Lab 7: DiD and TWFE

(slides taken from the legendary maestro of panel data Yiqing Xu)

2024-03-29

Plan

- DiD: a quick review
- 2WFE and its assumptions
- The weighting problem

DiD

- Two group (T, C), two periods* (t and t'), fixed treatment timing
- Functional form:

  $$Y_{it} = \tau_{it} D_{it} + \alpha_i + \zeta_t + \varepsilon_{it}$$

  or

  $$\begin{cases} Y_{it}(0) = \alpha_i + \zeta_t + \varepsilon_{it} \\ Y_{it}(1) = Y_{it}(0) + \tau_{it} \end{cases}$$

  where τ_{it} is the treatment effect for unit i at time t; $Y_{it}(0)$ is a combination of two additive fixed effects and idiosyncratic errors
- Parallel trends: $E[Y_{it}(0) - Y_{it}(0) | i \in T] = E[Y_{it'}(0) - Y_{it}(0) | j \in C]$
- Or equivalently, $E[\varepsilon_{it'} - \varepsilon_{it} | i \in T] = E[\varepsilon_{jt'} - \varepsilon_{jt} | j \in C]$
- ATT $= E[\tau_{it} | D_{it} = 1]$ can be non-parametrically identified if there are only two periods (or two treatment histories)

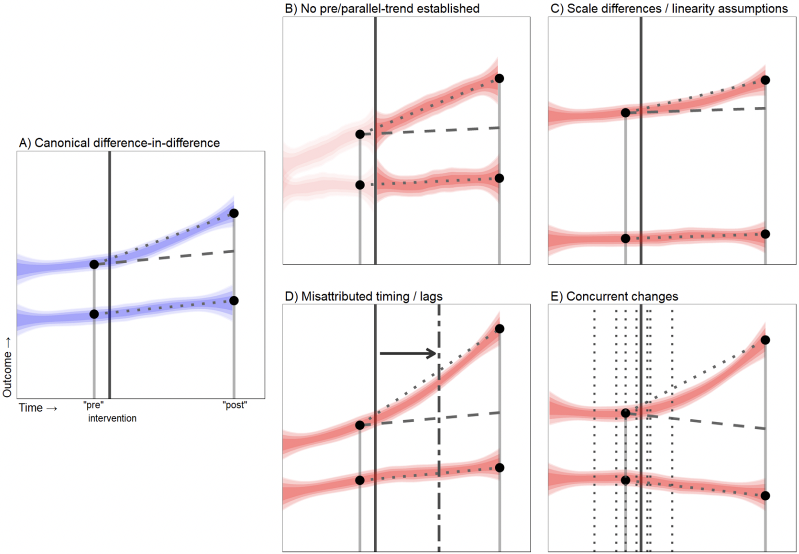
DiD

- Two group (T, C), two periods* (t and t'), fixed treatment timing
- Parallel trends: $E[Y_{it'}(0) - Y_{it}(0) | i \in T] = E[Y_{it'}(0) - Y_{it}(0) | j \in C]$
- Or equivalently, $E[\varepsilon_{it'} - \varepsilon_{it} | i \in T] = E[\varepsilon_{jt'} - \varepsilon_{jt} | j \in C]$
- ATT $= E[\tau_{it} | D_{it} = 1]$ can be non-parametrically identified if there are only two periods (or two treatment histories)

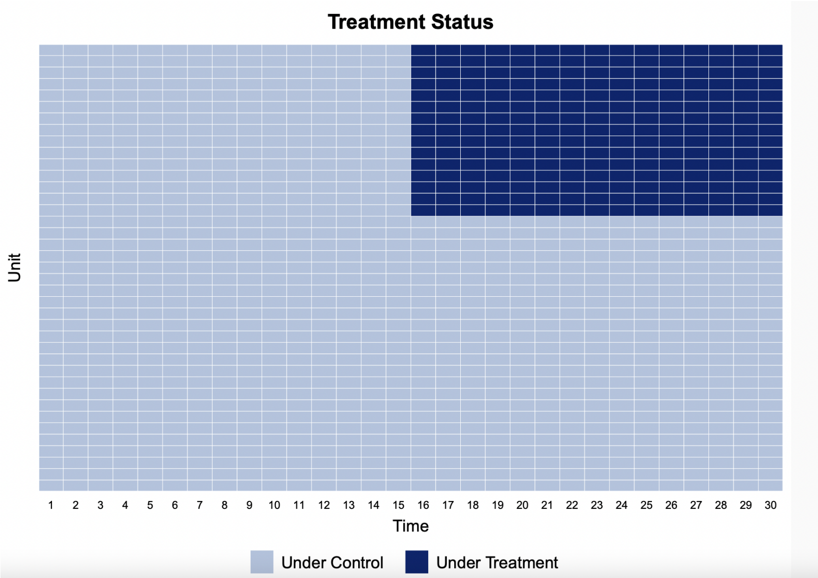
$$\begin{pmatrix} Y^0_{T,pre} & Y^1_{T,post} \\ Y^0_{C,pre} & Y^0_{C,post} \end{pmatrix}$$

