Causal Inference using Difference-in-Differences

Lecture 12: Reliable Estimators with Staggered Treatment Adoption

Pedro H. C. Sant'Anna
Emory University

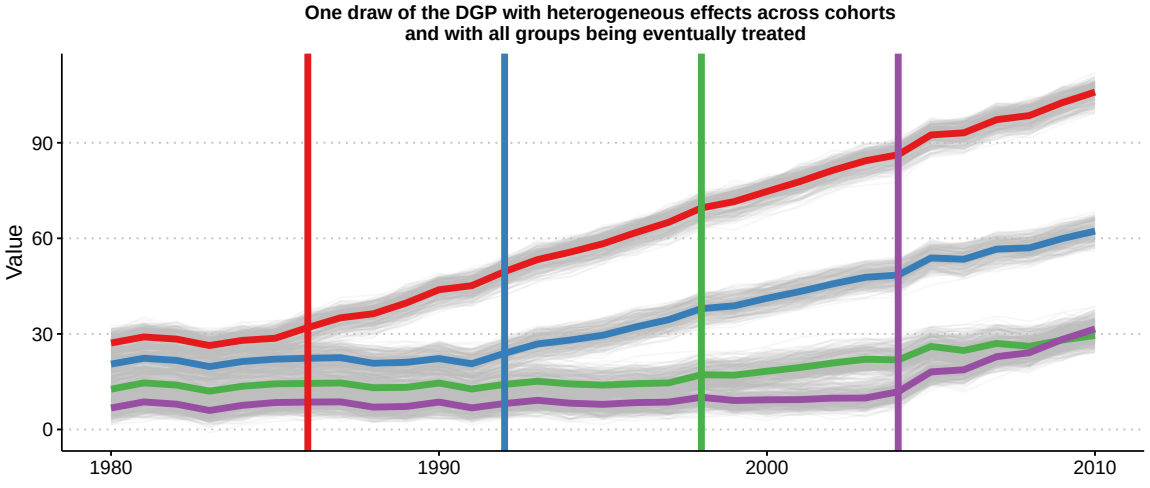
January 2025

Summary of previous lecture

Summary of previous lecture

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 the year treatment starts: $g = 1986, 1992, 1998, 2004$.
- Randomly assigns each state to a group.
- Outcome:

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

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

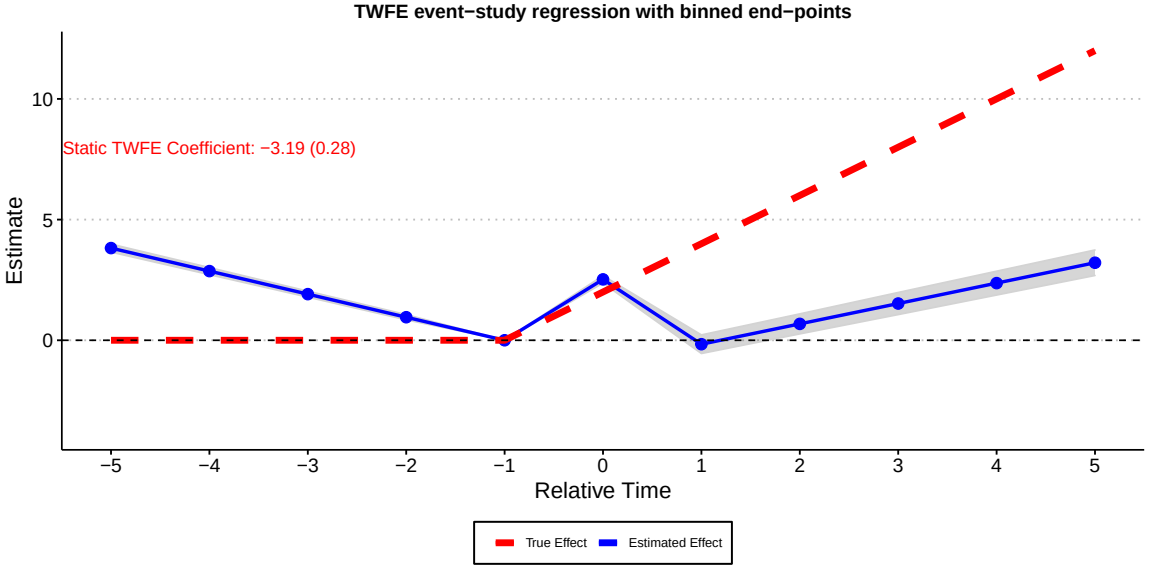
Traditional methods: TWFE event-study regression

