Recent Developments in Differences-in-differences

Scott Cunningham (Baylor)

Differences-in-differences

- DD is a very old and ingenious research design dating back at least to John Snow's effort to confirm that cholera was waterborne in the 19th century
- It was introduced into economics via Orley Ashenfelter in the late 1970s and then popularized through his student David Card (with Alan Krueger) in the 1990s
- It is now the single most popular research design in applied microeconomics, even more popular by a significant margin than RCTs

Bias in our go-to estimators

- The most common DD situation is one in which a treatment is adopted by different groups at different times
- And the most popular way to estimate the ATT in those situations is to use OLS with time and panel unit fixed effects – now commonly called the "twoway fixed effects (TWFE)" estimator
- But new econometric work of the last two years has shown that OLS models can be (and probably are in practice) severely biased
- I'll discuss the bias of TWFE, discuss a new solution, and fingers crossed a simulation if we have time

Population expectations for the simple 2x2

A 2x2 is a simple difference over time between a treated group (before treated and after treated) and an untreated group (same before and after)

$$\widehat{\delta}^{2x2}_{kU} \quad = \quad \bigg(E[Y_k|Post] - E[Y_k|Pre] \bigg) - \bigg(E[Y_U|Post] - E[Y_U|Pre] \bigg)$$

Non-parallel trends bias

With some simple algebra, we get

$$\begin{aligned}\widehat{\delta}^{2x2}_{kU} &= \underbrace{E[Y^1_k|Post] - E[Y^0_k|Post]}_{\text{ATT}} \\ &+ \underbrace{\Big[E[Y^0_k|Post] - E[Y^0_k|Pre]\Big] - \Big[E[Y^0_U|Post] - E[Y^0_U|Pre]\Big]}_{\text{Non-parallel trends bias in 2x2 case}}\end{aligned}$$

An unbiased estimate of the ATT with data needs that bias to be zero, which we get that with parallel trends.

But this is only the case for the simple 2x2

2x2 versus differential timing

- Parallel trends is **not enough** for TWFE to be unbiased when treatment adoption is described by differential timing
- Let's look at the paper by Goodman-Bacon (2018; 2019) which explains why
- But here's the problem that I hope to show you – TWFE with differential timing is flawed *because* it cannot help but use already-treated groups as controls

Decomposition Preview

- TWFE estimates a parameter that is interestingly enough a weighted average over all 2x2 in your sample
- Some of these 2x2 use already treated units act as both controls and treatment – and TWFE can't be stopped either!
- Also problematically, TWFE assigns weights that are a function of sample sizes of each "group" and the variance of the treatment dummies for those groups
- And none of that is even theoretically coherent, but TWFE does it anyway

Decomposition (cont.)

- TWFE needs two assumptions: that the variance weighted common trends are zero (far more parallel trends iow) and no dynamic treatment effects (not the case with 2x2)
- Under those assumptions, TWFE estimator estimates the variance weighted ATT is a weighted average of all possible ATTs

A simple example may help drive this home

- Suppose 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 later than l
- Denote $\overline{D}_k$ as the share of time each group spends in treatment status
- Denote $\widehat{\delta}_{ab}^{2x2,j}$ as the canonical 2×2 DD estimator for groups a and b where j is the treatment group
- How many 2×2 combinations are there? More than you think

How many 2x2?

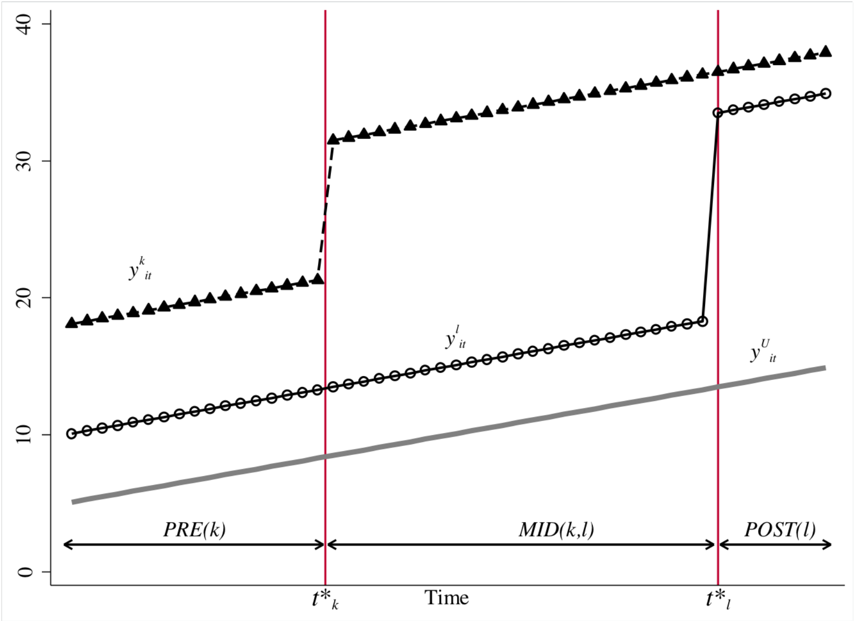
- When there's three groups, there are four 2x2s
- But typically, we have more than 3 groups making so the number of potential 2x2 even larger
- With K timing groups and one untreated group, you get K^2 distinct 2x2 DDs