$$\begin{pmatrix} Y^0_{T,pre} & ?? \\ Y^0_{C,pre} & Y^0_{C,post} \end{pmatrix}$$

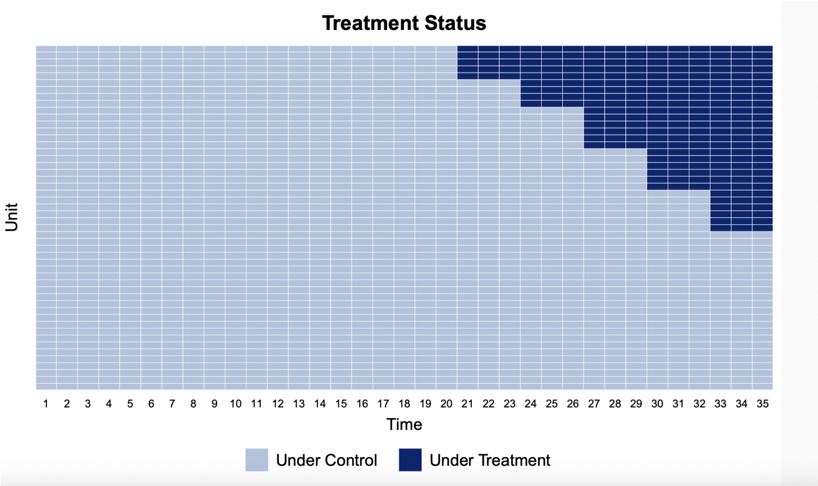
Threat to a DiD Design (Haber et al 2021)



DiD > 2 Periods: Classic



DiD > 2 Periods: Staggered Adoption



DiD from a Design-Based Perspective

Athey & Imbens (2018)

- $i \in \{1, \ldots, N\}, t \in \{1, \ldots, T\}$
- Treatment timing: $\mathcal{A} = \{1, \ldots, T, \infty\}$ (no treatment reversal)
- Treatment assignment: $A_i \in \mathcal{A}$, i.e., $(T+1)$ paths
- Treatment vector: $D_i \equiv \{0, \ldots, 0, 1, \ldots, 1\}$ with the subscript $A_i - 1$ under the first one
- Realized outcome: $Y_{it} \equiv Y_{it}(A_i)$
- All potential outcomes: $Y_{it} \equiv Y_{it}(\mathcal{A})$
- Average causal effect at time t from never getting treated to being treated at time a:

$$\tau_{t,0 \to a} = \frac{1}{N} \sum_{i=1}^{N} (Y_{it}(a) - Y_{it}(\infty))$$

DiD from a Design-Based Perspective

- Possible assumptions
  - Random assignment of A_i (stronger than parallel trends)
  - No anticipation: $Y_{it}(a) = Y_{it}(\infty)$ for any $t < a$
  - Invariance to history: $Y_{it}(a) = Y_{it}(1)$ for any $t \geq a$ (strong)
  - Constant treatment effect over units
  - Constant treatment effect over time
- Different causal quantities can be identified under different assumptions.
- In particular, under random assignment, randomization inference can be used; 2WFE is an unbiased estimator for a weighted average causal effect (more discussion below)

When the Parallel Trends Assumption is More Defensible?

Roth and Sant'Anna (2023)

- The parallel trends assumption is scale-dependent
- When is the assumption not sensitive to strictly monotonic transformation of the outcome?
- A "stronger parallel trends" for the entire distribution of $Y_{it}(0)$

$$F^{Y(0)}_{D=1,t=1}(y) - F^{Y(0)}_{D=1,t=0}(y) = F^{Y(0)}_{D=0,t=1}(y) - F^{Y(0)}_{D=0,t=0}(y), \text{ for all } y \in \mathbb{R}$$

- It holds when the population consists of:
  - A subgroup in which the treatment is as-if randomly assigned
  - A subgroup in which the distribution of $Y_{it}(0)$ is stable over time

Extension: Semi-parametric DiD

Abadie (2005)

- Assumption: non-parallel outcome dynamics between treated and controls caused by observed characteristics
- Two-step strategy:
  1. estimate the propensity score based on observed covariates; compute the fitted value
  2. run a weighted DiD model
- The idea of using pre-treatment variables to adjust trends is a precursor to synthetic control

