Causal Data Analysis and Difference-in-Differences

A Short Course

Daniel Halvarsson

The Ratio Institute, Stockholm

daniel.halvarsson@ratio.se

What we will cover

1.Stata excercise Load example data, Clean it, Calculate 2x2 Difference in diff 2.

Why we interested in difference-in-difference

- Because we want to understand the causal effect of some **treatment** e.g. a policy across different groups.
- What is the causal effect of some 'X' on 'Y'?
- We have access to observational data on 'X' and 'Y', and data on other characteristics for the different groups.
- The challenge...

The main challenge with causal analysis

- Different groups have different characteristics, which in turn may be correlated with the policy assignment.
- How then can we attempt to discover the causal effect without it being confounded by the different group characteristics?

The main challenge with causal analysis

- Different groups have different characteristics, which in turn may be correlated with the policy assignment.
- How then can we attempt to discover the causal effect without it being confounded by the different group characteristics?

Difference-In-Difference

First a word about causality?

- What do we mean by ”What is the causal effect of some ’X’ on ’Y’?”
- There are many ways to define causality.
- A good and simple way to think about it is...

  ”We can say that X causes Y if, were we to intervene and change the value of X, then the distribution of Y would also change as a result.”

  (The Effect, p.89, Huntington-Klein)

- By distribution, we are here thinking about probabilities
- If we intervene to change X, the value of Y does not actually have to change, but only the probability that Y occurs.

Effect heterogeneity

- The notion of a single causal effect in social science is perhaps best described as a fiction. Take the example of a new drug, while it may be very effective for someone, it has no or adverse effect for someone else, be it because of their e.g. gender, body chemistry, or age.
- This type of effect-variety is called **heterogeneous treatment effects**
- It's possible to estimate the distribution of these effects, e.g. to predict a specific effect that pertains to an individual with certain characteristics (personalized medicine).
- But in terms of estimating causal effects in social science, we are generally after some form of **average effect**.

Treatment-effect averages

- Having established the notion of heterogeneous treatment effects, we can consider the mean effect, i.e. **the average treatment effect** (ATE)
- Often times it is the ATE that we would like to know. If we implement a policy on everyone, what would be it's average effect?
- However, it's not always possible to estimate the ATE or sometimes not even desirable. The ATE for a new treatment of testicular or breast cancer is not desirable in determining it's effectualness.
- In running a medical trial, surely the participants are limited to either women or men. In this case, the average refers to the CATE, which stands for the conditional ATE, where the condition could be being male or female.

The average treatment effect of the treated (ATT)

- Because of research design (such as with DiD), we often end up estimating something else, which is the average effect for the group that gets treated, in other words the **average treatment effect of the treated** (ATT or ATET)
- However, should we have a large random sample there is rather likely that ATT = ATE (or CATE when it's applicable)
- As for the ATE. In a random control trial, when the sample is representative of the population the effect corresponds to the ATE. If it's not representative, it's conditional on being in the sample.

Difference-in-Difference: What is it?

- Difference-in-differences is a statistical method or technique popular in social science with observational data that mimics an experimental research design.
- It relies on two groups; one treated and one untreated control group. In DiD, the outcome is first measured over time for each group (the first difference) and the compared against one another (second difference). The result from these two differences can be a causal estimate.
- Importantly, the notion of a control group is a place holder for ”our best guess at what the treatment group would have been without treatment” (*The Effect*, *p.450*, Huntington-Klein).
- **The plan** with DiD is to use the change within some untreated control group to represent all changes within the treatment group that is not due to the treatment.

What does the DiD results actually mean and is the effect really causal?

- Since out plan was to use the control group to ”represent all changes **within the treatment group** that is not due to the treatment”, the DiD estimate the effect for the treated group.
- If certain conditions are satisfied, specifically the assumption of so called **parallel trends**, the DiD estimate represents a well defined causal parameter, namely the average treatment effect of the treated (ATT). **This is the connection between DiD and causality.**
- For the DiD estimate, therefore, we could not know if the treatment effect would be any different for other groups!

