Advanced Microeconometrics 6SSPP393

Topic 4: Advanced Difference-in-Differences

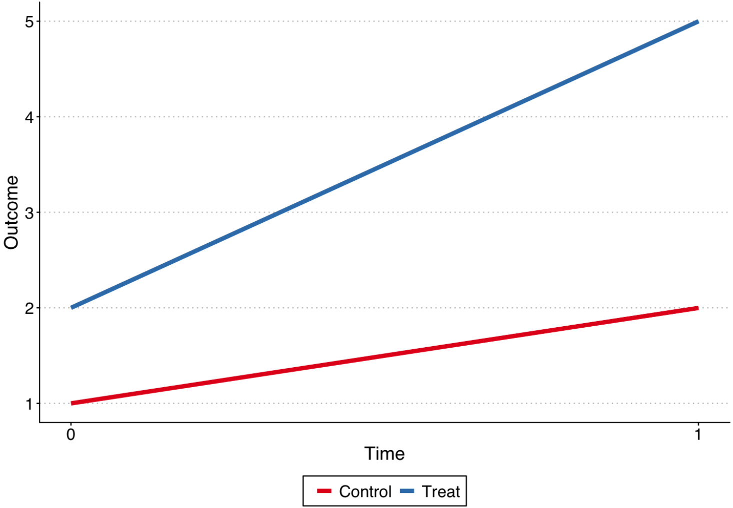
Yonatan Berman

Semester 2, 2023/24

Difference-in-Differences

- Difference-in-differences is one of the key designs for empirical research
- The basic approach is comparing 2 unit with 2 time periods
- One unit is treated, the other isn't; One time period before the treatment, the other after
- Key example – Card and Krueger: Minimum wage increase in New Jersey:
  - 1 unit (T) is treated, and receives treatment in the second period. The control unit (C) is never treated

Difference-in-Differences



Difference-in-Differences

- We can think of a simple 2x2 DiD as a fixed effects estimator
- Potential Outcomes
  - Y_{it}^1 – value of dependent variable for unit i in period t with treatment
  - Y_{it}^0 – value of dependent variable for unit i in period t without treatment
- The expected outcome is a linear function of unit and time fixed effects:

$$
\begin{aligned}
E\left[Y_{it}^0\right] &= \alpha_i + \alpha_t \\
E\left[Y_{it}^1\right] &= \alpha_i + \alpha_t + \delta D_{st}
\end{aligned}
$$

- Goal of DiD is to get an unbiased estimate of the treatment effect δ

Difference-in-Differences

- Difference in expectations for the *control* unit times t = 1 and t = 0:

$$
\begin{aligned}
E[Y^0_{C,1}] &= \alpha_1 + \alpha_C \\
E[Y^0_{C,0}] &= \alpha_0 + \alpha_C \\
E[Y^0_{C,1}] - E[Y^0_{C,0}] &= \alpha_1 - \alpha_0
\end{aligned}
$$

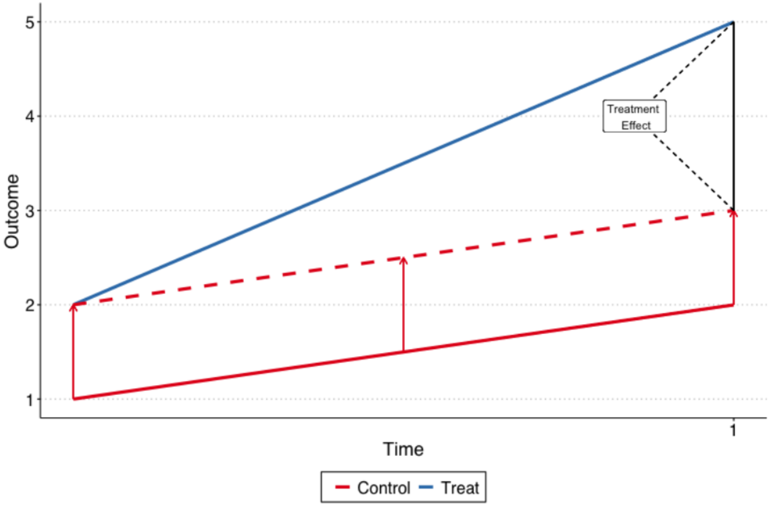
- Now do the same thing for the *treated* unit:

