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 of individuals start treatment in $t = 2$.
- And maybe many more cohorts of inidividuals start treatment at different times.

Cohorts and Relative Time

A **cohort** is a group of units that begin treatment at the same time. Often, there will be a variable in the dataset called cohort or first.treat, indicating the initial treatment time period of a unit.

Relative time R is a new type of variable - that indicates the time relative to initial treatment adoption for each cohort.

- No matter which time period a cohort begins treatment, the relative time is always $R = 0$.
- Pre-treatment periods are always negative $R \leq -1$.
- Post-treatment periods are always positive $R \geq 0$.

TWFE and Staggered Treatment

Two-way fixed effects is still possible in staggered treatment. The individual τ_{ATT} regression is still the same:

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

The dynamic treatment effects are now in terms of relative treatment periods (see previous slide):

$$\hat{Y}_{it} = \underbrace{\hat{\alpha}_i + \hat{\gamma}_t}_{\text{fixed effects}} + R_{it}\hat{\tau}_t + \mathbf{X}_{it}^{\top}\hat{\boldsymbol{\beta}}$$

- Relative time R_{it} is a **categorical** variable with a reference category of -1.

Issue with TWFE: Forbidden Comparisons

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). For staggered DiD, we should be making many of these comparisons - then finding the weighted average of these comparisons.
- Goodman-Bacon (2021) finds that TWFE's $\hat{\tau}_{\text{ATT}}$ in staggered treatment 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.

Thus, our causal estimate is a weighted average of all these smaller cohort DiD's.

- Logically, 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 $\hat{\tau}_{\text{ATT}}$ 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 interaction in the dynamic treatment effect TWFE regression:

$$\hat{Y}_{it} = \underbrace{\hat{\alpha}_i + \hat{\gamma}_t}_{\text{fixed effects}} + G_i R_{it} \hat{\tau}_{t,r} + \mathbf{X}_{it}^\top \hat{\boldsymbol{\beta}}$$

- G_i is a **categorical** variable indicating the cohort in which individual i belongs to.
- The interaction of G_i with relative time R_{it} means that there is a $\hat{\tau}_{t,r}$ estimate for every relative time period for every cohort.

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

Doubly-Robust (Callaway and Sant'Anna 2021)

The Doubly-Robust estimator does a very similar matching process to Interaction-Weighted to avoid forbidden comparisons.

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, IFEct, and MC (Liu, Xu, and Wang 2024).

(IFEct and MC are slightly more advanced and different).

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.