# Recent Advances in DiD Methods

## A selective (and personal) perspective

Pedro H. C. Sant'Anna
Microsoft and Vanderbilt University

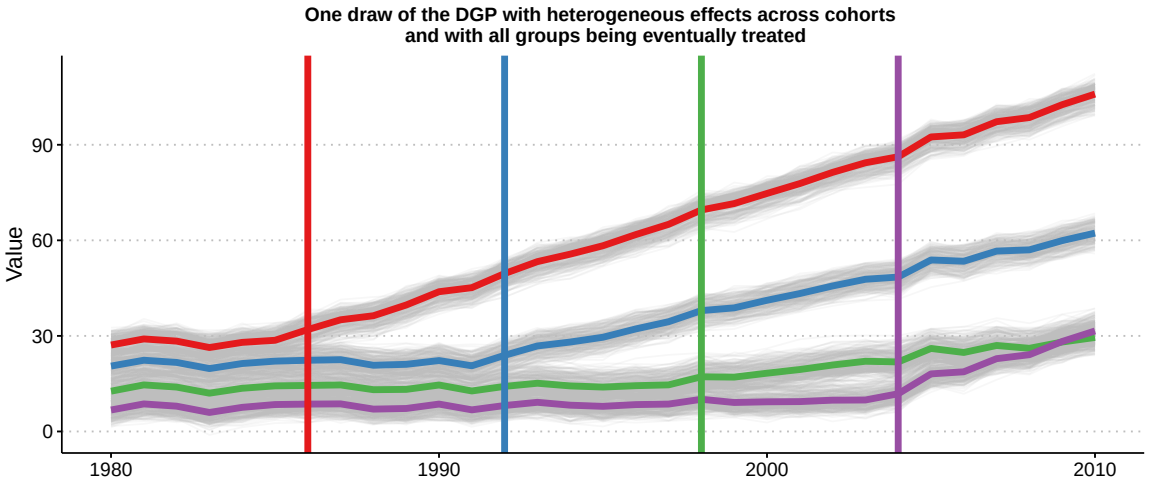
Brazilian Econometric Society (SBE) Meeting, December 2022

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

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

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

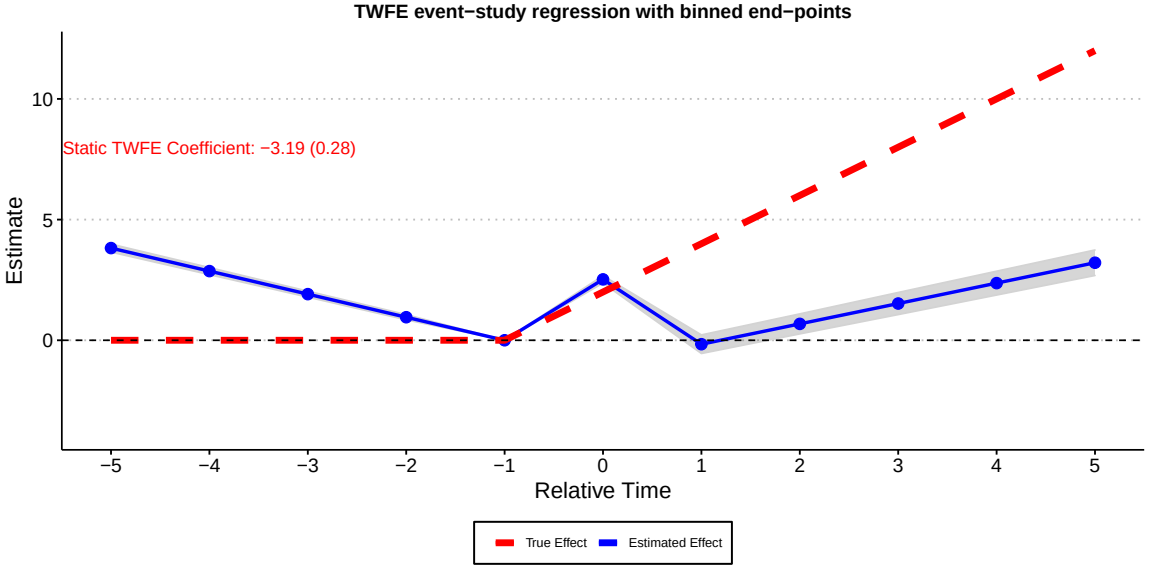
# Traditional methods: TWFE event-study regression

