing or $G \times T$

- Bacon Decomposition
- Callaway and Sant'Anna (CS)
- Calculating event study coefficients with CS

Checklists and My Online Dating Project

Alternative Estimators and Sensitivity Analysis

- Sun and Abraham (SA)
- de Chaisemartin and D'Haultfoeulle (dCDH)
- Honest DID

DDDiD

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

Bias of TWFE Event Study Specification

- Bacon only focused on the static specification, and that's where the biases due to dynamics revealed itself
- 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)

Event study specification with TWFE

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

We will focus on the coefficient $\widehat{\mu_g}$ when estimated with TWFE

Sun and Abraham 2021

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 Kolesar 2022)
3. SA presents an alternative estimator that is a version of CS only using the "last cohort" as the comparison 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 $CATT_l$ 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^∞)
- $Y^\infty_{i,t}$ 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
  1.      1   .3257218   1980   1    1
  2.      1   .3257218   1981   2    1
  3.      1   .3257218   1982   3    1
  4.      1   .3257218   1983   4    1
  5.      1   .3257218   1984   5    1
```

Definition 1

Definition 1: The cohort-specific ATT l periods from initial treatment date e is:

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

Fill out the second part of the Group-time ATT exercise together.

TWFE assumptions

- For consistent estimates of the coefficient leads and lags using TWFE model, we need three assumptions
- For SA and CS, we only need two
- Let's look then at the three

Assumption 1: Parallel trends

Assumption 1: Parallel trends in baseline outcomes:
$E[Y^{\infty}_{i,t} - Y^{\infty}_{i,s} | E_i = e]$ is the same for all $e \in supp(E_i)$ and for all s, t and is equal to $E[Y^{\infty}_{i,t} - Y^{\infty}_{i,s}]$

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).

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

Weight ($w^g_{e,l}$) 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:

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

Still biased under parallel trends and no anticipation

Proposition 3: If parallel trends holds and no anticipation holds for all $l < 0$ (i.e., no anticipatory behavior pre-treatment), then the population regression coefficient μ_g for g is a linear combination of post-treatment $CATT_{e,l'}$ for all $l' \geq 0$.

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

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 (2021) propose a 3-step interacted weighted estimator (IW) using last treated group as control group
- Callaway and Sant'anna (2021) 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 $\widehat{\delta}_{e,l}$

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e \notin C} \sum_{l \neq -1} \delta_{e,l}(1\{E_i = e\} \cdot D^l_{i,t}) + \varepsilon_{i,t}$$

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***	

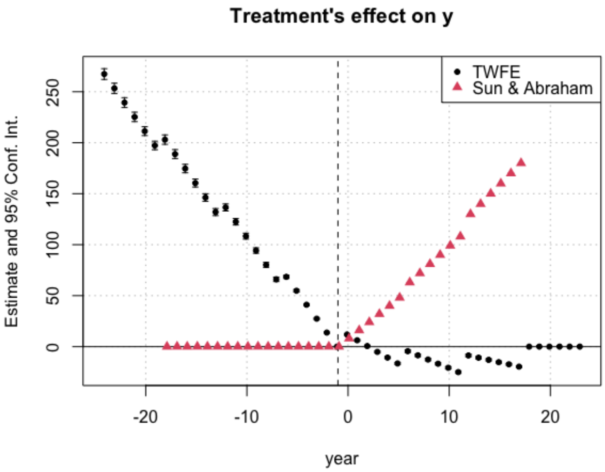
Computing relative event time leads and lags

	Truth				
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)	
1980	0	0	0	0	
1986	10	0	0	0	(10+8+6)/3 = 8
1987	20	0	0	0	(20+16+12)/3 = 16
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	

Relative time coefficients		
Leads	Truth	SA
t-2	0	0.02
t	8	8.01
t+1	16	16.00
t+2	24	24.00
t+3	32	31.99
