Difference-in-Differences methods

A brief guide to practice

Pedro H. C. Sant'Anna
Emory University

Instacart, September 23 2024

Why DiD is so popular?

Causality with Observational Data: What can I do?

- In many applications, we do not have access to experimental data.
- Without an experiment, we will rely on **observational data**.
- With **observational data**, we have no choice but rely on **assumptions** to talk about causal inference.
- Different methods rely on different assumptions.
- Our job as scientists is to assess the pros and cons of each method in their ability to answer the questions we (and the business) care about.

Causality with Observational Data: What can I do?

- DiD is very popular.
- **WHY?!**
- My guess: data requirements and availability of tools to assess the plausibility of assumptions.
- What are the main alternatives to DiD?

  1. Rely on unconfoundedness and leverage **regression, matching, re-weighting** or **double machine learning**.

     Drawback: Rule out selection on unobservables.
     We need to have data on everything that affects treatment timing and outcome of interest (unconfoundedness assumption).

Causality with Observational Data: What else?

- What are the other main alternative to DiD?

  2. Rely on **Pre-Post analysis**

     Drawback: Does not account for potential trends in outcomes.
     This is more reasonable if we study very short-run effects, but that is not usually the case.

The appeal of Difference-in-Differences

- DiD methods exploit variation in time (before vs. after) and across groups (treated vs. untreated) to recover causal effects of interest.
- **DiD combines previous approaches to avoid their pitfalls**.
- **Advantage: Allow for selection on unobservables and for time-trends.**
  We need to assume that, absent the treatment and conditional on covariates (features), the outcome of interest would grow similarly across groups/cohorts - **parallel trends assumption**.

We need to discuss why Parallel Trends is a plausible assumption in our application.

- **Data Requirements:** We need data from time periods before and after treatment to use DiD (and some periods where no unit is treated).

Structure of Today's Talk

Structure of the Talk

- My main goals for today are to

1. Expose everyone to the two-period DiD setup.

2. Discuss staggered treatment adoption setups

2.1 Problems with Two-Way-Fixed Effects (TWFE) linear regressions
Goodman-Bacon (2021), de Chaisemartin and D'Haultfœuille (2020), Sun and Abraham (2021).

2.2 Simple solutions to these problems
Callaway and Sant'Anna (2021), Sun and Abraham (2021), Wooldridge (2021), Borusyak, Jaravel and Spiess (2024)

Difference-in-Differences Checklist

1. Start plotting the treatment rollout (e.g., use panelView R package)
2. Document how many units are treated in each cohort.
3. Plot the evolution of average outcomes across cohorts.
4. Choose the comparison groups and the PT assumption carefully:
   Who decides treatment? What do they know? What type of selection is allowed?
5. Do event-study analysis without any covariates and assess if PT is plausible.
6. If unconditional PT is not plausible, incorporate covariates into the analysis.
7. When using covariates, check for overlap: If control groups are small, problems with overlap will probably arise. If you are OK with extrapolation, use regression adjustment DiD procedures.
8. Do event-study analysis after adjusting for covariates and assess if conditional PT is plausible.
9. Conduct some sensitivity analysis for violations of PT (e.g., use honestDiD R package).
10. If conditional PT is not plausible, look for other methods.

Let’s start with two-periods DiD

DiD basics

2x2 DiD Setup

Let's start our discussion using the simplest DiD setup known as the 2x2 case

- 2 time periods: $t = 1$ (before treatment) and $t = 2$ (after treatment)
- 2 groups: $G_i = 2$ (units treated at period 2) and $G_i = \infty$ (untreated by period 2)
- Some covariates X may be available.
- A large number of independent observations (or clusters) is available.

2x2 DiD Setup: Potential Outcomes and Target Parameters

- To formalize our causal analysis, we will introduce Potential Outcomes.
- $Y_t(g)$: Potential outcome at period t if units were exposed to treatment for the first time in period g.
- There are many different ways to define "Causal Effects".
- What causal parameter are we after?
- Main parameter of interest:
  Average Treatment Effect among Treated units in period $t = 2$,

$$ATT \equiv \underbrace{\mathbb{E}\left[Y_{t=2}\left(2\right)|G=2\right]}_{\text{estimable from the data}} - \underbrace{\mathbb{E}\left[Y_{t=2}\left(\infty\right)|G=2\right]}_{\text{counterfactual component}}$$

Causal Parameter of interest with multiple periods and multiple groups

- This is a good place to discuss a more general notation of causal treatment parameters when we have multiple periods and treatment groups.

$$ATT(g,t) \equiv \underbrace{\mathbb{E}\left[Y_t\left(g\right)|G=g\right]}_{\text{estimable from the data}} - \underbrace{\mathbb{E}\left[Y_t\left(\infty\right)|G=2\right]}_{\text{counterfactual component}}$$

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

- Effects can vary according to time of adoption g, time period t, and time since adoption $e = t - g$.