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

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

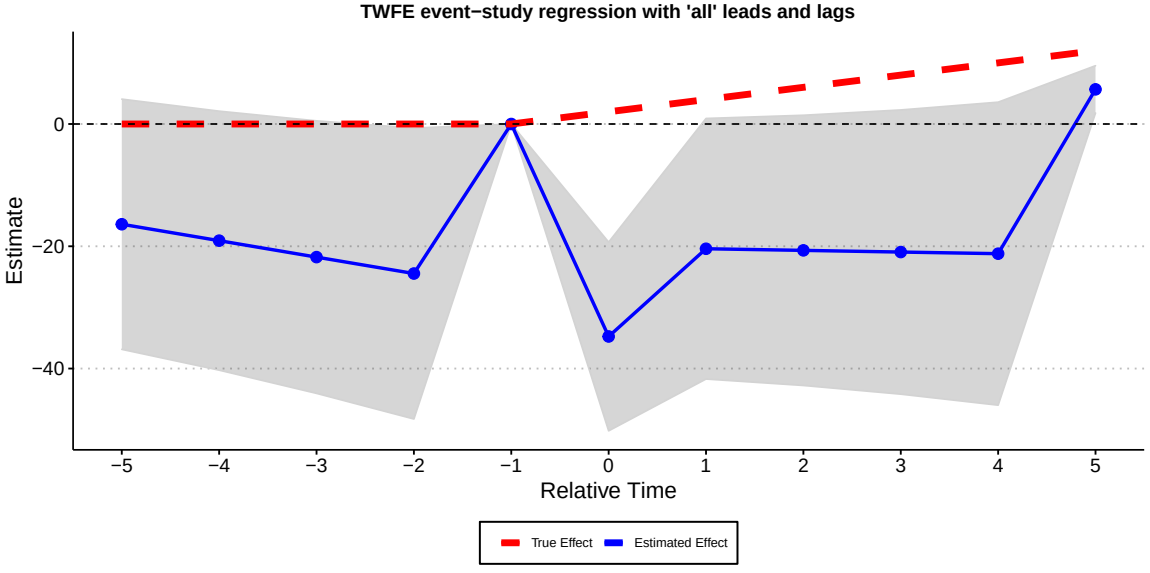
  with K and L to be equal to 5?

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

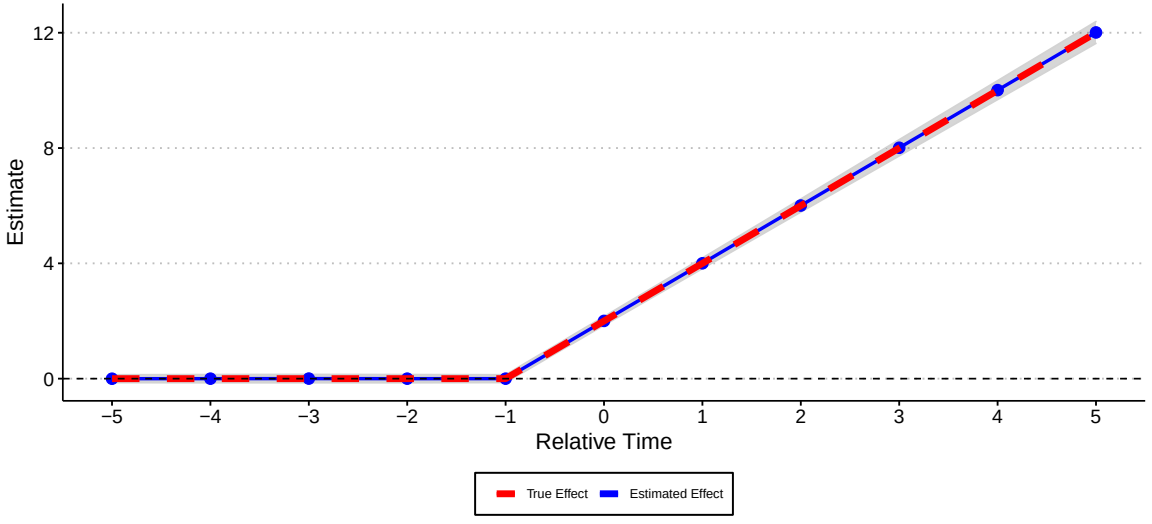

TWFE event−study regression with binned end−points
Static TWFE Coefficient: −3.19 (0.28)
Estimate
10
5
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
True Effect
Estimated Effect

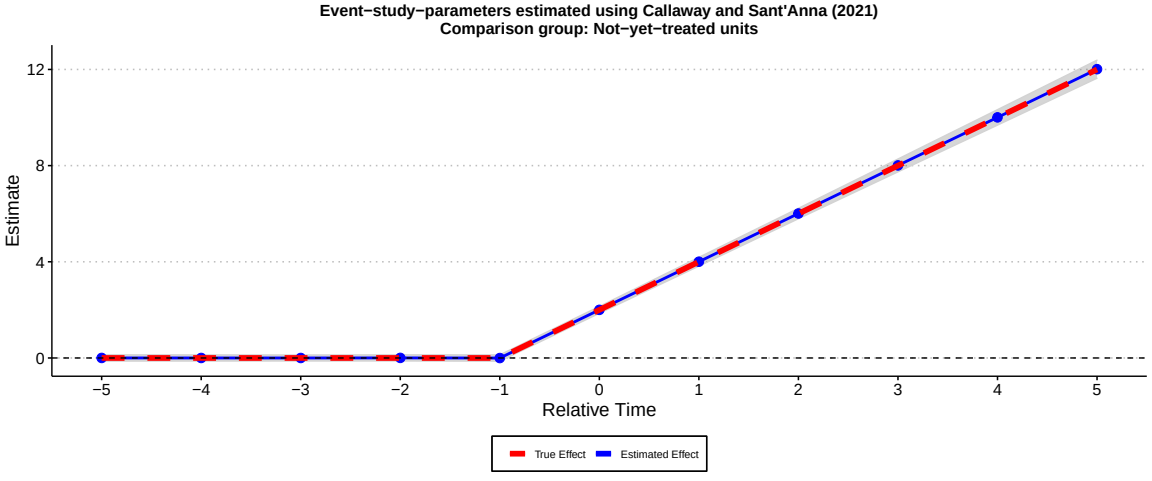
Traditional methods: TWFE event-study regression

- 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 ?