Strezhnev (2018) extends this approach to incorporate pre-treatment outcomes

Assumptions for 2WFE

$$Y_{it} = \tau D_{it} + X_i' \beta + \alpha_i + \zeta_t + \varepsilon_{it}$$

in which D_{it} is dichotomous

1. Functional form
   - Additive fixed effect
   - *Constant and contemporaneous treatment effect*
   - Linearity in covariates
2. Strict exogeneity

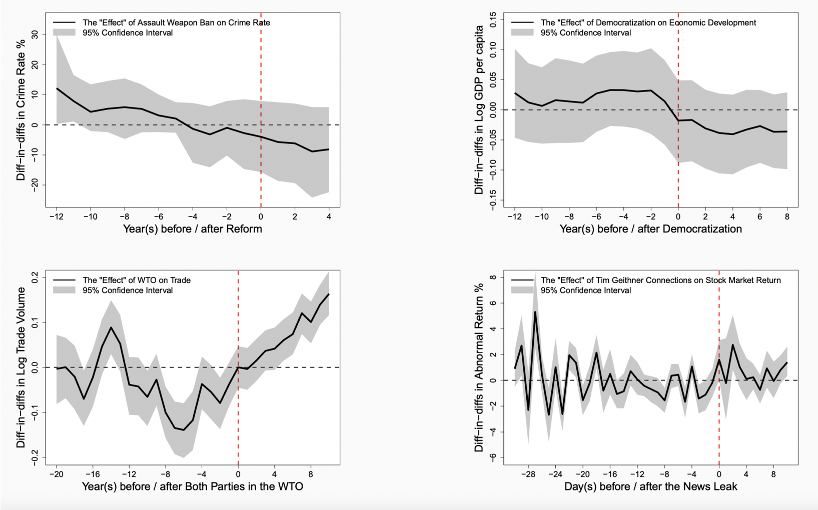
$$\varepsilon_{it} \perp\!\!\!\perp D_{js}, X_{js}, \alpha_j, \zeta_s \quad \forall i, j, t, s$$

$$\Rightarrow \{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp D_{js} | X, \alpha, \zeta \quad \forall i, j, t, s$$

if only two groups, parallel trends:

$$\Rightarrow E[Y_{it}(0) - Y_{it'}(0) | X] = E[Y_{jt}(0) - Y_{jt'}(0) | X] \quad i \in T, j \in C, \forall t, t'$$

Failure of Parallel Trends



What We Don't Know about 2WFE

Goodman-Bacon (2021)

- Most panel applications diverge from this 2x2 set up, because treatments
- We know relatively little about 2WFE when treatment timing varies:
  - Rely on general descriptions of the identifying assumption like random interventions
  - Do not know precisely how it compares mean outcomes across groups
  - Limited understanding of the treatment effect parameter
  - Often cannot evaluate how and why alternative specifications change estimates
- Many related papers recently, e.g., Chernozhukov et al (2017), Borusyak & Jaravel (2017), Strezhnev (2018), Callaway & Sant'Anna (2020), de Chaisemartin & D'Haultfœuille (2020) Imai and Kim (2020)

Goodman-Bacon (2021): 2WFE Decomposition

- The 2WFE estimator under staggered adoption is a weighted average of all possible 2x2 DiD estimators that compare timing groups to each other

$$Y_{it} = \beta^{2WFE} D_{it} + \alpha_i + \zeta_t + \varepsilon_{it}$$

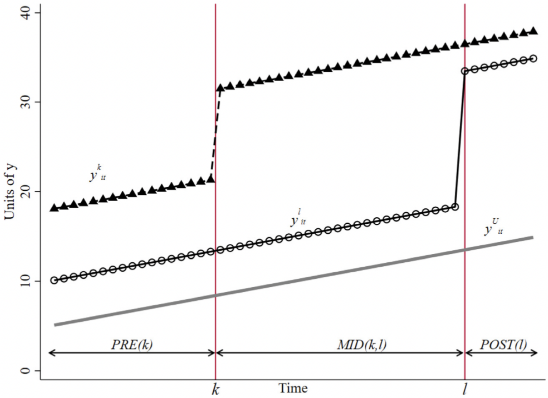
- The weights on the 2x2 DiDs are proportional to timing group sizes and the variance of the treatment dummy in each pair, which is highest for units treated in the middle of the panel.
- Source of biases:

$$\text{plim}_{N \to \infty} \beta^{2WFE} = VWATT + VWCT - \Delta ATT$$

  - VWATT: variance weighted ATT
  - VWCT: variance weighted common trends
  - ΔATT: change in treatment effects over time