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

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

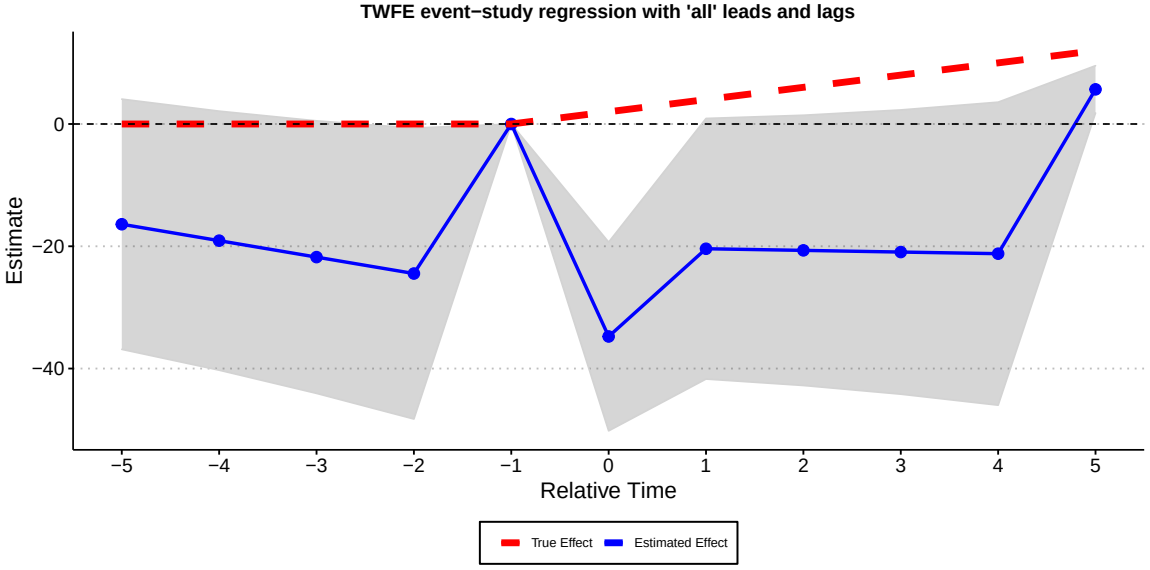
  with $K$ and $L$ to be equal to 5 ?

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

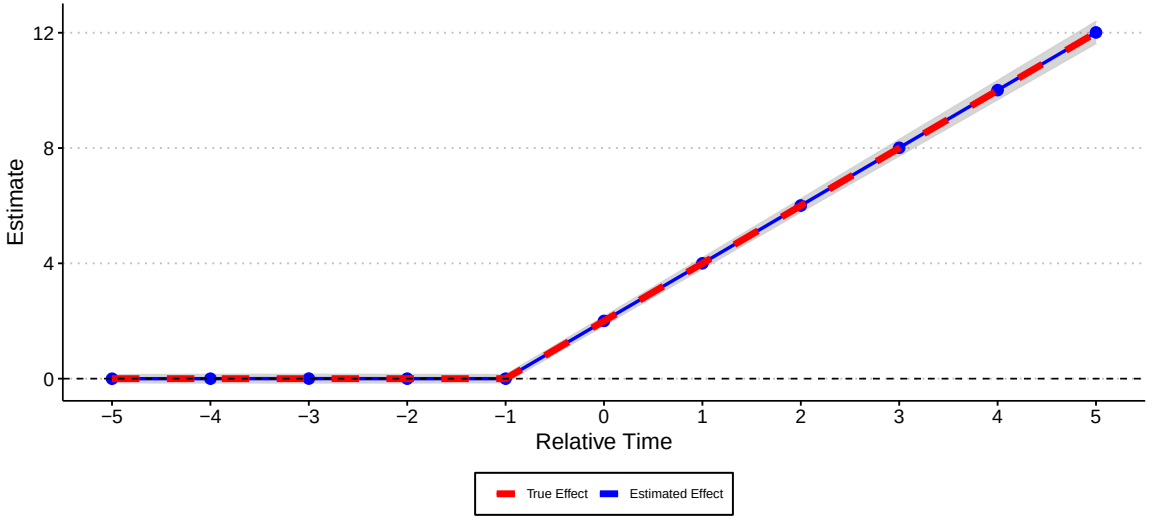

TWFE event−study regression with binned end−points
Static TWFE Coefficient: −3.19 (0.28)
Estimate
10
5
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
True Effect
Estimated Effect

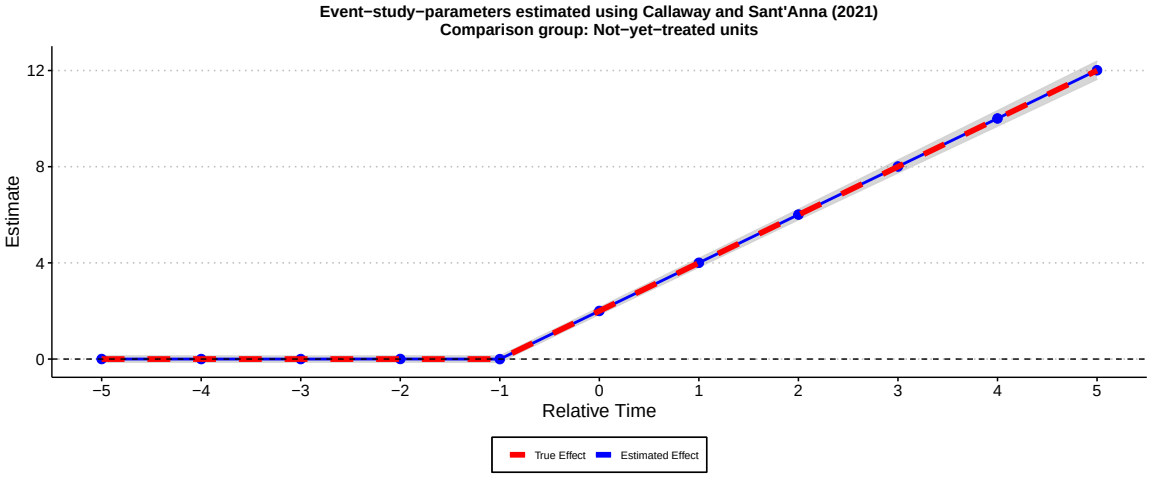
# Traditional methods: TWFE event-study regression

- 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 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

# Main take-way message

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

# Callaway and Sant'Anna (2021) in practice

Can be implemented via the R package did.