The history of DiD goes back more than 150 years

- Today somewhat of a cliché, but still very instructional.
- In 1855, John Snow (like an Aegon Targaryen) demonstrated for the first time and for the world that cholera was spread by contaminated water and not through the air.
- The culprit, water intake downstream the Thames contained all sorts of sewage waste.
- Between 1849 and 1854 a policy was implemented that required the Lambeth Company move their water intake upstream outside of London.
- The results...

What happened with death rate in areas served by the Lambeth company?

Region Supplier	Death Rates 1849	Death Rates 1854
Non-Lambeth Only (Dirty)	134.9	146.6
Lambeth + Others (Mix Dirty and Clean)	130.1	84.9

Death rates are deaths per 10,000 1851 population.

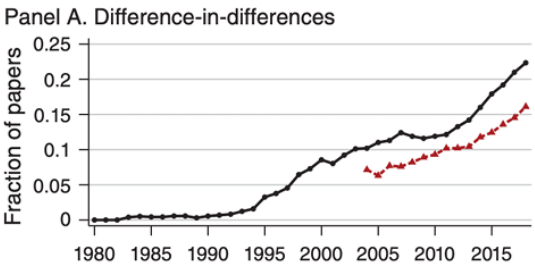
(*The Effect*, *p.438*, Huntington-Klein)

Table 18.1: London Cholera Deaths per 10,000 from Snow (1855)

- Let's compare the differences in death rates between 1854 and 1849 for Lambeth and non-Lambeth
- For Lambeth we get $84.9 - 130.1 = -45.2$ and for non-Lambeth $146.6 - 134.9 = 11.7$
- How much do the difference differ? $-45.2 - 11.7 = -57.1$
- By moving the water pump, the cholera mortality rate decreased by 57.1 deaths per $10,000$.

Difference-in-Difference is still very much relevant in empirical social science research

- In econmics, Difference-in-Difference (or DiD) is today arguably the most widely used method for estimating causal effects in non-experimental settings.
- An early and influential study is David Card (1990), *The Impact of the Mariel boatlift on the Miami labor market*, IRL Review 43(2):245-257.



Let's work through an example using real data

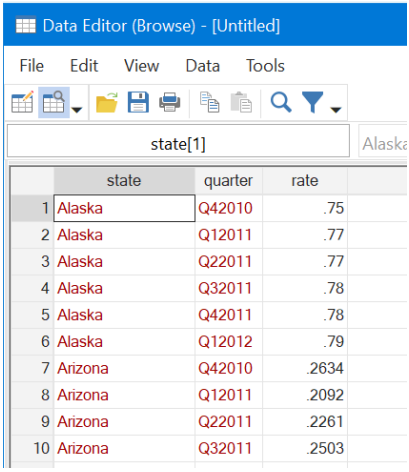
- We are going to use the data from the paper by Kessler and Roth (2014) called *Don't take "no" for an answer: And experiment with actual organ donor registrations*
- Most US states have so called "opt-in" (as opposed to "opt out") organ donation programs. When you get your drivers license you can choose to partake in the organ donation program.
- In July 2011, California implemented a policy to switch from an "opt-in" program to one with "active choice", of either yes or no.
- Did the policy positively affect the donation rate in California, as believed by the policy makers?

