Difference-in-Differences III: Staggered DiD

Lecture 5 - Introduction to Causal Inference

Kevin Li

Staggered Treatment

So far, we have assumed that all units in the treated group $G = 1$ start to get treated in the same time period $t = 0$.

But this is often not the case. For example, in the United States, individual states often adopt policies at different times.

Staggered DiD allows for units to adopt treatment at different times.

- Perhaps one **cohort**/group of individuals start treatment in $t = 0$. Another cohort/group starts in $t = 2$. And so on...

We should adjust the relative time variables R for each cohort, such that the initial treatment year is always $R = 0$.

TWFE Decomposition

Goodman-Bacon (2021) proves that TWFE in staggered DiD is actually a weighted average of different "comparisons" between different cohorts/groups:

$$\hat{\tau}_{\text{TWFE}} = a_1\beta_1 + a_2\beta_2 + a_3\beta_3 + \cdots + a_k\beta_k$$

Where a are weights and β are comparison estimates.

Goodman-Bacon finds that you can summarise these comparisons into 3 types to get this:

$$\hat{\tau}_{\text{TWFE}} = a_1\beta_{\text{early vs. late}} + a_2\beta_{\text{late vs. early}} + a_3\beta_{\text{treat vs. control}}$$

Issue with TWFE: Forbidden Comparison

The three types of comparions that Goodman-Bacon finds are:

1. Earlier treated units who are treated, compared with later treated units, who have yet to be treated.
2. Later treated units who are treated, compared with earlier treated units who are treated.
3. Treated units vs. units in the never-treated group.

Look at comparison 2: it is actually comparing units that are treated (the later-treated ones) with units that are **already treated** (earlier treated units).

But in DiD, we only compare treated units to control, or control units to control (for trends). So comparison 2 is a **forbidden comparison** that should not be in DiD, but TWFE includes it.

Issue with TWFE: Negative Weighting

The comparisons in TWFE are weighted together with weights $a_1, a_2, ...$

Logically, these comparison weights should be determined by how many observations are in each comparison. In TWFE however, comparison weights are determined by treatment timing.

- Earlier treated units and later treated units get very small (sometimes negative) weights.
- Units treated in the middle get the largest weights.

Did you see the issue: **negative weights** are possible. Under no circumstance does this make any sense. Thus, TWFE can be biased.

Solution: Matching and Reweighting

So there are two problems with TWFE in staggered DiD: forbidden comparisons, and nonsensical weighting.

How do we solve this? By matching and reweighting:

1. We first "match" the proper comparisons, ensuring no forbidden comparisons occur.
2. The estimates of these comparisons are then properly weighted by the number of observations in each comparison.

Three "modern" DiD estimators do this:

- Interaction-Weighted (Sun and Abraham 2021)
- Doubly-Robust (Callaway and Sant'Anna 2021)
- DIDMultiple (De Chaisemartin and D'Haultfœuille 2024)

Interaction-Weighted (Sun and Abraham 2021)

The interaction-weighted (IW) estimator first “matches” the correct comparisons by including interactions in TWFE:

$$Y_{itgr} = \hat{\alpha}_i + \hat{\delta}_t + \sum_{g} \sum_{r \neq -1} 1\{G_{ig} = g\} \cdot 1\{R_{itgr} = r\} \cdot \hat{\tau}_{g,r} + \hat{\varepsilon}_{it}$$

- $\sum_r$ is the same as dynamic treatment effects in TWFE.
- $\sum_g$ tells us to do \$\sum_r \$ for all different cohorts/groups g.
- Basically, we are estimating dynamic treatment effects for each cohort separately.

These numerous $\hat{\tau}_{g,r}$ are aggregated into either a singular τ_{ATT}, or dynamic treatment effects across groups.

Doubly-Robust (Callaway and Sant'Anna 2021)

The Doubly-Robust estimator does a very similar matching process of running dynamic treatment effects for each cohort group separately.

However, instead of relying solely on regression, Doubly-Robust relies on both interacted regression and inverse probability weighting (not important to know what this is).

Then, these comparisons are aggregated together into either a singular τ_{ATT}, or dynamic treatment effects.

Since inverse probability weighting is non-parametric (i.e. it does not assume a linear relationship between confounders X and outcome Y), the Doubly-Robust estimator can handle conditional parallel trends more flexibly.

Imputation Estimators

An alternative approach other than matching and reweighting to solve the issues with TWFE is **imputation**.

Recall our causal inference problem in DiD: we cannot observe counterfactual $Y_{it}(0)$ for treated units in post-treatment periods.

Why don't we estimate it for every treated unit? This is called imputation. Several estimators use this method:

- 2-Stage DiD (Gardner 2021).
- DiD Imputation (Borusyak, Jaravel, and Spiess 2024).
- FEct (Liu, Xu, and Wang 2024).

Estimating Counterfactuals

The TWFE model looks like this:

$$Y_{it} = \hat{\alpha}_i + \hat{\delta}_t + D_{it}\hat{\tau} + \hat{\varepsilon}_{it}$$

If we plug in $D_{it} = 0$, then we can estimate the value of Y_{it} if a unit had no treatment, which is the missing $Y_{it}(0)$:

$$\hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\delta}_t$$

Imputation estimators use untreated observations $D_{it} = 0$ to estimate $\hat{\alpha}_i$ and $\hat{\delta}_t$. We can also include covariates.

Then, they use the above equation to predict the missing counterfactual $Y_{it}(0)$ for treated units, allowing us to directly calculate treatment effects.