Causal Inference using Difference-in-Differences

Lecture 11: The Problems of TWFE with Staggered Treatment Adoption

Pedro H. C. Sant'Anna
Emory University

January 2025

Summary of previous lecture

DiD procedures with multiple periods

- We discuss how we can learn about treatment effect dynamics: $ATT(g,t)$'s
- We discussed estimator and inference procedures for them.
- Discuss how we can assess the plausibility of our assumptions by looking at pre-treatment periods.
- Parallel re-treatment trends: not necessary nor sufficient for post-treatment PT.
- Failing to reject pre-trends differs greatly from having evidence favoring it.
  - There are some important power issues, see, e.g., Roth (2022).
- Discussed how to relax No-Anticipation to Limited Anticipation.
- All this in the context of 2 groups and multiple time periods.

What if we have variation in treatment timing?

Does TWFE "work" in setups with variation in treatment timing?

Traditional methods: TWFE regressions

- We know that, in the 2x2 case,

$$Y_{i,t} = \alpha_0 + \gamma_0 1\{G_i = 2\} + \lambda_0 1\{T_i = 2\} + \underbrace{\beta_0^{twfe}}_{\equiv ATT} (1\{G_i = 2\} \cdot 1\{T_i = 2\}) + \varepsilon_{i,t},$$

- It is tempting to "extrapolate" from this setup and use variations of the following TWFE specification to estimate causal effects:

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

  where dummies $D_{i,t} = 1\{t - G_i \geq 0\}$, where G_i indicates the period unit i is first treated (Group).

- For simplicity, let's assume that treatment is "irreversible": once a unit is treated, it is forever treated - aka staggered design

Does TWFE “work” in setups with variation in treatment timing?

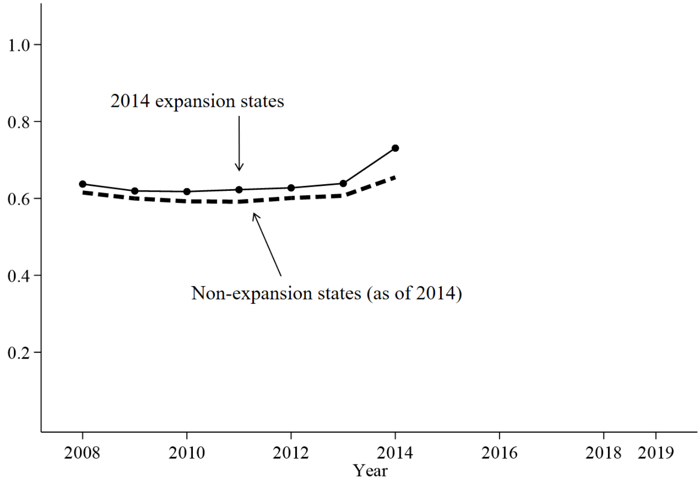
Example: Effect of ACA Medicaid expansion on health insurance rate

Empirical Example: Medicaid Expansion

- To motivate our problem, let's look at a classic example: Medicaid Expansion
- We want to analyze its effect on health insurance rates among low-income, childless adults aged 25-64.

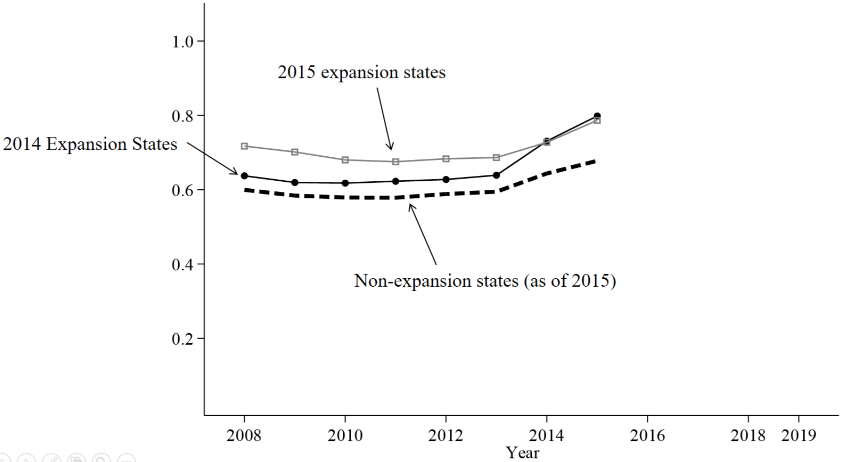
ACA Medicaid expansion circa 2014

Figure 1: Health Insurance Rate (low-income Childless Adults Aged 25-64)



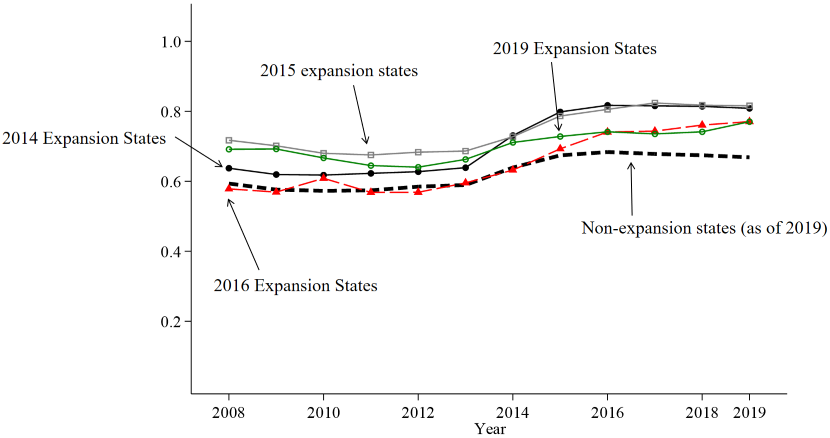
ACA Medicaid expansion circa 2015

Figure 2: Health Insurance Rate (low-income Childless Adults Aged 25-64)