t+4	40	40.00
t+5	48	48.01
t+6	63	62.99
t+7	72	72.00
t+8	81	80.99
t+9	90	89.98
t+10	99	99.06
t+11	108	108.01
t+12	130	129.92
t+13	140	139.92
t+14	150	150.01
t+15	160	159.97
t+16	170	169.99
t+17	180	179.98

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 reverse engineering and forward engineering
- 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 (but assumptions and control group needed might surprise you)
- 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):

$$\Delta^g_{i,t} = Y^1_{i,t} - Y^\infty_{i,t}$$

  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).
- Very important theorem – established the "no sign flip property" for OLS with differential timing in the canonical static specification

OLS Weighting

- 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
- The weights in OLS all come out of the model itself, *not the economic question*

OLS Weighting

- What makes something a good vs a bad weight?
- Not being negative is the absolute minimal requirement
- But that's the minimum – we mainly are trying to weight to the target parameter, not justify the use of a model
- It is also not a good sign if you can't really explain the weights

dCdH Solution

- dCdH propose an alternative that doesn't have the problems of TWFE – both avoiding negative weights and improving interpretability
- Their model can handle reversible treatments, but in the context of differential timing is equivalent to CS and SA with a particular choice of weights
- For diagnostic purposes, they recommend reporting the number/fraction of group-time ATTs that receive negative weights, as well as the degree of heterogeneity in treatment effects that would be necessary for the estimated treatment effect to have the "wrong sign"

DID_{M} Estimator – Introduction

- DID_{M} estimator from dCDH (de Chaisemartin and D'Haultfoeulle, 2020) estimates treatment effects around each treatment transition.
- Separately captures effects for:
  - $\rightarrow$ "Joiners" (entering treatment)
  - $\rightarrow$ "Leavers" (exiting treatment)
- Defined as weighted average:

$$DID_M = \text{weighted average of } \text{DID}_{+,t} \text{ and } \text{DID}_{-,t}$$

- Avoids TWFE negative weighting problem.

Estimating $\text{DID}_{+,t}$ ("Turning On")

- For units that begin treatment at time t:

$$DID_{+,t} = \underbrace{(Y_t^{newly\ treated} - Y_{t-1}^{newly\ treated})}_{\text{Change for joiners}} - \underbrace{(Y_t^{untreated} - Y_{t-1}^{untreated})}_{\text{Change for untreated}}$$

- Compares outcomes of "joiners" to those never treated.
- Similar conceptually to Callaway & Sant'Anna and Sun & Abraham in scenarios where treatment turns on.

Estimating DID$_{-,t}$ (”Turning Off”)

- For units exiting treatment at time t, their estimator identifies the effect of ”stopping treatment”

$$\text{DID}_{-,t} = \underbrace{(Y_t^{\text{leavers}} - Y_{t-1}^{\text{leavers}})}_{\text{Change for leavers}} - \underbrace{(Y_t^{\text{continuously treated}} - Y_{t-1}^{\text{continuously treated}})}_{\text{Change for continuously treated}}$$

- You are now missing Y^1 in this new causal effect, so you need a control group whose outcome is treated (Y^1)
- Whatever treatment state the ”exiting group” had been at in baseline, the control group must be too (well defined treatment statuses again)

Combining $DID_{+,t}$ and $DID_{-,t}$

- DID_M combines these into a single estimate:

$$DID_M = \sum_t (\text{weights}_t \cdot DID_{+,t}) + \sum_t (\text{weights}_t \cdot DID_{-,t})$$

