Causal Inference using Difference-in-Differences

Lecture 8: Learning about Treatment Effect Dynamics via Event Studies

Pedro H. C. Sant'Anna
Emory University

January 2025

Overview of previous lectures

DiD procedures with covariates

- We can include covariates into DiD to allow for covariate-specific trends.
- Covariates should not be post-treatment variables.
- There are several “correct” ways of implementing conditional DiD:
  - Regression adjustments;
  - Inverse probability weighting;
  - Doubly Robust (augmented inverse probability weighting).
- TWFE, though, can be severely biased (depends on specification!!!)
- DR DiD is my preferred method:
  - More robust against model misspecifications
  - Can be semiparametrically efficient (confidence intervals are tighter)

What if we have multiple time periods?

DiD with multiple time periods

DiD setup with multiple time periods

- So far, we have considered the 2x2 DiD setup:
  - 2 time periods: $t = 1$ (before treatment) and $t = 2$ (after treatment)
  - 2 groups: $G = 2$ (treated at period 2) and $G = \infty$ (untreated by period 2)
- Now, let's consider the case where we have more periods, but treatment can happen at a fixed point in time (so we still have two groups):
  - T time periods: $t = 1, 2, \ldots, T$.
  - Treatment may happen at a given time, say g.
  - Pre-treatment periods: $t = 1, 2, \ldots, g - 1$.
    Post-treatment periods: $t = g, g + 1, \ldots, T$.
  - 2 groups: $G = g$ (treated at period g) and $G = \infty$ (untreated by period T)

What are the parameters of interest?

Parameters of interest

- We need to discuss what we want to learn in this setup.
- Building on the 2x2 setup, it is natural to focus on ATT-type parameters.
- But now, we have multiple post-treatment periods, so we will talk about time (and group) specific ATT's:

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

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

- Here, we have only one $g \neq \infty$, so we will only vary t.

Parameters of interest

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

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

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

- Here, we have only one $g \neq \infty$, so we will only vary t.

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

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

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

- This is just a change of variable, right?!

Parameters of interest

- We can also further aggregate these ATT's across periods:

$$\theta_s^{agg} = \frac{\sum_{e=0}^{s} ATT(g, g+e)}{s+1}, \quad s \geq 0.$$

- If you want to "discount" more distant periods, that is easy, as long as these $w_e's$ are estimable or known:

$$\theta_{s,w}^{agg} = \frac{\sum_{e=0}^{s} w_e \cdot ATT(g, g+e)}{\sum_{e=0}^{s} w_e}, \quad s \geq 0.$$

- Advantage of these aggregations: one-dimensional summary parameters.

Identification

Multi-period DiD setup: Assumptions

Identification of the ATT(g, t)'s is achieved via two main assumptions:
No-Anticipation and Parallel trends (we are taking SUTVA for granted now).

Assumption (No-Anticipation)

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

- The No-Anticipation implies that $ATT(g,t) = 0$ for all pre-treatment periods $t < g$.
- We will play with this assumption later, but let's keep it simple for now.

Multi-period DiD setup: Assumptions

Identification of the ATT(g, t)'s is achieved via two main assumptions:
No-Anticipation and Parallel trends (we are taking SUTVA for granted now).

Assumption (Parallel Trends Assumption)

For all $t \geq g$,

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

The parallel trends (PT) assumption states that, in the absence of treatment, the evolution of the outcomes among the treated units is, on average, the same as the evolution of the outcomes among the untreated units, **in all post-treatment periods**.

How does these PT help us?

Parallel Trends and the ATT(g,t)

When $t = g$, we are essentially in the 2x2 case, so we know that

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

(check slide 28 at Lecture 2)

We also have that

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

Does anything here get your attention?

Let's look at the effect at time $t = g + 1$

Parallel Trends and the ATT(g,g+1)

1) First, recall the PT assumption at time period $t = g + 1$:

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

2) By simple manipulation, we can write it as

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

3) Now, exploiting No-Anticipation and SUTVA:

$$\mathbb{E}\left[Y_{i,g+1}(\infty)|G_i = g\right] = \underbrace{\mathbb{E}\left[Y_{i,g}(\infty)|G_i = g\right]}_{\text{this is post-treatment}} + \underbrace{\left(\mathbb{E}\left[Y_{i,g+1}|G_i = \infty\right] - \mathbb{E}\left[Y_{i,g}|G_i = \infty\right]\right)}_{\text{by SUTVA}}$$

