Causal Inference with Panel Data

Lecture 1: Difference-in-Differences and Fixed Effects Models

Yiqing Xu (Stanford University)

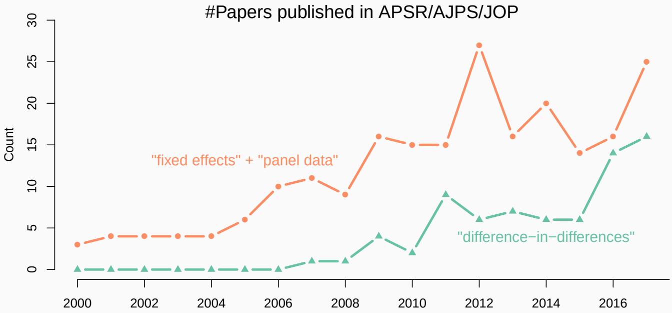
Washington University in St. Louis

23 August 2021

Motivation

Motivations

- Causal inference is challenging with observation data
- Panel data often come to the rescue
- Among panel data methods, difference-in-differences (DiD) and two-way fixed effects (2WFE) are the most popular



Motivations

DiD and 2WFE have benefits:

- Accounting for unobserved unit and time heterogeneity
- Accommodate many types of data structure
- Easy to estimate and interpret

However, lots of reflections recently

- Identification assumptions (e.g., parallel trends, no feedback)
- Modeling choices, esp. treatment effect heterogeneity

This Short “Course”: What We Try to Answer

1. What are the key differences between DiD and 2WFE?
2. What are the main identification regimes with panel data?
3. How to address failure of the parallel trends assumption?
4. How to fix the weighting problem when treatment effects are heterogeneous?
5. How can we do better than the synthetic control method (SCM) in comparative case studies?

What's Special about Panel Data?

- The fundamental problem of causal inference (Holland 1986)

$$\tau_i = Y_{1i} - Y_{i0}$$

- A statistical solution makes use of others' information

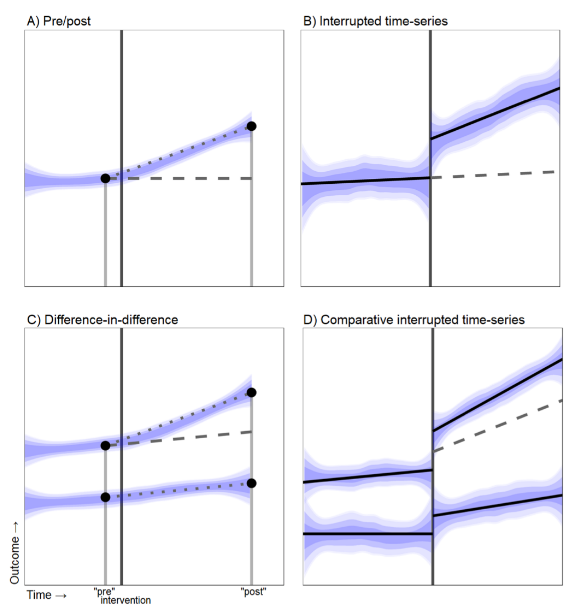
$$\text{e.g. } ATE = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$$

- A scientific solution exploits homogeneity or invariance assumptions

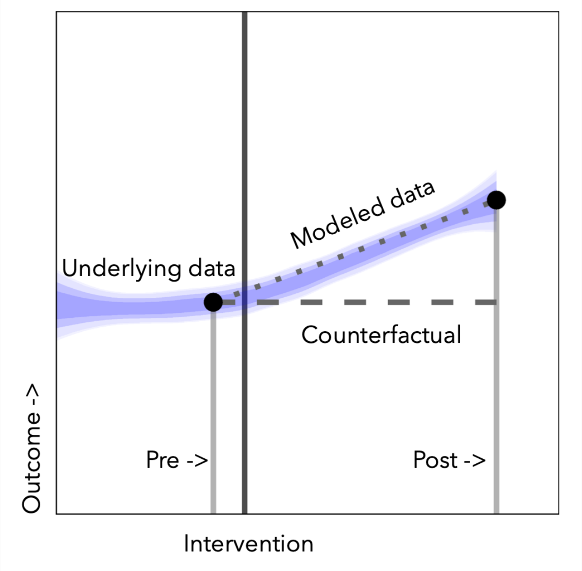
e.g. The long-run growth rate of the US economy is 2.5%.