ACA Medicaid expansion circa 2019

Figure 3: Health Insurance Rate (low-income Childless Adults Aged 25-64)



EMORY UNIVERSITY

ACA Medicaid expansion circa 2019

- 23 states expanded circa 2014 - 4 did it earlier (ACA is effectively relabeled), we drop them.
- 3 states expanded circa 2015
- 2 states expanded circa 2016
- 1 states expanded circa 2017
- 2 states expanded circa 2019
- 16 states haven't expanded by 2019

OLS estimate of β

- Let $\widehat{\beta}$ be the OLS estimator of the following TWFE regression specification:

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

- What is $\widehat{\beta}$?
- Goodman-Bacon (2021) shows that we can answer this question following these three steps:
  1. Remove unit means

$$D_{i,t} - \overline{D}_i$$

  2. Remove time means of $(D_{i,t} - \overline{D}_i)$:

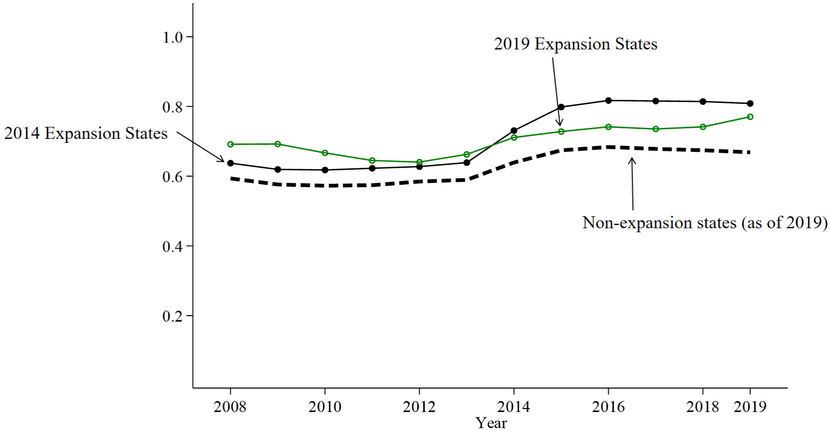
$$\widetilde{D}_{it} = (D_{i,t} - \overline{D}_i) - (\overline{D}_t - \overline{\overline{D}})$$

  3. Calculate univariate regression of $Y_{i,t}$ on $\widetilde{D}_{it}$:

$$\widehat{\beta} = \frac{(nT)^{-1} \sum_{i,t} Y_{i,t} \cdot \widetilde{D}_{it}}{(nT)^{-1} \sum_{i,t} \widetilde{D}_{it}^2}$$

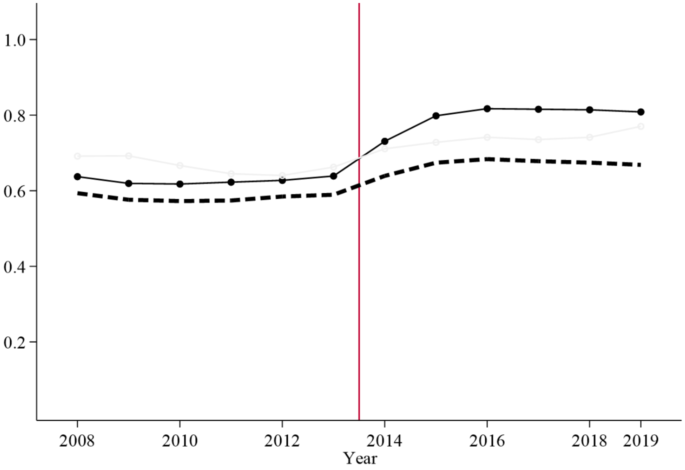
Three groups example

Figure 4: Health Insurance Rate (low-income Childless Adults Aged 25-64)



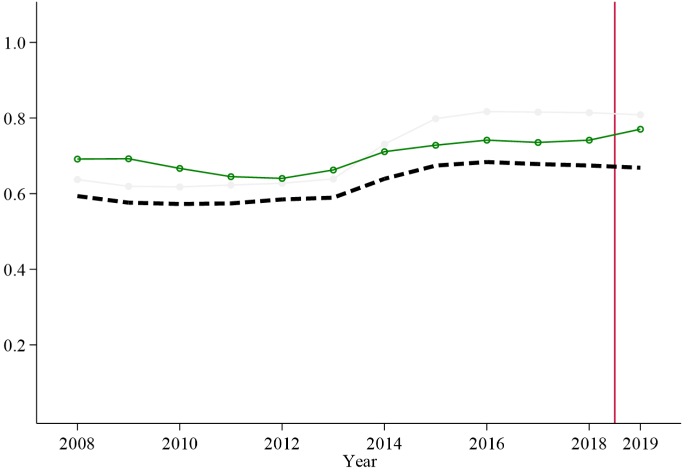
Treated in 2014 vs. Never-Treated

Figure 5: Health Insurance Rate (low-income Childless Adults Aged 25-64)