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
Estimate
12
8
4
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
True Effect
Estimated Effect


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

Recent Boom of New DiD Methods: Solutions to the TWFE problems

- The problems associated with using standard TWFE specifications are evident.
- OLS is variational hungry but causal inference is variational cautious!
- How to solve the TWFE problem in DiD setups?
- Ensure that you only make the comparisons you want to
- **Callaway and Sant'Anna (2021)** propose a guided and transparent way to do this!
  - Allow for covariates, different comparison groups, panel and repeated cross-sections.
  - **Separate the analysis into identification, aggregation, and estimation/inference.**

Addressing the TWFE problems

Recent Boom of New DiD Methods: Solutions to the TWFE problems

- Callaway and Sant'Anna (2021) is not the only game in town:
  - **Sun and Abraham (2021)**: Their proposed estimator coincides with CS when there are no covariates and uses the never-treated/last-treated cohort as a comparison group. However, this paper has many other results about the pitfalls of TWFE that are not in CS. No treatment of covariates. A great complement to CS.
  - **Gardner (2021)**, **Borusyak, Jaravel and Spiess (2024)** and **Wooldridge (2021)**: Propose "imputation"/regression based methods to recover cohort-time ATT's . These three papers do not nest nor are nested by CS, but identification assumptions are sometimes stronger. Benefit: Sometimes (but not always), they can get you more precise estimates when these (additional) assumptions are correct. Unlike CS, the handling of covariates in Gardner (2021) and Borusyak et al. (2024) is restrictive (i.e., rule out heterogeneous treatment effects); Wooldridge (2021) is more flexible than G and BJS in this regard.
  - **Wooldridge (2023)**: Propose estimators that are suitable for nonlinear models. It relies on alternative types of parallel trend assumptions, e.g., 'ratio-in-ratios" if exponential model. If use canonical link functions, standard errors can be easily estimated.

Recent Boom of New DiD Methods: Solutions to the TWFE problems (cont.)

- Callaway and Sant'Anna (2021) is not the only game in town:
  - **de Chaisemartin and D'Haultfœuille (2020, 2024)**: Their proposed estimator coincides with CS when there are no covariates, uses not-yet-treated units as the comparison group, and treatment is staggered. However, these two papers allow treatment to turn on and off, which is not allowed in CS. However, to allow treatment to turn on and off and still get an interesting/easy-to-interpret parameter of interest, they impose restrictions on treatment effect dynamics. In fact, de Chaisemartin and D'Haultfœuille (2020) completely rules out dynamic treatment effects.

    These papers only briefly talk about covariates. When they are available, their proposal do not nest nor are nested by CS. However, the way covariates are allowed is arguably restrictive and rules out interesting treatment effect heterogeneity (e.g., they do not allow the ATT to vary according to age).

Main take-way message from all DiD proposed solutions

Clearly separate identification, aggregation, and estimation/inference steps!

In what follows, I will focus on Callaway and Sant'Anna's (2021) approach.

In the future, I plan to add more discussions about the other methods.

Digging into Callaway and Sant'Anna (2021)

Callaway and Sant'Anna (2021) in practice

Can be implemented via the R package did.

Can be implemented via the Stata packages csdid, csdid2, hdidregress and xthdidregress

Can be implemented via the Pythgon package csdid

Main take-way message

Clearly separate identification, aggregation, and estimation/inference steps!

Identification

Let’s talk about identification

Identification

Building block of the analysis

- If the sample size was not a limitation (we have all the data in the world), what kind of question would we like to answer?
- 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.

Identifying Assumptions: No-Anticipation

- Given that we never observe $Y(\infty)$ in post-treatment periods among units that have been treated, we need to make assumptions to identify $ATT(g,t)$'s
- **No-Anticipation Assumption**: For all i,t and $t < g, g'$, $Y_{i,t}(g) = Y_{i,t}(g')$.
- Unit treatment effects are zero before treatment takes place.
- Exactly the same content as in the 2x2 case.

Parallel trend assumption based on a “never treated” group

Assumption (Parallel Trends based on a “never-treated”)

For each $t \in \{2, \ldots, T\}$, $g \in \mathcal{G}$ such that $t \geq g$,

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G_g = 1] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G = \infty]$$

Parallel Trends based on not-yet treated groups

Assumption (Parallel Trends based on "Not-Yet-Treated" Groups)

For each $(s,t) \in \{2,\dots,T\} \times \{2,\dots,T\}$, $g \in \mathcal{G}$ *such that* $t \geq g$, $s \geq t$

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | G_g = 1] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | D_s = 0, G_g = 0].$$

ATT(g,t) Estimand: “never-treated” as comparison group

- Under no-anticipation and PT based on “never-treated”, we have

$$ATT^{nev}_{unc}(g,t) = \mathbb{E}[Y_t - Y_{g-1}|G_g = 1] - \mathbb{E}[Y_t - Y_{g-1}|G = \infty].$$

- This looks very similar to the two periods, two-groups DiD result without covariates.
- The difference is now we take a “long difference”.
- Same intuition carries, though!
- This result appears in Callaway and Sant’Anna (2021) and Sun and Abraham (2021).