- Weights typically based on group size or variance.
- Simplifies to weighted average of $DID_{+,t}$ when no units revert (staggered adoption without exit).

Key Assumption: No Carryover

- Important assumption: treatment effects disappear immediately after treatment stops (not treated if not treated)
- If treatment effects linger, estimator will underestimate the true effect of exiting treatment.
- Practical consideration: Is turning off treatment truly reverting to baseline, or is it moving into a different treatment state?

Control Group Must Be Treated Continuously

- When treatment is "switching off" (i.e., $D_i = 1$ to $D_i = 0$), you're going to need as your control group "always treated units"
- This is again because when treatment is turned off, the unit has gone from $Y = Y^1$ at the new baseline to $Y = Y^0$
- Means that for any treatment effect, the missing potential outcome will be Y^1, and so you'll need at both periods units "always treated"
- Needs parallel trends in Y^1 therefore – which means they cannot be "the first to leave treatment"

Comparison to Callaway & Sant'Anna and Sun & Abraham

- CS and SA primarily handle switching on, but switching off
- DID_M extends naturally to reversible treatments (on and off transitions)
- But you have to be sure that parallel trends holds for the comparison group (who is continuously treated) – again, you're not using all the data
- All three methods avoid negative weighting problems inherent in traditional TWFE.

Sensitivity Analysis

- Assume the worst – use the absolute worst gap in pre-trends and imagine PT broke by that much post-treatment
- How bad does that have to get before your treatment effect coefficient covers zero?
- Called `honestdid` by Rambachan and Roth (2023)
- Don't think of it as rejecting PT – it's just saying how dependent on it you are

Sensitivity Analysis

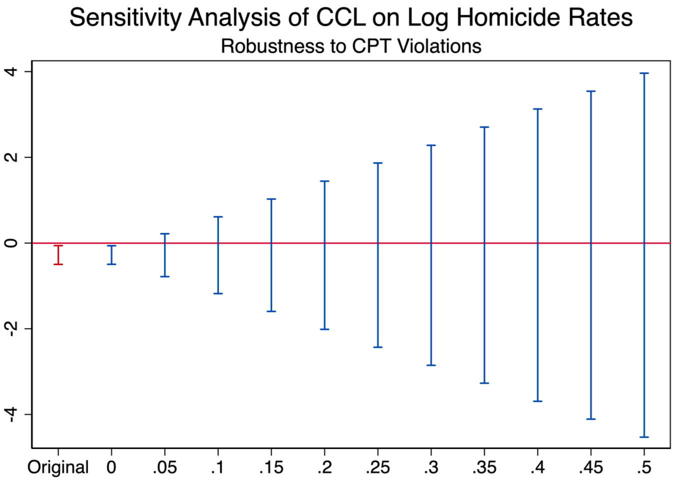


Figure 97: Robustness of Estimated Simple ATT to Parallel Trends Violations Using the Rambachan and Roth [2023] `honestdid` Bounding Approach

Roadmap

Differential timing or $G \times T$

- Bacon Decomposition
- Callaway and Sant'Anna (CS)
- Calculating event study coefficients with CS

Checklists and My Online Dating Project

Alternative Estimators and Sensitivity Analysis

- Sun and Abraham (SA)
- de Chaisemartin and D'Haultfoeulle (dCDH)
- Honest DID

DDDiD

Final Step is to Do DDDiD

- DDDiD is a powerful new estimator by Guido Imbens that you use when parallel trends doesn’t hold

Final Step is to Do DDDiD

- DDDiD is a powerful new estimator by Guido Imbens that you use when parallel trends doesn't hold
- "Don't Do Diff-in-Diff"

Goal was never to use diff-in-diff though

- Chainsaws are amazing but that doesn’t mean you should try to use them to sharpen pencils

Goal was never to use diff-in-diff though

- Chainsaws are amazing but that doesn't mean you should try to use them to sharpen pencils
- If you simply do not believe parallel trends assumption holds in your data, for whatever reason, then diff-in-diff is the wrong estimator
- Some things can be good for some stuff but not other things and that's okay
- Besides – our goal was never to use diff-in-diff
- Our goal was to get good, believable answers to good questions and then tell people about them in truthful, careful, non-confusing ways
- Kyle is going to talk about imputation estimators as alternatives to diff-in-diff!