Topics in Applied Econometrics for Public Policy

TA Session 7

Sergi Quintana

Barcelona School of Economics

June 15, 2023

Difference-in-differences

We are interested in estimating the causal effect of a treatment on an outcome of interest.

In a very simple way we could do that with:

- **Before vs. after comparisons**. The problem here is that it might not account for trends in the outcome of interest.
- **Treated vs. untreated comparisons**. The problem here is that it rules out selection on unobservables.

Difference-in-differences methods exploit variation in time (before vs. after) and across groups (treated vs. untreated) to recover causal effects of interest.

Difference-in-differences

The parameter of interest is always the average treatment effect for a particular group.

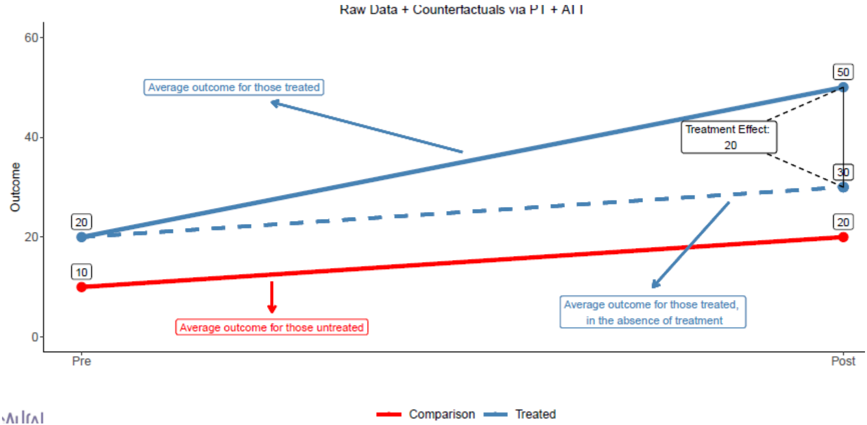
- **ATT**
  The Average Treatment Effect on the Treated at time period $t = 2$ is
  $$ATT = \mathbb{E}\left[Y_{i,t=2}(2) - Y_{i,t=2}(\infty) | G_i = 2\right]$$

- **ATU**
  The Average Treatment Effect on the Untreated at time period $t = 2$ is
  $$ATU = \mathbb{E}\left[Y_{i,t=2}(2) - Y_{i,t=2}(\infty) | G_i = \infty\right]$$

- **ATE**
  The (overall) Average Treatment Effect at time period $t = 2$ is
  $$ATE = \mathbb{E}\left[Y_{i,t=2}(2) - Y_{i,t=2}(\infty)\right]$$

Difference-in-differences

Graphically:



Difference-in-differences

Mathematically:

- **The canonical 2×2 DiD estimator** is given by

$$\widehat{\theta}_n^{DiD} = \left(\overline{Y}_{g=treated,t=post} - \overline{Y}_{g=treated,t=pre}\right) - \left(\overline{Y}_{g=untreated,t=post} - \overline{Y}_{g=untreated,t=pre}\right),$$

  where $\overline{Y}_{g=d,t=j}$ is the sample mean of the outcome Y for units in group d in time period j,

$$\overline{Y}_{g=d,t=j} = \frac{1}{N_{g=d,t=j}} \sum_{i=1}^{N_{all}} Y_i 1\{G_i = d\} 1\{T_i = j\},$$

  with

$$N_{g=d,t=j} = \sum_{i=1}^{N_{all}} 1\{G_i = d\} 1\{T_i = j\},$$

  G_i and T_i are group and time dummy, respectively, and Y_i is the "pooled" outcome data.

Difference-in-differences

But we will need some assumptions to identify the effect.

SUTVA. Stable Unit Treatment Value Assumption.

Assumption (SUTVA)

Observed outcomes at time t are realized as

$$Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g).$$

- Implicitly implies that potential outcomes for unit i are not affected by the treatment of unit j.
  - Rules out interference across units
  - Rules out spillover effects
  - Rules out general equilibrium effects

Difference-in-differences

Assumption (No-Anticipation)

For all units i, $Y_{i,t}(g) = Y_{i,t}(\infty)$ for all groups in their pre-treatment periods, i.e., for all $t < g$.