ATT(g,t) Estimand: not-yet treated as comparison group

- If one wants to use the units that have not yet been exposed to treatment by time t, we have a different estimand:

$$ATT^{ny}_{unc}(g,t) = \mathbb{E}[Y_t - Y_{g-1}|G_g = 1] - \mathbb{E}[Y_t - Y_{g-1}|D_t = 0, G_g = 0].$$

- This looks similar to the two periods, two-groups DiD result without covariates, too.
- The difference is that we now take a "long difference" and the comparison group changes over time.
- Same intuition carries, though!
- This result appears in Callaway and Sant'Anna (2021) and de Chaisemartin and D'Haultfœuille (2020), though de Chaisemartin and D'Haultfœuille (2020) focus exclusively on instantaneous treatment effects, i.e., the case with $g = t$.

What if we want to allow for covariate-specific trends?

Parallel trend assumption based on a “never treated” group

Assumption (Conditional Parallel Trends based on a “never-treated”)

For each $t \in \{2, \ldots, T\}$, $g \in \mathcal{G}$ *such that* $t \geq g$,

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | X, G_g = 1] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty) | X, G = \infty] \ a.s..$$

Parallel Trends based on not-yet treated groups

Assumption (Conditional Parallel Trends based on "Not-Yet-Treated" Groups)

For each $(s,t) \in \{2,\dots,T\} \times \{2,\dots,T\}$, $g \in \mathcal{G}$ *such that* $t \geq g$, $s \geq t$

$$\mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty)|X, G_g = 1] = \mathbb{E}[Y_t(\infty) - Y_{t-1}(\infty)|X, D_s = 0, G_g = 0] \text{ a.s..}$$

Identification results - never treated as comparison group

- Under these assumptions, Callaway and Sant'Anna (2021) proved that, for all g and t such that $g \in \mathcal{G} \equiv \mathcal{G} \cap \{2, 3, \ldots, T\}$, $t \in \{2, \ldots, T\}$ and $t \geq g$, $ATT(g,t)$ is nonparametrically identified by the DR estimand

$$ATT_{dr}^{nev}(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)\left(Y_t - Y_{g-1} - m_{g,t}^{nev}(X)\right)\right].$$

where $m_{g,t}^{nev}(X) = \mathbb{E}[Y_t - Y_{g-1} | X, G = \infty]$.

- Extends Heckman, Ichimura and Todd (1997), Abadie (2005) and Sant'Anna and Zhao (2020).

Identification results - never treated as comparison group

- Under these assumptions, Callaway and Sant'Anna (2021) proved that, for all g and t such that $g \in \mathcal{G} \equiv \mathcal{G} \cap \{2, 3, \ldots, T\}$, $t \in \{2, \ldots, T\}$ and $t \geq g$, $ATT(g,t)$ is nonparametrically identified by the DR estimand

$$ATT_{dr}^{nev}(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)\left(Y_t - Y_{g-1} - m_{g,t}^{nev}(X)\right)\right].$$

where $m_{g,t}^{nev}(X) = \mathbb{E}[Y_t - Y_{g-1} | X, G = \infty]$.

- Extends Heckman et al. (1997), Abadie (2005) and Sant'Anna and Zhao (2020).

Identification results - never treated as comparison group

- Under these assumptions,Callaway and Sant'Anna (2021) proved that, for all g and t such that $g \in \mathcal{G} \equiv \mathcal{G} \cap \{2, 3, \ldots, T\}$, $t \in \{2, \ldots, T\}$ and $t \geq g$, $ATT(g,t)$ is nonparametrically identified by the DR estimand

$$ATT_{dr}^{nev}(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)\left(Y_t - Y_{g-1} - m_{g,t}^{nev}(X)\right)\right].$$

where $m_{g,t}^{nev}(X) = \mathbb{E}[Y_t - Y_{g-1} | X, G = \infty]$.

- Extends Heckman et al. (1997), Abadie (2005) and Sant'Anna and Zhao (2020).

Identification results - never treated as comparison group

- Under these assumptions, Callaway and Sant'Anna (2021) proved that, for all g and t such that $g \in \mathcal{G} \equiv \mathcal{G} \cap \{2, 3, \ldots, T\}$, $t \in \{2, \ldots, T\}$ and $t \geq g$, $ATT(g,t)$ is nonparametrically identified by the DR estimand

$$ATT_{dr}^{nev}(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)\left(Y_t - Y_{g-1} - m_{g,t}^{nev}(X)\right)\right].$$

where $m_{g,t}^{nev}(X) = \mathbb{E}[Y_t - Y_{g-1} | X, G = \infty]$.

- Extends Heckman et al. (1997), Abadie (2005) and Sant'Anna and Zhao (2020).