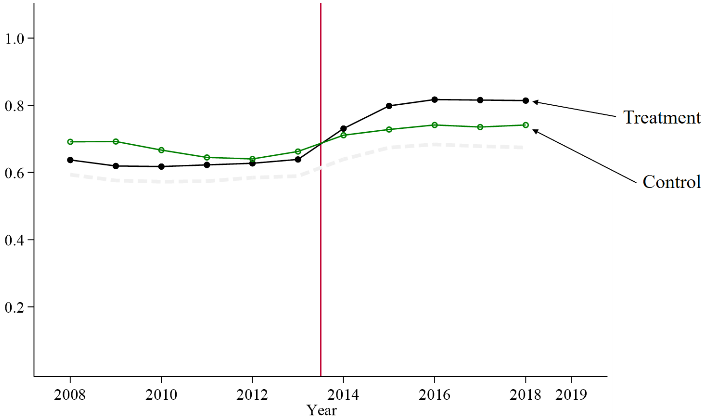
Treated in 2019 vs. Never-Treated

Figure 6: Health Insurance Rate (low-income Childless Adults Aged 25-64)



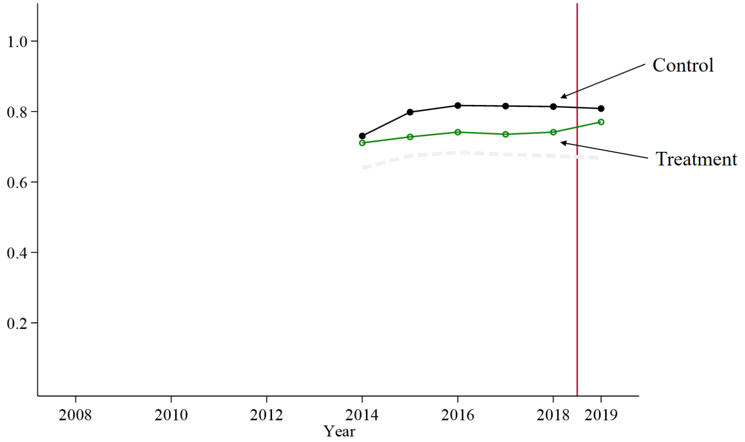
Treated in 2014 vs. Treated in 2019 ($t < 2019$)

Figure 7: Health Insurance Rate (low-income Childless Adults Aged 25-64)



Treated in 2019 vs. Treated in 2014 ($t \geq 2014$)

Figure 8: Health Insurance Rate (low-income Childless Adults Aged 25-64)



OLS estimate of β

- OLS is "variational hungry" and exploit all these 2x2 comparisons.
- But how does OLS aggregate them?
- Main result of Goodman-Bacon (2021) is the Bacon-Decomposition:

$$\widehat{\beta} = s_{k,U} \cdot \widehat{\beta}_{k,U} + s_{\ell,U} \cdot \widehat{\beta}_{\ell,U} + \left[s_{k,\ell} \cdot \widehat{\beta}_{k,\ell} + s_{\ell,k} \cdot \widehat{\beta}_{\ell,k} \right]$$

- In our example:
  - $k = 2014$
  - $\ell = 2019$
  - $U =$ never-treated

Does TWFE "work" in setups with variation in treatment timing?

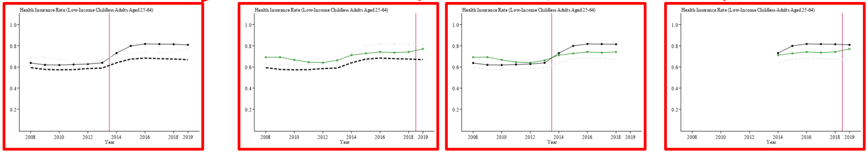
Bacon decomposition

Bacon decomposition

- Main result of Goodman-Bacon (2021) is the Bacon-Decomposition:

Figure 9: Bacon-Decomposition: The 2x2 $\widehat{\beta}$

$$\hat{\beta}^{DD} = s_{kU}\boxed{\hat{\beta}^{DD}_{kU}} + s_{\ell U}\boxed{\hat{\beta}^{DD}_{\ell U}} + [s^{k}_{k\ell}\boxed{\hat{\beta}^{DD,k}_{k\ell}} + s^{\ell}_{k\ell}\boxed{\hat{\beta}^{DD,\ell}_{k\ell}}]$$



Bacon decomposition

- Main result of Goodman-Bacon (2021) is the Bacon-Decomposition:

Figure 10: Bacon decomposition: The weights

$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}^{DD}_{kU} + s_{\ell U}\hat{\beta}^{DD}_{\ell U} + \left[s^{k}_{k\ell}\hat{\beta}^{DD,k}_{k\ell} + s^{\ell}_{k\ell}\hat{\beta}^{DD,\ell}_{k\ell}\right]$$

$$s_{kU} = \frac{(n_k + n_U)^2 n_{kU}(1 - n_{kU})\bar{D}_k(1 - \bar{D}_k)}{V(\tilde{D}_{it})}$$

Sample size^2

$$s^{k}_{k\ell} = \frac{\left((n_k + n_\ell)(1 - \bar{D}_\ell)\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{1 - \bar{D}_\ell}\frac{1 - \bar{D}_k}{1 - \bar{D}_\ell}}{V(\tilde{D}_{it})}$$

$$s^{\ell}_{k\ell} = \frac{\left((n_k + n_\ell)\bar{D}_k\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{\bar{D}_k}\frac{\bar{D}_\ell}{\bar{D}_k}}{V(\tilde{D}_{it})}$$