- Common assumption in duration analysis (Abbring and van den Berg, 2003; Sianesi, 2004).
- This assumption says that unit-specific treatment effects are zero in all **pre-treatment periods**.
- It does not restrict treatment effect heterogeneity in **post-treatment periods**.
- This is plausible in many setups, especially if treatment is not announced in advance.
- But it is not innocuous (Malani and Reif, 2015).

Difference-in-differences

We need more assumptions!

Problem:
Comparison of outcomes at $t = 2$ between the treated and the untreated units do not usually give the right answer.

$$
\begin{aligned}
\mathbb{E}\left[Y_{i,t=2}|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}|G_i = \infty\right] &= \mathbb{E}\left[Y_{i,t=2}(2)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right] \\
&= \mathbb{E}\left[Y_{i,t=2}(2) - Y_{i,t=2}(\infty)|G_i = 2\right] \\
&\quad + \left(\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right]\right) \\
&= \text{ATT} + \text{Selection Bias}
\end{aligned}
$$

- Selection Bias term unlikely to be zero in most applications.
- Selection into treatment is often associated with the potential outcomes.

Difference-in-differences

Since a simple comparison of means at time $t = 2$ does not recover a parameter of interest (ATT), we can take a different route.

Assumption (Parallel Trends Assumption)

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = 2\right] = \mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = \infty\right]$$

The parallel trends (PT) assumption states that, in the absence of treatment, the evolution of the outcome among the treated units is, on average, the same as the evolution among the untreated units.

Difference-in-differences

- We will start from the perspective that the *ATT* at time $t = 2$ is the target parameter.
- From the definition of the ATT and SUTVA, we have

$$
\begin{aligned}
ATT &\equiv \mathbb{E}\left[Y_{i,t=2}(2)\,|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}(\infty)\,|G_i = 2\right] \\
&= \underbrace{\mathbb{E}\left[Y_{i,t=2}|G_i = 2\right]}_{\text{by SUTVA}} - \mathbb{E}\left[Y_{i,t=2}(\infty)\,|G_i = 2\right]
\end{aligned}
$$

- Green object is estimable from data (under SUTVA).
- Red object still depends on potential outcomes, and our goal is to find ways to "impute" it.
- This is where PT comes into play!

Difference-in-differences

1) First, recall the PT assumption:

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = 2\right] = \mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = \infty\right].$$

2) By simple manipulation, we can write it as

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] = \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = 2\right] + \left(\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = \infty\right]\right)$$

3) Now, exploiting No-Anticipation and SUTVA:

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] = \underbrace{\mathbb{E}\left[Y_{i,t=1}(2)|G_i = 2\right]}_{by\ No-Anticipation} + \left(\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = \infty\right]\right)$$

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] = \underbrace{\mathbb{E}\left[Y_{i,t=1}|G_i = 2\right] + \left(\mathbb{E}\left[Y_{i,t=2}|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}|G_i = \infty\right]\right)}_{by\ SUTVA}$$

Difference-in-differences

- Combining these results together, we have that, under SUTVA + No-Anticipation + PT assumptions, it follows that

$$
\begin{aligned}
\textbf{ATT} &= \mathbb{E}\left[Y_{i,t=2}|G_i = 2\right] - \left(\mathbb{E}\left[Y_{i,t=1}|G_i = 2\right] + \left(\mathbb{E}\left[Y_{i,t=2}|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}|G_i = \infty\right]\right)\right) \\
&= \left(\mathbb{E}\left[Y_{i,t=2}|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=1}|G_i = 2\right]\right) - \left(\mathbb{E}\left[Y_{i,t=2}|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}|G_i = \infty\right]\right)
\end{aligned}
$$

- This is “the birth” of the DiD estimand!

Difference-in-differences

- The TWFE specification is given by

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

  where $\mathbb{E}[\varepsilon_{i,t}|G,T] = 0$ *almost surely*.