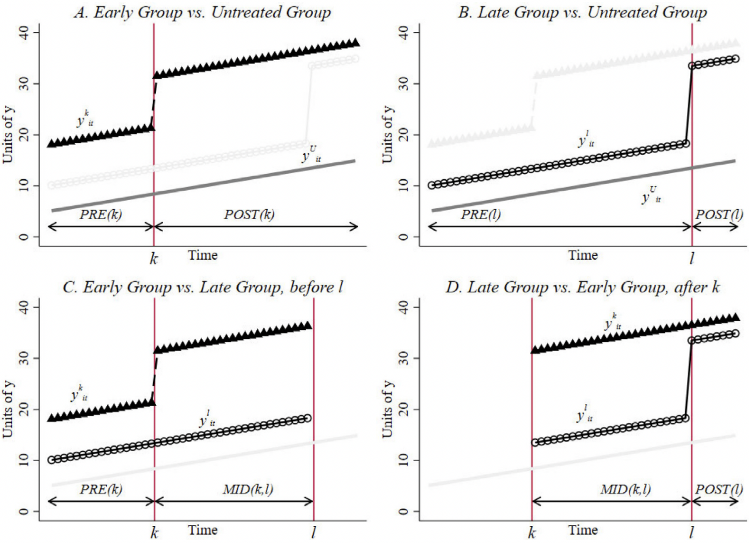
2WFE Decomposition

- An early treatment group k, which receives a binary treatment at $t_i = k$
- A late treatment group ℓ, $t_i = \ell > k$
- An untreated group U, $t_i = \infty$.



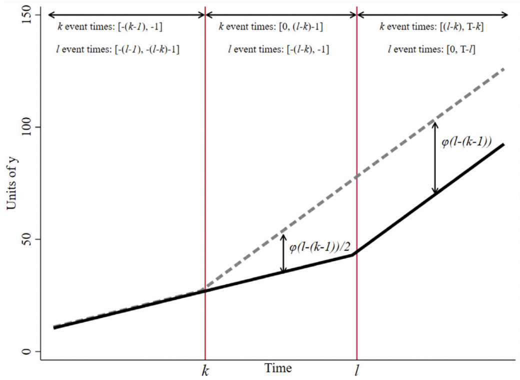
2WFE Decomposition

Four simple (2x2) DiD estimates in the three group case:



The Consequence of Time-Varying Treatment Effects

- Change ATT across all 2x2 DiDs
- Bias estimates away from VWATT because $\Delta ATT \neq 0$
- Recall $\text{plim}_{N\to\infty}\beta^{2WFE} = VWATT + VWCT - \Delta ATT$



So. . .

- The 2WFE estimator (under staggered adoption) only has a meaningful causal interpretation under strong assumptions on treatment effects, i.e., $VWCT = 0$, $\Delta ATT = 0$
    - How to test whether $VWCT = 0$?
    - How to test whether $\Delta ATT = 0$, or avoid this problem all together?
- Even then, it converges to $VWATT$, which may not be what researchers are interested in
- Let's investigate this from a slightly different perspective

The Negative Weighting Problem

de Chaisemartin and D'Haultfœuille (2020)

- Denote π_{it} the treatment effect for unit i at time t
- 2WFE converges to a weighted average of π_{it}

$$E[\beta^{2WFE}] = E\left[\sum_{D_{it}=1} W_{it}\pi_{it}\right]$$

in which $W_{it} = \frac{\hat{\varepsilon}_{it}}{\sum_{D_{it}=1}\hat{\varepsilon}_{it}}$ and $\hat{\varepsilon}_{it}$ is residuals from running D on the fixed effects.

- Smaller weights to periods where more units are treated, and to units with more treated periods
- If staggered adoption, proportion is non-increasing in time. Later periods have smaller (and even negative) weights
- **Problem:** W_{it} can be negative. As a result, even all π_{it} are positive, β^{2WFE} can be negative.

The Negative Weighting Problem

	t = 1	t = 2	t = 3
i = 1	0	0	1
i = 2	0	1	1

$$\beta^{2WFE} = 0.5E[\tau_{13}] + E[\tau_{22}] - 0.5E[\tau_{23}]$$

- if τ_{23} is very large, β^{2WFE} can be negative even if all $\tau_{it} > 0$
- Intuition (Goodman-Bacon): using early adopters as control for late adopters; estimated effect can be affected by over-time changes in treated effects
- Measure of robustness: smallest amount of heterogeneity needed for conditional ATT/ATE to be opposite sign as 2WFE estimand
  - If small, then even very little heterogeneity can be problematic
  - If large, then 2WFE likely robust to realistic levels of heterogeneity (possible efficiency gains from using 2WFE).

What We've Learned So Far

- The parallel trends assumption involves function-form requirements; it is not a weak assumption from a design-based perspective
- 2WFE models require stronger assumptions than we normally admit
- 2WFE estimates can be biased due to (1) presence of time-varying confounders (well-known); (2) feedback from past outcome (known, but often ignored); (3) heterogeneous treatment effects (often completely ignored)
- Robust causal inference using panel data needs to address these issues or relies different identification assumptions, e.g. sequential ignorability