Causal Inference using Difference-in-Differences

Lecture 9: TWFE with multiple periods

Pedro H. C. Sant'Anna
Emory University

January 2025

Summary of previous lectures

DiD procedures with multiple periods

- We have now covered the more fun and slightly more complex setup with multiple periods and two groups.
- We discuss how we can learn about treatment effect dynamics: $ATT(g,t)$'s
- We maintained the No-anticipation assumption.
- We extend the Parallel Trends assumption to hold for **all post-treatment** time periods.
  - Implication: Long-run effects are "harder" to learn than short-run effects
- Estimation: Subset the data to look like a 2x2 setup.
- Inference: Make sure you account for multiple testing (rely on simultaneous confidence bands).

Parameters of interest

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

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

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

- 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 periods after ($e \geq 0$) / before ($e < 0$) treatment started.

- These allow us to talk about dynamics!

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 $t < g$.

Assumption (Parallel Trends Assumption)

For all $t \geq g$,

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

Identification result for ATT(g,t)'s

- We have shown that, under our assumptions, for every $t \geq g$,

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

“TWFE” DiD estimator for ATT(g,t)

- First, subset your data to have data only for time periods t and $g-1$, for $t \geq g$.
- In this subset of the data, run the TWFE regression using the following spec:

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

  where Y_i is the “poolled” outcome data.
- We can leverage the regression to make (pointwise) inference about the $ATT(g,t)$ (But be careful with the problem with multiple testing).
- Better to use simultaneous confidence intervals to avoid the multiple-testing issues (already implemented in did R package and csdid Stata pacakge).

“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.

What if we do not subset the data?

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 to have data only for 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.
- I have two questions:
  1. What does β^{twfe} recovers?
  2. What is the implicit parallel trends assumption here?

"TWFE" DiD estimator without subsetting

- The answer to the first question is not very hard to get.
- The key is to realize that we can replace unit and time FE with group and post-treatment dummies:

$$Y_j = \alpha_0 + \gamma_0 1\{G_j = 2\} + \lambda_0 1\{Time_j \geq g\} + \beta^{twfe} D_j + \varepsilon_j,$$

  where now we have pooled all the data into the "long format" (so each unit j is an (i,t)-pair).
- By making simple comparisons of means, we have:

$$\beta^{twfe} = (\mathbb{E}[Y|G=g, t \geq g] - \mathbb{E}[Y|G=g, t < g]) - (\mathbb{E}[Y|G=\infty, t \geq g] - \mathbb{E}[Y|G=\infty, t < g])$$

"TWFE" DiD estimator without subsetting: Assumptions

$$\beta^{twfe} = \left(\mathbb{E}[Y_j|G=g, t \geq g] - \mathbb{E}[Y_j|G=g, t < g]\right) - \left(\mathbb{E}[Y_j|G=\infty, t \geq g] - \mathbb{E}[Y_j|G=\infty, t < g]\right)$$

- Some remarks on (implicit) assumptions:
  - β^{twfe} implicly uses all available pre-treatment periods.
  - So far, we have assumed parallel trends only for post-treatment periods, $t \geq g$.
  - Thus, at least implicitly, the TWFE regressions rely on a "different" type of PT assumption!
  - A PT version compatible with TWFE and "DiD-by-hand" is that PT holds for all time periods (both pre-and post-treatment).
  - But what about the "PT only restricts post-treatment potential outcomes " type of folk wisdon?!

"TWFE" DiD estimator without subsetting: Interpretation

$$\beta^{twfe} = \left(\mathbb{E}[Y_j|G=g, t \geq g] - \mathbb{E}[Y_j|G=g, t < g]\right) - \left(\mathbb{E}[Y_j|G=\infty, t \geq g] - \mathbb{E}[Y_j|G=\infty, t < g]\right)$$

- Some remarks on interpretation:
  - What type of summary parameter does β^{twfe} represent when we no-anticipation and PT for all time periods hold?
  - Under these stronger assumptions, we can show that

$$\beta^{twfe} = \frac{\sum_{s=g}^{T} ATT(g,s)}{T-g+1} = \frac{\sum_{e=0}^{T-g} ATT(g,g+e)}{T-g+1}.$$

  - **Problem Set Question:** Prove the above result.

What if we include leads and lags?

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?

TWFE with dynamics

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

- When one fully saturates the model, i.e., include all possible treatment leads and lags, all the γ's should coincide with the "DiD-by-hand" estimators for the $ATT(g, g+e)$'s.
- Intuition: model is fully nonparametric (under our original assumptions), with no over-identifying restrictions.
- Now, if you "bin" the endpoints, you are implicitly changing the modeling assumptions and are imposing additional restrictions.
- In these latter cases, results should not be numerically the same.

TWFE with dynamics

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

- **Recommendation:** when you care about some event-times but not others, estimate them all and report the ones you care!
- This is usually more transparent.

What if we want to leverage more pre-treatment periods?