- Now, let's play with its terms:

$$\begin{aligned}
\mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 1] &= \alpha_0 \\
\mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 2] &= \alpha_0 + \lambda_0 \\
\mathbb{E}[Y_{i,t}|G_i = 2, T_i = 1] &= \alpha_0 + \gamma_0 \\
\mathbb{E}[Y_{i,t}|G_i = 2, T_i = 2] &= \alpha_0 + \gamma_0 + \lambda_0 + \beta_0^{twfe}
\end{aligned}$$

Difference-in-differences

- Set of moment restrictions:

$$\begin{aligned}\mathbb{E}[Y_{i,t}|G_i=\infty,T_i=1] &= \alpha_0\\ \mathbb{E}[Y_{i,t}|G_i=\infty,T_i=2] &= \alpha_0+\lambda_0\\ \mathbb{E}[Y_{i,t}|G_i=2,T_i=1] &= \alpha_0+\gamma_0\\ \mathbb{E}[Y_{i,t}|G_i=2,T_i=2] &= \alpha_0+\gamma_0+\lambda_0+\beta_0^{twfe}\end{aligned}$$

- These imply that

$$\mathbb{E}[Y_{i,t}|G_i=2,T_i=2]-\mathbb{E}[Y_{i,t}|G_i=2,T_i=1]=\lambda_0+\beta_0^{twfe}$$

  and that

$$\mathbb{E}[Y_{i,t}|G_i=\infty,T_i=2]-\mathbb{E}[Y_{i,t}|G_i=\infty,T_i=1]=\lambda_0.$$

- Thus, we can clearly see that β_0^{twfe} is equal to the DiD estimand.

Multiple periods

Until now we have seen the most basic set-up. What if we have multiple periods?

- But now, we have multiple post-treatment periods so we will talk about time (and group) specific ATTs:

$$ATT(g,t) \equiv \mathbb{E}\left[Y_t(g) - Y_t(\infty) | G = g\right] = \mathbb{E}\left[Y_t(g) | G = g\right] - \mathbb{E}\left[Y_t(\infty) | G = g\right]$$

Average Treatment Effect among units treated at time g, at time t.

- Sometimes, we may re-express the ATT(g,t) in "event-time" e:

$$ATT(g,g+e) \equiv \mathbb{E}\left[Y_{g+e}(g) - Y_{g+e}(\infty) | G = g\right] = \mathbb{E}\left[Y_{g+e}(g) | G = g\right] - \mathbb{E}\left[Y_{g+e}(\infty) | G = g\right]$$

Average Treatment Effect among units treated at time g, e periods after ($e \geq 0$) / before ($e < 0$) treatment started.

We are now interested in multiple treatment effects.

Multiple Periods

- What if we do not subset the data and use the following TWFE specification?

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

where $D_{i,t}$ is a treatment dummy if unit i is treated by time t.

- What type of summary parameter does β^{twfe} represent when we no-anticipation and PT for all time periods hold?
- Under these stronger assumptions, we can show that

$$\beta^{twfe} = \frac{\sum_{s=g}^{T} ATT(g,s)}{T-g+1} = \frac{\sum_{e=0}^{T-g} ATT(g,g+e)}{T-g+1}.$$

Variation in treatment timing

Now we are under a framework like the one on the problem set, in which we observe multiple time periods and variation in treatment timing.

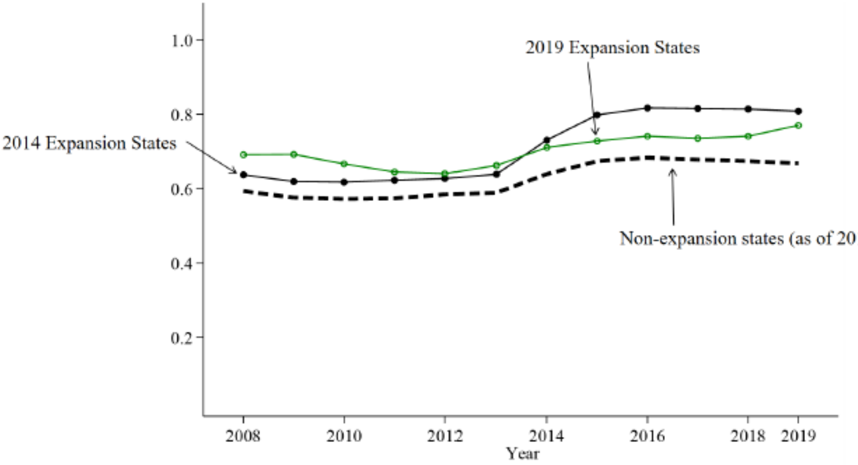
What is β_{twfe} capturing?