Can be implemented via the Stata package csdid.

Can be implemented via the Python package differences.

# Identification

Let’s talk about identification

# Identification

# Building block of the analysis

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

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

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

- 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 $t \geq g$,*

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

# 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

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

- 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 an 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 now we take a "long difference" , and that 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 in 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* $t \in \{2, \dots, T\}$, $g \in \mathcal{G}$ *such that* $t \geq g$,

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

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

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

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

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

- 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 $t \geq g$,

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

  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,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 better understand treatment effect heterogeneity.
- 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 $\beta$ associated with the treatment.
  - Run a TWFE event-study regression and focus on $\beta$ 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

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

- 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 $t$

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

# 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

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

- 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 hands, 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 $\psi_{gt}$ 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”.

# ACA Medicaid Expansion Example

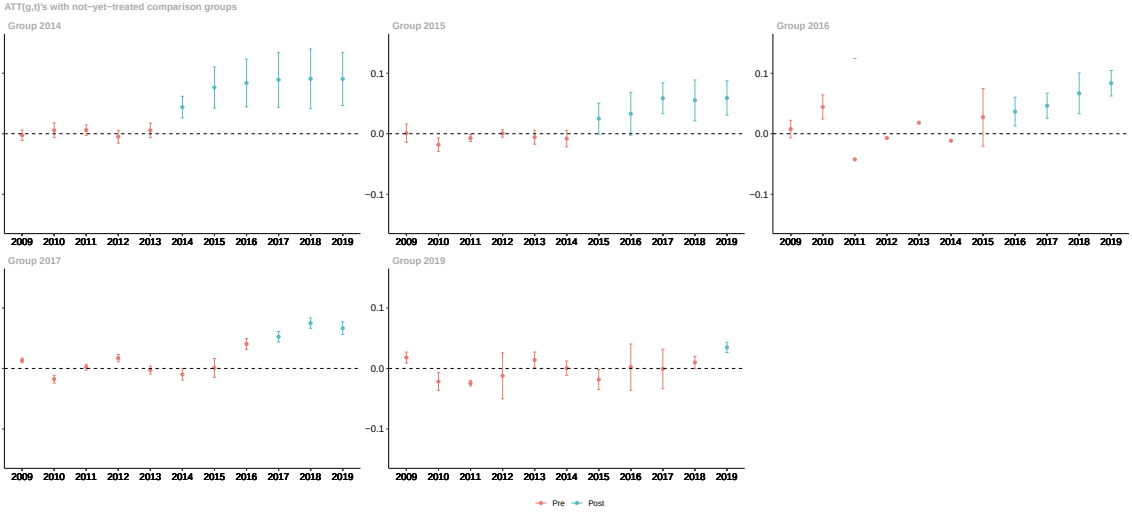
# Let’s go back to the ACA Medicaid Expansion Example

# 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

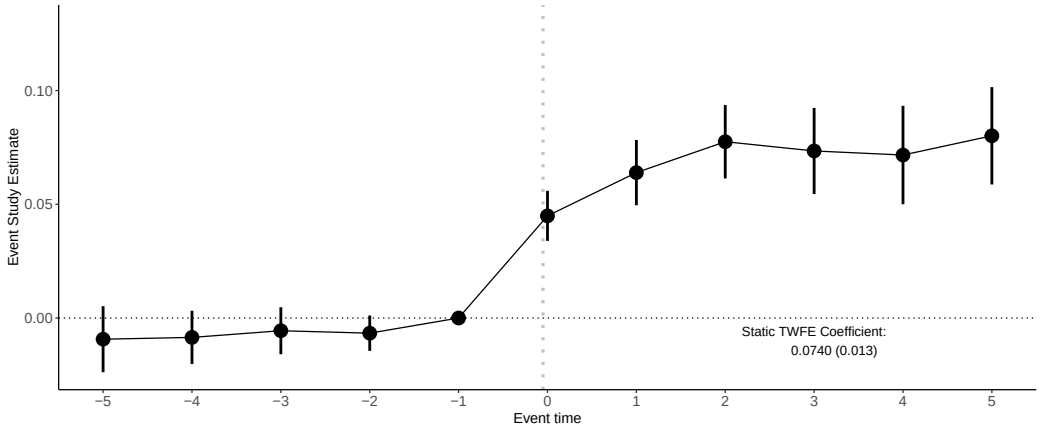
Challenge setup to make inference on ATT(g,t)'s per se