- Panel data allow us to construct treated counterfactuals using information from both the past and the others with the caveat that treatment assignment mechanism may be complicated
- Panel data is also difficult because of all kinds of interferences (SUTVA violations)

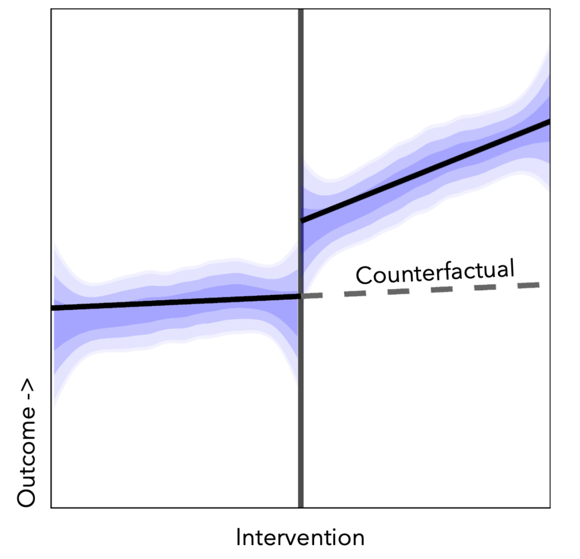
Common Panel Designs (Haber et al 2021)



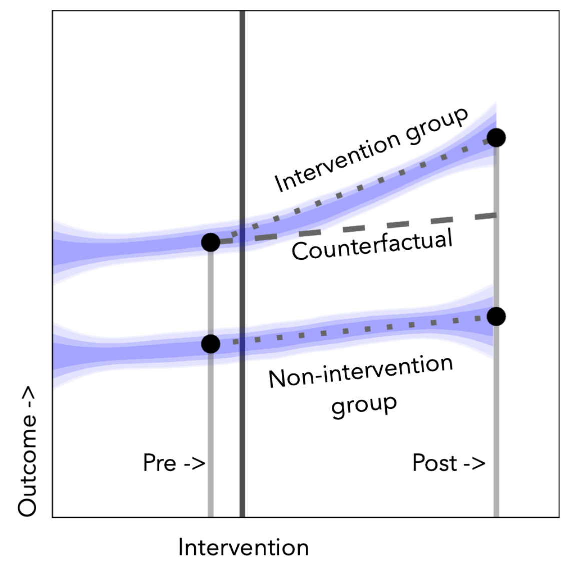
Pre/Post


Underlying data
Modeled data
Counterfactual
Pre ->
Post ->
Outcome ->
Intervention

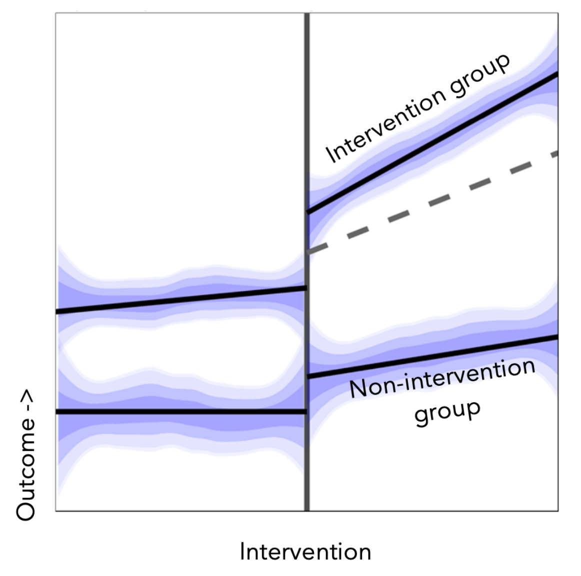
Interrupted Time-Series



DiD


Intervention group
Counterfactual
Non-intervention group
Pre ->
Post ->
Outcome ->
Intervention

Comparative Interrupted Time-Series



Causal Inference with Panel Data

- “Scientific” solution: modeling (but all models are wrong...)
- Statistical solution: extend the conventional ignorability assumption (based on selection on observables)
- Panel data make both easier
  - “Free” testing data (e.g. pretreatment data) → more information for modeling
  - The additional dimension helps relax conventional ignorability
- And we can do more... e.g., taking advantage of the matrix/tensor structure

This Workshop Series

- Lecture 1 – DiD and 2WFE Models
- Lecture 2 – The SCM and Its Extensions
- Lecture 3 – Factor-Augmented Methods
- Lecture 4 – Matching/Reweighting & Hybrid Methods