Organ donation rates in the US

- Data over US organ donation rates:

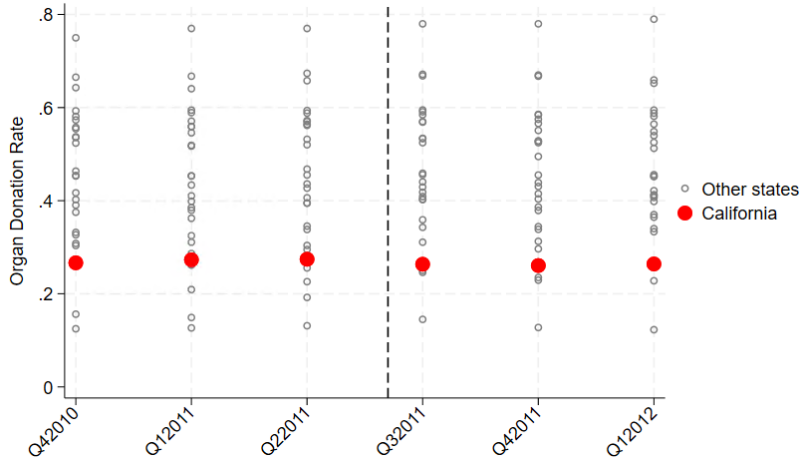
https://github.com/DanielHalvarsson/IntroductionDiD/

- Download the data manually using the above link. And open the data file from within stata.
- Alternatively, the data can be retrieve directly from within Stata by linking directly to the 'organ_donations.dta' file.
- To follow along, you need the following Stata packages: *reghdfe*



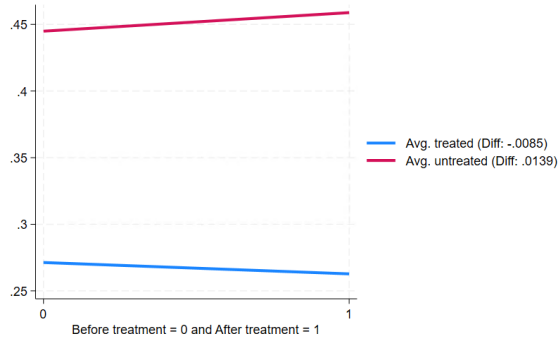
	state	quarter	rate
1	Alaska	Q42010	.75
2	Alaska	Q12011	.77
3	Alaska	Q22011	.77
4	Alaska	Q32011	.78
5	Alaska	Q42011	.78
6	Alaska	Q12012	.79
7	Arizona	Q42010	.2634
8	Arizona	Q12011	.2092
9	Arizona	Q22011	.2261
10	Arizona	Q32011	.2503

Visualizing the organ donation rates over time



Did the policy causally affect the donation rate in California?

- To answer that question, we calculate the 2×2 Difference-in-Difference.
- Just like with the cholera example we are in the business of comparing averages.
- The DiD effect amounts to a -2.24 percentage point reduction in organ donation rates, following the policy in California.



Difference-In-Difference estimate: $-0.0085 - 0.0139 = -0.0224$

Is the DiD estimate really the causal effect?

- It depends...
- For DiD to work, we obviously need an untreated control group. Specifically, we need an untreated control group that satisfies the **parallel trends assumption**.
- According to the parallell trends assumption:

  "if no treatment had occurred, the difference between the treated group and the untreated group would have stayed the same in the post-treatment period as it was in the pre-treatment period."

  (The Effect, p.441, Huntington-Klein)

- Importantly, the parallel trends assumption can in no way be tested as it concerns the counterfactual, i.e. how things would have turned out, had there been no treatment.

The paralell trends assumption

- Recall that the plan with DiD was to use the change in the untreated group to represent all non-treated related changes in the treated group.
- To see what this means consider the difference in outcome for the treated group before and after the treatment as $\mathit{TretmentEffect} + \mathit{OtherTreatedGroupChanges}$
- For the untreated group, the difference in outcome before and after the treatment is given by $\mathit{OtherTreatedGroupChanges}$
- The DiD estimate is therefore given by

$$\mathit{TretmentEffect} + \mathit{OtherTreatedGroupChanges} - \mathit{OtherTreatedGroupChanges} \quad (1)$$

- For the DiD to identify only the treatment effect $\mathit{OtherTreatedGroupChanges} = \mathit{OtherTreatedGroupChanges}$, precisely as required by the parallel trends assumption.