$$
\begin{aligned}
E[Y^1_{T,1}] &= \alpha_1 + \alpha_T + \delta \\
E[Y^1_{T,0}] &= \alpha_0 + \alpha_T \\
E[Y^1_{T,1}] - E[Y^1_{T,0}] &= \alpha_1 - \alpha_0 + \delta
\end{aligned}
$$

- If we assume the linear structure of DiD, then unbiased estimate of δ is:

$$
\delta = \left(E[Y^1_{T,1}] - E[Y^1_{T,0}]\right) - \left(E[Y^0_{C,1}] - E[Y^0_{C,0}]\right)
$$

Difference-in-Differences



Difference-in-Differences

The DiD can be estimated through linear regression of the form:

$$y_{it} = \alpha + \beta_1 TREAT_i + \beta_2 POST_t + \delta(TREAT_i \cdot POST_t) + \epsilon_{it} \tag{1}$$

The coefficients from the regression estimate in (1) recover the same parameters as the double-differencing performed above:

$$\begin{aligned}
\alpha &= E[y_{it}|i = C, t = 0] = \alpha_0 + \alpha_C \\
\beta_1 &= E[y_{it}|i = T, t = 0] - E[y_{it}|i = C, t = 0] \\
&= (\alpha_0 + \alpha_T) - (\alpha_0 + \alpha_C) = \alpha_T - \alpha_C \\
\beta_2 &= E[y_{it}|i = C, t = 1] - E[y_{it}|i = C, t = 0] \\
&= (\alpha_1 + \alpha_C) - (\alpha_0 + \alpha_C) = \alpha_1 - \alpha_0 \\
\delta &= (E[y_{it}|i = T, t = 1] - E[y_{it}|i = T, t = 0]) - \\
&\quad (E[y_{it}|i = C, t = 1] - E[y_{it}|i = Ct = 0]) = \delta
\end{aligned}$$

Difference-in-Differences

- Advantage of regression DiD - it provides both estimates of δ and standard errors for the estimates.
- Angrist & Pischke (2008):
  - "It's also easy to add additional (units) or periods to the regression setup... [and] it's easy to add additional covariates."
- Two-way fixed effects estimator:

$$y_{it} = \alpha_i + \alpha_t + \delta^{DD} D_{it} + \epsilon_{it}$$

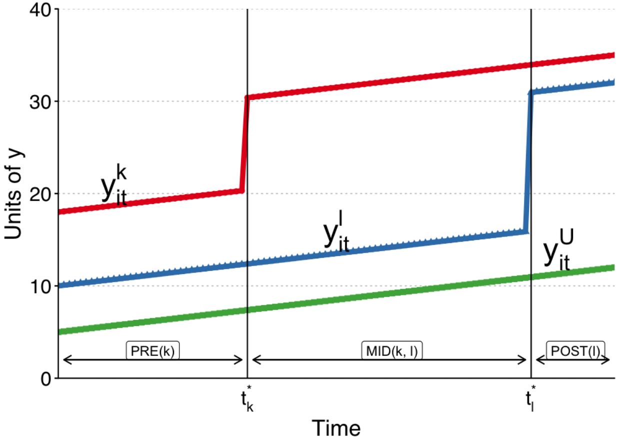
  - α_i and α_t are unit and time fixed effects, D_{it} is the unit-time indicator for treatment.
  - $TREAT_i$ and $POST_t$ now subsumed by the fixed effects.
  - can be easily modified to include covariate matrix X_{it}, time trends, dynamic treatment effects estimation, etc.

Difference-in-Differences – so where do things go wrong?