Identification results - not-yet treated as comparison group

- If one invokes the Conditional PTA based on "not-yet-treated" units, Callaway and Sant'Anna (2021) proved that, for all g and t such that $g \in \mathcal{G}$, $t \in 2, \ldots, T$ and $t \geq g$,

$$ATT_{dr}^{ny}(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_{g,t}(X)(1-D_t)}{1-p_{g,t}(X)}}{\mathbb{E}\left[\frac{p_{g,t}(X)(1-D_t)}{1-p_{g,t}(X)}\right]}\right)\left(Y_t - Y_{g-1} - m_{g,t}^{ny}(X)\right)\right].$$

  where $m_{g,t}^{ny}(X) = \mathbb{E}[Y_t - Y_{g-1}|X, D_t = 0, G_g = 0]$ . .

- Extends Heckman et al. (1997), Abadie (2005) and Sant'Anna and Zhao (2020).

Identification results - not-yet treated as comparison group

- If one invokes the Conditional PTA based on "not-yet-treated" units, Callaway and Sant'Anna (2021) proved that, for all g and t such that $g \in \mathcal{G}$, $t \in 2, \ldots, T$ and $t \geq g$,

$$ATT^{ny}_{dr}(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_{g,t}(X)(1-D_t)}{1-p_{g,t}(X)}}{\mathbb{E}\left[\frac{p_{g,t}(X)(1-D_t)}{1-p_{g,t}(X)}\right]}\right)\left(Y_t - Y_{g-1} - m^{ny}_{g,t}(X)\right)\right],$$

  where $m^{ny}_{g,t}(X) = \mathbb{E}[Y_t - Y_{g-1}|X, D_t = 0, G_g = 0]$. .

- Extends Heckman et al. (1997), Abadie (2005) and Sant'Anna and Zhao (2020).

Identification results - not-yet treated as comparison group

- If one invokes the Conditional PTA based on "not-yet-treated" units, Callaway and Sant'Anna (2021) proved that, for all g and t such that $g \in \mathcal{G}$, $t \in 2, \ldots, T$ and $t \geq g$,

$$ATT^{ny}_{dr}(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_{g,t}(X)(1-D_t)}{1-p_{g,t}(X)}}{\mathbb{E}\left[\frac{p_{g,t}(X)(1-D_t)}{1-p_{g,t}(X)}\right]}\right)\left(Y_t - Y_{g-1} - m^{ny}_{g,t}(X)\right)\right],$$

  where $m^{ny}_{g,t}(X) = \mathbb{E}[Y_t - Y_{g-1} | X, D_t = 0, G_g = 0]$. .

- Extends Heckman et al. (1997), Abadie (2005) and Sant'Anna and Zhao (2020).

Identification results - not-yet treated as comparison group

- If one invokes the Conditional PTA based on "not-yet-treated" units, Callaway and Sant'Anna (2021) proved that,we prove that, for all g and t such that $g \in \mathcal{G}$, $t \in 2, \ldots, T$ and $t \geq g$,

$$ATT_{dr}^{ny}(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_{g,t}(X)(1-D_t)}{1-p_{g,t}(X)}}{\mathbb{E}\left[\frac{p_{g,t}(X)(1-D_t)}{1-p_{g,t}(X)}\right]}\right)\left(Y_t - Y_{g-1} - m_{g,t}^{ny}(X)\right)\right],$$

  where $m_{g,t}^{ny}(X) = \mathbb{E}[Y_t - Y_{g-1}|X, D_t = 0, G_g = 0]$. .

- Extends Heckman et al. (1997), Abadie (2005) and Sant'Anna and Zhao (2020).

Aggregation

Second step: Aggregation

Summarizing ATT(g,t)

- $ATT(g,t)$ are very useful parameters that allow us to understand treatment effect heterogeneity better.
- We can also use these to summarize the treatment effects across groups, time since treatment, and calendar time.
- Practitioners routinely attempt to pursue this avenue:
  - Run a TWFE "static" regression and focus on the β associated with the treatment.
  - Run a TWFE event-study regression and focus on β associated with the treatment leads and lags.
  - Collapse data into a 2 x 2 Design (average pre and post-treatment periods).

Summarizing ATT(g,t)

- We propose taking weighted averages of the $ATT(g,t)$ of the form:

$$\sum_{g=2}^{T}\sum_{t=2}^{T}\mathbf{1}\{g \leq t\}w_{gt}ATT(g,t)$$

- The two simplest ways of combining $ATT(g,t)$ across g and t are, assuming no-anticipation,

$$\theta_M^O := \frac{2}{T(T-1)}\sum_{g=2}^{T}\sum_{t=2}^{T}\mathbf{1}\{g \leq t\}ATT(g,t) \tag{1}$$

and

$$\theta_W^O := \frac{1}{\kappa}\sum_{g=2}^{T}\sum_{t=2}^{T}\mathbf{1}\{g \leq t\}ATT(g,t)P(G=g|X \neq 1) \tag{2}$$

- Problem: They "overweight" units that have been treated earlier

Summarizing ATT(g,t): Cohort-heterogeneity

- More empirically motivated aggregations do exist!
- Average effect of participating in the treatment that units in group g experienced:

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

Summarizing ATT(g,t): Calendar time heterogeneity

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

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

Summarizing ATT(g,t): Event-study / dynamic treatment effects

- The effect of a policy intervention may depend on the length of exposure to it.
- Average effect of participating in the treatment for the group of units that have been exposed to the treatment for exactly e time periods

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

- This is perhaps the most popular summary measure currently adopted by empiricists.