Parallel Trends and the ATT(g,g+1)

3) Now, exploiting No-Anticipation and SUTVA:

$$\mathbb{E}\left[Y_{i,g+1}(\infty)\,|G_i=g\right] = \underbrace{\mathbb{E}\left[Y_{i,g}(\infty)\,|G_i=g\right]}_{\textit{this is post-treatment}} + \underbrace{\left(\mathbb{E}\left[Y_{i,g+1}|G_i=\infty\right]-\mathbb{E}\left[Y_{i,g}|G_i=\infty\right]\right)}_{\textit{by SUTVA}}$$

4) Now, exploring that we have already identified $\mathbb{E}\left[Y_{i,g}(\infty)\,|G_i=g\right]$ in (1):

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

5) Now, simplifying our formula (canceling cross-terms):

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

Parallel Trends and the ATT(g,g+1)

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

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

- This is “the birth” of the “long-difference” approach to DiD!

Parallel Trends and the ATT(g,t),

- Following the same steps as above, we can easily show that, for every $t \geq g$, under our assumptions,

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

Estimation and inference

“Brute force” DiD estimator for the ATT(g,t)

- Canonical DiD Estimator:

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

where $\overline{Y}_{a,b}$ is the sample mean of the outcome Y for units in group a in time period b,

$$\overline{Y}_{a,b} = \frac{1}{N_{a,b}} \sum_{i=1}^{N \cdot T} Y_i 1\{G_i = a\} 1\{T_i = b\},$$

with

$$N_{a,b} = \sum_{i=1}^{N \cdot T} 1\{G_i = a\} 1\{T_i = b\},$$

G_i and T_i are group and time dummy, respectively, and Y_i is the “poolled” outcome data.

“TWFE” DiD estimator for ATT(g,t)

- It is very easy to get this via TWFE regressions.
- First, subset your data to have data only for periods t and $g-1$, for $t \geq g$.
- In this subset of the data, run the TWFE regression using the following linear specification:

$$Y_i = \alpha_0 + \gamma_0 1\{G_i = g\} + \lambda_0 1\{T_i = t\} + \underbrace{\beta_{0,gt}^{twfe}}_{\equiv ATT(g,t)} (1\{G_i = g\} \cdot 1\{T_i = t\}) + \varepsilon_i,$$

  where Y_i is the “poolled” outcome data.

- We can leverage the regression to make (pointwise) inference about the $ATT(g,t)$.

What if we want to include covariates?

Including covariates

Conditional Parallel Trends Assumption

- In order to "relax" the PTA, we can assume that it holds **only after** conditioning on a vector of observed pre-treatment covariates

Assumption (Conditional Parallel Trends Assumption)

For all $t \geq g$,

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

The conditional PT assumption states that, in the absence of treatment, conditional on X, the evolution of the outcome among the treated units is, on average, the same as the evolution of the outcome among the untreated units, **in all post-treatment periods**.

Strong overlap

- When covariates are available, we will introduce an additional assumption stating that every unit has a strictly positive probability of being in the untreated group.

Assumption (Strong Overlap Assumption)

The conditional probability of belonging to the treatment group, given observed characteristics X, is uniformly bounded away from 1.

That is, for some $\epsilon > 0$, $\mathbb{P}[G = g|X] < 1 - \epsilon$ *almost surely.*

- The covariates X here are the same as those used to justify the conditional PT assumption!
- For identification purposes, we can take $\epsilon = 0$. For (standard) inference, though, we would have problems without relying on "extrapolation"; see, e.g., Khan and Tamer (2010).

We now can use all the DiD estimators suitable for 2x2 setups!

Regression adjustment procedure with Panel Data

- Very easy to show that, under our assumptions,

$$ATT(g,t) \quad = \quad \mathbb{E}\left[Y_t - Y_{g-1} | G = g\right] - \mathbb{E}\left[m_{g-1,t}^{G=\infty}(X) \,|G = g\right]$$

- Only have to model one conditional expectation:

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

- Choose your favorite method to estimate this regression (as long as it is "smooth enough")!

Inverse probability weighted estimator with Panel

We can also get (normalized) IPW estimators.