- Recent literature, starting around 2018 observed important issues with the standard TWFE (two-way fixed effects) DiD
- The problem arises in the context of *staggered treatment timing*
- This means that different units receive treatment at different periods in time
- This has become the most common use of DiD, as it allows to consider more realistically real-world situations, compared with the standard 'pre' and 'post' setting
- So why is there a problem?
  - With staggered treatment δ is a weighted average of different treatment effects
  - Standard TWFE can lead to biased results due to the weighted average creating various issues
  - This turns out to have a large effect in practice

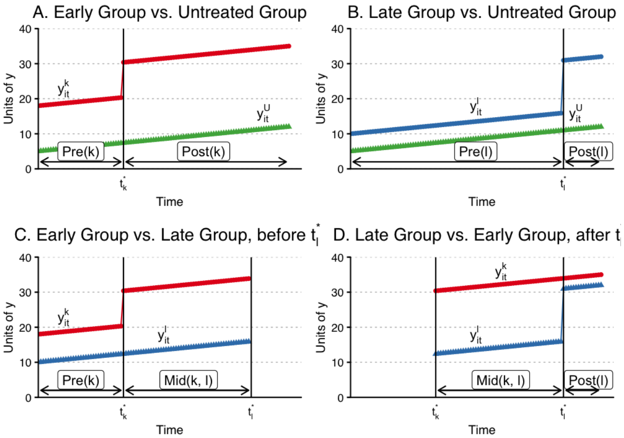
Staggered treatment

- Goodman-Bacon (2021) provides a clear graphical intuition for the bias. Assume three treatment groups - never treated units (U), early treated units (k), and later treated units (l)



Staggered treatment

- We can form four different 2x2 groups in this setting, where the effect can be estimated using the simple regression DiD in each group:



Staggered treatment

- Important Insights:
  - δ is just the weighted average of the four 2x2 treatment effects (*i.e.*, $\delta = \sum_{i=1}^{4} w_i \delta_i$)
  - The weights are a function of the size of the subsample, relative size of treatment and control units, and the timing of treatment in the sub sample
  - Already-treated units act as controls even though they are treated
  - Given the weighting function, panel length alone can change the DiD estimates substantially, even when each δ_i does not change
  - Groups treated closer to middle of panel receive higher weights than those treated earlier or later

Simulations

- Do demonstrate how staggered treatments works we can use simulated data
- Simulated data would mean that we randomly draw values according to pre-determined model
- Repeating this many times creates an “artificial sample”
- We can then use the artificial sample and run our estimator on it, to test various things
- In particular, because we pre-determine the model, we could test whether the estimation produces biased or unbiased estimates because we know the “true” value of the model parameters
- This is sometimes called Monte-Carlo simulations

Simulations

- Consider two sets of DiD estimates – one where the treatment occurs in one period, and one where the treatment is staggered
- The data generating process is linear: $y_{it} = \alpha_i + \alpha_t + \delta_{it} + \epsilon_{it}$
  - $\alpha_i,\ \alpha_t \sim N(0,1)$
  - $\epsilon_{it} \sim N\left(0, \left(\frac{1}{2}\right)^2\right)$
- We will consider two different treatment assignment set ups for δ_{it}

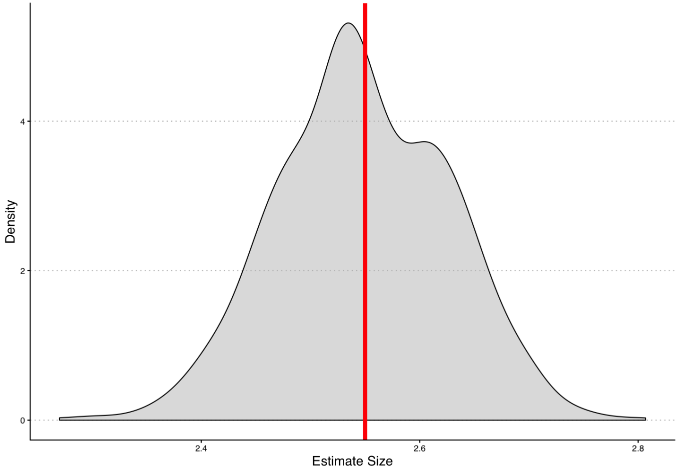
Simulation 1 – one period treatment

- There are 40 states s, and 1000 units i randomly drawn from the 40 states.
- Data covers years 1980 to 2010, and half the states receive "treatment" in 1995.
- For every unit incorporated in a treated state, we pull a unit-specific treatment effect from $\mu_i \sim N(0.3, (1/5)^2)$.
- Treatment effects here are trend breaks rather than unit shifts: the accumulated treatment effect δ_{it} is $\mu_i \times (year - 1995 + 1)$ for years after 1995.
- We then estimate the average treatment effect as $\hat{\delta}$ from:

