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

Issues with TWFE

We can still use TWFE in the same way as generalised DiD. However, TWFE has two issues which make it **biased (bad)** for estimating causal effects in staggered DiD.

- DiD involves comparing treated to untreated units, or untreated to untreated units (for trends).
- Goodman-Bacon (2021) finds that TWFE's $\hat{\tau}_{\text{ATT}}$ in staggered DiD also compares **treated** units from earlier cohorts with **treated** units from later cohorts. (Comparing treated to treated).
- This comparison of treated to treated should not be in DiD. Thus, we consider this a "**forbidden comparison**" that can cause bias in our TWFE $\hat{\tau}_{\text{ATT}}$ estimate.

Issue with TWFE: Negative Weighting

In Staggered DiD, we are essentially running a bunch of smaller generalised DiD's for each cohort, and then combining them together into one causal estimate.

- The weight of each cohort DiD should be based on how large that cohort is (how many observations are in the cohort).
- But TWFE does not weight like this. It weights based on initial treatment period - earlier and later treated cohorts are given less (sometimes negative) weights.

This weighting (especially negative weighting) makes no sense, and makes our TWFE $\hat{\tau}_{\mathsf{ATT}}$ estimates incorrect.

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 estimator first "matches" the correct comparisons by including interactions in TWFE:

$$\hat{Y}_{it} = \underbrace{\hat{\alpha}_i + \hat{\gamma}_t}_{\text{fixed effects}} + \sum_g \sum_{r \neq -1} I_{itgr} \hat{\tau}_{g,r} + \mathbf{X}_{it}^\top \hat{\boldsymbol{\beta}}$$

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

These numerous $\hat{\tau}_{g,r}$ for each group are then aggregated together 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.

DIDmultiple (De Chaisemartin and D'Haultfœuille 2024)

DIDmultiple is an estimator that focuses on **switchers** - those units who change their treatment status between two time periods.

The estimator compares the change Δ in Y between switchers and non-switchers in that specific two-time period window.

$$\tau_t = \mathbb{E}[\Delta Y|\text{switchers}] - \mathbb{E}[\Delta Y|\text{non-switchers}]$$

These τ_t are the properly weighted together for a singular ATT or dynamic treatment effects.

The advantage of focusing on switchers: switchers can be generalised to **continuous** treatment variables D_{it}, making this estimator very versatile.

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:

$$\hat{Y}_{it} = \hat{\alpha}_i + \hat{\gamma}_t + D_{it}\hat{\tau}_{\mathsf{ATT}} + \mathbf{X}_{it}^\top \hat{\boldsymbol{\beta}}$$

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 $Y_{it}(0)$:

$$\hat{Y}_{it}(0) = \hat{\alpha}_i + \hat{\gamma}_t + \mathbf{X}_{it}^\top \hat{\boldsymbol{\beta}}$$

Imputation estimators use untreated observations $D_{it} = 0$ to estimate $\hat{\alpha}_i$, $\hat{\gamma}_t$, and $\hat{\boldsymbol{\beta}}$.

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

Matching + Reweighting or Imputation

Which type of modern DiD estimator should we use in staggered DiD?

- Matching + Reweighting is more established in the econometrics literature (especially Doubly-Robust, the most common method).
- Imputation estimators are better for checking pre-treatment parallel trends. They also allow for a more granular look at causal effects (since we can estimate individual causal effects - but these can be noisy).
- It is quite common to run all of them, and report all of the results.