Review day 3:

1. Covariates
   1. Illustrated the role of various “heterogeneities in the potential outcomes” in determining whether or not OLS specifications could identify the ATT under what assumptions.
      1. We need conditional parallel or unconditional parallel trends
         1. Data generating process about average Y(0) for treatment and control either entirely (unconditional parallel trends) or sub-populations (conditional parallel trends).
         2. Subpopulations are represented with covariates.
         3. In a large enough sample, then the mean Y(0) across the dimensions of those covariates are expected to be the same over time for treatment and control if conditional parallel trends holds
         4. OLS can handle that
      2. But canonical/traditional OLS specification **also needed** two more assumptions that the other methods we reviewed did not need:
         1. Constant covariate trends (over time). OLS needs that, but DR, IPW and RA don’t.
         2. Here’s an example in Stata code of “**constant covariate trends**”:  
              
            \* Canonical OLS specification needs this one:

bys county\_id: gen **trend** = .

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==1

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==2

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==3

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==4

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==5

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==6

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==7

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==8

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==9

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==10  
  
bys county\_id: gen **y0** = 15 + county\_fe + **trend**\*(year) + u  
  
3. And here’s an example in Stata of “heterogenous covariate trends”.

* Canonical OLS specification *cannot handle* this one, even though it can handle conditional parallel trends.

bys county\_id: gen **trend** = .  
replace **trend** = -0.15\*urban + 0.1\*(1-urban) if year==1

replace **trend** = -0.25\*urban + 0.1\*(1-urban) if year==2

replace **trend** = -0.15\*urban + 0.1\*(1-urban) if year==3

replace **trend** = -0.1\*urban + 0.1\*(1-urban) if year==4

replace **trend** = -0.1\*urban + 0.1\*(1-urban) if year==5

replace **trend** = -0.5\*urban + 0.1\*(1-urban) if year==6

replace **trend** = -1\*urban + 0.1\*(1-urban) if year==7

replace **trend** = -1.5\*urban + 0.1\*(1-urban) if year==8

replace **trend** = -2\*urban + 0.1\*(1-urban) if year==9

replace **trend** = -2.5\*urban + 0.1\*(1-urban) if year==10  
  
Both of those trends are entering into the Y(0) data generating process:  
  
bys county\_id: gen **y0** = 15 + county\_fe + **trend**\*(year) + u  
  
What’s the difference between conditional parallel trends and constant/heterogenous coviarate trends?  
  
Conditional parallel trends just means whatever the relationship is between average Y(0) and your covariates over time, it’s the same for treatment and control. Only refers to *trends* (not levels) but just means that the relationship between Change in E[Y(0)] is the same for both treatment and control. And as you can see, since neither the **trend** variable in Stata above nor the **y0** variable are at all related to the treatment.

* + - 1. But the other thing that the standard/canonical OLS specification in diff-in-diff needs is “homogenous treatment effects” which is expressed, not in Y(0) data generating process, but in Y(1). And this was the example of homogenous treatment effects using Stata code:  
           
          \* Homogenous treatment effects

gen y1 = y0

qui replace y1 = y0 + 1000 if year == 1991

\* Covariate-based **treatment effect** **heterogeneity**

gen y1 = y0

qui replace y1 = y0 + 1000 **+ 100 \* age + 500 \* gpa** if year == 1991

Lesson was the standard OLS specification:  
  
Y = a + b1 Post + b2 D + delta (Post x D) + beta X + e

(Additive covariates)

* Conditional PT
* No covariate specific trends
* Homogenous treatment effects

If it didn’t have all three, then depending on the magnitudes and their relationship to treatment, then delta hat would not equal the ATT.

But the other diff-in-diff methods we used:

* Inverse probability weighting (IPW) by Abadie 2005
* Regression adjustment, or “outcome regression”, from Heckman, Ichimura and Todd (1997)
* Double robust (DRDID) from Sant’Anna and Zhao (2020)

They only need:

* Conditional parallel trends

But this is part of this broader movement, it’s almost scientific-philosophical, **towards methods that are agnostic about the Y(0) and Y(1) terms**. This has been called “unrestricted heterogenous treatment effects” or “arbitrary heterogenous treatment effects”.

Diff-in-diff Practitioner’s Guide by Baker, et al. (2025)

<https://arxiv.org/abs/2503.13323>

Instrumental Variables with Unobserved Heterogeneity in Treatment Effects by Mogstad and Torgovitsky (2024)

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4952488>

It started with Imbens and Angrist’s classic 1994 paper on instrumental variables and the LATE (though maybe it’s actually as far back as Roy 1951, and a lot of Jim Heckman’s writings in the 1970s). It really gets emphasized with this adoption of potential outcomes because potential outcomes is not theoretical. It’s just notation. And so because the notation is not connected to a social science, it is not connected to any restrictions on what treatment effects can be, or the underlying potential outcomes. So as a result of it, it almost was inevitable that this “unrestricted heterogenous treatment effects” would encroach upon our models and force a re-evaluation of difference-in-differences.

