Canonical Research Designs I: Difference-in-Differences III:

Paul Goldsmith-Pinkham

March 28, 2023

Today's Topics

- Complications that arise in DiD:
  1. continuous treatments
  2. multiple treatments
  3. random timing
  4. Covariates + time-varying covariates

- Checklist of what you need to consider

Continuous treatment

- Recall that estimating treatment effects with random assignment and continuous treatments was data hungry, but doable
- Without random assignment, and unconstrained heterogeneity, the challenge is more complicated
- First, consider what the estimand of interest is. Note that with potential outcomes, we now have $Y_i(D_i)$ with D_i multivalued.
- So what is the contrast of interest?

$$ATT(d|d) = E(Y_i(d) - Y_i(0)|D_i = d) \tag{1}$$

$$ACRT(d|d) = \frac{\partial E(Y_i(l)|D_i = d)}{\partial l}\bigg|_{l=d} \tag{2}$$

Continuous treatment – Estimand

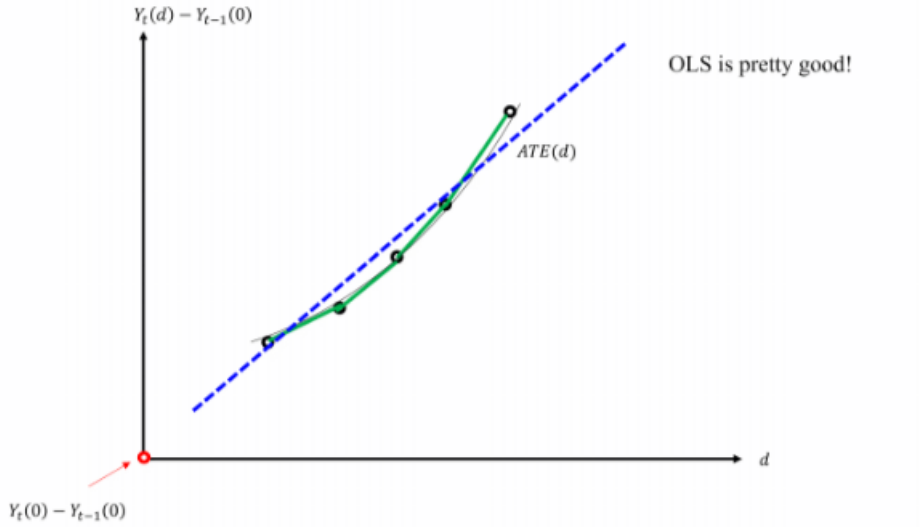
- How do these differ? Consider their interpretations:
  - ATT: effect relative to baseline of zero effect
  - ACRT: effect of marginally increasing treatment, evaluted at a given point (can integrate too!)

- Key point in DiD vs. RCT: these are *conditional* on treatment status, which is not random!

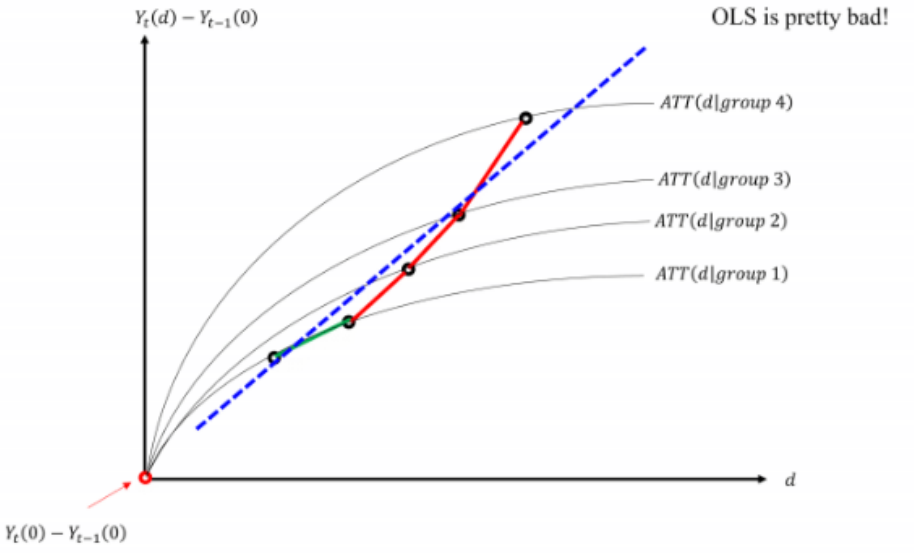
- Can we move between these? Consider comparing two ATTs in 2 period diff-in-diff, assuming standard parallel trends:

$$\begin{aligned} ATT(d|d) - ATT(d'|d') &= (E[\Delta Y|D=d] - E[\Delta Y|D=0]) \\ &\quad - (E[\Delta Y|D=d'] - E[\Delta Y|D=0]) \\ &= E[\Delta Y|D=d] - E[\Delta Y|D=d'] \\ &= E[\Delta Y(d) - \Delta Y(d')|D=d] + ATT(d'|d) - ATT(d'|d') \end{aligned}$$