Let’s go back to the 2x2 case to solidify things

Assumptions in 2x2 DiD Setups

Assumption (SUTVA - No spillovers or interference)

Observed outcomes at time t are realized as

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

That is, for units who are treated in period $t = 2$, we observe $Y_{i,t}(2)$.

For units who remain untreated in period $t = 2$, we observe $Y_{i,t}(\infty)$.

Assumptions in 2x2 DiD Setups

Assumption (No-Anticipation)

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

That is, units that are treated do not change their behavior before treatment starts in anticipation of what is coming next.

Assumption (Strong Overlap Assumption)

For some $\epsilon > 0$, $\mathbb{P}[G = 2|X] < 1 - \epsilon$ almost surely.

That is, we can find units in the control group that have the same covariate values X as those units in the treatment group.

Intuitively, if you tell me X, I cannot say if that a unit is treated with 100% confidence.

Assumptions in 2x2 DiD Setups

Assumption (Conditional Parallel Trends Assumption)

$$\mathbb{E}\left[Y_{t=2}(\infty)|G=2,X\right]-\mathbb{E}\left[Y_{t=1}(\infty)|G=2,X\right]=\mathbb{E}\left[Y_{t=2}(\infty)|G=\infty,X\right]-\mathbb{E}\left[Y_{t=1}(\infty)|G=\infty,X\right] \quad a.s.$$

That is, in the absence of treatment, within each covariate-strata, the average evolution of outcome Y among units treated in period 2 is the same as the average evolution of outcome Y among units that remain untreated.

Identification without covariates

- Under parallel trends and no anticipation, can show that

$$\tau_{ATT} = \underbrace{(E[Y_{i,t=2}|G_i = 2] - E[Y_{i,t=1}|G_i = 2])}_{\text{Change for treated group}} - \underbrace{(E[Y_{i,t=2}|G_i = \infty] - E[Y_{i,t=1}|G_i = \infty])}_{\text{Change for comparison group}},$$

  a "difference-in-differences" of population means.

- This can be easily estimated by

$$\widehat{ATT}_n^{DiD} = \left(\overline{Y}_{g=2,t=2} - \overline{Y}_{g=2,t=1}\right) - \left(\overline{Y}_{g=\infty,t=2} - \overline{Y}_{g=\infty,t=1}\right),$$

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

“TWFE” specification without covariates

- In practice, most of us would rely on the following regression specification:

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

- With balanced panel data, that above “simpler” regression is equivalent to the TWFE regression

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

- The “simpler” spec also works for unbalanced panels and repeated cross-section data.

- **Inference:** clustered standard errors are valid as the number of clusters is “large”.

What if I want to add covariates?

Being inspired by the recent developments in Causal ML

- In the last 10 years or so, we have been seeing a lot of advances in Causal ML.
  - Belloni, Chernozhukov and Hansen (2014)
  - Farrell (2015)
  - Belloni, Chernozhukov, Fernández-Val and Hansen (2017),
  - Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2017)
  - Athey and Wager (2018)
  - Athey, Tibshirani and Wager (2019)
  - Chernozhukov, Demirer, Duflo and Fernández-Val (2022).

- All these papers propose estimators that are Doubly Robust/Neyman Orthogonal.

- These ideas have been explored in DiD setups only recently;
  see, e.g., Sant'Anna and Zhao (2020); Chang (2020); Callaway, Drukker, Liu and Sant'Anna (2023).

Doubly Robust DiD procedure with Panel

Sant'Anna and Zhao (2020) proposed the following DR DiD estimand:

$$ATT^{dr,p} = \mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{\frac{p(X)(1-D)}{1-p(X)}}{\mathbb{E}\left[\frac{p(X)(1-D)}{1-p(X)}\right]}\right)\left(\Delta Y - m_{\Delta}^{G=\infty}(X)\right)\right],$$

where

$$\Delta Y = Y_{t=2} - Y_{t=1}, \quad D_i = 1\{G_i = 2\}, \quad m_{\Delta}^{G=\infty}(X) = \mathbb{E}[\Delta Y | X, G = \infty], \quad p(X) = \mathbb{P}[G = 2|X].$$

- This is similar to cross-sectional DR formulation but with outcomes measured as "post - pre" instead of "post" (and the focus is on ATT not ATE).

DiD basics

What if we have variations in treatment timing?

What if we have staggered treatment timing?

- What if we have staggered treatment adoption?
- This is, what if units can select into treatment at different points in time?
- In the 2-period case this was not possible, but this is realistic in many applications!

