Difference-in-Differences with Multiple Time Periods

Brantly Callaway
University of Georgia

Pedro H. C. Sant'Anna
Vanderbilt University

Forthcoming at the Journal of Econometrics, December 2020

Callaway and Sant'Anna (2020) in a nutshell

- We study average treatment effects in DiD setups with:
  1. Multiple time periods;
  2. Variation in treatment timing (but with staggered treatment adoption);
  3. Parallel trends assumption holds after conditioning on observed covariates;
- We want to better understand treatment effect heterogeneity:
  - Group-time average treatment effects:

$$ATT(g,t) = \mathbb{E}\left[Y_{i,t}(g) - Y_{i,t}(0) \,|\, G_{i,g} = 1\right].$$

- Discuss how to summarize these causal effects (e.g. event-study analysis).

Clearly separate identification, aggregation and estimation/inference steps!

Callaway and Sant'Anna (2020) in practice

Proposed tools are suitable for both panel and repeated cross-section data.

Can be implemented via the `R` package `did`.

What is the empirical relevance of our proposal?

Currie, Kleven and Zwiers (2020), AEA P&P.

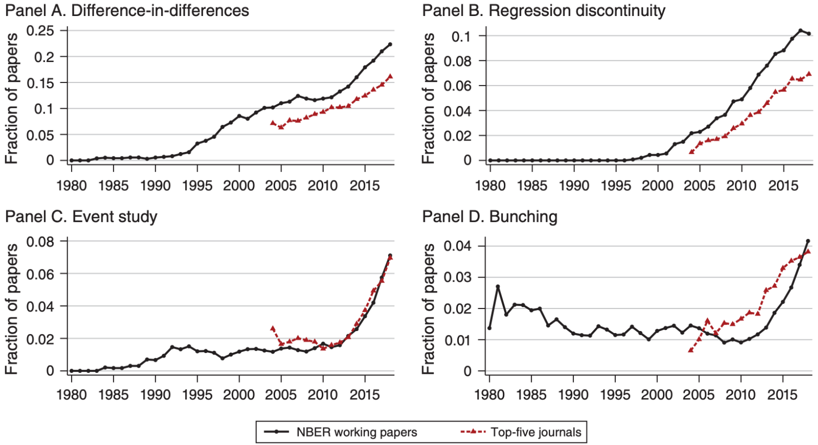


FIGURE 4. QUASI-EXPERIMENTAL METHODS

Notes: This figure shows the fraction of papers referring to each type of quasi-experimental approach. See Table A.I for a list of terms. The series show five-year moving averages.

Difference-in-Differences

- Difference-in-Differences (DiD) is one of the most popular designs for causal inference.
- Canonical format:
  - 2 groups: $G = 0$ and $G = 1$;
  - 2 times periods: $t = 0$ and $t = 1$.
- Parameter of interest:

$$ATT \equiv \mathbb{E}\left[Y_{i,1}(1)\,|G_i = 1\right] - \mathbb{E}\left[Y_{i,1}(0)\,|G_i = 1\right]$$

- **Parallel Trends Assumption:**

$$\mathbb{E}\left[Y_1(0) - Y_0(0)\,|G = 1\right] \;=\; \mathbb{E}\left[Y_1(0) - Y_0(0)\,|G = 0\right]$$

Difference-in-Differences as a Regression

- Canonical DiD:

$$\widehat{ATT_n} = \mathbb{E}_n \left[Y_1 - Y_0 | G = 1 \right] - \mathbb{E}_n \left[Y_1 - Y_0 | G = 0 \right].$$

- We can use the regression to estimate β, the ATT:

$$Y_{i,t} = \alpha + \gamma G_i + \lambda 1 \left\{ t = 1 \right\} + \underbrace{\beta}_{\equiv ATT} \left(G_i \cdot 1 \left\{ t = 1 \right\} \right) + \varepsilon_{i,t}.$$

- We can leverage its regression representation to conduct asymptotically valid inference.