$$y_{it} = \hat{\alpha}_i + \hat{\alpha}_t + \hat{\delta}\, D_{it}$$

- Simulate this data 1,000 and plot the distribution of estimates $\hat{\delta}$ and the true effect (red line).

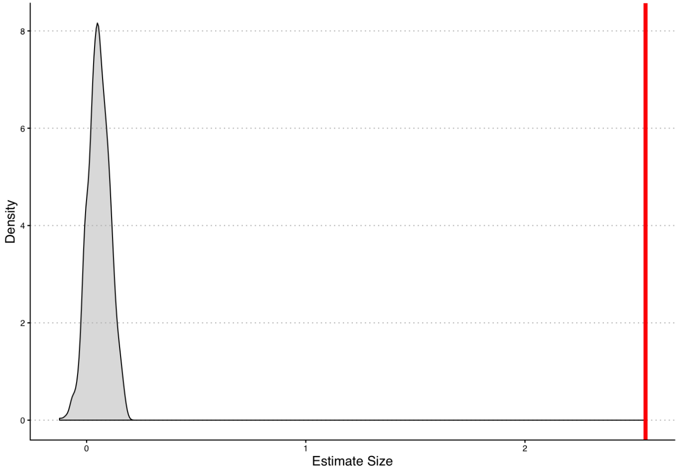
Simulation 1 – one period treatment – DiD working pretty well



Simulation 2 – staggered treatment

- Run similar analysis with staggered treatment.
- The 40 states are randomly assigned into four treatment cohorts of size 250 depending on year of treatment assignment (1986, 1992, 1998, and 2004)
- DGP is identical, except that now δ_{it} is equal to $\mu_i \times (year - \tau_g + 1)$ where τ_g is the treatment assignment year.
- Estimate the treatment effect using TWFE and compare to the analytically derived true δ (red line).

Simulation 2 – staggered treatment – DiD not working so well



Simulation 2 – staggered treatment – DiD not working so well

- Why hasn't it been working well?
  - Main problem – we use prior treated units as controls
  - When the treatment effect is "dynamic", *i.e.*, takes more than one period to be incorporated into your dependent variable, we are subtracting the treatment effects from prior treated units from the estimate of future control units
  - This biases your estimates towards zero when all the treatment effects are the same

Simulation 3 – staggered treatment – dynamic treatment effect

- Can we actually get estimates for δ that are of the *wrong sign*? Yes, if treatment effects for early treated units are larger (in absolute magnitude) than the treatment effects on later treated units.
- Here firms are randomly assigned to one of 50 states. The 50 states are randomly assigned into one of 5 treatment groups G_g based on treatment being initiated in 1985, 1991, 1997, 2003, and 2009.
- All treated firms incorporated in a state in treatment group G_g receive a treatment effect $\delta_i \sim N(\delta_g, .2^2)$.
- The treatment effect is cumulative or dynamic - $\delta_{it} = \delta_i \times (year - G_g)$.