Let $D = 1\{G = g\}$ and $p(X) = \mathbb{E}[D|X]$. Then,

$$
\begin{aligned}
ATT_{std}^{ipw,p}(g,t) &= \mathbb{E}\left[\left(w_{G=g}^{p}(D) - w_{G=\infty}^{p}(D,X;p)\right)(Y_t - Y_{g-1})\right] \\
&= \mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{\frac{p(X)(1-D)}{1-p(X)}}{\mathbb{E}\left[\frac{p(X)(1-D)}{1-p(X)}\right]}\right)(Y_t - Y_{g-1})\right],
\end{aligned}
$$

where

$$
w_{G=g}^{p}(D) = \frac{D}{\mathbb{E}[D]}, \quad \text{and} \quad w_{G=\infty}^{p}(D,X;g) = \frac{g(X)(1-D)}{1-g(X)} \Big/ \mathbb{E}\left[\frac{g(X)(1-D)}{1-g(X)}\right]
$$

Doubly robust DiD procedure with Panel

DR estimator (Sant'Anna and Zhao, 2020; Callaway and Sant'Anna, 2021)

$$
\begin{aligned}
ATT^{dr,p} &= \mathbb{E}\left[\left(w^{p}_{G=g}(D) - w^{p}_{G=\infty}(D,X;p)\right)\left((Y_t - Y_{g-1}) - \left(m_t^{G=\infty}(X) - m_{g-1}^{G=\infty}(X)\right)\right)\right] \\
&= \mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{\frac{p(X)(1-D)}{1-p(X)}}{\mathbb{E}\left[\frac{p(X)(1-D)}{1-p(X)}\right]}\right)\left((Y_t - Y_{g-1}) - \left(m_t^{G=\infty}(X) - m_{g-1}^{G=\infty}(X)\right)\right)\right],
\end{aligned}
$$

where

$$
w^{p}_{G=g}(D) = \frac{D}{\mathbb{E}[D]}, \quad \text{and} \quad w^{p}_{G=\infty}(D,X;g) = \frac{g(X)(1-D)}{1-g(X)} \Big/ \mathbb{E}\left[\frac{g(X)(1-D)}{1-g(X)}\right]
$$

All these are implemented in Stata's csdid package

All these are implemented in R's did package

Remark on pre-treatment period

Remark on pre-treatment periods

- So far, we have focused on the case where $t \geq g$.
- But we can "fix" all the formulas above and also consider cases where $t < g$.
- This may allow us to "pre-test" for the reliability of parallel trends in this context.
- Rationale is: if pre-trends are parallel, post-treatment trends may be more likely to be parallel.
- Very easy to do.

Remark on inference

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 (and everything we have discussed so far in this lecture) ignores the dependence across g and t, and "multiple-testing" problems.

Simultaneous Inference

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

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

where

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

Simultaneous confidence intervals

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

$$\widehat{ATT}^*_{g\leq t} = \widehat{ATT}_{g\leq t} + \mathbb{E}_n\left[V \cdot \widehat{\Psi}_{g\leq t}(\mathcal{W})\right]. \tag{2}$$

Multiplier Bootstrap procedure

1. Draw a realization of $\{V_i\}_{i=1}^n$.
2. Compute $\widehat{ATT}^*_{g\leq t}$ as in (2), 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".

But what do standard TWFE regressions recover?

“TWFE” DiD estimator without subsetting

- So far, we have shown how you can tweak regressions to respect our assumptions from the get-go.
- That involved subsetting the data o have data only for time periods t and $g-1$, for $t \geq g$.
- What if we do not subset the data and use the following TWFE specification?

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

  where $D_{i,t}$ is a treatment dummy if unit i is treated by time t.
- What does β^{twfe} recovers?
- What is the implicit parallel trends assumption here?
- I want you to try to answer this question in the next couple of days!

TWFE with dynamics?

- In practice, it is also common to use a TWFE dynamic regression spec:

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

  with the event study dummies $D_{i,t}^k = 1\{t - G_i = k\}$.

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

Do we know what type of causal effect γ's actually recover?
Do the $\widehat{\gamma}$'s coincide with our "by-hand" estimators for the ATT(g,t)'s?

What is next?

What is next?

- How do we assess the validity of our assumptions?
- What are the dangers of pre-testing?
- Can we allow for anticipation behavior? If so, of what type?
- What if we want to use more information about pre-treatment periods?
- **All this, next class!**

References

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

Khan, Shakeeb and Elie Tamer, "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 2010, *78* (6), 2021–2042.

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