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

Traditional methods: TWFE regressions

- We know that, in the 2x2 case,

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

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

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

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

- $D_{i,t}$ is an indicator for unit i being treated by period t.

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

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

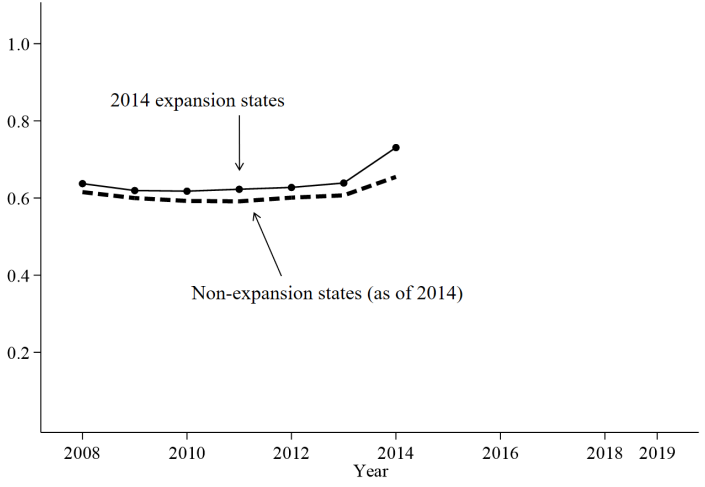
Example: Effect of ACA Medicaid Expansion on Health Insurance rate

Empirical Example: Medicaid Expansion

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

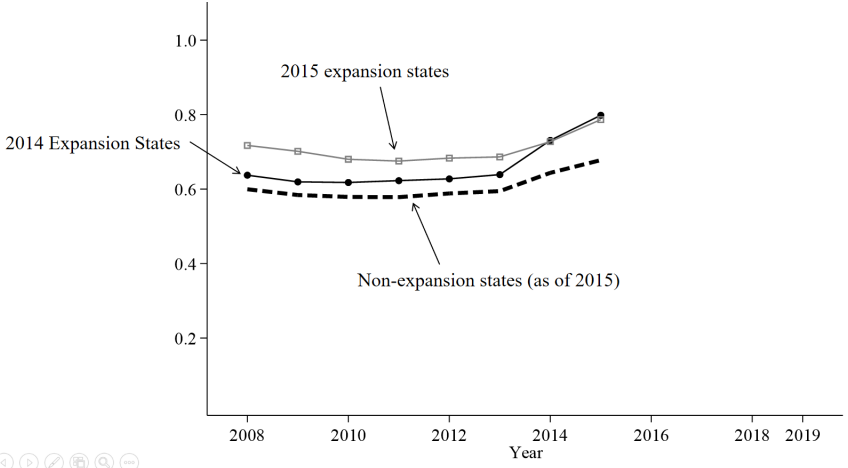
ACA Medicaid Expansion Circa 2014

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



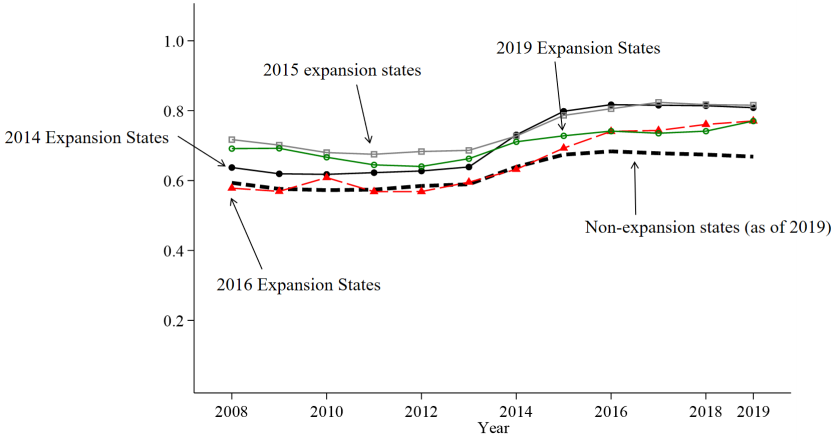
ACA Medicaid Expansion Circa 2015

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



ACA Medicaid Expansion Circa 2019

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



ACA Medicaid Expansion Circa 2019

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

OLS estimate of β

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

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

- What is $\widehat{\beta}$?

- Goodman-Bacon (2021) shows that we can answer this question following these three steps:
  1. Remove unit means

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

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

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

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

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

TWFE computes weighted-averages of 2x2 DiD's

- $\widehat{\beta} = 0.074$ in the empirical application.
- Goodman-Bacon (2021) highlights that:
- **OLS weights use sample size and variance**
- Is that what you really want?