Summarizing ATT(g,t): Event-study

- When we compare $\theta_D(e)$ across two relative times e_1 and e_2, we have that

$$
\begin{aligned}
&\theta_D(e_2) - \theta_D(e_1) \\
&= \sum_{g=2}^{T} \mathbf{1}\{g + e_1 \leq T\} \underbrace{\left(ATT(g, g+e_2) - ATT(g, g+e_1)\right)}_{\text{dynamic effect for group } g} P(G = g | G + e_1 \leq T) \\
&\quad + \sum_{g=2}^{T} \mathbf{1}\{g + e_2 \leq T\} ATT(g, g+e_2) \underbrace{\left(P(G = g | G + e_2 \leq T) - P(G = g | G + e_1 \leq T)\right)}_{\text{differences in weights}} \\
&\quad - \sum_{g=2}^{T} \underbrace{\mathbf{1}\{T - e_2 \leq g \leq T - e_1\}}_{\text{different composition of groups}} ATT(g, g+e_2) P(G = g | G + e_2 \leq T)
\end{aligned}
$$

- Balance sample in "event time" to avoid compositional changes that complicate comparisons across e.

Third step: Estimation and Inference

Estimation and Inference

Estimation

- Identification results suggest a simple two-step estimation procedure.
- Estimate the generalized propensity score $p_g(X)$ by $\widehat{p}_g(X)$.
- Estimate outcome regression models for the comparison group, $m^C_{g-1}(X)$ and $m^C_t(X)$, by $\widehat{m}^C_{g-1}(X)$, and $\widehat{m}^C_t(X)$, respectively.
- With these estimators on hand, estimate the $ATT(g,t)$ using the plug-in principle (you can use IPW, OR, or DR estimands!).
- Callaway and Sant'Anna (2021) provides high-level conditions that these first-step estimators have to satisfy.
  - Similar to Chen, Linton and Van Keilegom (2003) and Chen, Hong and Tarozzi (2008)

Inference

- Under relatively weak regularity conditions,

$$\sqrt{n}\left(\widehat{ATT}(g,t) - ATT(g,t)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{gt}(\mathcal{W}_i) + o_p(1)$$

- From the above asymptotic linear representation and a CLT, we have

$$\sqrt{n}\left(\widehat{ATT}(g,t) - ATT(g,t)\right) \xrightarrow{d} N(0, \Sigma_{g,t})$$

  where $\Sigma_{gt} = \mathbb{E}[\psi_{gt}(\mathcal{W})\psi_{gt}(\mathcal{W})']$.

- Above result ignores the dependence across g and t, and "multiple-testing" problems.

Simultaneous Inference

- Let's simplify and ignore anticipation issues for the moment.
- Let $ATT_{g \leq t}$ and $\widehat{ATT}_{g \leq t}$ denote the vector of $ATT(g,t)$ and $\widehat{ATT}(g,t)$, respectively, for all $g = 2, \ldots, T$ and $t = 2, \ldots, T$ with $g \leq t$.
- Analogously, let $\Psi_{g \leq t}$ denote the collection of ψ_{gt} across all periods t and groups g such that $g \leq t$.
- Hence, we have

$$\sqrt{n}(\widehat{ATT}_{g \leq t} - ATT_{g \leq t}) \xrightarrow{d} N(0, \Sigma)$$

where

$$\Sigma = \mathbb{E}[\Psi_{g \leq t}(\mathcal{W}) \Psi_{g \leq t}(\mathcal{W})'].$$

Simultaneous confidence intervals

- How to construct simultaneous confidence intervals?
- We propose the use of a simple multiplier bootstrap procedure.
- Let $\widehat{\Psi}_{g\leq t}(\mathcal{W})$ denote the sample-analogue of $\Psi_{g\leq t}(\mathcal{W})$.
- Let $\{V_i\}_{i=1}^n$ be a sequence of *iid* random variables with zero mean, unit variance, and bounded third moment, independent of the original sample $\{\mathcal{W}_i\}_{i=1}^n$
- $\widehat{ATT}^*_{g\leq t}$, a bootstrap draw of $\widehat{ATT}_{g\leq t}$, via

$$\widehat{ATT}^*_{g\leq t} = \widehat{ATT}_{g\leq t} + \mathbb{E}_n\left[V \cdot \widehat{\Psi}_{g\leq t}(\mathcal{W})\right]. \tag{3}$$

Multiplier Bootstrap procedure

1. Draw a realization of $\{V_i\}_{i=1}^n$.
2. Compute $\widehat{ATT}^*_{g\leq t}$ as in (3), denote its (g,t)-element as $\widehat{ATT}^*(g,t)$, and form a bootstrap draw of its limiting distribution as
$$\hat{R}^*(g,t) = \sqrt{n}\left(\widehat{ATT}^*(g,t) - \widehat{ATT}(g,t)\right)$$
3. Repeat steps 1-2 B times.
4. Estimate $\Sigma^{1/2}(g,t)$ by
$$\widehat{\Sigma}^{1/2}(g,t) = \left(q_{0.75}(g,t) - q_{0.25}(g,t)\right) / \left(z_{0.75} - z_{0.25}\right)$$
5. For each bootstrap draw, compute $t-test^*_{g\leq t} = \mathbf{max}_{(g,t)} \left|\hat{R}^*(g,t)\right| \widehat{\Sigma}(g,t)^{-1/2}$.
6. Construct $\widehat{c}_{1-\alpha}$ as the empirical $(1-a)$-quantile of the B bootstrap draws of $t-test^*_{g\leq t}$.
7. Construct the bootstrapped simultaneous confidence intervals for $ATT(g,t)$, $g \leq t$, as
$$\widehat{C}(g,t) = [\widehat{ATT}(g,t) \pm \widehat{c}_{1-\alpha} \cdot \widehat{\Sigma}(g,t)^{-1/2} / \sqrt{n}].$$