Bacon decomposition

- Main result of Goodman-Bacon (2021) is the Bacon-Decomposition:

Figure 11: Bacon decomposition: The weights

$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}^{DD}_{kU} + s_{\ell U}\hat{\beta}^{DD}_{\ell U} + \left[s^{k}_{k\ell}\hat{\beta}^{DD,k}_{k\ell} + s^{\ell}_{k\ell}\hat{\beta}^{DD,\ell}_{k\ell}\right]$$

$$s_{kU} = \frac{(n_k + n_U)^2 n_{kU}(1 - n_{kU})\bar{D}_k(1 - \bar{D}_k)}{V(\tilde{D}_{it})}$$

$$s^{k}_{k\ell} = \frac{\left((n_k + n_\ell)(1 - \bar{D}_\ell)\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{1 - \bar{D}_\ell}\frac{1 - \bar{D}_k}{1 - \bar{D}_\ell}}{V(\tilde{D}_{it})}$$

$$s^{\ell}_{k\ell} = \frac{\left((n_k + n_\ell)\bar{D}_k\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{\bar{D}_k}\frac{\bar{D}_\ell}{\bar{D}_k}}{V(\tilde{D}_{it})}$$

If you did TWFE on this subsample, what would the variance of $\tilde{D}_{it}$ be?

EMORY UNIVERSITY

Bacon decomposition: General case

Theorem (Goodman-Bacon (2021) decomposition)

Assume that there are $k = 1, \ldots, K$ groups of treated units ordered by treatment time t_k^* and one "never-treated" group, U, which does not receive treatment in the data. The share of units in group k is n_k, and the share of periods that group k spends under treatment is $\overline{D}_k$. The regression estimate from a two-way fixed effects model is a weighted average of all two-group DiD estimators:

$$\widehat{\beta} = \sum_{k \neq U} \left(s_{k,U} \cdot \widehat{\beta}_{k,U} \right) + \sum_{k \neq U} \sum_{\ell > k} \left(s_{k,\ell} \cdot \widehat{\beta}_{k,\ell} + s_{\ell,k} \cdot \widehat{\beta}_{\ell,k} \right),$$

where the weights are given by

$$s_{k,U} = \frac{(n_k + n_U)^2 \widehat{V}_{k,U}}{\widehat{V}\left(\widetilde{D}_{i,t}\right)}, \quad s_{k,\ell} = \frac{\left((n_k + n_\ell)\left(1 - \overline{D}_\ell\right)\right)^2 \widehat{V}_{k,\ell}}{\widehat{V}\left(\widetilde{D}_{i,t}\right)}, \quad s_{\ell,k} = \frac{\left((n_k + n_\ell)\,\overline{D}_k\right)^2 \widehat{V}_{\ell,k}}{\widehat{V}\left(\widetilde{D}_{i,t}\right)},$$

such that $\sum_{k \neq U} s_{k,U} + \sum_{k \neq U} \sum_{\ell > k} \left(s_{k,\ell} + s_{\ell,k}\right) = 1.$

What does this mean to TWFE regressions?

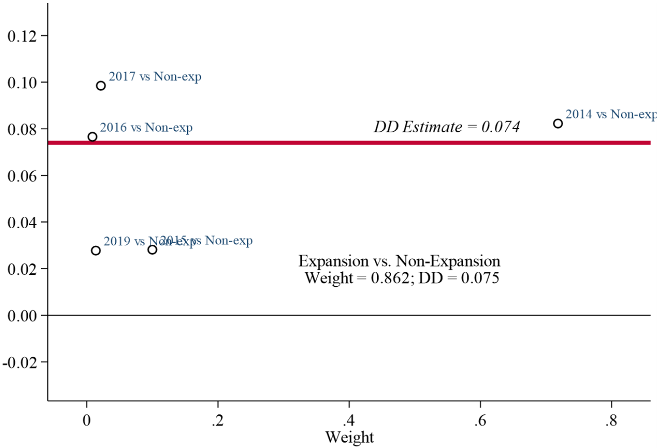
TWFE computes weighted-averages of 2x2 DiD's

- $\widehat{\beta} = 0.074$ in the empirical application.
- OLS weights use sample size and variance
- Is that what you really want?

- TWFE exploits all 2x2 DiD comparisons
  - Treated vs. "Never-treated"
  - Early-treated vs. Later-treated
  - Later-treated vs. Already-treated
- Are all these comparisons "reasonable" to attach a causal interpretation to $\widehat{\beta}$?

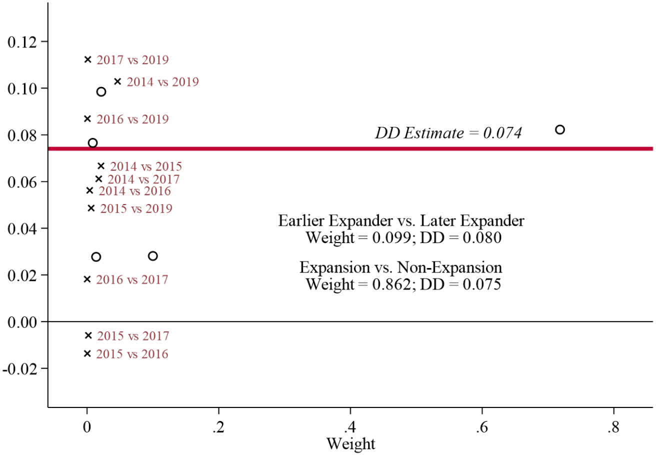
Bacon-Decomposition: Treated vs. Never-Treated