Today's Plan

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

DiD: A Quick Review

DiD

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

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

or

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

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

DiD

- Two group $(\mathcal{T}, \mathcal{C})$, two periods* ($t$ and t'), fixed treatment timing

- Parallel trends: $\mathbb{E}[Y_{it'}(0) - Y_{it}(0) | i \in \mathcal{T}] = \mathbb{E}[Y_{jt'}(0) - Y_{jt}(0) | j \in \mathcal{C}]$

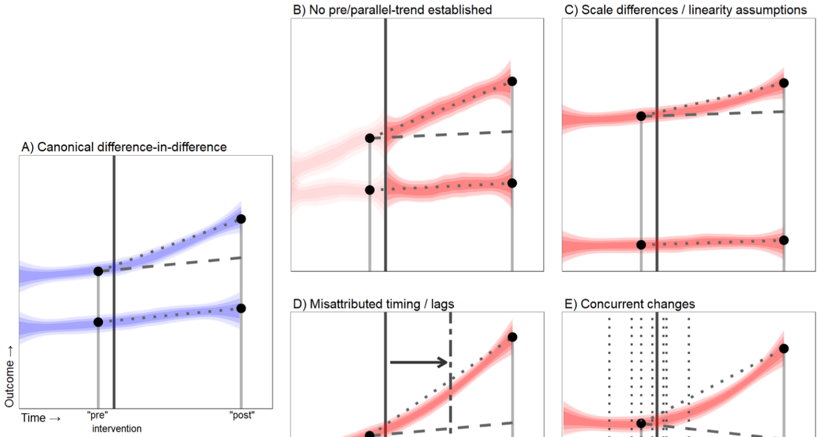
- Or equivalently, $\mathbb{E}[\varepsilon_{it'} - \varepsilon_{it} | i \in \mathcal{T}] = \mathbb{E}[\varepsilon_{jt'} - \varepsilon_{jt} | j \in \mathcal{C}]$

- $ATT = \mathbb{E}[\tau_{it} | D_{it} = 1]$ can be non-parametrically identified if there are only two periods (or two treatment histories)

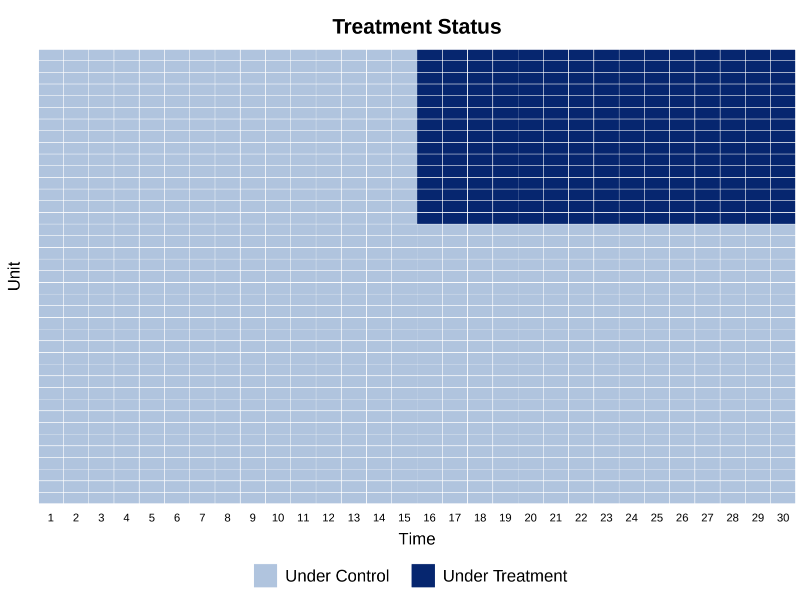
$$\begin{pmatrix} Y^0_{\mathcal{T},pre} & Y^1_{\mathcal{T},post} \\ Y^0_{\mathcal{C},pre} & Y^0_{\mathcal{C},post} \end{pmatrix}$$