Continuous treatment – challenge



Continuous treatment – challenge



Continuous treatment – stronger assumption about heterogeneity

- "Strong" parallel trends assumption: $E(\Delta Y_i(d)) = E(\Delta Y_i(d)|D_i = d)$
- No selection bias "on average"
- Effectively assuming some kind of homogeneity in the treatment across groups

Multiple Treatments in Diff-in-diff

- Hull (2018) and deChaisemartin and D'hautefeuille (2022) are relevant cites

- One "easy" context that this shows up (Hull 2018) is "mover" designs
  - I move from city i to city j – if my move was random, can we use it to understand the effect of city j?
  - Without strong additional assumptions, challenging to interpret coefficients from standard TWFE

$$Y_{it} = \alpha_i + \tau_t + \sum_{j \neq 0} \beta_j D_{ijt} + \epsilon_{it} \tag{3}$$

- Consider 2 period case, and use first differences:

$$\Delta Y_i = \tau + \sum_{j \neq 0} \beta_j \Delta D_{ij} + \Delta \epsilon_i \tag{4}$$

- Recall Goldsmith-Pinkham, Hull and Kolesar (2022) – multiple treatments (ΔD_{ij}) is potentially contaminated and negative weighted.

Random Timing in Diff-in-Diff

- Athey and Imbens (2022) take design-based approach to diff-in-diff
  - What does this mean? Can consider a well-defined propensity scores for the treatment
  - Focus on staggered adoption with binary treatment
  - Since treamtent is absorbing, this simplifies problem into "when did I get the treatment (if ever)?"

- E.g. define a propensity score: $Pr(A_i = a)$ over when the assignment occurs.

- Key question: who is the relevant counterfactual group?

Three key assumptions (not testable!)

1. Random assignment (can make this happen by design)
2. No anticipation (we assume this already)
3. Invariance to history ($1_{a \leq t}(Y_{it}(a) - Y_{it}(1)) = 0$) (no causal effect of an early adotpion vs. a later adoption on outcome as long as adoption occured before or on period t)

Efficiency in estimating staggered random roll-out

- Roth and Sant'anna (2023) show that it is much more efficient to condition on lagged outcomes over using standard diff-in-diff in staggered roll-outs

- E.g. Consider the class of estimators:

$$\hat{\theta}_\beta = (\overline{Y}_{22} - \overline{Y}_{2\infty}) - \beta(\overline{Y}_{12} - \overline{Y}_{1\infty}) \tag{5}$$

  Did is the special case of $\beta = 1$.

- We want to put more weight on this setting when the lagged outcomes are more predictive, and less when it is not!
  - Intuition is similar to synthetic controls

Covariates

- With time-invariant covariates that treatment does not affect, controlling for covariates is conditional parallel trends assumption

$$E(Y_{i,2}(0) - Y_{i,1}(0)|D_i = 1, X_i) = E(Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0, X_i) \tag{6}$$

- However, this isn't enough to satisfy this TWFE regression approach in two periods:

$$Y_{it} = \alpha_i + \phi_t + 1(t = 2)D_i\beta + X_i 1(t = 2)\gamma + \epsilon_{it} \tag{7}$$

Why? (consider age)

- Even worse, if covariates affected by treatment, this is a form of collider bias!
  - Should be very careful about time-varying covariates

Checklist

- First, consider the treatment features of your setup
  - What is your treatment? Binary or continuous?
  - Single or multiple?
  - Absorbing or not?
  - How many treated units? Single or many?
  - How many interventions? Single or staggered timing?
- Second, consider the relevant identifying assumptions for your condition
- Third, what tests are feasible in your setting?
- Fourth, what estimates do you want?
- Fifth, what relaxations of assumptions can you make?

Treatment features

- If treatment is binary or only a few values, easy to consider the simple case
  - Particularly helps if it's absorbing!
- If continuous and/or not absorbing, need to worry about many more homogeneity assumptions
- First order of business – *define your estimand*.
  - What treatment do you care about?
- Can you simplify the problem in some way if it's more complex?

Assumptions

- Write out carefully in math and then words what assumptions you need
  - Parallel trends (with or without covariates)?
  - No anticipation?
  - Others?

- If you know your estimand, much easier to know what you need

- Do you need to construct your estimator using lagged outcomes, e.g. synthetic control?

Testable assumptions

- What is testable? Pre-trends can be useful, but underpowered
- If using random timing, not much is testable
  - But could look for balance across timing groups in exogeneous features

What estimates do you want?

- Do you need a long run effect? Short-run?
  - Long-run effect relies *heavily* on parametric extrapolation. Is it plausible?
- Are you studying ATTs? CATTs? ACRTs?

Relaxation of assumptions?

- What group comparaisons are you making for parallel trends?
- Could you assume random timing instead of parallel trends?
- Could you condition on lagged outcomes?