Figure 12: Bacon-Decomposition: The weights



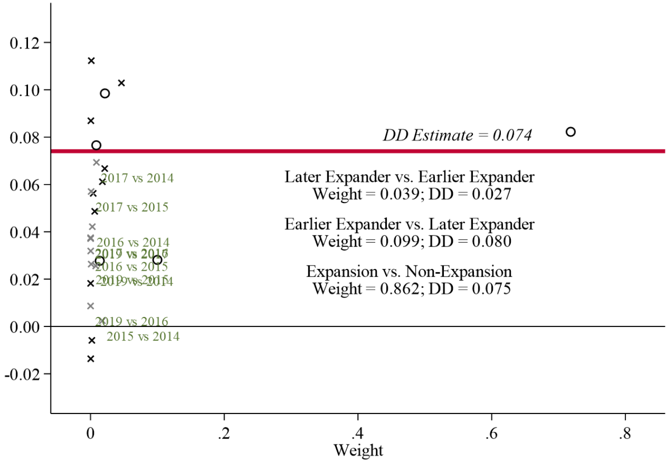
Bacon-Decomposition: Early-Treated vs. Later-treated

Figure 13: Bacon-Decomposition: The weights



Bacon-Decomposition: Later-treated vs. Early-Treated

Figure 14: Bacon-Decomposition: The weights



EMORY UNIVERSITY

TWFE regressions, in general,

do not recover an easy-to-interpret

causal parameter of interest,

unless we rule out TE heterogeneity/dynamics

How do we know this?

TWFE, Identifying Assumptions, and Causal Effects

- Goodman-Bacon (2021) decomposition is "atheoretical" in that it does not rely on causal assumptions.
- To endow the decomposition with a causal interpretation, we need to make some assumptions - PT and no-anticipation - or restrict assignment mechanisms.
- It is also worth stressing that Goodman-Bacon (2021) decomposition is not "unique".
- If you choose a different "building block" than the "time-averaged" 2x2 DiD estimates, you get a different decomposition.
- Two alternative characterizations worth mentioning are those of Athey and Imbens (2022) and de Chaisemartin and D'Haultfœuille (2020).
- Let's zoom into de Chaisemartin and D'Haultfœuille (2020), as they impose additional assumptions to get causal effects interpretation

Does TWFE “work” in setups with variation in treatment timing?

de Chaisemartin and D’Haultfœuille (2020) decomposition

de Chaisemartin and D'Haultfœuille (2020)

- de Chaisemartin and D'Haultfœuille (2020) consider a setup where treatment may turn on and off across time.
- For simplicity and easy-of-interpretation, we will focus on the staggered case (treatment is "irreversible").
- My notation will also impose a random sampling setup, which differs from what they do in their paper.
- However, it greatly simplifies the exposition.

de Chaisemartin and D'Haultfœuille (2020)

- Let us introduce the unit-specific treatment effect

$$\Delta^g_{i,t} = Y_{i,t}(g) - Y_{i,t}(\infty)$$

- Let $\epsilon_{i,t}$ be the error of the following TWFE specification:

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

- Consider the weights

$$w_{i,t} = \frac{\epsilon_{i,t}}{N_1^{-1} \sum_{i,t:D_{i,t}=1} \epsilon_{i,t}},$$

where $N_1 = \sum_{i,t} D_{i,t}$

- **Strong unconditional PTA:** Assume that for every time period t and every group g, g',

$$\mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) \middle| G = g \right] = \mathbb{E}\left[Y_t(\infty) - Y_{t-1}(\infty) \middle| G = g' \right]$$

de Chaisemartin and D'Haultfœuille (2020) - decomposition results

Theorem (de Chaisemartin and D'Haultfœuille (2020) decomposition)

Suppose SUTVA, No-anticipation, and the Strong unconditional PT hold. Let β be TWFE estimand associated with

$$Y_{i,t} = \alpha_i + \alpha_t + \beta \cdot D_{i,t} + \varepsilon_{i,t}.$$

Then, it follows that

$$\beta = \mathbb{E}\left[\sum_{i,t:D_{i,t}=1} \frac{1}{N_1} w_{i,t} \cdot \Delta^g_{i,t} \right],$$

where $\sum_{i,t:D_{i,t}=1} \frac{w_{i,t}}{N_1} = 1$, but $w_{i,t}$ can be negative.

- Weights are non-convex and can be negative
- Goodman-Bacon (2021) clarified why: we are using already-treated units as comparison groups to "later treated" units; see also Borusyak and Jaravel (2017).

Do we have negative weights in our application?

- In our application, we do not have negative weights, though.
- This is expected, as most states got treated in 2014, and we have a relatively big "never-treated" group.
- Does this mean that TWFE "worked"?
- Weights being non-negative is a **very minimal** requirement.
- The fact that we do not understand the weights attached to each ATT makes TWFE **unattractive**.

What happens when we consider a TWFE event-study specification?

Event-Study via TWFE specifications

Event-Study via TWFE specifications

- One of the main attractive features of observing multiple time periods is that we can attempt to "learn" about treatment effect dynamics.