K^2 distinct DDs

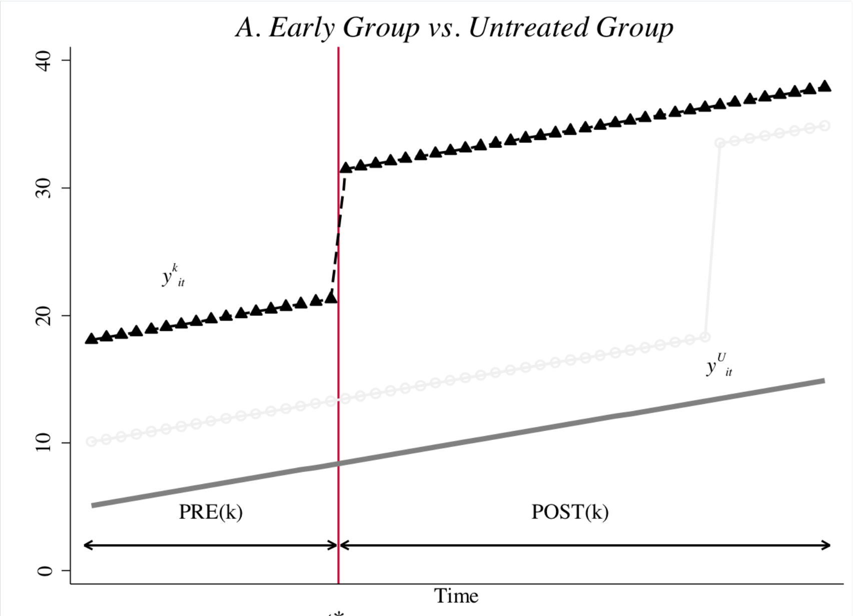
Assume 3 timing groups (a, b and c) and one untreated group (U). Then there should be 9 2x2 DDs. Here they are:

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

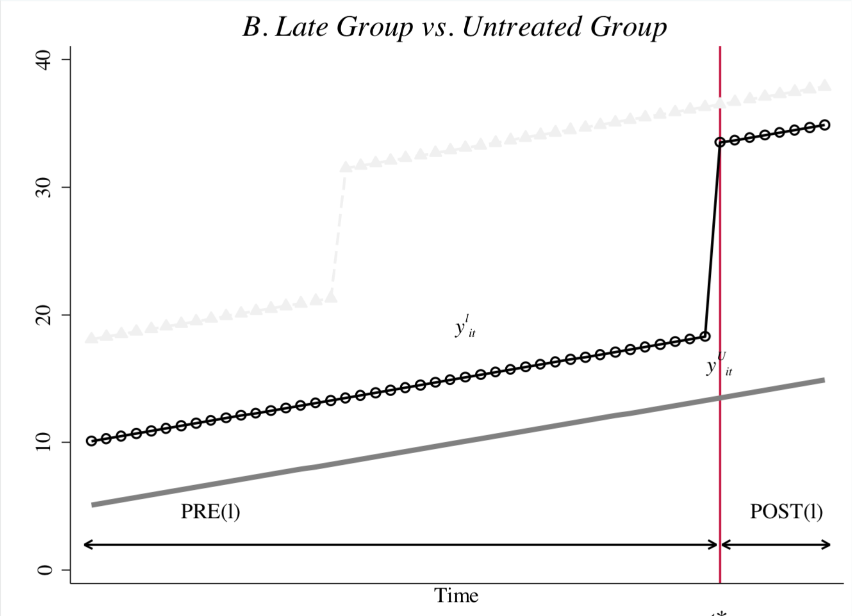
Let's return to our simpler example with k group treated at t_k^* and l treated at t_l^* plus the U untreated group