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

TWFE regressions, in general,

do not recover an easy-to-interpret

causal parameter of interest,

unless we rule out TE heterogeneity and dynamics

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

Event-Study via TWFE specifications

Event-Study via TWFE specifications

- One of the main attractive features of observing multiple time periods is that we can attempt to "learn" about treatment effect dynamics.
- Status-quo in the literature is to consider variants of the TWFE event-study regression

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

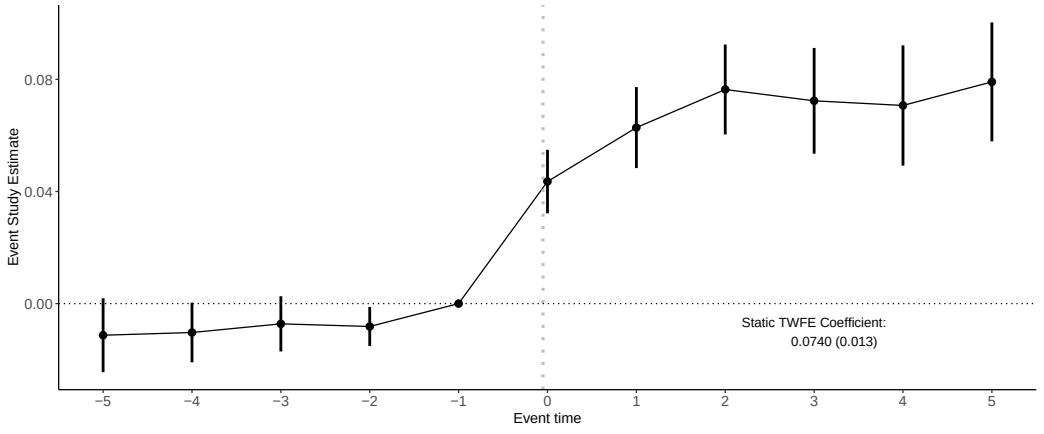
  with the event study dummies $D_{i,t}^k = 1\{t - G_i = k\}$, where G_i indicates the period unit i is first treated (Group).

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

Does this strategy “work”?

ACA Medicaid Expansion: TWFE Event-study specification

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



Event-Study via TWFE specifications

Sun and Abraham (2021)

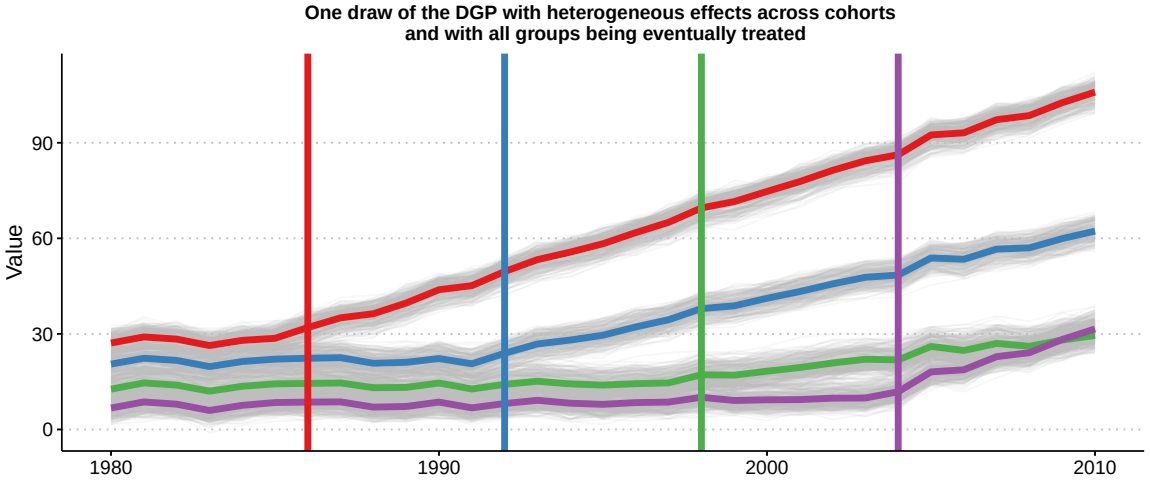
Problem with Event-Study via TWFE specifications: Sun and Abraham (2021)