The paralell trends assumption

if no treatment had occurred, the difference between the treated group and the untreated group would have stayed the same in the post-treatment period as it was in the pre-treatment period.

- To unpack this. Let's imagine the counterfactual world where no treatment has or will ever happened.
- If the parallel trends assumption holds, then the outcome for both the treated and untreated groups should experience the same change over time.
- But if they experience the same change over time, it must be also the case that $OtherTreatedGroupChanges = OtherTreatedGroupChanges$.
- Since we don't live in the counterfactual world, we can never know whether this any of this is true.

Three good reasons for why your DiD design is believable

- Even if the parallel trends assumption can't be tested, there are goods signs to look for.
  1. We can not think of a good reason for why the outcome in the untreated group would suddenly change at the time of treatment.
  2. The characteristics of the treated and untreated groups generally similar.
  3. Looking at time periods leading up to the treatment date, the dependent variable evolves similarly for the treated and untreated group.
- What it often times comes down to in practice is to pick a control group such that 1-3 is believable.

Additional implications of the parallel trends assumption

- The parallel trends assumption is not only an assumption about causality, but also about the gap (or difference) that is supposed to remain constant.
- Therefore, if the parallel trends assumption holds for some 'Y' it does not hold for transformations of 'Y', which includes '$\log Y$'
- It is easy to see that if the gap before the treatment was e.g. $20 - 10 = 10$ and the counterfactual difference after the treatment was $25 - 15 = 10$, in logs we get $\ln 20 - \ln 10 \approx 0.3$ and $\ln 25 - \ln 15 \approx 0.22$.
- Finally, there is a risk of pre-testing, as it can cause statistical problems by potentially contaminating your design.

How can we check if our untreated group is appropriate?

- Even if we can't test the parallel trends assumption, we can and should test for common pre-trends (good sign nr. 3), **IF** we have more than two time periods.
- The second test we can do is a *placebo test*. It means that we throw out all the dates for which the treatment was in effect and pick different periods before the actual treatment to see if the same DiD analysis gives a zero effect.
- If the placebo test would show up significant, we can count that as evidence that something can be awry with the paralell trends assumption.
- However, with more than two time periods we should move DiD into a regression framework.

Difference-in-Difference in a regression setting

- For the 2×2 DiD analysis with two groups and two periods, we can reach exactly the same DiD estimate by using a regression.
- Instead of the tedious comparing of means, we can simply estimate the following regression model:

$$\begin{aligned} Y = &\beta_0 + \beta_1 \mathit{AfterTreatment} + \beta_2 \mathit{TreatedGroup} \\ &+ \beta_3 \mathit{AfterTreatment} \times \mathit{TreatedGroup} + \epsilon \end{aligned} \tag{2}$$

- with *AfterTreatment* being a dummy variable that takes the value of one the period after treatment and zero the period before treatment.
- and where *TreatedGroup* is a dummy variable that takes the value of one for the trewated group and the and zero for the untreated group.
- *AfterTreatment* $\times$ *TreatedGroup* is an interaction term that takes the value of one for the treated group after treatment, and zero otherwise.

Difference-in-Difference in a regression setting

- Exactly the same DiD estimate as before is here given by $\hat{\beta}_3$.
- This is fairly easy to see. Considering the average outcomes across the four different groups:
  - *TreatedGroup* $= 0$ and *AfterTreatment* $= 0$ gives $Y = \beta_0$
  - *TreatedGroup* $= 0$ and *AfterTreatment* $= 1$ gives $Y = \beta_0 + \beta_1$
  - *TreatedGroup* $= 1$ and *AfterTreatment* $= 0$ gives $Y = \beta_0 + \beta_2$
  - *TreatedGroup* $= 1$ and *AfterTreatment* $= 1$ gives $Y = \beta_0 + \beta_1 + \beta_2 + \beta_3$
- Now take the difference for the untreated group before and after treatment
  $\beta_0 + \beta_1 - \beta_0 = \beta_1$
- Take the same difference but for the treated group
  $\beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 - \beta_2 = \beta_1 + \beta_3$
- To get the DiD, subtract the untreated difference from the treated difference to get
  DiD $= \beta_1 + \beta_3 - \beta_1 = \beta_3$