- Status-quo in the literature is to consider variants of the TWFE event-study regression

$$Y_{i,t} = \alpha_i + \alpha_t + \gamma_k^{-K} D_{i,t}^{<-K} + \sum_{k=-K}^{-2} \gamma_k^{lead} D_{i,t}^{k} + \sum_{k=0}^{L} \gamma_k^{lags} D_{i,t}^{k} + \gamma_k^{L+} D_{i,t}^{>L} + \varepsilon_{i,t}$$

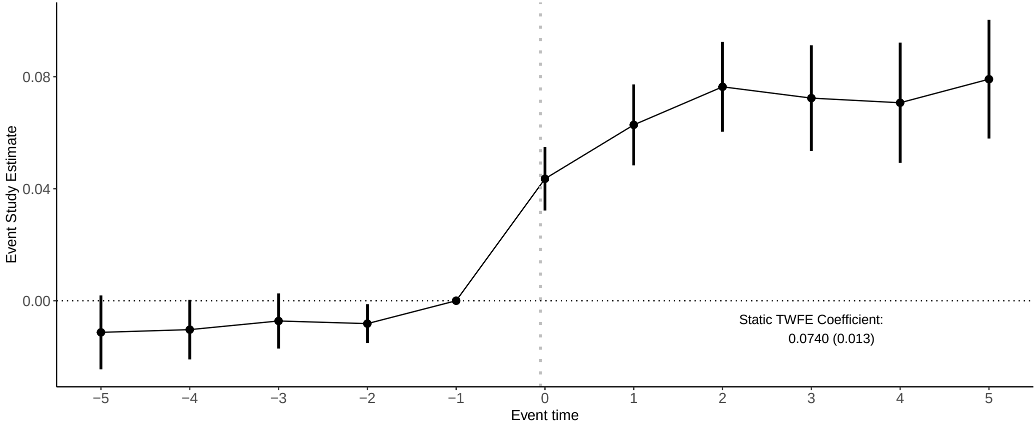
  with the event study dummies $D_{i,t}^{k} = 1\{t - G_i = k\}$, where G_i indicates the period unit i is first treated (Group).

- $D_{i,t}^{k}$ is an indicator for unit i being k periods away from initial treatment at time t.

Does this strategy “work”?

ACA Medicaid expansion: TWFE event study specification

Figure 15: Health Insurance Rate (low-income Childless Adults Aged 25-64



Event study via TWFE specifications

- Can we (a priori) “trust” these results?
- What treatment effect parameter is reported in this event-study?
- What kind of assumptions are we implicitly relying on?
- What kind of comparisons are being made “behind the scenes”?
- These are important questions!

Event-Study via TWFE specifications

Sun and Abraham (2021)

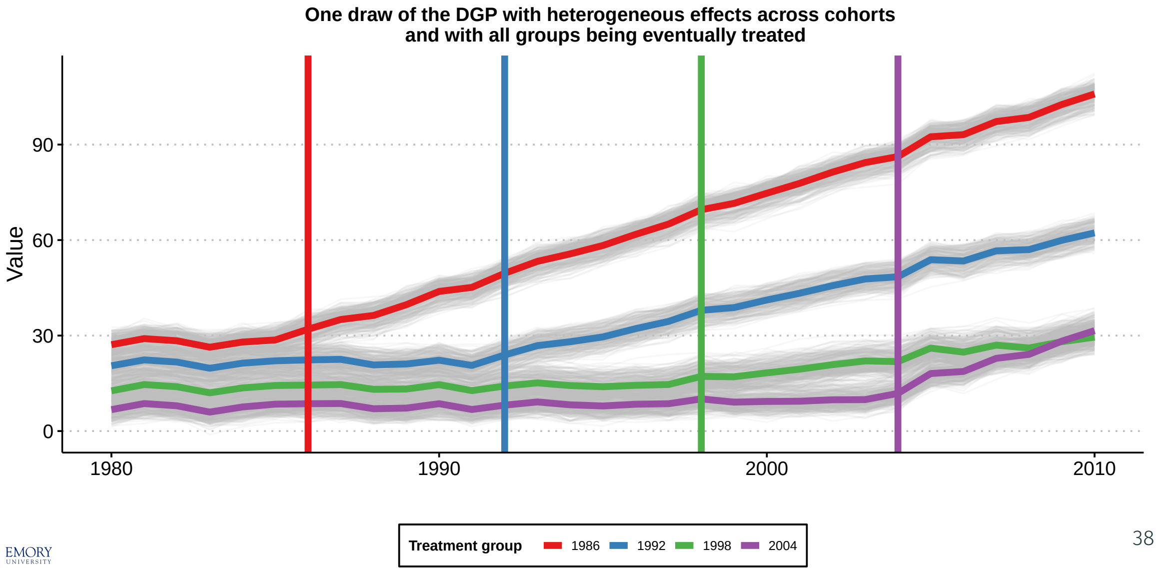
Problem with event study via TWFE specifications: Sun and Abraham (2021)

- Sun and Abraham (2021) bring "bad" news, once again!
- Even when we impose the Strong unconditional parallel trends and the no-anticipation assumption, the OLS coefficients of the TWFE ES specification are, in general, very hard to interpret.
- Coefficient on a given lead or lag can be contaminated by effects from other periods
- Pre-trends can arise solely from treatment effects heterogeneity!
- Even under treatment effect homogeneity across cohorts (they all share the same dynamics in event-time), the OLS coefficients can still be contaminated by treatment effects from the excluded periods.