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



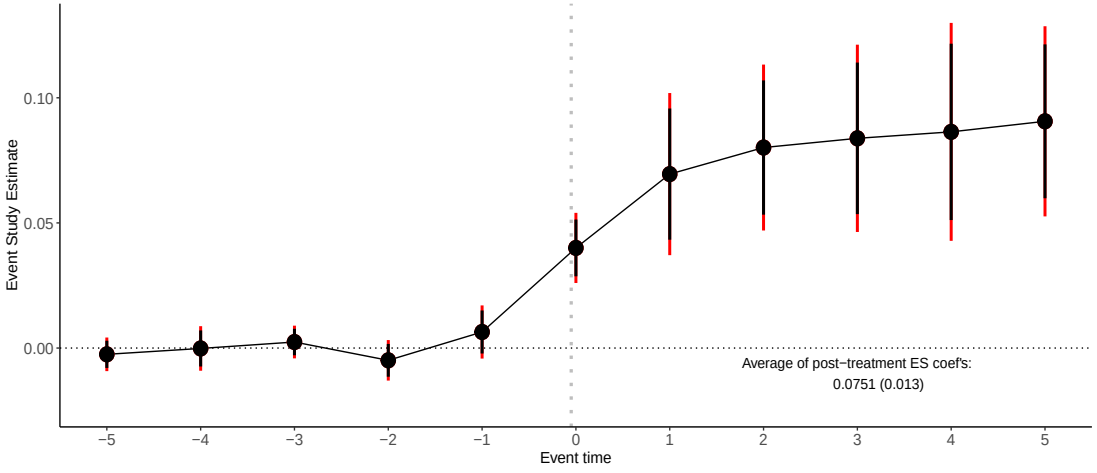
# ACA Medicaid Expansion: TWFE Event-study specification

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



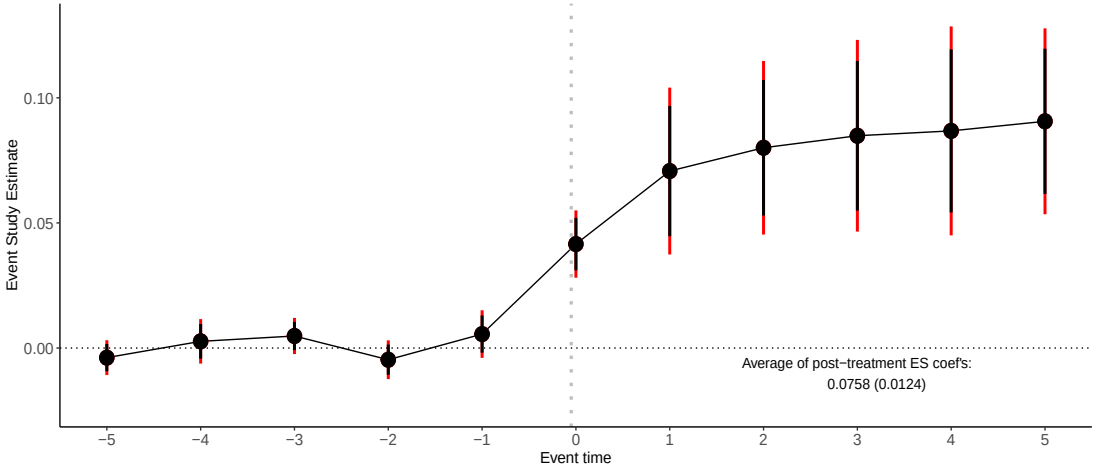
# ACA Medicaid Expansion: CS Event-study specification

**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



# Minimum Wage Illustration

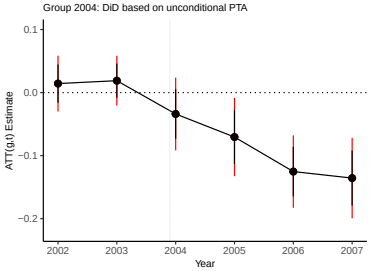
# Effect of minimum wage on teen employment

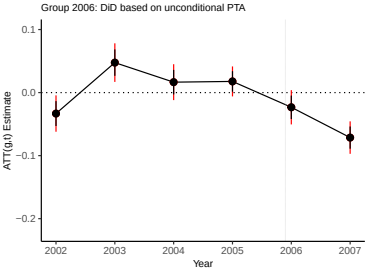
- Standard economic theory suggests that the wage floor should result in lower employment
- However, many studies find that increases in the minimum wage do not lead to disemployment effects
  - e.g. Card and Krueger (1994), Dube, Lester and Reich (2010)
- Not everyone agrees with those empirical results
  - e.g., Neumark and Wascher (2000), Neumark, Salas and Wascher (2014)
- Let's apply our proposed tools to revisit this debate.
- Treatment: MW above federal MW (we ignore how much higher it is, though).