Difference-in-Differences in Practice

- Many DiD empirical applications, however, deviate from the canonical DiD setup
  - Availability of covariates X
  - More than two time periods
  - Variation in treatment timing

Traditional methods: TWFE event-study regression

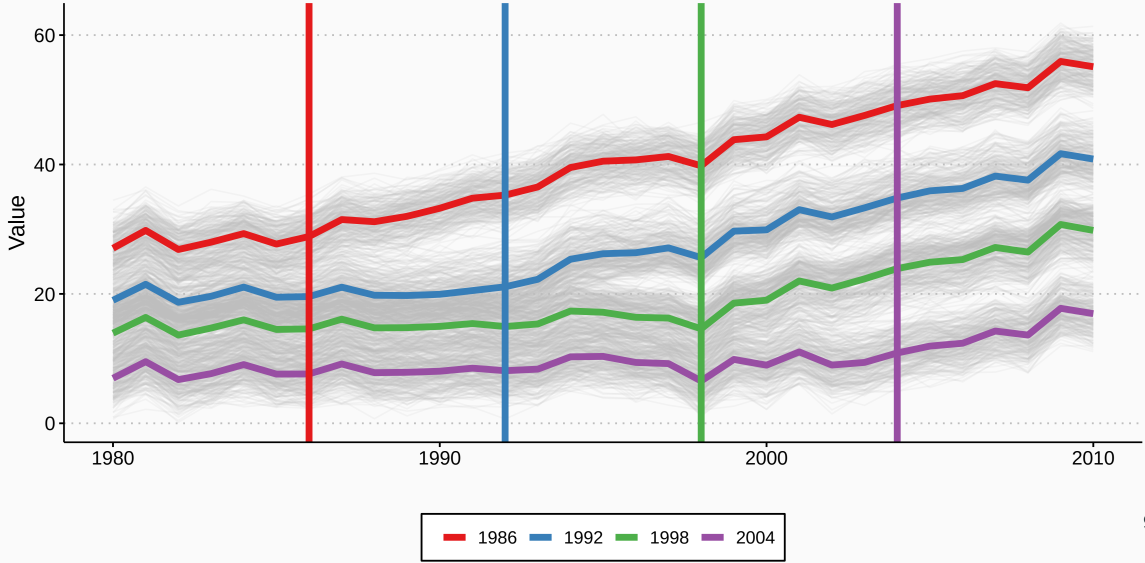
- It is tempting to "extrapolate" from the canonical DiD setup and use variations of following TWFE specification to estimate causal effects:

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

  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.

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:

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

- ATT at the first treatment period is 1, at the second period since treatment is 2, etc.

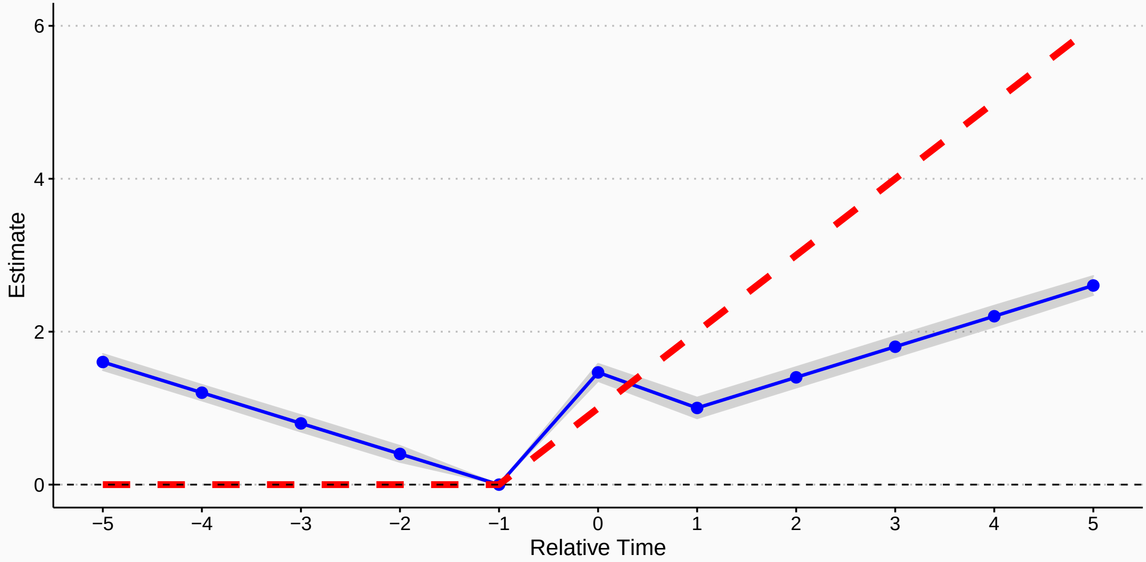
Traditional methods: TWFE event-study regression