Event-Study via TWFE specifications

Stylized example using simulated data

Stylized example using simulated data



Stylized example using simulated data

- 1000 units ($i = 1, 2, \ldots, 1000$) from 40 states ($state = 1, 2, \ldots, 40$).
- Data from 1980 to 2010 (31 years).
- 4 different groups based on year that treatment starts: $g = 1986, 1992, 1998, 2004$.
- Randomly assign each state to a group.
- Outcome:

$$Y_{i,t} = \underbrace{(2010 - g)}_{\text{cohort-specific intercept}} + \underbrace{\alpha_i}_{N(\frac{state}{5},1)} + \underbrace{\alpha_t}_{\frac{(t-g)}{10}+N(0,1)} + \underbrace{\tau_{i,t}}_{\mu_g \cdot (t-g+1) \cdot 1\{t \geq g\}} + \underbrace{\varepsilon_{i,t}}_{N\left(0,\left(\frac{1}{2}\right)^2\right)}$$

- $\mu_{1986} = \mu_{2004} = 3,\ \mu_{1992} = 2,\ \mu_{1998} = 1$
- ATT for group g at the first treatment period is μ_g, at the second period since treatment is $2 \cdot \mu_g$, etc.

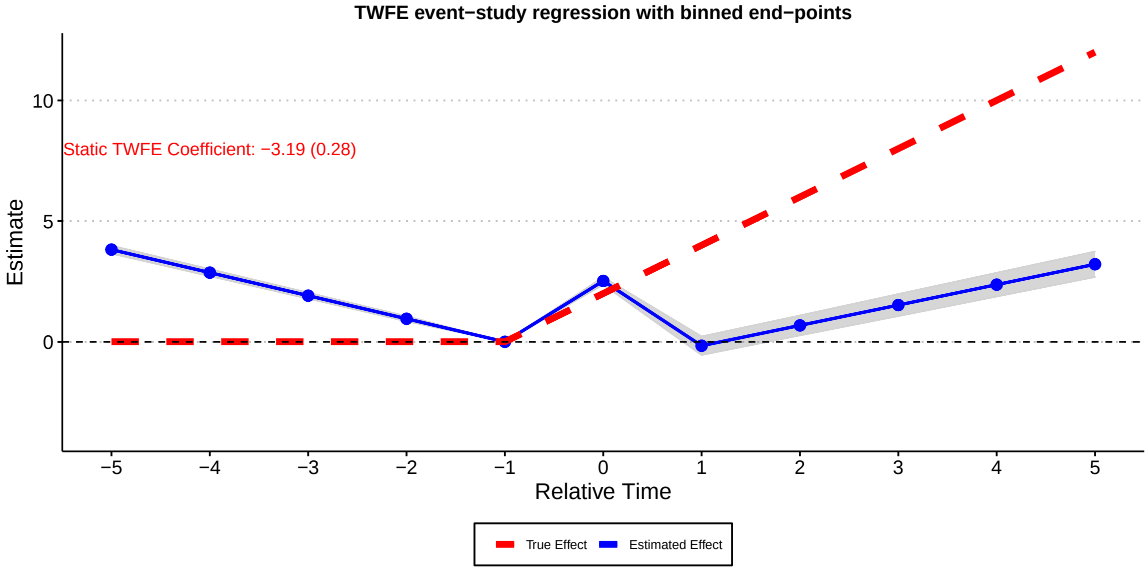
Traditional methods: TWFE event-study regression

- What if we tried to estimate the treatment effects using traditional TWFE event-study regressions,

$$Y_{i,t} = \alpha_i + \alpha_t + \gamma_k^{-K} D_{i,t}^{<-K} + \sum_{k=-K}^{-2} \gamma_k^{lead} D_{i,t}^{k} + \sum_{k=0}^{L} \gamma_k^{lags} D_{i,t}^{k} + \gamma_k^{L+} D_{i,t}^{>L} + \varepsilon_{i,t},$$

  with K and L to be equal to 5?

- Simulate data and repeat 1,000 times to compute bias and simulation standard deviations.

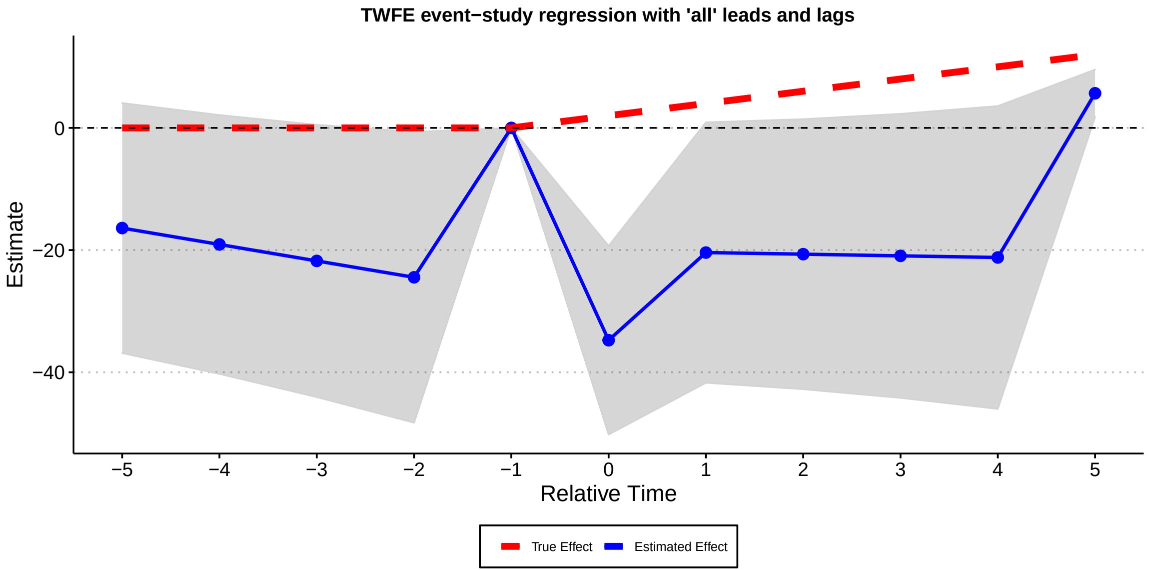


EMORY UNIVERSITY

Traditional methods: TWFE event-study regression

- What if we include all possible leads and lags in the TWFE event study specification, i.e., to set K and L to the maximum allowable in the data, making the inclusion of $D_{i,t}^{<-K}$ and of $D_{i,t}^{>L}$ unnecessary ?