Simulation 3 – staggered treatment – dynamic treatment effect

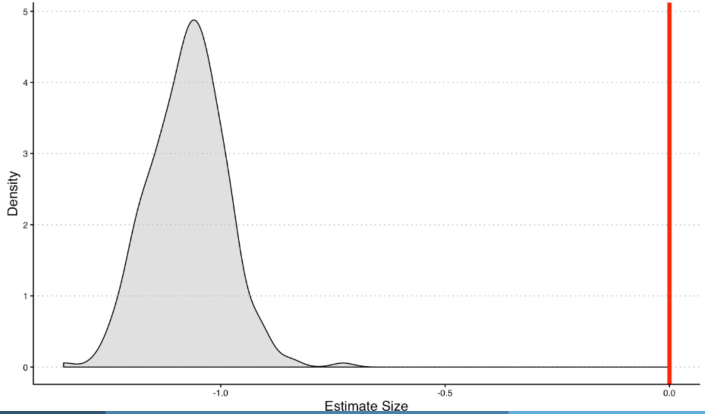
- The average treatment effect multiple decreases over time:

Treatment Effect Averages

`$G_g$`	`$\delta_g$`
1985	0.5
1991	0.4
1997	0.3
2003	0.2
2009	0.1

Simulation 3 – staggered treatment – dynamic treatment effect

- First let's look at the distribution of δ^{DD} using TWFE estimation with this simulated sample:



Staggered treatment

- Before getting into possible solution to these issues we need to take one step back
- How do we properly define such problems?
- In order to do so we need to think more carefully about notation
- The basic structure of an event study estimation is:

$$Y_{it} = \alpha_i + \lambda_t + \sum_l \mu_l D_{it}^l + v_{it} \,,$$

where D_{it}^l is an indicator for l periods relative to i's initial treatment ($l = 0$ is the period of initial treatment)

Staggered treatment

- Two comments about

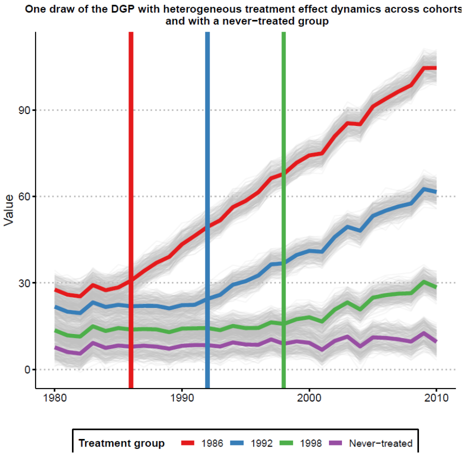
$$Y_{it} = \alpha_i + \lambda_t + \sum_{l} \mu_l D_{it}^{l} + v_{it} \ ,$$

  - First, as we saw – the problem here is that if different units are treated at different times, there is no real definition of treated and control groups
  - Note that it is different from the definition we had initially – the effects μ are not defined for every point in time t, but for a relative time difference from treatment l – so to practically estimate the TWFE DiD coefficients, the needs to be transformed

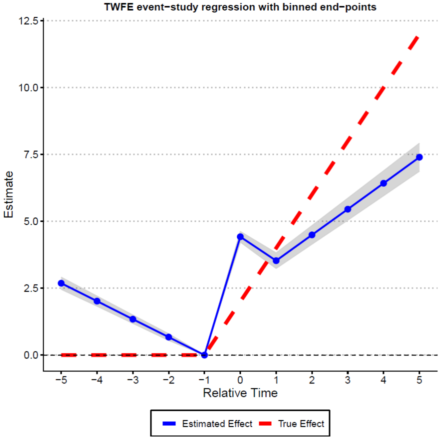
Staggered treatment – another example

- Let's think of another scenario
- Suppose we want to estimate the impact of new Tube stations on rent prices in some neighborhood
- We consider 3 new stations opening at different times and 40 neighborhoods where we randomly sample 250 rents by square meters each year for instance (repeated cross section)
- We generate a model with homogeneous treatment effect (*i.e.*, all treated units receive the same treatment value), but dynamic (because not all treated units receive treatment at the same time)
- We also generate a model where later treated units have small treatment effects
- We estimate the model using TWFE and compare with the true effect