To adress such a question, we will decompose it's effect using:

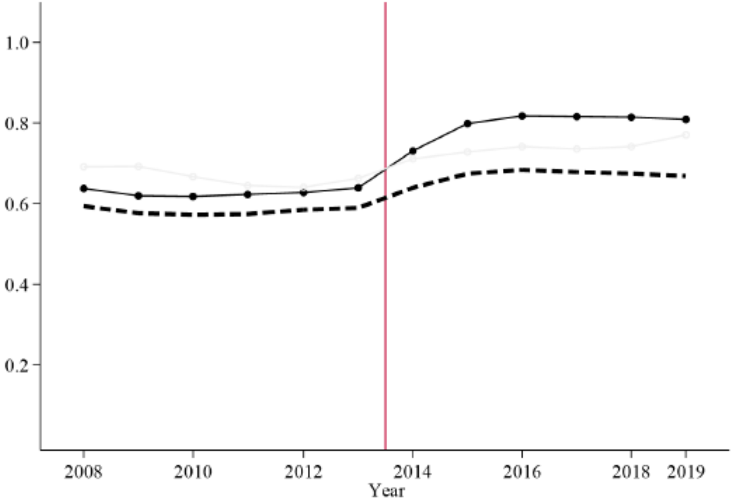
- Bacon-Decomposition.
- de Chaisemartin and D'Haultfoeuille (2021) decomposition.

Bacon-Decomposition

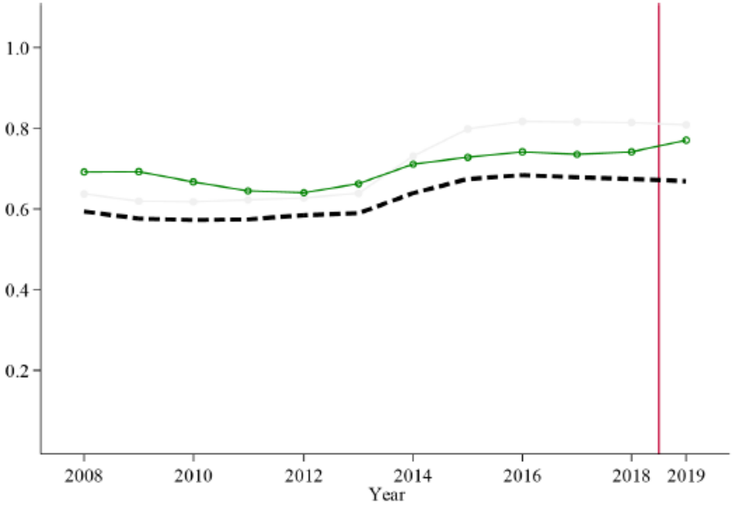
Example of U.S. health insurance expansion.