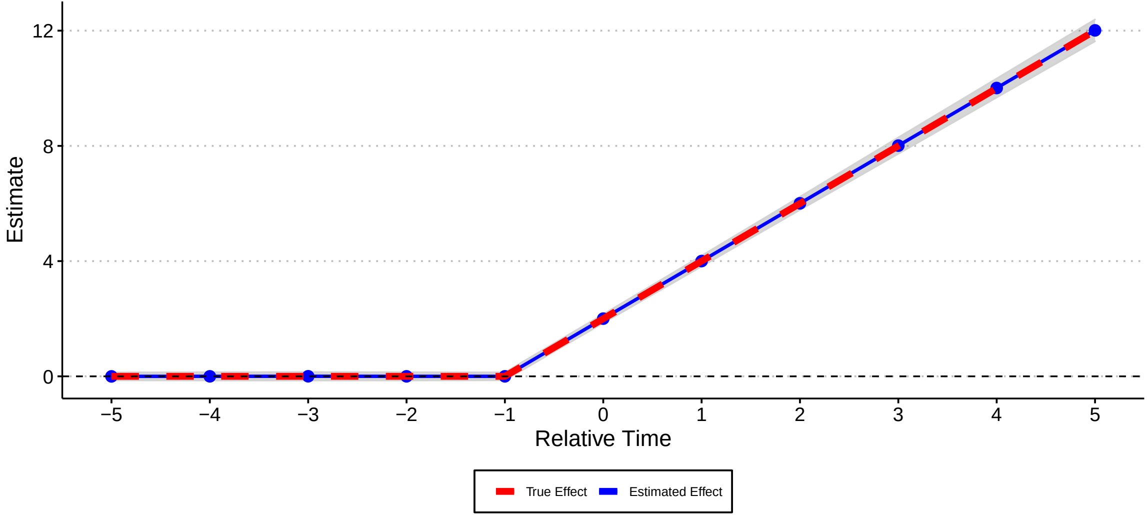
EMORY UNIVERSITY

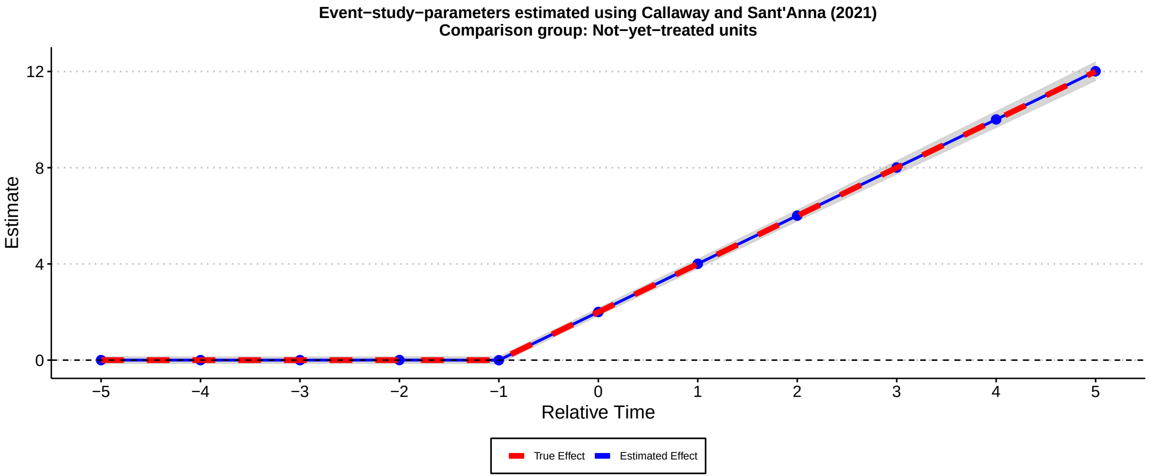


TWFE event−study regression with 'all' leads and lags
Estimate
0
−20
−40
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
True Effect
Estimated Effect

Is there hope?

Event−study−parameters estimated using Callaway and Sant'Anna (2021)
Comparison group: Last−treated−Cohort units




Event−study−parameters estimated using Callaway and Sant'Anna (2021)
Comparison group: Not−yet−treated units
Estimate
12
8
4
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
True Effect
Estimated Effect

References

Athey, Susan and Guido Imbens, "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 2022, *226* (1), 62–79.

Borusyak, Kirill and Xavier Jaravel, "Revisiting Event Study Designs," SSRN Scholarly Paper ID 2826228, Social Science Research Network, Rochester, NY August 2017.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

Goodman-Bacon, Andrew, "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, 2021, *225* (2).

Roth, Jonathan, "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends," *American Economic Review: Insights*, 2022, *4* (3), 305–322.

Sun, Liyan and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021, *225* (2).