Staggered treatment – another example



Staggered treatment – another example



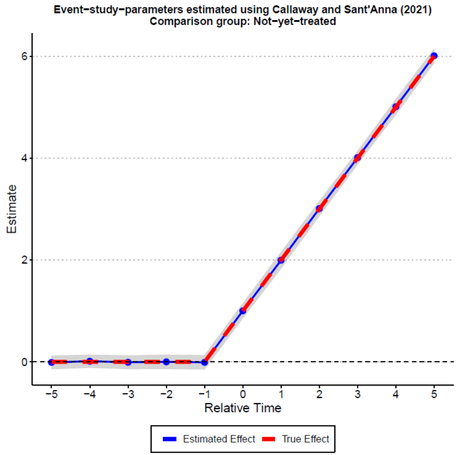
Staggered treatment – another example

- The results above show that these TWFE event-study type estimates are severely biased for the true treatment effects
- Furthermore, using the estimates of coefficient of treatment leads as a way to find evidence of pre-trends is very problematic
- Now, putting it simply, the results above highlight that such TWFE linear regression should not be used to highlight treatment effect dynamics!

Staggered treatment – solutions

- Several solutions have been presented since 2020: Sun & Abraham (2021); Callaway and Sant'Anna (2021); Borusyak *et al.*(2021); Dube *et al.*(2023)
- Each of the papers discusses other estimators for solving the problem of heterogeneous and dynamic treatment effects
- The easiest conceptually is Callaway and Sant'Anna (2021): The idea is simple – compute every ATT (average treatment effect) for each group at g each date t and then aggregate the ATTs into a single weighted ATT
- New Stata commands (from version 18 only): `hdidregress` and `xthdidregress`

Staggered treatment – solutions



Real world example – Facebook and mental health

- Braghieri, Levy & Makarin (2022) wanted to study the effect of social media on mental health
- They used the fact that Facebook was introduced over a relatively long period in different US university campuses
- They found that the rollout of Facebook at a college had a negative impact on student mental health
- It also increased the likelihood with which students reported experiencing impairments to academic performance due to poor mental health

Real world example – Facebook and mental health

- Braghieri, Levy & Makarin (2022) wanted to study the effect of social media on mental health
- They used the fact that Facebook was introduced over a relatively long period in different US university campuses
- They found that the rollout of Facebook at a college had a negative impact on student mental health
- It also increased the likelihood with which students reported experiencing impairments to academic performance due to poor mental health

Real world example – Facebook and mental health

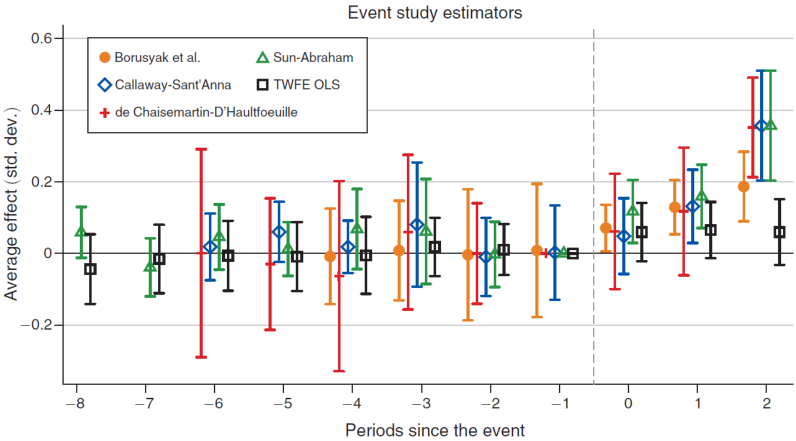


Figure 2. Effects of Facebook on the Index of Poor Mental Health Based on Distance to/from Facebook Introduction

Takeaway points

- DiD is one of the most commonly used empirical designs
- If treatment timing and magnitude varies across groups our standard estimators could be biased and pre trends could potentially be wrongly identified
- This is because in such staggered treatment we not only compare treated and non-treated, but also treated to 'to-be' treated, and to 'never-to-be' treated
- Because many real-world examples are staggered – solutions are important, and they have been introduced in the past few years
- Importantly – this is still an evolving literature!

For next week

- Read Braghieri, Levy & Makarin (2022)
- Exercise C7 in JW Chapter 14 exercises (on Keats – note that numbering can be different in different versions)