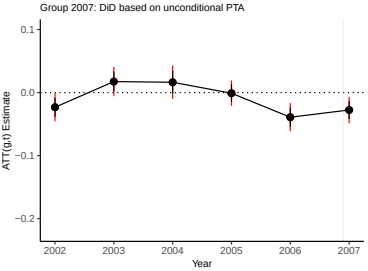
# Data

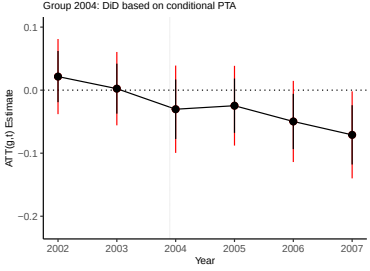
- County level data on youth employment and other county characteristics from 2001 - 2007
  - Federal minimum wage from 1999 until July 2007: \$5.15
  - In July 2007: increase to \$5.85
- We will exploit raises in state minimum wage before July 2007.
- 29 states whose minimum wage was equal to the federal minimum wage
- $Y_{i,t}$ : log teen first-quarter employment in county $i$ at year $t$.
- $X_i$ : Region, population, population squared, median income, median income squared, fraction of white, fraction with a high school education, poverty rate.
- No evidence of pscore misspecification: Sant'Anna and Song (2019)

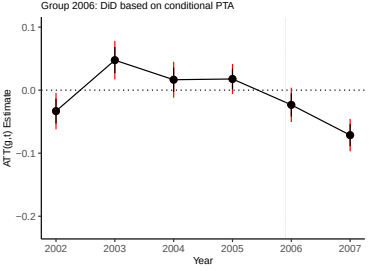
# Min. wage results using “never-treated” as comp. group

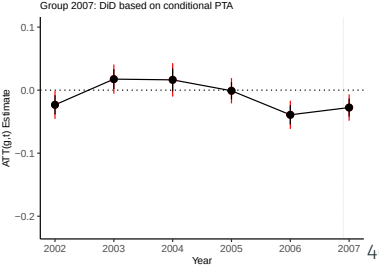


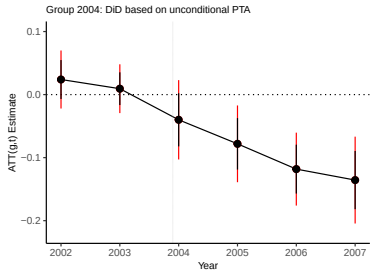



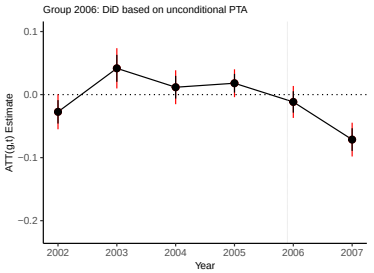



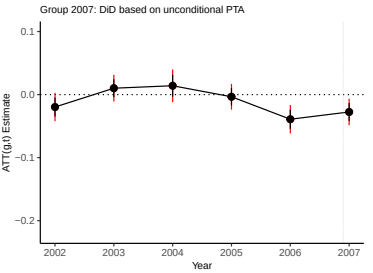



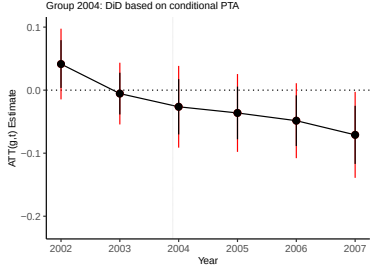



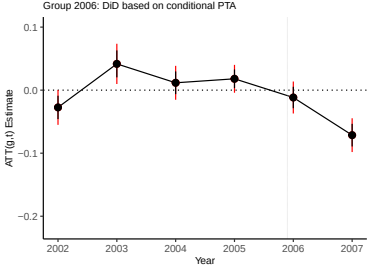
# Min. wage results using “not-yet-treated” as comp. group

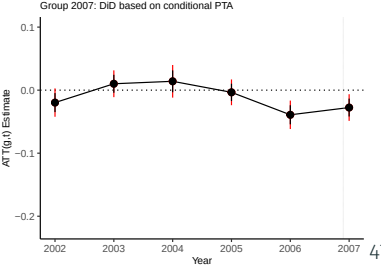


# Summary measures based on “never treated”

**(b) Conditional Parallel Trends**

| | Partially Aggregated | | | | Single Parameters |
|---|---|---|---|---|---|
| TWFE | | | | | -0.008 |
| | | | | | (0.006) |
| Simple Weighted Average | | | | | -0.033 |
| | | | | | (0.007) |
| Group-Specific Effects | g=2004 | g=2006 | g=2007 | | |
| | -0.044 | -0.029 | -0.029 | | -0.031 |
| | (0.020) | (0.008) | (0.008) | | (0.007) |
| Event Study | e=0 | e=1 | e=2 | e=3 | |
| | -0.024 | -0.041 | -0.050 | -0.071 | -0.046 |
| | (0.006) | (0.009) | (0.022) | (0.026) | (0.013) |
| Calendar Time Effects | t=2004 | t=2005 | t=2006 | t=2007 | |
| | -0.030 | -0.025 | -0.030 | -0.049 | -0.033 |
| | (0.022) | (0.021) | (0.009) | (0.007) | (0.012) |
| Event Study | e=0 | e=1 | | | |
| w/ Balanced Groups | -0.016 | -0.041 | | | -0.028 |
| | (0.010) | (0.009) | | | (0.008) |

# Take-way messages

# DiD procedures multiple time periods

- With multiple time periods and variation 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.