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

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

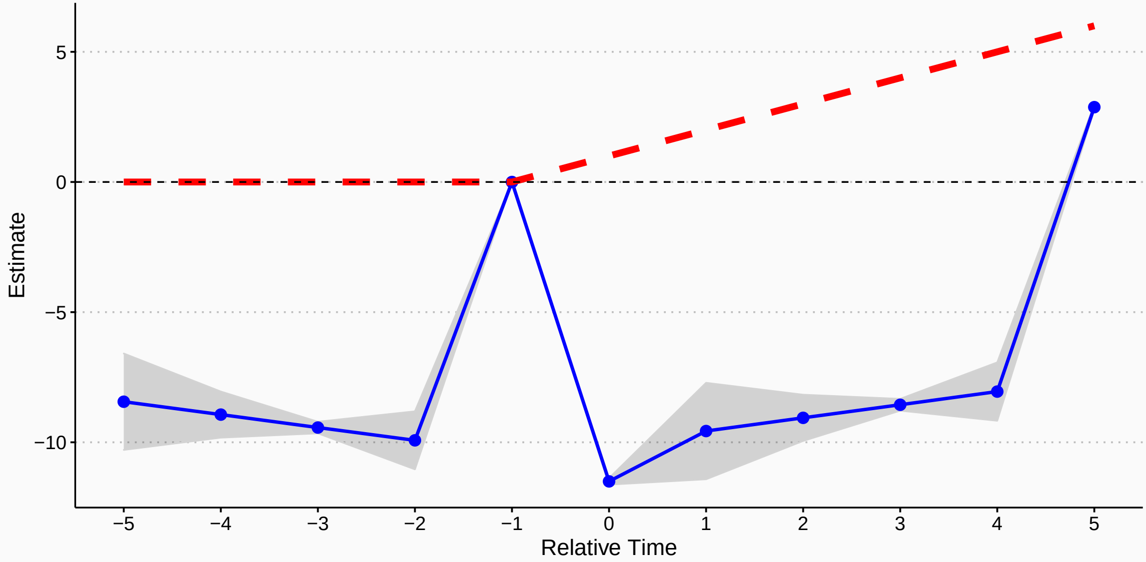
with K and L to be equal to 5 ?

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


Estimate
6
4
2
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time

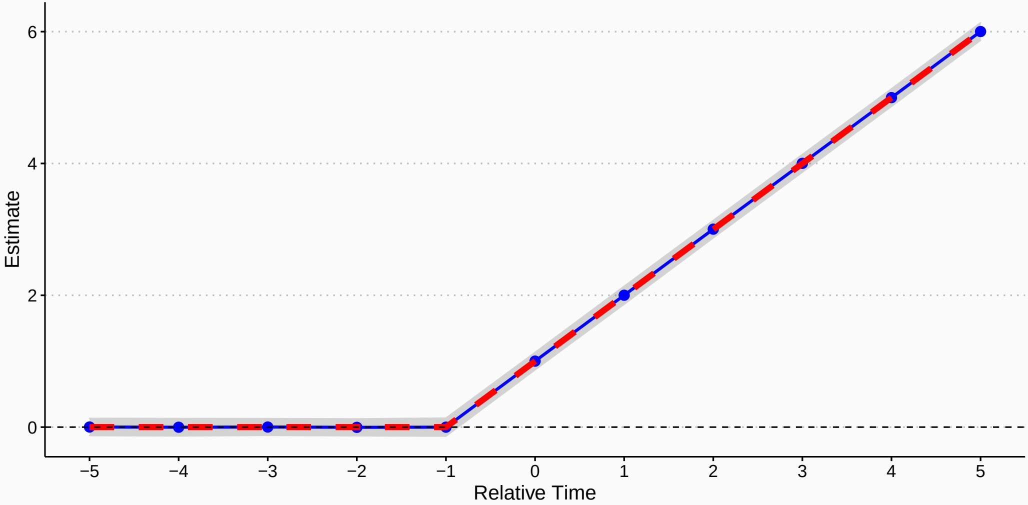
Traditional methods: TWFE event-study regression

- 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 ?