TWFE with dynamics

- The TWFE specification with all leads and all lags is given by

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

- Since γ_e^{lags} is equivalent to $ATT(g, g+e)$, for $e \geq 0$, we know that this specification is using data from period $t = g - 1$ as baseline.

- If we were willing to accept that PT hold in all periods, this specification would not use all the pre-treatment information to estimate the $ATT(g, g+e)$s parameters.

- How can we modify the above TWFE ES specification to leverage more pre-treatment data to estimate the parameters of interest?

Modified TWFE with dynamics

- This is the idea behind Borusyak, Jaravel and Spiess (2024), Wooldridge (2021) and Gardner (2021)!

- They (implicitly) considered the modified specification

$$Y_{i,t} = \alpha_i + \alpha_t + \sum_{k=0}^{L} \widetilde{\gamma}_k^{lags} D_{i,t}^{k} + \gamma_k^{L+} D_{i,t}^{>L} + \varepsilon_{i,t}.$$

- Now, OLS estimators of $\widetilde{\gamma}_k^{lags}$ are not equivalent to $\widehat{ATT}(g,t)$.

- The difference is that $\widetilde{\gamma}_k^{lags}$ (implicit) uses the average of all pre-treatment outcomes as the baseline period.

EMORY UNIVERSITY

Modified TWFE with dynamics

- More precisely, you should be able to show that, for $e \geq 0$,

$$\begin{aligned}
\tilde{\gamma}_e^{lags} &= \frac{1}{g-1}\sum_{t'=1}^{g-1} \mathbb{E}\left[Y_{g+e} - Y_{t'} | G = g\right] - \mathbb{E}\left[Y_{g+e} - Y_{t'} | G = \infty\right] \\
&= \left(\mathbb{E}\left[Y_{g+e} | G = g\right] - \mathbb{E}\left[\frac{1}{g-1}\sum_{t'=1}^{g-1} Y_{t'} \middle| G = g\right]\right) \\
&\qquad - \left(\mathbb{E}\left[Y_{g+e} | G = \infty\right] - \mathbb{E}\left[\frac{1}{g-1}\sum_{t'=1}^{g-1} Y_{t'} \middle| G = \infty\right]\right).
\end{aligned}$$

- Thus, when PT hold in all periods (both pre and post-treatment), as well as the other identification assumptions hold,

$$\tilde{\gamma}_e^{lags} = ATT(g, g+e).$$

- Are OLS estimators of $\tilde{\gamma}_e^{lags}$ more efficient than $\widehat{ATT}(g,t)$, as it uses more data?

Comparing across TWFE ES specifications

Comparing across TWFE ES specifications

- As formalized by Chen, Sant'Anna and Xie (2024), when PT hold in all periods, the DiD model is nonparametrically over-identified.
- This has interesting consequences for comparing different specifications, as their estimators have different (asymptotic) efficiency properties.
- When one is willing to impose strong assumptions on treatment effect heterogeneity (by assuming homoskedasticity) and serial correlation (by assuming "error terms" are independent over time), Borusyak et al. (2024) and Wooldridge (2021) have shown that OLS estimators for $\widetilde{\gamma}_e^{lags}$ are asymptotic efficient (in a Gauss-Markov sense).
- However, these conditions are not realistic in most applications: if we believe errors were uncorrelated, we would never cluster our standard errors.
- Without these strong conditions, it is generally not possible to rank OLS estimators for $\widetilde{\gamma}_e^{lags}$ and $\widehat{ATT}(g,t)$ in terms of the length of confidence intervals (precision).

Comparing across TWFE ES specifications

- Chen et al. (2024) discuss how one can fully leverage the empirical content of PT holding in all periods to form estimators for $ATT(g, g+e)$ that are asymptotically efficient.
- Their proposed efficient estimators weigh observations from different pre-treatment periods differently, so it explores the correlation structure of the outcome evolution between the treatment and comparison groups.
- Their estimator does not make strong assumptions about spherical error terms (homoskedastic and zero serial correlation) or impose additional time-series restrictions beyond those used in the identification assumptions.
- Their estimator dominates the other available estimators regarding asymptotic efficiency.

What if I also want to add covariates into my TWFE ES?

TWFE ES specifications with covariates

- All the discussion so far focused on DiD and TWFE specifications without covariates.
- Although it is always easy to linearly add covariates into a TWFE specification, it is not easy to guarantee that the OLS coefficients from these specifications recover meaningful average treatment effects of interest.
- Indeed, all the discussion and the equivalences we have established in this lecture are only valid in setups without covariates.
- If you want to add covariates, I strongly recommend using an alternative estimation procedure.
- See Caetano and Callaway (2023) for a more through discussion.

References

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

Caetano, Carolina and Brantly Callaway, "Difference-in-Differences with Time-Varying Covariates in the Parallel Trends Assumption," 2023. arXiv:2202.02903.

Chen, Xiaohong, Pedro H. C. Sant'Anna, and Haitian Xie, "Efficient Difference-in-Differences and Event Study Estimators," *Working Paper*, 2024.

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

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