www.causal-solutions.com

For any questions/comments, you can contact me via

pedrosantanna@causal-solutions.com
pedro.h.santanna@vanderbilt.edu
psantanna@microsoft.com

@pedrohcgs

# Violations of Parallel Trends

# Recent Boom of New DiD Methods: Violations of PT

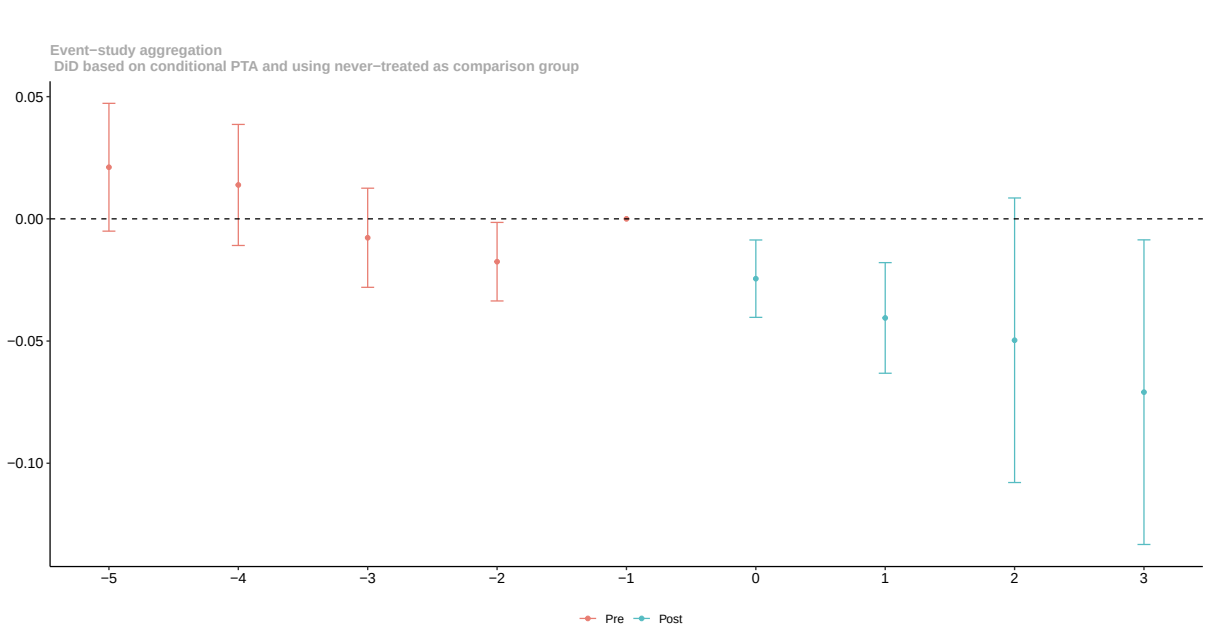
- What if treatment Parallel Trends Assumption is violated?
- **Rambachan and Roth (2022):** Shows how you can use pre-trends to bound ATT's when PT are violated.
- Build on Manski and Pepper (2015) but provide new and practically relevant uniformly valid inference procedures. New rationale for violations of PT, too!
- Can be easily combined with Callaway and Sant'Anna (2021) - *https://github.com/pedrohcgs/CS_RR*.
- **This is my favorite paper of this "batch" of new DiD papers.**

## Why do I like this paper so much?

- Currently common practice on pre-test has limitations with important practical consequences.
- However, as a good econometrician, instead of sitting in our Ivory Tower, we need to seek several practical, easy-to-use tools that can alleviate some of these problems.
- This is what Rambachan and Roth (2022) do!
- In my view, the sensitivity analysis procedures in Rambachan and Roth (2022) are fundamental to improving the reliability and transparency of DiD procedures.
- Let's briefly show this using the did and HonestDiD R packages, which implements Callaway and Sant'Anna (2021) and Rambachan and Roth (2022), respectively.