Estimate
5
0
−5
−10
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time

Event-study plot using CS proposed estimator



Recent related literature

- Recent and emerging literature on heterogeneous treatment effects in DiD with variation in treatment timing.
- The papers closest to ours are Athey and Imbens (2018), Borusyak and Jaravel (2017), de Chaisemartin and D'Haultfouille (2020), Goodman-Bacon (2019) and Sun and Abraham (2020)
- All these papers present "negative" results about using TWFE, which we do not have.
  - Sun and Abraham (2020) has results for the event-study TWFE regressions that rationalize the bad results shown in the previous simulation slides.

Recent related literature

- On the other hand, our paper has some unique features on it:
  - We attempt to make minimal parallel trends assumptions to identify the $ATT(g,t)$;
  - We allow for covariates in a flexible form
  - We propose different estimation procedures based on outcome regression, IPW and doubly robust methods;
  - We discuss different aggregation schemes to further summarize the effects of the treatment;
  - We cover both panel and (stationary) repeated-cross section cases.

Let me explain the building blocks of CS

Framework for the panel data case

- Consider a random sample

$$\{(Y_{i,1}, Y_{i,2}, \dots, Y_{i,\mathcal{T}}, D_{i,1}, D_{i,2}, \dots, D_{i,\mathcal{T}}, X_i)\}_{i=1}^{n}$$

  where $D_{i,t} = 1$ if unit i is treated in period t, and 0 otherwise
- $G_{i,g} = 1$ if unit i is first treated at time g, and zero otherwise ("Treatment start-time dummies")
- $C = 1$ is a "never-treated" comparison group
- Staggered treatment adoption: $D_{i,t} = 1 \implies D_{i,t+1} = 1, \textit{ for } t = 1, 2, \dots, \mathcal{T}$.

Framework for the panel data case (cont.)

- Limited Treatment Anticipation: There is a known $\delta \geq 0$ s.t.

$$\mathbb{E}[Y_t(g)|X, G_g = 1] = \mathbb{E}[Y_t(0)|X, G_g = 1] \; a.s..$$

for all $g \in \mathcal{G}, t \in 1, \ldots, \mathcal{T}$ such that $\underbrace{t < g - \delta}_{\text{“before effective starting date”}}$.

- Generalized propensity score uniformly bounded away from 1:

$$p_{g,t}(X) = P\left(G_g = 1 | X, G_g + (1 - D_t)(1 - G_g) = 1\right) \leq 1 - \epsilon \; a.s..$$

Parameter of interest

- Parameter of interest:

$$ATT\left(g,t\right)=\mathbb{E}\left[Y_t\left(g\right)-Y_t\left(0\right)\left|G_g=1\right.\right],\ \textit{for } t\geq g-\delta.$$

Parallel trend assumption based on a "never treated" group

Assumption (Conditional Parallel Trends based on a "never-treated")
For each $t \in \{2, \ldots, \mathcal{T}\}$, $g \in \mathcal{G}$ *such that* $t \geq g - \delta$,

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

Parallel Trends based on not-yet treated groups

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

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

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

Identification results - never treated as comparison group

- Under these assumptions, we prove that, for all g and t such that $g \in \mathcal{G}_\delta \equiv \mathcal{G} \cap \{2+\delta, 3+\delta, \ldots, \mathcal{T}\}$, $t \in \{2, \ldots \mathcal{T}-\delta\}$ and $t \geq g-\delta$, $ATT\left(g,t\right)$ is nonparametrically identified by the DR estimand

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

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

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

Identification results - never treated as comparison group

- Under these assumptions, we prove that, for all g and t such that $g \in \mathcal{G}_\delta \equiv \mathcal{G} \cap \{2+\delta, 3+\delta, \ldots, \mathcal{T}\}$, $t \in \{2, \ldots \mathcal{T} - \delta\}$ and $t \geq g - \delta$, $ATT\left(g, t\right)$ is nonparametrically identified by the DR estimand

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

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

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

Identification results - never treated as comparison group

- Under these assumptions, we prove that, for all g and t such that $g \in \mathcal{G}_\delta \equiv \mathcal{G} \cap \{2+\delta, 3+\delta, \ldots, \mathcal{T}\}$, $t \in \{2, \ldots \mathcal{T}-\delta\}$ and $t \geq g-\delta$, $ATT(g,t)$ is nonparametrically identified by the DR estimand