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

Event-Study via TWFE specifications

Stylized example using simulated data

Stylized example using simulated data



Stylized example using simulated data

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

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

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

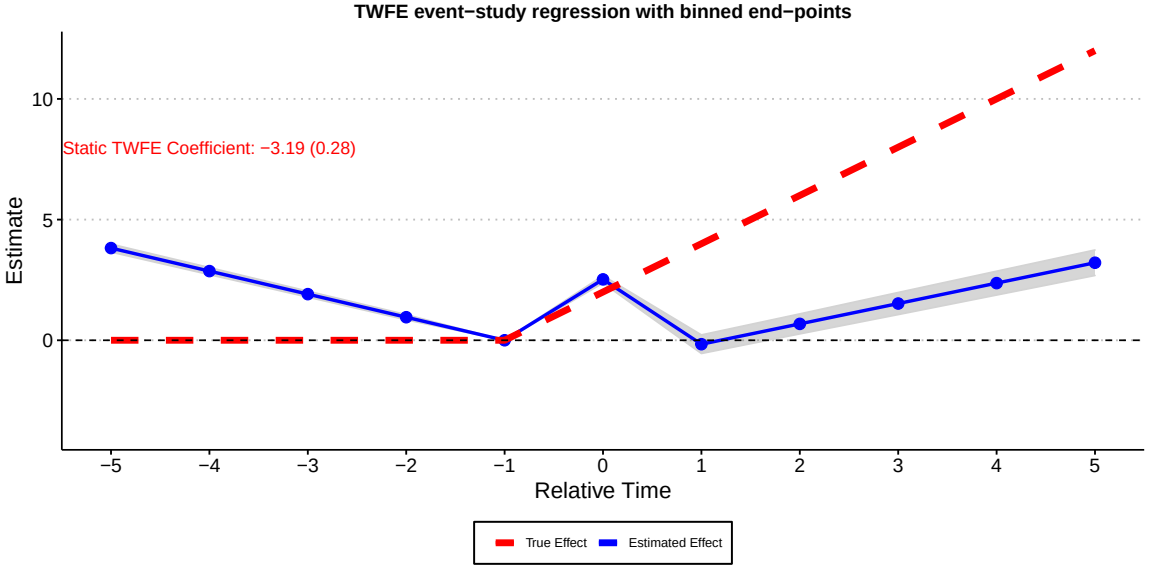
Traditional methods: TWFE event-study regression

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

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

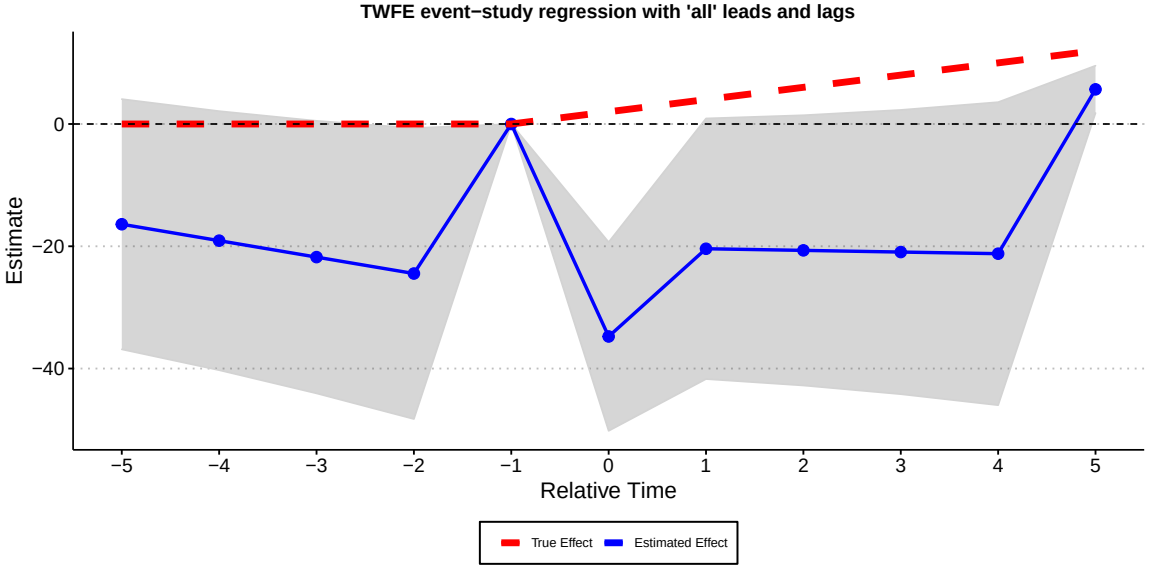
  with K and L to be equal to 5 ?