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

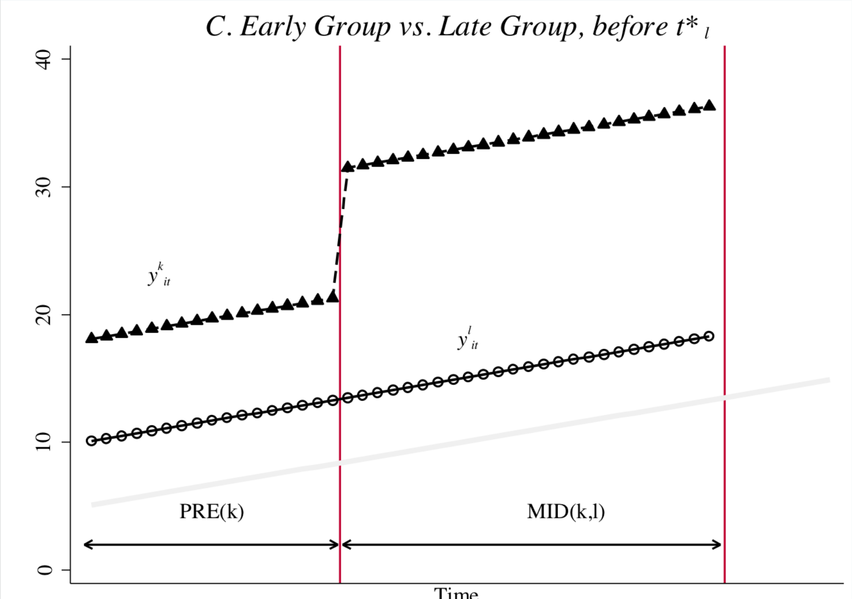
$$\widehat{\delta}^{2x2}_{kU} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)}\right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)}\right)$$



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

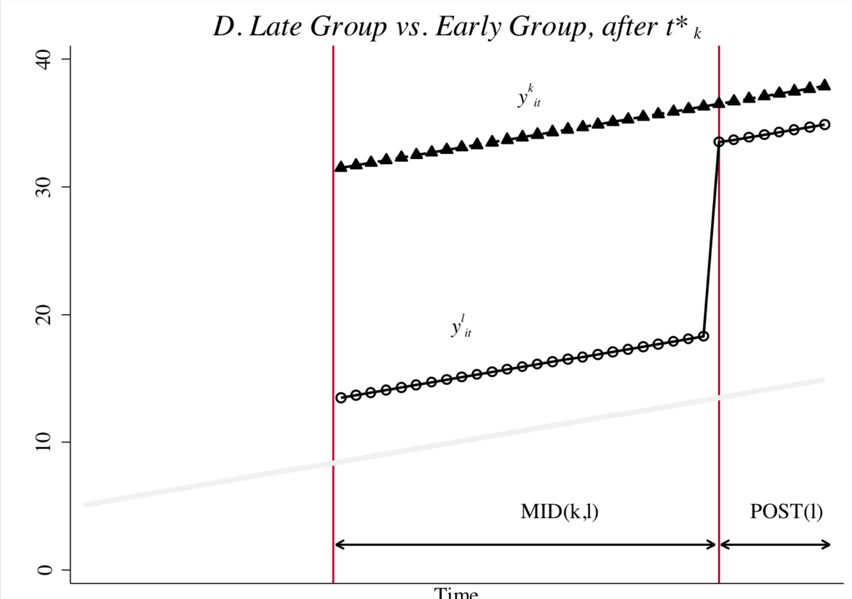


$$\delta_{kl}^{2x2,k} = \left(\overline{y}_k^{MID(k,l)} - \overline{y}_k^{Pre(k,l)}\right) - \left(\overline{y}_l^{MID(k,l)} - \overline{y}_l^{PRE(k,l)}\right)$$



C. Early Group vs. Late Group, before t^*_l

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



D. Late Group vs. Early Group, after t^*_k

Bacon decomposition

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

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

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 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 pop out of these weights
  - "Group" variation matters more than unit-level variation. A group is if two states got treated in 1995. They are the 1995 group. More units in a group, the bigger that 2x2 is practically
  - Within-group *treatment* variance matters a lot.
- Think about what causes the treatment variance to be as big as possible. Let's think about the s_{ku} weights.
  1. $\overline{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
  2. $\overline{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
  3. $\overline{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$
- 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 $0.52 \times 0.48 = 0.2496$.

TWFE and centralities

- Groups in the middle of the panel weight up their respective 2x2s via the variance weighting
- This is the first thing about TWFE that should give us pause, as not all estimators do this, and it's not theoretically clear why we should care either
- Highlights the strange role of panel length – should you start at 5 years before first treatment or 10? What about post-treatment?
- Different choices about panel length, which maybe we didn't give much thought to, change not only the 2x2 *but also* the weights based on variance of treatment

Moving from 2x2s to causal effects and bias terms

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

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

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:

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

- The first part is the ATT we are looking for
- The second part is the bias from non-parallel trends from mid to post period
- The third is new: a heterogeneity bias if the ATT for *k* is *dynamic*. If not, then it just zeroes out.