Simultaneous cluster-robust confidence intervals

- Sometimes, one wishes to account for clustering.
- This is straightforward to implement with the multiplier bootstrap described above.
- Example: allow for clustering at the state level
  - draw a scalar U_s S times – where S is the number of states
  - set $V_i = U_s$ for all observations i in state s
- This procedure is justified, provided that the number of clusters is "large".

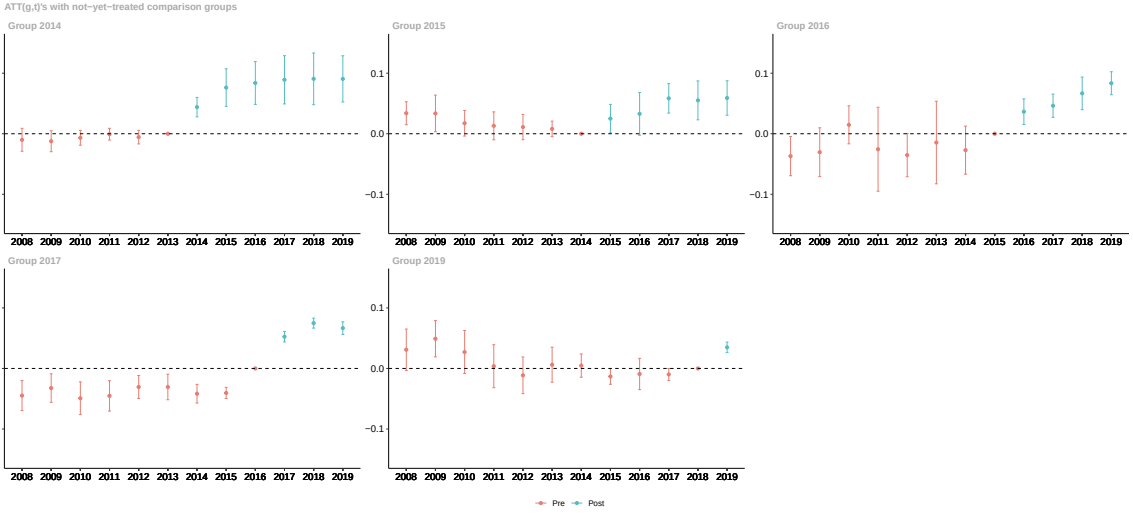
Let’s go back to the ACA Medicaid Expansion Example

ACA Medicaid Expansion

- 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

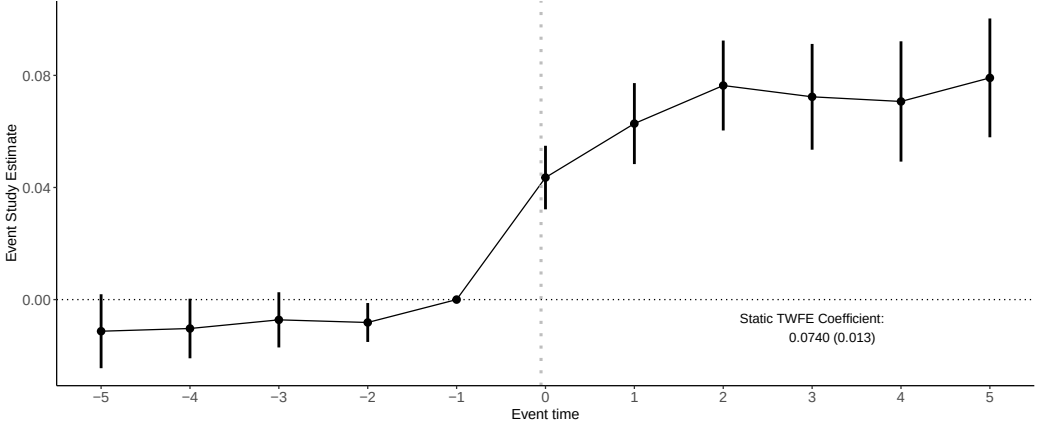
Challenge setup to make inference on ATT(g,t)'s per se