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

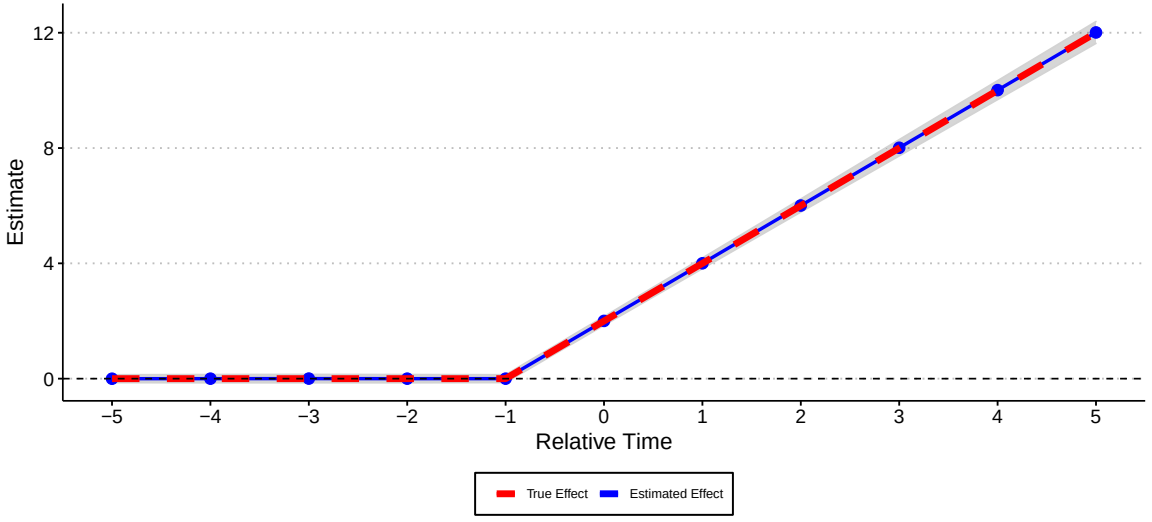

TWFE event−study regression with binned end−points
Static TWFE Coefficient: −3.19 (0.28)
Estimate
10
5
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
True Effect
Estimated Effect

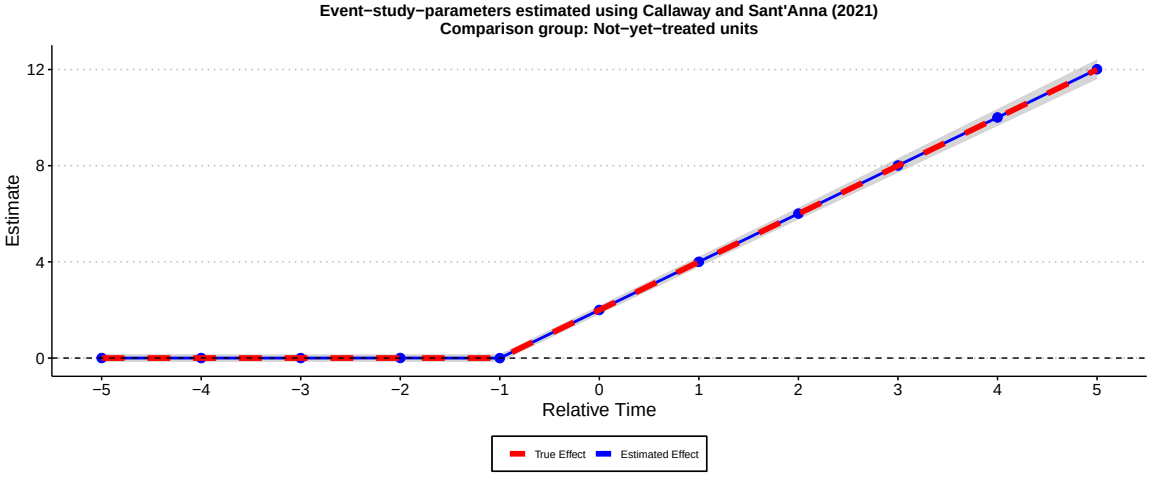
Traditional methods: TWFE event-study regression

- 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 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

Callaway and Sant'Anna (2021)

Main take-way message

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

Identification

Let’s talk about identification

Callaway and Sant'Anna (2021)

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 $ATT(g,t)$

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

- 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* $t \geq g$,

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

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!

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!

Callaway and Sant'Anna (2021)

Aggregation

Second step: Aggregation

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.

Third step: Estimation and Inference

Callaway and Sant'Anna (2021)

Estimation and Inference

Estimation

- Identification results suggest a simple plug-in estimation procedure.
- Replace population expectations with their empirical analogues.
- Callaway and Sant'Anna (2021) allows for covariates and provides high-level conditions that first-step estimators have to satisfy.

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.

- **Solution:** Use bootstrap to do simultaneous inference.

- Details are on the paper (and also on slides available on my webpage).