# Combining Callaway and Sant'Anna (2021) and Rambachan and Roth (2021)

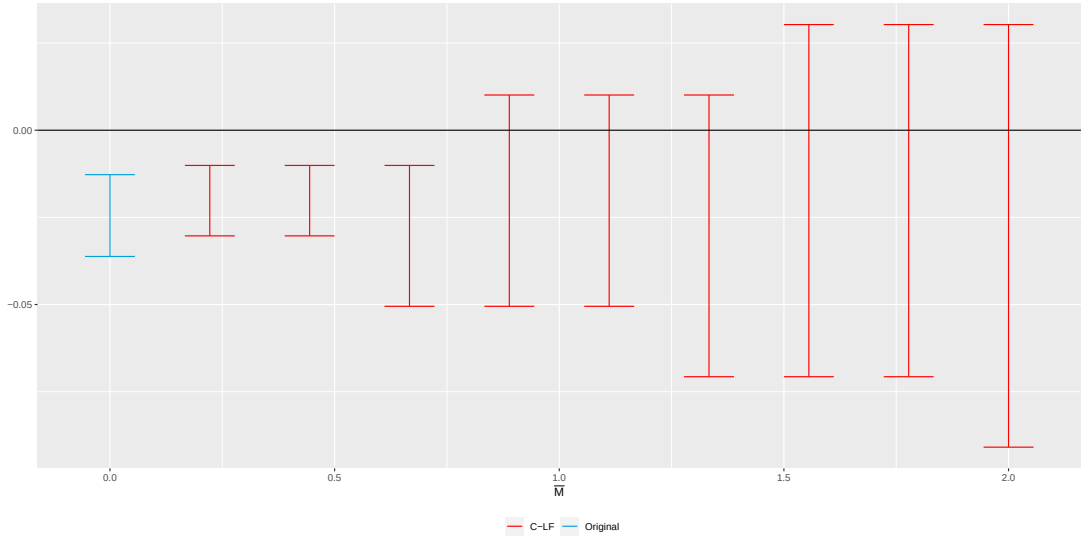
```
# Install the packages (I used the Github versions)
devtools::install_github("bcallaway11/did");
devtools::install_github("asheshrambachan/HonestDiD")

#Load the packages
library(did); library(HonestDiD); library(dplyr); library(here)

# Load data used by Callaway and Sant'Anna (2021)
min_wage <- readRDS((here("data",'min_wage_CS.rds')))

#-------------------------------------------------------------------------------------------
# Formula for covariates
xformla <- ~ region + (medinc + pop ) + I(pop^2) + I(medinc^2)  + white + hs  + pov
#-------------------------------------------------------------------------------------------

# Estimate ATT(g,t)'s using DR DiD with never-treated as comparison group
CS_never_cond <- did::att_gt(yname="lemp", tname="year", idname="countyreal", gname="first.treat",
                             xformla = xformla, control_group="nevertreated", data = min_wage,
                             panel = TRUE, base_period="universal", bstrap = TRUE, cband = TRUE)

# compute event-study aggregation
CS_es_never_cond <- aggte(CS_never_cond, type = "dynamic", min_e = -5, max_e = 5)
ggdid(CS_es_never_cond,
      title = "Event-study aggregation \n DiD based on conditional PTA and using never-treated as comparison group ")
```


Event−study aggregation
DiD based on conditional PTA and using never−treated as comparison group
0.05
0.00
−0.05
−0.10
−5
−4
−3
−2
−1
0
1
2
3
Pre
Post

# Rambachan and Roth (2021) after Callaway and Sant'Anna (2021)

```
# Brant has written a wrapper for HonestDiD that allows one to use aggte did outputs as inputs

# Here we apply the wrapper, and use the ``relative magnitude'' type of sensitivity analysis

# Doing it for instantaneous treatment effect, e = 0
hd_cs_rm_never <- honest_did(CS_es_never_cond,
                             e = 0,
                             type="relative_magnitude")

# Plot results
cs_HDiD_relmag <- createSensitivityPlot_relativeMagnitudes(hd_cs_rm_never$robust_ci,
                                        hd_cs_rm_never$orig_ci)
cs_HDiD_relmag
```

# Sensitivity Analysis based on “relative magnitude” restrictions



# When is PT sensitive to functional form?

# Recent Boom of New DiD Methods: When is PT sensitive to functional form?

- When is PT sensitive to functional form?
- **Roth and Sant'Anna (2022):** Provide necessary and sufficient conditions for DiD estimators to be insensitive to functional form restrictions.
- This holds if and only if PT holds in a distributional sense.
- This is testable - can cast it as a test of monotonicity!
- Also provides some "microfoundations" of how PT can hold in this particular distributional sense.

# Non-binary treatments

# Recent Boom of New DiD Methods: Continuous and Multi-valued Treatments

- What if treatment is multi-valued or continuous?
- **Callaway, Goodman-Bacon and Sant'Anna (2021):** Make some advances on this problem (still in progress).
- We can measure treatment effect "in levels":

$$ATT(a|b) = \mathbb{E}[Y_t(a) - Y_t(0)|D = b] \quad \text{and} \quad ATE(d) = \mathbb{E}[Y_t(d) - Y_t(0)].$$

- But we can also measure treatment effects in "increments":

$$ACRT(d|d) = \left.\frac{\partial \mathbb{E}[Y_t(l)|D = d]}{\partial l}\right|_{l=d} \quad \text{and} \quad ACR(d) = \frac{\partial \mathbb{E}[Y_t(d)]}{\partial d}.$$

or

$$ACRT(d_j|d_j) = \mathbb{E}[Y_t(d_j) - Y_t(d_{j-1})|D = d_j] \quad \text{and} \quad ACR(d_j) = \mathbb{E}[Y_t(d_j) - Y_t(d_{j-1})].$$

- Discuss problems with TWFE and how to fix some of these (more to come soon!)

# Importance of being careful about parameter of interest

- With binary treatments and staggered adoption, the literature has somehow stressed the pitfalls of using variants of the TWFE regression

$$Y_{it} = \alpha_i + \lambda_t + \beta D_{it} + \varepsilon_{it}.$$

- Issue is that, under some assumptions,

$$\beta = \sum_{t,g} w_{t,g} \cdot ATT(g,t),$$

but the weights $w_{t,g}$ are not guaranteed to be convex, i.e., they can be negative; see, e.g., Athey and Imbens (2021), Borusyak, Jaravel and Spiess (2021), de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), Sun and Abraham (2021).