Generalization of DiD to many periods

- For now we only consider one common effect for the treated group. Later we will allow for dynamic treatment effects.
- Even with more periods, it's always possible to estimate the 2×2 DiD by combining time periods or by aggregating data.
- But first, note a powerful way of rewriting the regression

$$\begin{aligned} Y =& \beta_0 + \beta_1 \mathit{AfterTreatment} + \beta_2 \mathit{TreatedGroup} \\ &+ \beta_3 \mathit{AfterTreatment} \times \mathit{TreatedGroup} + \epsilon. \end{aligned} \tag{3}$$

- By exploiting the fact that *TreatedGroups* and *AfterTreatment* can be regarded as fixed effects α_g and α_t, we can write the same 2 × 2 DiD regression as follows

$$Y = \alpha_g + \alpha_t + \beta_3 \mathit{AfterTreatment} \times \mathit{TreatedGroup} + \epsilon. \tag{4}$$

Leveraging DiD by using fixed-effects regression

- In the DiD specification with fixed effects

$$Y = \alpha_g + \alpha_t + \beta_3 AfterTreatment \times TreatedGroup + \epsilon \tag{5}$$

  it is clear that DiD uses so called ”within” variation to identify the effect.
- What ever *unobservable* variation that does not change over the period for each group or time period is controlled for in the model and can not contaminate the ”effect”.
- But why stop there, if the treated and untreated groups are comprised of individual workers, firms, or states, we can shift the level of fixed effects from α_g to α_i, where the index *i* stands for individual units.
- This model is called the **Two Way Fixed Effects** or **TWFE model**, and is the work horse in many DiD applications.
- For individuals α_i would capture characteristics such as personality, upbringing, ability, gender to the extent they are correlated with Y or the treatment

Let's return to the Kessler and Roth data and run some regressions!

1. We start by estimating

$$\begin{aligned} Y =& \beta_0 + \beta_1 \mathit{AfterTreatment} + \beta_2 \mathit{TreatedGroup} \\ &+ \beta_3 \mathit{AfterTreatment} \times \mathit{TreatedGroup} + \epsilon. \end{aligned} \tag{6}$$

2. Verify the that the coefficients corresponds to the different group means.
   - $\mathit{TreatedGroup} = 0$ and $\mathit{AfterTreatment} = 0$ gives $Y = \beta_0$
   - $\mathit{TreatedGroup} = 0$ and $\mathit{AfterTreatment} = 1$ gives $Y = \beta_0 + \beta_1$
   - $\mathit{TreatedGroup} = 1$ and $\mathit{AfterTreatment} = 0$ gives $Y = \beta_0 + \beta_2$
   - $\mathit{TreatedGroup} = 1$ and $\mathit{AfterTreatment} = 1$ gives $Y = \beta_0 + \beta_1 + \beta_2 + \beta_3$
3. We also estimate the fixed-effects version of the regression in (6).

$$Y = \alpha_g + \alpha_t + \beta_3 \mathit{AfterTreatment} \times \mathit{TreatedGroup} + \epsilon. \tag{7}$$

4. and it's extension

$$Y = \alpha_i + \alpha_t + \beta_3 \mathit{AfterTreatment} \times \mathit{TreatedGroup} + \epsilon. \tag{8}$$

Are the DiD design supportive of the parallel trends assumption?