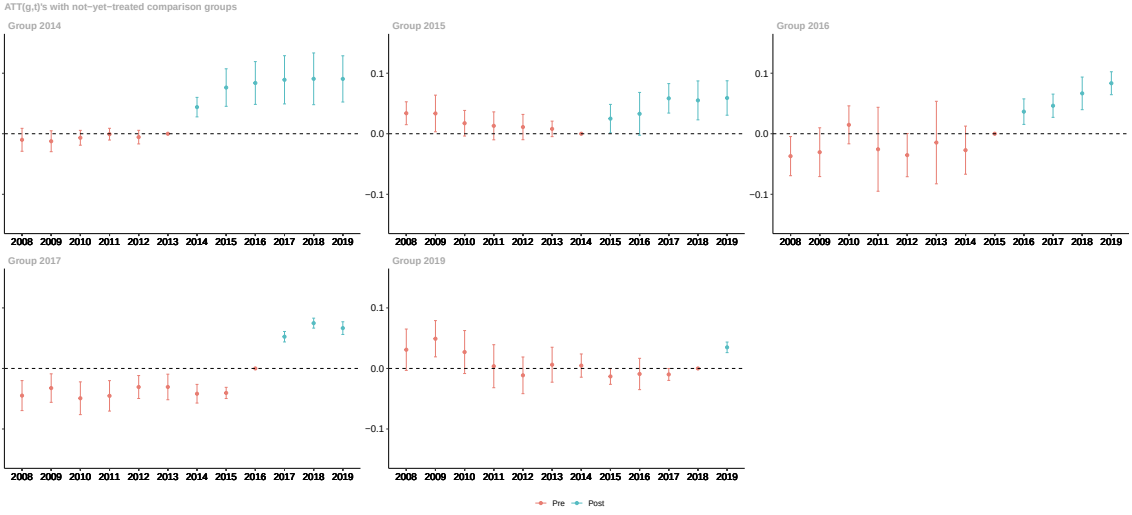
Let’s go back to the ACA Medicaid Expansion Example

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

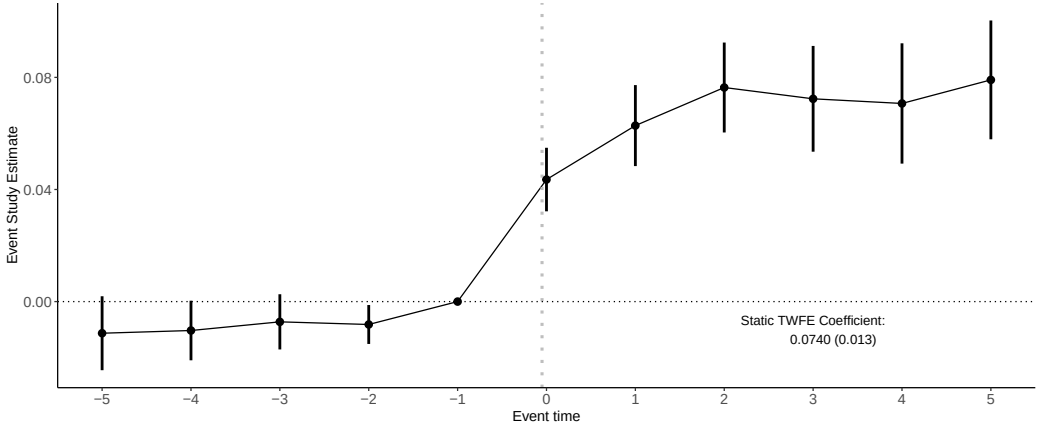
Challenge setup to make inference on ATT(g,t)'s per se

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



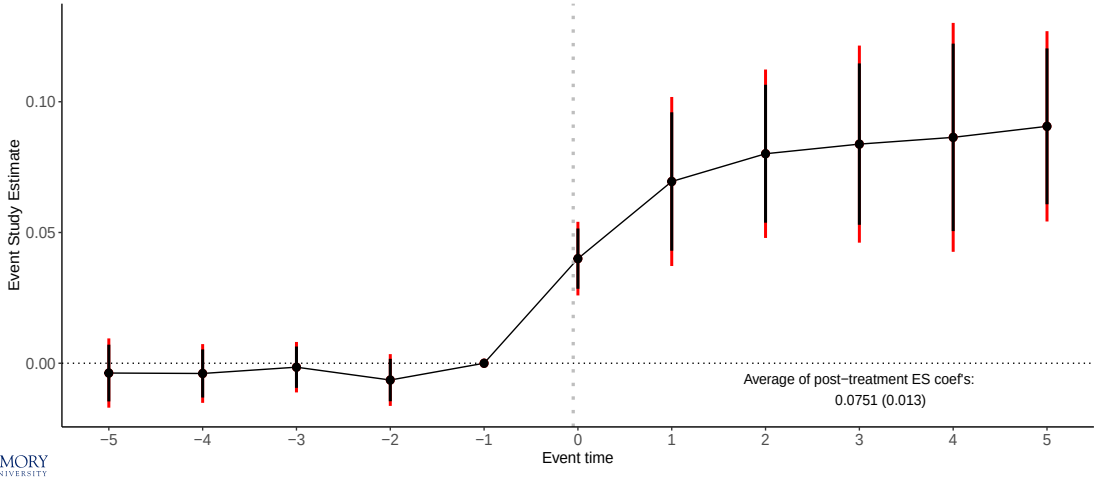
ACA Medicaid Expansion: TWFE Event-study specification