- What if the weights were convex? Would this be "fine"?

- LATE and MTE IV literature have been debating this issue for the last 20 years: **What is the causal question of interest? That should help us pick "good" weights.**

# What if treatment is continuous?

- With continuous treatments, this becomes even more important, as discussed in Callaway et al. (2021)

- Even with two periods, with no units being treated in period $t = 1$, some units remaining untreated at period $t = 2$, and the others receiving different dosages $d$, the $\beta$ from the TWFE regression

$$Y_{it} = \alpha_i + \lambda_t + \beta D_{it} + \varepsilon_{it}$$

  can have **very different** causal interpretations!

# What if treatment is continuous?

- Under a "strong parallel trends" assumption, we have:
  - If we were to use "slope effects" as "building blocks":

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1(l) ACR(l)\, dl + w_0 \frac{ATE(d_L)}{d_L},$$

    where $ACR(d) = \frac{\partial \mathbb{E}[Y_t(d)]}{\partial d}$, and all weights are non-negative and integrate to one.
  - If we were to use "level effects" as "building blocks":

$$\beta^{twfe} = \int_{\mathcal{D}_+} w_1^{\text{alt}}(l) \frac{ATE(l)}{l}\, dl$$

    where $ATE(d) = \mathbb{E}[Y_t(d) - Y_t(0)]$, and the weights integrate to one but are non-convex (i.e., can be negative).

- Same estimator and same assumptions, but sharply different interpretations!

In my view, whenever it is possible, we should be clear about the causal parameter of interest from the very beginning!

# Recent Boom of New DiD Methods: Continuous and Multi-valued Treatments

- What about fuzzy DiD setups?
- de Chaisemartin and D'Haultfœuille (2018): fantastic paper showing how one can handle setups where treatment is binary (say at unit level) but one is willing to impose parallel trends at a more aggregate level (say state level).
- The aggregation step leads to non-binary "treatments", and potentially all "clusters" are exposed to treatment in all periods (but with different intensity).
- de Chaisemartin and D'Haultfœuille (2018) shows that the "Wald-estimand" has a LATE interpretation when the effect of the treatment is stable over time, and if the effect of the treatment is the same in the treatment and in the control group.
- Since these assumptions are strong, the authors also propose alternative estimators that build on Athey and Imbens (2006) and do not rely on these assumptions.

# Inference with few treated clusters

# Recent Boom of New DiD Methods: What if we have a handful of clusters only?

- What if we have a handful of clusters only?
- The literature has tackled this question using different restrictions on potential outcomes and/or treatment effect heterogeneity.
- This is a hard problem, especially when we do not want to impose restrictions on the time-dependency of the potential outcomes.
- Most of the literature adopts a “regression view” of the problem, which, in my view, hides important implications of the required assumptions for these solutions to work.
- For the interest of time, I refer to Section 5 of Roth, Sant’Anna, Bilinski and Poe (2021) for more details.

# References

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

**Athey, Susan and Guido Imbens**, "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 2006, *74* (2), 431–497.

**_ and _**, "Design-Based Analysis in Difference-In-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 2021, (Forthcoming).

**Borusyak, Kirill, Xavier Jaravel, and Jann Spiess**, "Revisiting Event Study Designs: Robust and Efficient Estimation," 2021.

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

**_, Andrew Goodman-Bacon, and Pedro H.C. Sant'Anna**, "Difference-in-Differences with a Continuous Treatment," *arXiv:2107.02637*, 2021.

**Card, David and Alan Krueger**, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 1994, *84* (4), 772–793.

**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**, "Fuzzy Differences-in-Differences," *The Review of Economic Studies*, April 2018, *85* (2), 999–1028.

**_ and _** , "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

**Dube, Arindrajit, T William Lester, and Michael Reich**, "Minimum Wage Effects across State Borders: Estimates Using Contiguous Counties," *The Review of Economics and Statistics*, 2010, *92* (4), 945–964.

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

**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.

**Manski, Charles F and John V Pepper**, "How Do Right-To-Carry Laws Affect Crime Rates? Coping With Ambiguity Using Bounded-Variation Assumptions," Working Paper 21701, National Bureau of Economic Research November 2015. Series: Working Paper Series.

**Neumark, David and William Wascher**, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment," *American Economic Review*, December 2000, *90* (5), 1362–1396.

**_ , J. M. Ian Salas, and William Wascher**, "Revisiting the Minimum Wage—Employment Debate: Throwing Out the Baby with the Bathwater?," *ILR Review*, 2014, *67* (3).

**Rambachan, Ashesh and Jonathan Roth**, "A More Credible Approach to Parallel Trends," *The Review of Economic Studies*, 2022, *Forthcoming*.

**Roth, Jonathan and Pedro H. C. Sant'Anna**, "When Is Parallel Trends Sensitive to Functional Form?," *Econometrica*, 2022, *Forthcoming*.

**_ , _ , Alyssa Bilinski, and John Poe**, "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature," *arXiv:2201.01194*, 2021.

**Sant'Anna, Pedro H.C. and Xiaojun Song**, "Specification tests for the propensity score," 2019, *210* (2), 379–404.

**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).