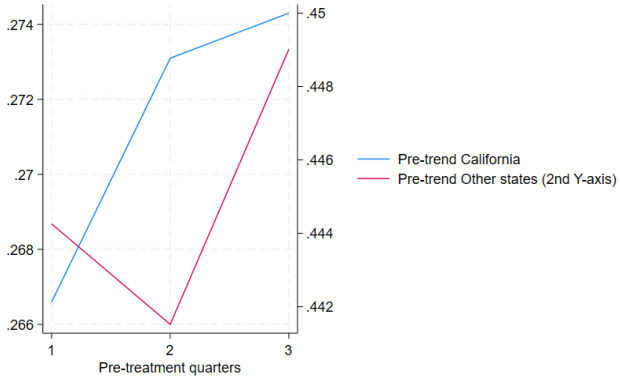
- With more than two periods in the data, we can have a look at pretrends and run placebo tests.
  1. Let's plot the pretrends
  2. Test for diverging pretrends. Specifically, we estimate the model with possibly different linear time trends before treatment.

$$Y = \alpha_g + \beta_1 \mathit{TimePeriods} + \beta_2 \mathit{TimePeriods} \times \mathit{TreatmentGroup} + \epsilon \tag{9}$$

where *TimePeriods* is a count variable for the periods leading up to the treatmentperiod. In the model β_1 is the slope estimate of the pretrend for untreated group. For the treated group the slope estimate for the pretrend is given by $\beta_1 + \beta_2$.

Thus, if we can't reject H_0 that $\beta_2 = 0$, then we we can draw the conclusion that the best linear approximation of the average pre-trends for the treated and untreated groups are probably the same.

Are the DiD design supportive of the parallel trends assumption?



3. We also run two placebo tests. Instead of using 2011Q3 as the treatment date, we reestimate the models pretending that treatment occured 2011Q1 and 2011Q2 using data only up until 2011Q2.

What to do when the parallel assumption is likely violated?

- In the basic DiD setting, you generally wont fix a violation in the parallel trends assumption by adding a bunch of covariates.
- Why?
- Because time invariant covariates are already controlled for by the fixed effect version of the DiD.
- Adding time-varying covariates can destroy identification if they themselves are caused by the treatment.
- A common option is to look for a better control group by using some form of **statistical matching**, e.g. propensity score matching.
- If matching is successful, there is a good chance that the parallel trends assumption is more likely to hold.

Dynamic DiD: long-term effects

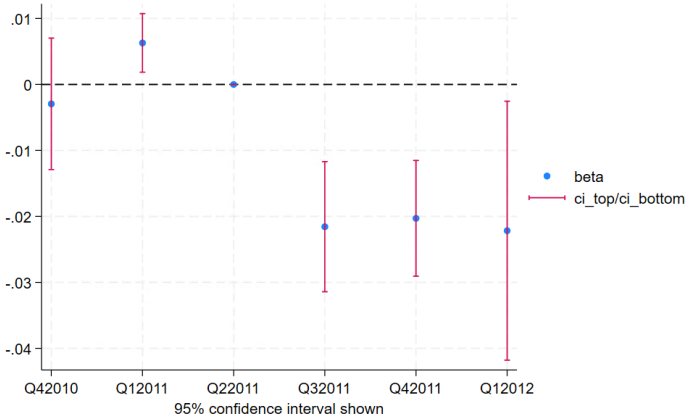
- So far, even with access to multiple periods, we have lumped the periods together in the period before and after treatment
- However, by only looking at one potential effect, useful information about how the effect changes over time may be lost.
- For example, it may take some time for an effect to materialize or it may taper off.
- Starting from the fixed effect DiD regression, the dynamic version replaces the *AfterTreatment* dummy in the interaction term with separate dummies for each year, both before and after treatment.
- For example, a design with six periods and treatment in period four, the dynamic DiD can be written as,

$$Y = \alpha_i + \alpha_t + (\beta_1 D_{-3} + \beta_2 D_{-2} + \beta_3 D_0 + \beta_4 D_1 + \beta_5 D_2 + \beta_6 D_3) \times \mathit{AfterTreatment} + \epsilon. \tag{10}$$

- Often times, calender time is replaced with relative-treatment time.