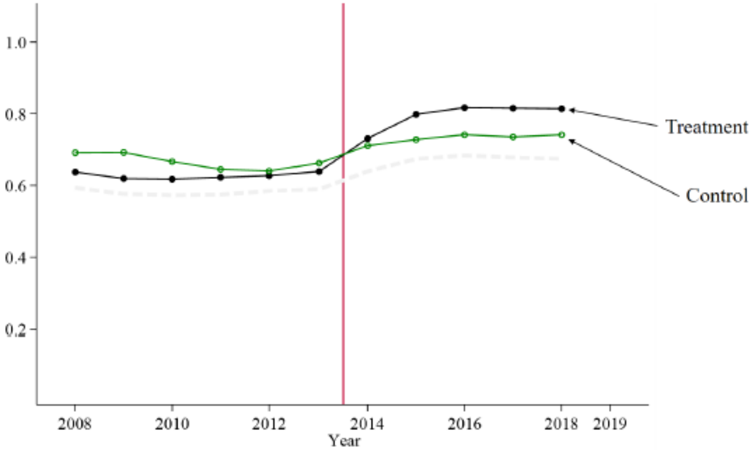
Bacon-Decomposition



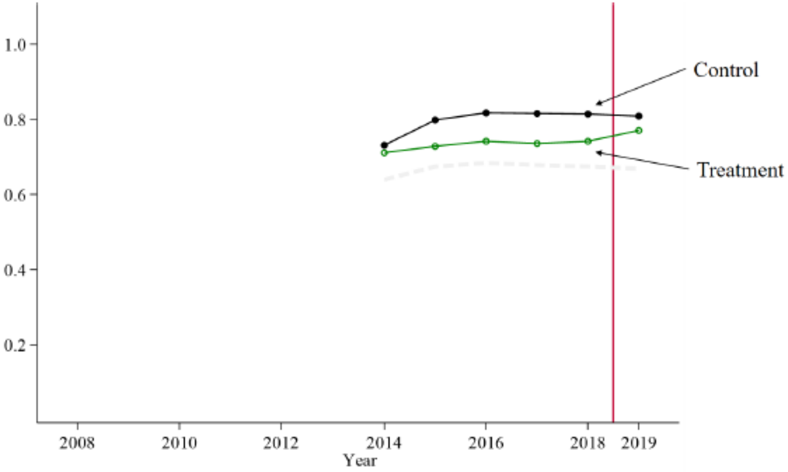
Bacon-Decomposition



Bacon-Decomposition



Bacon-Decomposition



Bacon-Decomposition

The paper shows that the TWFE is a weighted average of all 2x2 DiD comparisons:

- Treated vs Never-Treated.
- Early-Treated vs. Later-Treated.
- Later-Treated vs. Already-Treated. This is known as the forbiden comparison.

Bacon-Decomposition

- 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

Important, the weights depend on:

- Sample size of each group.
- Variance.

De Chaisemartin and D'Haultfoeuille (2021)

de Chaisemartin and D'Haultfoeuille (2021) decompose the TWFE as a combination of individual treatment effects.

- 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'

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

De Chaisemartin and D'Haultfoeuille (2021)

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).

Conclusion

The TWFE coefficient reports difficult to interpret effects. We will now revise some of the methods to overcome this problem:

- Callaway and Sant'Anna 2021
- Sun and Abraham 2021
- Borusyak et al. 2021

Callaway and Sant'Anna 2021

We can compute many 2x2 DiD

- In staggered setups, a parameter that is interesting and has clear economic interpretation is the $ATT(g,t)$

$$ATT(g,t) = \mathbb{E}\left[Y_t(g) - Y_t(\infty) \,|G_g = 1\right], \text{ for } t \geq g.$$

- Average Treatment Effect at time t of starting treatment at time g, among the units that indeed started treatment at time g.

And then aggregate in our preferred way:

- Average effect of participating in the treatment that units in group g experienced:

$$\theta_S(g) = \frac{1}{T-g+1}\sum_{t=2}^{T} 1\{g \leq t\} ATT(g,t)$$

Callaway and Sant'Anna 2021

- Average effect of participating in the treatment in time period t for groups that have participated in the treatment by time period t

$$\theta_C(t) = \sum_{g=2}^{T} 1\{g \leq t\} ATT(g,t) P(G = g | G \leq t, G \neq \infty)$$

- Average effect of participating in the treatment for the group of units that have been exposed to the treatment for exactly e time periods

$$\theta_D(e) = \sum_{g=2}^{T} 1\{g + e \leq T\} ATT(g, g+e) P(G = g | G + e \leq T, G \neq \infty)$$

Sun and Abraham 2021

It can be understood as a specific case of Callaway and Sant'Anna 2021 for event-study studies.