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

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

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

Identification results - never treated as comparison group

- Under these assumptions, we prove that, for all g and t such that $g \in \mathcal{G}_\delta \equiv \mathcal{G} \cap \{2+\delta, 3+\delta, \ldots, \mathcal{T}\}$, $t \in \{2, \ldots \mathcal{T} - \delta\}$ and $t \geq g - \delta$, $ATT(g,t)$ is nonparametrically identified by the DR estimand

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

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

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

What if the identifying assumptions hold unconditionally?

- In the case where covariates do not play a major role into the DiD identification analysis, these formulas simplify to

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

- 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!

Identification results - not-yet treated as comparison group

- If one invokes the Conditional PTA based on "not-yet-treated" units, we prove that, for all g and t such that $g \in \mathcal{G}_\delta$, $t \in 2, \ldots \mathcal{T} - \delta$ and $t \geq g - \delta$,

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

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

- Extends Heckman, Ichimura and Todd (1997), Abadie (2005), 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, we prove that, for all g and t such that $g \in \mathcal{G}_\delta$, $t \in 2, \ldots \mathcal{T} - \delta$ and $t \geq g - \delta$,

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

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

- Extends Heckman, Ichimura and Todd (1997), Abadie (2005), 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, we prove that, for all g and t such that $g \in \mathcal{G}_\delta$, $t \in 2, \ldots \mathcal{T} - \delta$ and $t \geq g - \delta$,

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

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

- Extends Heckman, Ichimura and Todd (1997), Abadie (2005), 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, we prove that, for all g and t such that $g \in \mathcal{G}_\delta$, $t \in 2, \ldots \mathcal{T} - \delta$ and $t \geq g - \delta$,

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

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

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

What if the identifying assumptions hold unconditionally?

- In this simpler case, the identifying results simplify to

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

- 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!

Summarizing the $ATT(g,t)$'s

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, calendar time.
- Empiricist 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:

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

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

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

and

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

- 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:

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

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

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

- Very informally, this is akin to asking:
  "How many lives have we saved until time t by adopting the shelter-at-home policy?"

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

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

- 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

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

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

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, $m_{g-1}^{C}(X)$ and $m_t^{C}(X)$, by $\widehat{m}_{g-1}^{C}(X)$, and $\widehat{m}_t^{C}(X)$, respectively.
- With these estimators on hands, estimate the $ATT(g,t)$ using the plug-in principle (you can use IPW, OR or DR estimands!).
- In the paper, we provide 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 $g = 2, \dots, \mathcal{T}$ and $t = 2, \dots, \mathcal{T}$ 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 $t-test_{g\leq t}^{*} = \max_{(g,t)}\left|\hat{R}^{*}(g,t)\right|\widehat{\Sigma}(g,t)^{-1/2}$.
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".

Empirical Illustration

Effect of minimum wage on teen employment

- Standard economic theory suggests that wage floor should result in lower employment
- However, many studies find that increases in the minimal 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
  - Neumark and Wascher (1992, 2000, 2007, 2008), Neumark, Sala 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)