Let's code the dynamic DiD

1. Using the organ donation data, let's code the dynamic version of the DiD.
2. What is the interpretation?



Interpreting the dynamic DiD results

- Near zero effects during the pretreatment period (compared to Q22011)
- One badly behaving pre-treatment effect
- Yet, three distinct negative effects immediately following the policy.
- In accordance with the 2×2 DiD design, the dynamic DiD finds a -2.2 percentage point decrease in organ donation rates in California as a result of active choice.

Insights from the dynamic specification

- In addition to the additional insights that about treatment-effect dynamics
- The dynamic DiD provides a direct test for parallel **pre**trends (a type of placebo test).
- For each of the pretreatment gaps, they should not be larger or smaller compared to the gap before the treatment takes place, i.e. the reference.
- With many pre-periods, however, there can be significant effects due to pure chance.

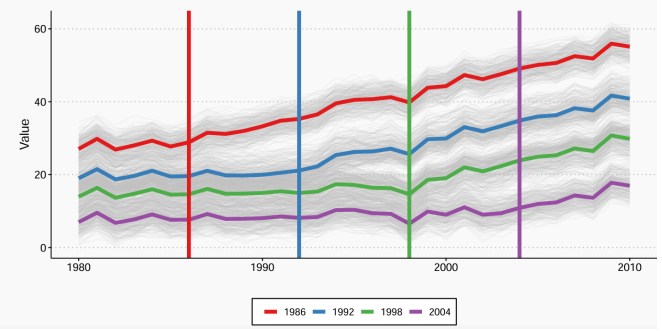
Insights from the dynamic specification cont.

- However, as three are less data devoted to estimating each of the effects, expect less precise estimate.
- Even if the individual effects is significant, the average overall effect can still be significant!
- However, when there are many treated groups with varying treatment timing, dynamic DiD can not be trusted.

Roll-out adoption (staggered treatment) design

- Expands the dynamic DiD designs to allow for *multiple groups* with *different treatment adoption dates*.

- Example is a policy with a regional roll-out such that the policy is adopted in different areas at different dates.



The problem with TWFE and ”the secrete shame of econometrics”

- The TWFE model is well understood and commonly applied to the basic 2×2 setup and to the dynamic (event-study) setup.

- However, for a roll-out treatment design, with different groups getting treated at different times, the TWFE model no longer works.

- **The problem** can be quite serious. With DiD estimates displaying a negative average effect despite the true effect being positive for everyone in the sample.

”For decades researchers were basically unaware of this problem and used two-way fixed effects anyway.”

(The Effect, p.458, Huntington-Klein)

The problem

- The problem is somewhat complex.
- TWFE relies on **within variation** for comparing treated and controls. It means that units that *remains untreated* during the period end up as controls, but the same goes for units that *remains treated* from earlier roll-outs.
- **IF** treatment effect varies with treatment time e.g. effects grow larger over time, earlier treated units in the control group will have an increasing trend that is distinct from just-now-treated units,
- **THEN** parallel trends assumption breaks and identification fails.

The State of the DiD literature

- Things are moving fast
- This literature has had a certain amount of upheaval over the past 5-6 years.
- With the upheaval there is a **tension** for how people currently and historically have used DiD.
- The modern literature has pointed out many issues but has provided solutions to almost all of the them.
- Good tools are now readily available (including Stata), so nothing to prevent you from using a DiD with staggered analysis.

When TWFE is NOT a problem in staggered adoption design

- However, despite the severity of the problem, as emphasized at the beginning, there are situations where TWFE estimates is reliable.
- If the treatment effect is the same across all treated groups over time the dynamic TWFE works just fine.
- However, we can never know whether this is true by only estimating the dynamic TWFE model.