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

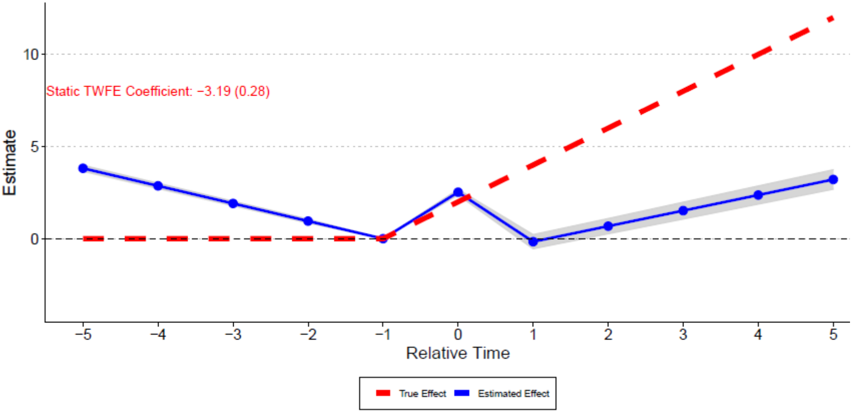
$$Y_{i,t} = \alpha_i + \alpha_t + \gamma_k^{-K} D_{i,t}^{<-K} + \sum_{k=-K}^{-2} \gamma_k^{lead} D_{i,t}^{k} + \sum_{k=0}^{L} \gamma_k^{lags} D_{i,t}^{k} + \gamma_k^{L+} D_{i,t}^{>L} + \varepsilon_{i,t}$$

  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).

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.

Sun and Abraham 2021



Sun and Abraham 2021

$$Y_{i,t} = \alpha_i + \alpha_t + \sum_{l=-K}^{-2} \gamma_l D_{i,t}^l + \sum_{l=0}^{L} \gamma_l D_{i,t}^l + \epsilon_{i,t}$$

The coefficient for each relative period can be decomposed as:

$$E[\gamma_l] = E\left[\sum_g w_{g,l} TE_g(l) + \sum_{l' \neq l} \sum_g w_{g,l'} TE_g(l') + \sum_g w_{g,l_e} TE(l_e)\right]$$

where:

- $w_{g,l}$ are the weights associated to each group g at relative time l. They sum to 1.
- $w_{g,l'}$ are the weights associated to each gropu g at relative time l'. They sum up to 0.
- $w_{g,l'}$ are the weights associated to each group$g$ at the excluded time l_e. They sum up to -1.

Sun and Abraham 2021

Therefore the event-study is clearly contaminated. To address this problem Sun and Abraham propose the Interaction Weighted Estimator.

- It uses never treated or last-treated as control. Unlike Callway and Sant'Anna which also uses the not yet treated.
- **Step 1.** Estimate an event study regression with the interaction of relative time and group/cohort. This will consistently estimate an event study for each cohort.
- **Step 2.** Compute the share of each group at each relative time.
- **Step 3.** Compute the weighted average using the estimated shares.

Borusyak et al. 2021

The authors propose two numerically equivalent methods to estimate the effects.

Method 1.

- Estimate a TWFE regression of the outcome on group and time fixed effects, and fixed effects for every treated (g, t) cell.
- The estimtated coefficients are the most efficient by Gauss-Markov.
- Then aggregate the cells to recover the parameter of interest.

Borusyak et al. 2021

Method 2.

- Fit a regression of $Y_{i,t}$ on individual and time fixed effects on the untreated sample.
- Input a counterfactual output for each individual.
- Compute the average using all the individual effects to get the parameter of interest.

Similarities and Differences

It is important to understand that all these methods where developed contemporaneously.

- Under parallel trends, Borusyak et al. 2021 is the most efficient if Gauss-Markov assumptions are satisfied.
- Another difference between these approaches is that Borusyak et al. (2021) impose parallel trends for every group and between every pair of consecutive time periods. Callaway and Sant'Anna (2021), on the other hand, impose a weaker parallel trends assumption: from period c onwards, cohort c must be on the same trend as the never-treated groups, but before that cohort c may have been on a different trend