Figure 1: Minimum Wage Results using “never-treated” as a comparison group

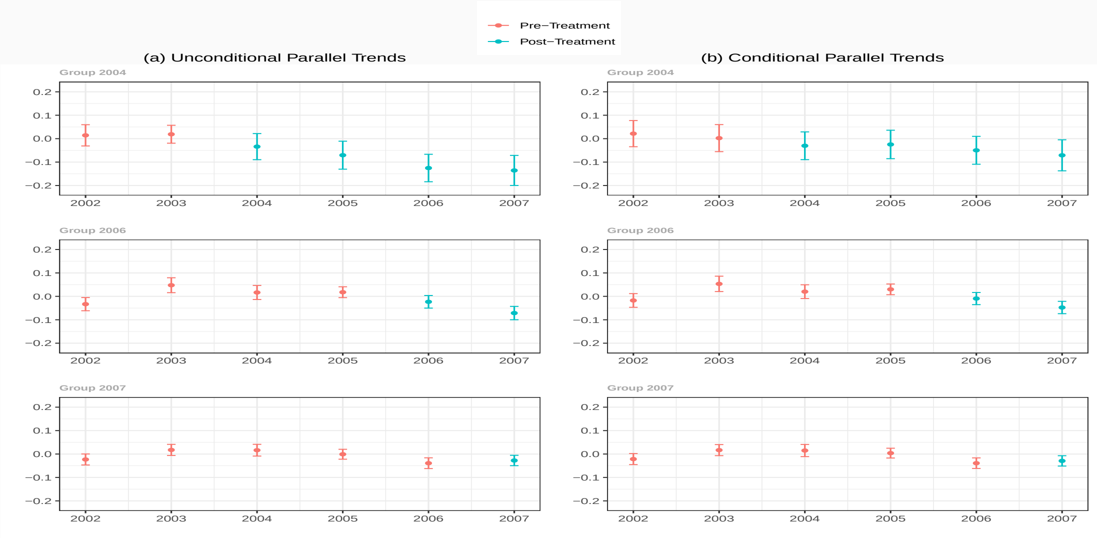
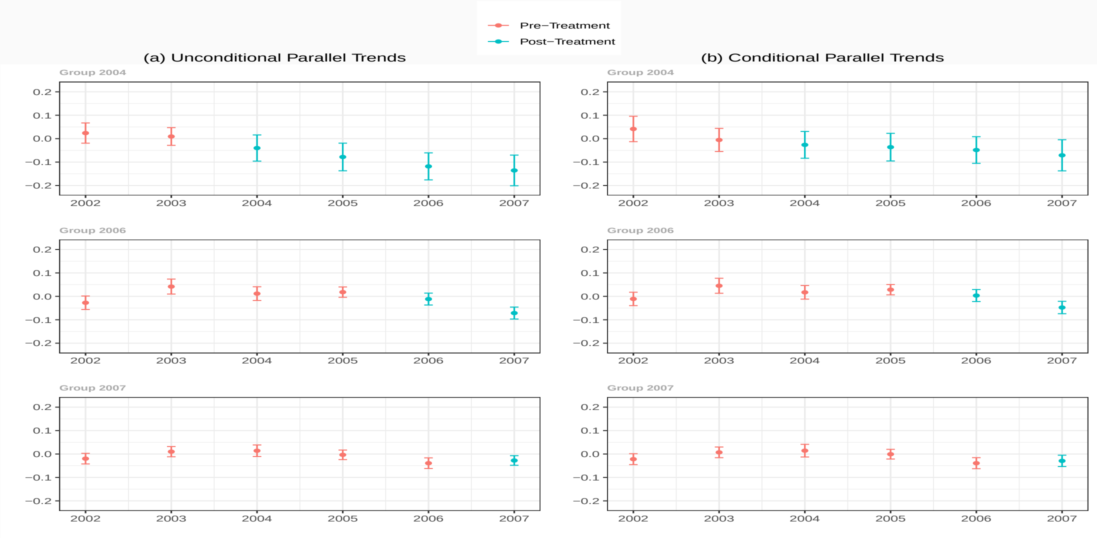


Figure 2: Minimum Wage Results using "not-yet-treated" as comparison groups



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)

Can we relax the common trend assumption?

- **Parallel Trends Assumption:** for all $t = 2, \ldots, \mathcal{T}$, $g = 2, \ldots, \mathcal{T}$, such that $g \leq t$,

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

- **Can we relax it to a inequality to get bounds?**
  - For all $t = 2, \ldots, \mathcal{T}$, $g = 2, \ldots, \mathcal{T}$, such that $g \leq t$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \,|X, G_g = 1\right] \geq \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \,|X, C = 1\right] \ a.s.$$

- This identifying assumption then implies that

$$\mathbb{E}\left[Y_t(0) \,|X, G_g = 1\right] \geq \mathbb{E}\left[Y_{t-1}(0) \,|X, G_g = 1\right] + \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \,|X, C = 1\right] \ a.s.$$

- Then $\widehat{ATT}(g, t)$ could be then interpret as an upper bound.

Conclusion

Conclusion

- We proposed a semi-parametric DiD estimators when there are multiple time-periods and variation in treatment timing.
- We provided valid inference procedures to assess the effectiveness of the policy.
- Applied these tools to revisit the debate about the effect of minimum wage on teen employment