Roll-out design as many different sub-experiments

- For each treated group in a roll-out design, the causal effect is just the same as in dynamic DiD when compared to an untreated group.
- Plenty of causal effect estimates: What to do with them?
  - In a small design, it could be be beneficial to analyze each of the sub-experiments separately to gain insight into treatment effect heterogeneity, although it may be statistically inefficient.
  - In a larger study, a separate analysis may not simply be feasible (nor desirable).
- For both cases it's often desirable to average or aggregate many estimates into a single causal effect.
- But for this purpose, the TWFE model don't correspond to a causal effect, without imposing strong and quite artificial assumptions.

Goodman-Bacon Decomposition

- An influential paper, Goodman-Bacon (2021) showed that TWFE estimators in a staggered setting can be expressed as a weighted average of the of all underlying DiDs.

- **Their contribution** was the insight that some of the DiDs are actually confounded despite them individually satisfying the common trend + no anticipation assumptions!

- There are 4 types of DiD comparisons between treated and control implicitly made in the TWFE model
  1. Treated (early and late) vs. Never treated DiDs
  2. Early vs. Late DiDs
  3. Late vs. Early DiDs ☠

  The potentially problematic comparison being the 'Late vs. Early' where a treatment group is compared to a group that is already treated! .

- How large is the problem: Use **bacondecomp** in Stata for Goodman-Bacon decomposition.

What to do with staggered timing in DiD?

"What to do then, when we have a nice roll-out design? Don't use two-way fixed effects, but also don't despair. **You're not out of luck, you're just moving into the realm of what the pros do**." (The effect)

- There's really no reason to use the baseline TWFE in staggered timings
  - A perfect example wherein the estimator does not generate an estimate that maps to a meaningful estimand
- There are different approaches proposed in the literature that are just as good!
  - Let's have a loo at Sun and Abraham (2020), which extends the dynamic DiD model to account for the different treatment effects for different groups.

Modified event-study design: Sun and Abraham, 2020

- Starting from the basic event-study design, it can be modified to include many groups with a staggered roll-out.

- First, each of the sub-experiments are **centered** relative to their own treatment start.
  - It keeps track on the already treated groups so that they don't get included in comparisons.

- Second, Sun and Abraham then propose that the relative time dummies are **interacted** with group dummies such that each treated $group \times relative - time - dummies$ get their own effect.

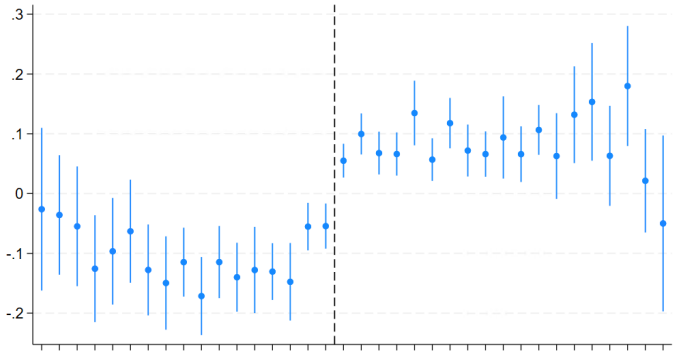
- It's then up to you to avoid making bad comparisons when averaging coefficients to get e.g. the time-varying treatment effect (see example below).

- Each effect is either compared to the group of (i) never treated observation or (ii) not yet treated observations (from the last treated group(s)).

Modified event-study design: Sun and Abraham, 2020

- Sun and Abraham (2020) can be accessed in Stata using the **eventstudyintereact** package.

1. Let's try it out!



Interpreting the results from the union application

- The gap between treated and untreated in the pretreatment period is smaller compared to the gap in the period before treatment.
- Why?
- Possibly because the gap suddenly increases just before becoming union member.
- It's difficult to speculate without looking closer at the data, but the pattern would agree with a situation where wages in the treated group shoots up the year before the join the union.
- Regardless if this is the case, it would be difficult to convince someone that the parallel trends assumption is satisfied.