ACA Medicaid Expansion: Not-yet-treated as comparison group



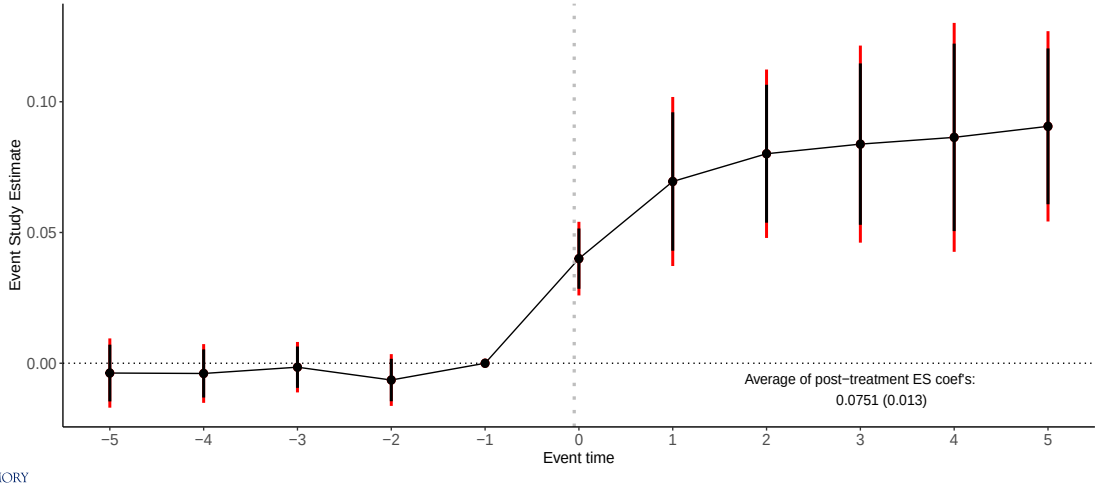
ACA Medicaid Expansion: TWFE Event-study specification

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



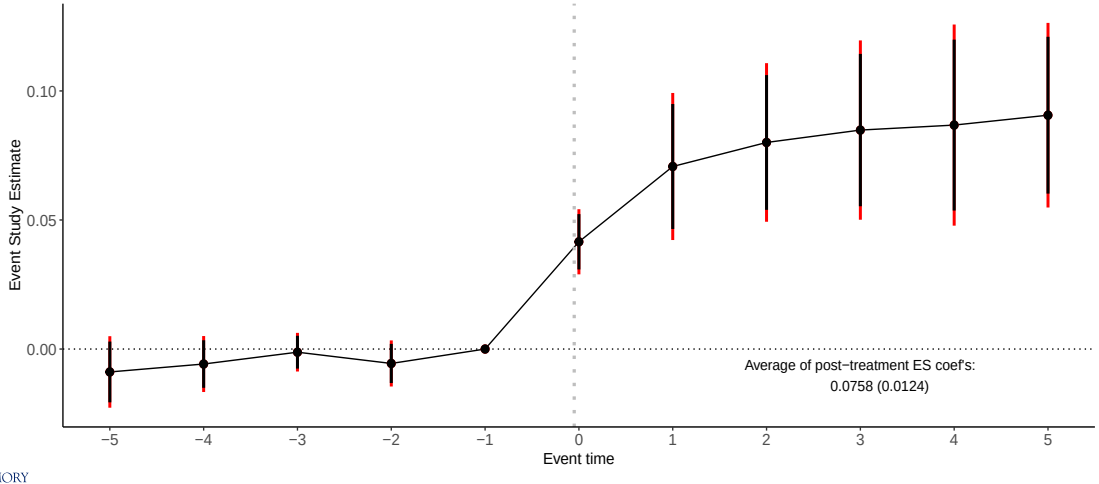
ACA Medicaid Expansion: CS Event-study specification

Figure 2: Results using "never-treated" as a comparison group



ACA Medicaid Expansion: CS Event-study specification

Figure 3: Results using "not-yet-treated" as comparison groups



Take-way messages

DiD procedures multiple time periods

- With multiple time periods and variations in treatment timing, TWFE does not respect our assumptions:
  - OLS is “variational hungry” and makes many comparisons of means
  - Some of these comparisons are bad: use already-treated units as a comparison group to “later-treated” groups
  - This can lead to “negative weighting” problems.
- Solution to the TWFE problem is simple
  - Separate the identification, aggregation, and estimation/inference parts of the problem
- Use $ATT(g,t)$ as a building block so we can transparently see how things are constructed
- Many different aggregation schemes are possible: they deliver different parameters!
- Can allow for covariates via regressions adjustments, IPW, and DR.

Empirical Application

Empirical Application

- Let's switch to R/Stata so we can see how to do all these things!

References

Abadie, Alberto, "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 2005, *72* (1), 1–19.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," *Review of Economic Studies*, 2024, *Forthcoming*.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi, "Semiparametric efficiency in GMM models with auxiliary data," *The Annals of Statistics*, apr 2008, *36* (2), 808–843.

_ , Oliver Linton, and Ingrid Van Keilegom, "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 2003, *71* (5), 1591–1608.

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

_ and _ , "Difference-in-Differences Estimators of Intertemporal Treatment Effects," *The Review of Economics and Statistics*, 2024, *Forthcoming*, 1–45.

Gardner, John, “Two-Stage Difference-in-Differences,” Technical Report, Working Paper 2021.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd, “Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme,” *The Review of Economic Studies*, October 1997, *64* (4), 605–654.

Sant’Anna, Pedro H. C. and Jun Zhao, “Doubly robust difference-in-differences estimators,” *Journal of Econometrics*, November 2020, *219* (1), 101–122.

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

Wooldridge, Jeffrey M, “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators,” *Working Paper*, 2021, pp. 1–89.

Wooldridge, Jeffrey M., “Simple approaches to nonlinear difference-in-differences with panel data,” *Econometrics Journal*, 2023, *Forthcoming*.