Substitute all this stuff into the decomposition formula

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

where we will make these substitutions

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

Notice all those potential sources of biases!

Potential Outcome Notation

$$\begin{aligned} p\ lim\ \widehat{\delta}^{DD}_{n\to\infty} &= \delta^{DD} \\ &= VWATT + VWCT - \Delta ATT \end{aligned}$$

- 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!
- Let's look at each of these three parts more closely

Variance weighted ATT

$$\begin{aligned} VWATT &= \sum_{k \neq U} \sigma_{kU} ATT_k(Post(k)) \\ &+ \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \Big[\mu_{kl} ATT_k(MID) + (1 - \mu_{kl}) ATT_l(POST(l)) \Big] \end{aligned}$$

where σ is like s only population terms not samples.

- Weights sum to one.
- Note, if all the ATT are identical, then the weighting is irrelevant.
- But otherwise, it's basically weighting each of the individual sets of ATT we have been discussing, where weights depend on group size and variance

Variance weighted common trends

$$
\begin{aligned}
VWCT &= \sum_{k \neq U} \sigma_{kU} \Big[\Delta Y_k^0(Post(k), Pre) - \Delta Y_U^0(Post(k), Pre) \Big] \\
&+ \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \Big[\mu_{kl} \{ \Delta Y_k^0(Mid, Pre(k)) - \Delta Y_l^0(Mid, Pre(k)) \} \\
&+ (1 - \mu_{kl}) \{ \Delta Y_l^0(Post(l), Mid) - \Delta Y_k^0(Post(l), Mid) \} \Big]
\end{aligned}
$$

This is new. That's a lot of parallel trends we need equalling zero, and this was only with **two treatment groups**!

Heterogeneity bias

$$\Delta ATT = \sum_{k \neq U} \sum_{l > k} (1 - \mu_{kl}) \Big[ATT_k(Post(l) - ATT_k(Mid)) \Big]$$

Now, if the ATT is constant over time, then this difference is zero, but what if the ATT is not constant? Then TWFE is biased, and depending on the dynamics and the VWATT, may even flip signs

Alternatives to TWFE

- New papers are coming out focused on the issues that we are seeing with TWFE
- I'll discuss one though by Callaway and Sant'anna (2019) (currently R&R at Journal of Econometrics) due to time constraints (call it CS)
- If we have time, I'll run through a simulation illustrating both the bias of TWFE and the unbiased estimation of this CS estimator

When might you use this estimator

Probably in the very situations plaguing your own study

1. When treatment effects heterogenous by time of adoption
2. When treatment effects change over time
3. When shortrun effects more pronounced than longrun effects
4. When treatment effect dynamics differ if people are first treated in a recession relative to expansion years

Preliminary

CS considers identification, estimation and inference procedures for ATT in DD designs with

1. multiple time periods
2. variation in treatment timing (i.e., differential timing)
3. parallel trends only holds after conditioning on observables

Group-time ATT is the parameter of interest in CS

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

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
- Their method will calculate an ATT per group/time which yields *many* individual ATE estimates
- Group-time ATT estimates are not determined by the estimation method one adopts (first difference or FE)
- Does not directly restrict heterogeneity with respect to observed covariates, timing or the evolution of treatment effects over time
- Provides a way to aggregate over these to get a single ATT
- Inference is the bootstrap

Notation

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

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

Assumptions

Assumption 1: Sampling is iid (panel data)

Assumption 2: Conditional parallel trends

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

Assumption 3: Irreversible treatment

Assumption 4: Common support (propensity score)

CS Estimator

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

CS stops us from using already-treated as controls as that is a sin!

Remarks about “staggered adoption” with universal coverage

Proof.

Remark 1: In some applications, eventually all units are treated, implying that C is never equal to one. In such cases one can consider the “not yet treated” ($D_t = 0$) as a control group instead of the “never treated?” ($C = 1$). □

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"
- Also allows for estimating pre-treatment coefficients as TWFE also does those badly (Sun and Abraham (2020))
- Inference from a bootstrap

Stata simulation

Let's now review a simulation in Stata which can be downloaded from my github repo called `baker.do`

Concluding remarks on DD

- Prediction: You're going to write a DD paper, and you probably will want to use TWFE because it's *easy to do*. Do not.
- Goodman-Bacon (2018, 2019) shows the bias of TWFE. It suffers from attenuation bias, and it's theoretically possible that it flips the sign!
- CS is an alternative that doesn't suffer from any of those TWFE biases. R package available from github.
- But there are others such as Athey, et al (2018) matrix completion for panel data, the stacked method by Cengiz, et al (2019), and several papers by de Chaisemartin and D'Haultfoeuille
- Remember: never use already-treated units as a control as that's a sin and that's the source of TWFE bias