What we learned is that OLS “exogeneity” imposed a lot of homogeneity on treatment effects *and* just all the potential outcomes. The only assumption that people tended to even know about was parallel trends, but in fact OLS had additional “hidden assumptions” like the ones I just named.

1. Differential Timing

First, we decomposed what the twoway fixed effects “standard/canonical panel fixed effects estimator” was equal to using Bacon’s decomposition which was based on the application of a theorem in econometrics going back to the 30s/40s called “Frisch-Waugh-Lovell”. Lovell is much later and he cleans up the proof. Bacon shows **two things** with that coefficient:

* 1. He shows that the TWFE coefficient was equal to a **positively weighted average** of “all 2x2 diff-in-diffs in the data”. With K treatment timing groups and one untreated group, there are in fact not just K diff-in-diffs, but in fact “K-squared” diff-in-diffs. And the weights were based on two things: sample shares and “variance in treatment” and the latter “weighted up” treatment groups’ diff-in-diffs treated in the middle of the panel / “weighted down” treatment groups diff-in-diffs treated at the start of the panel or the end of the panel. OLS creates its own weights based on the variance in the treatment timing inside your panel length.
  2. But since diff-in-diff equations are how we derive biases and average treatment effects by substituting potential outcomes for realized outcomes, Bacon then does that with his decomposition. He substitutes the potential outcomes into those diff-in-diffs and shows that TWFE needs more than just parallel trends. And that is because in order to minimize the sum of squared residuals, OLS is taking advantage of *all variation in the data*, which in this case included taking diff-in-diffs using “earlier treated groups as controls”. And that introduces a second bias other than parallel trends requiring that the earlier treated groups’ treatment effects not change over time. It’s in this *second step* of the Bacon decomposition that we see for the first time “the negative weights”. But the negative weights are not on the diff-in-diffs, which is what he showed in the part (a). In Part (a), the Bacon decomposition is variance weights, which are positive, on **diff-in-diffs**. But once he substituted potential outcomes, that’s where you see negative weights. The negative weights are on the **dynamic treatment effects**, not the diff-in-diffs.   
       
     TWFE coefficient = Variance Weighted ATTs + Variance Weighted Parallel Trends **minus change in ATT over time for the comparison groups**  
       
     TWFE delta hat = VWATT + VWPT **minus Delta ATT**

It's just another version of the earlier ‘heterogeneity biases’ in the OLS specification, very similar to what we talked about with the standard 2x2. See the “minus sign” is on the Delta ATT and the minus sign is “the negative weight” in Bacon’s decomposition.

There’s other papers. There were other people writing at the exact same time as Bacon and they were all grad students or assistant professors. All these authors are young, not in a professional network yet, and don’t appear to be interacting with one another, but they approach the same or similar TWFE specifications and come to the same conclusions, but they’ll use different decomposition methods. So for instance, Bacon uses FWL and that’s why his decomposition looks the way it does. And you see the negative weighting on the dynamics and it's super intuitive. But de Chaisemartin and D’Haulfoueille have a different approach and their negative weights come in through the error term of the TWFE equation. Sophie Sun and Sarah Abraham they use the “omitted variable bias formula” that is very popular with Josh Angrist (who is at MIT with them) and they only focused on the TWFE dynamic regression with leads and lags (“event study”) and they find a different decomposition but also negative weights coming from this “using already treated as controls”.

There is a degree of redundancy though across the papers which is because they all are showing similar biases. But you have to remember, there’s simultaneous discoveries happening.

Stockley’s “production function” for writing papers:

Y = f(inputs)

Y = (X1)(X2)(X3) … (XK)

Inputs none of them are about being smart. Grit stuff. “X1 is oming up with good ideas” X2 “being pretty good at writing”, “X3 knowing when to stop writing”.

All the inputs multiply against each other to make a paper. Versus this production function:  
  
Y = X1 + X2 + X3 + … + XK

Notice that if they multiply, then you have to non-zero in each one. But if they add, you don’t.

* 1. Introduced a method that DID NOT depend on homogeneity, and it was Callaway and Sant’anna. Didn’t depend on any restrictions on either Y(0) or Y(1) except for conditional parallel trends.

Every diff-in-diff has one major restriction on potential outcomes: Conditional parallel trends or unconditional parallel trends.

Parallel trends: (with or without covariates)

( E[Y(0)|D=1, Post] – E[Y(0)|D=1, Pre] ) **–** ( E[Y(0)|D=0, Post] – E[Y(0)|D=0, Pre] ) = 0

Since the first one is unobserved, it therefore is a “restriction” on what it can be. And it has to be a number such that those four averages and three subtractions equal zero.

All of diff-in-diff is “hard” if you think conditional parallel trends is a “hard assumption” (since it’s unseen). The estimators are not hard **at all**. They are versions of “four averages and three subtractions” – it’s like impossible for that to be hard because it’s subtraction and division.

Because we only have imperfect detections of conditional PT. “We see through glass dimly” for PT because we are missing the first average and there are so many ways the event study won’t give you the right answer.