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



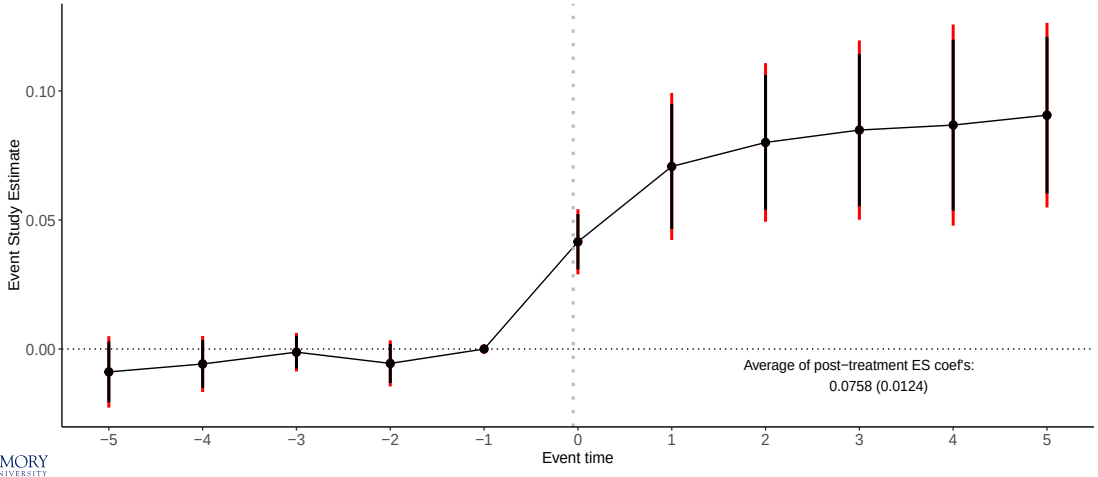
ACA Medicaid Expansion: CS Event-study specification

Figure 6: Results using "never-treated" as a comparison group



ACA Medicaid Expansion: CS Event-study specification

Figure 7: Results using "not-yet-treated" as comparison groups



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.
- The solution to the TWFE problem is simple
  - Separate the identification, aggregation and estimation/inference parts of the problem
- Use $ATT(g,t)$ as building blocks 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.

Difference-in-Differences Checklist

1. Start plotting the treatment rollout (e.g., use panelView R package)
2. Document how many units are treated in each cohort.
3. Plot the evolution of average outcomes across cohorts.
4. Choose the comparison groups and the PT assumption carefully:
   Who decides treatment? What do they know? What type of selection is allowed?
5. Do event-study analysis without any covariates and assess if PT is plausible.
6. If unconditional PT is not plausible, incorporate covariates into the analysis.
7. When using covariates, check for overlap: If control groups are small, problems with overlap will probably arise. If you are OK with extrapolation, use regression adjustment DiD procedures.
8. Do event-study analysis after adjusting for covariates and assess if conditional PT is plausible.
9. Conduct some sensitivity analysis for violations of PT (e.g., use honestDiD R package).
10. If conditional PT is not plausible, look for other methods.

References

Athey, Susan and Stefan Wager, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228 – 1242.

_ , Julie Tibshirani, and Stefan Wager, "Generalized random forests," *The Annals of Statistics*, 2019, *47* (2), 1148 – 1178.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, apr 2014, *81* (2), 608–650.

_ , _ , Iván Fernández-Val, and Christian Hansen, "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 2017, *85* (1), 233–298.

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.

_ , **David Drukker, Di Liu, and Pedro H. C. Sant'Anna**, "Difference-in-Differences via Machine Learning," *Working Paper*, 2023.

Chang, Neng-Chieh, "Double/debiased machine learning for difference-in-differences models," *The Econometrics Journal*, 2020, *23* (2), 177––191.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, jun 2017, pp. 1–71.

_ , **Mert Demirer, Esther Duflo, and Iván Fernández-Val**, "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments ," *arXiv:1712.04802*, 2022.

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

Farrell, Max H., "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 2015, *189* (1), 1–23.

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

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.