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

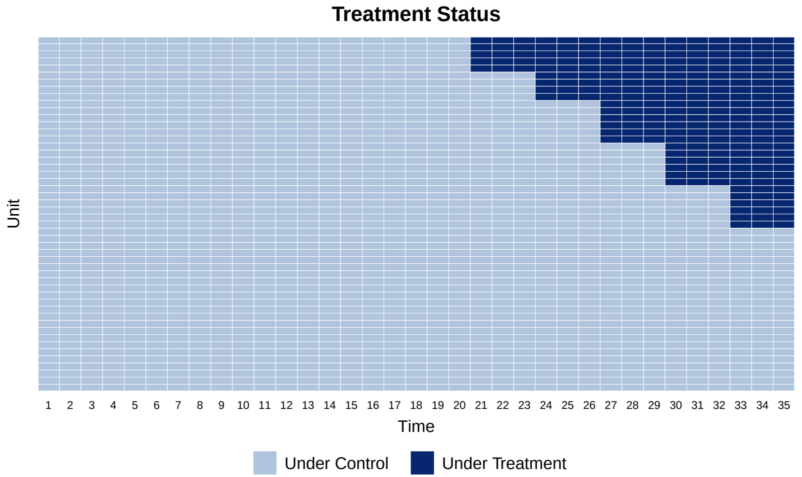
Threat to a DiD Design (Haber et al 2021)



DiD > 2 Periods: Classic



DiD > 2 Periods: Staggered Adoption



DiD from a Deign-Based Perspective

Athey & Imbens (2018)

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

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

DiD from a Deign-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 (2021)

- 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 $Y_{it}(0)$

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

- It holds when the population are 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

2WFE and Its Assumptions

Assumptions for 2WFE

$$Y_{it} = \tau D_{it} + X'\beta + \alpha_i + \xi_t + \varepsilon_{it}$$

in which D_{it} is dichotomous

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

2. Strict exogeneity
$$\begin{aligned} & \varepsilon_{it} \perp\!\!\!\perp D_{js}, X_{js}, \alpha_j, \xi_s \qquad \forall i, j, t, s \\ \Rightarrow \quad & \{Y_{it}(0), Y_{it}(1)\} \perp\!\!\!\perp D_{js} | \boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\xi} \qquad \forall i, j, t, s \end{aligned}$$

   if only two groups, parallel trends:
$$\Rightarrow \quad \mathbb{E}[Y_{it}(0) - Y_{it'}(0)|\boldsymbol{X}] = \mathbb{E}[Y_{jt}(0) - Y_{jt'}(0)|\boldsymbol{X}] \quad i \in \mathcal{T},\ j \in \mathcal{C}, \forall t, t'$$

Shortcomings of 2WFE

$$Y_{it} = \tau D_{it} + X'\beta + \alpha_i + \xi_t + \varepsilon_{it}$$

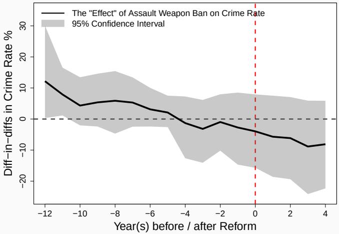
Assumptions

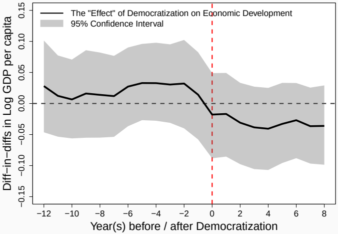
1. Functional form
2. Strict exogeneity

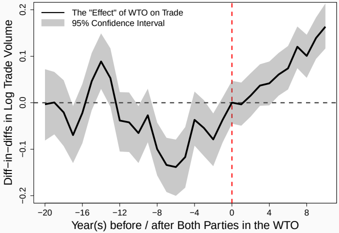
Challenges

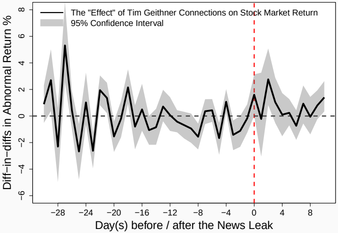
1. Treatment effect heterogeneity leads to bias (more to follow)
2. Prevalent parallel trends failure
3. Strict exogeneity means a lot more than what you think
4. A deeper question: what does fixed effects approach imply from a design-based inference perspective?

Failure of Parallel Trends

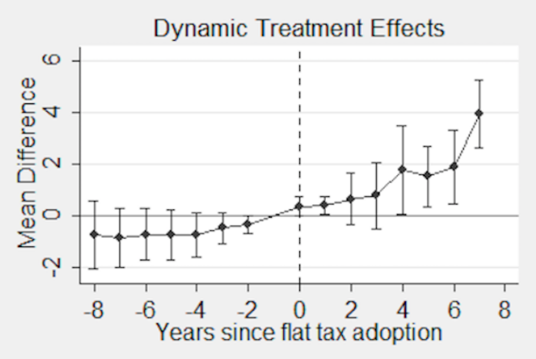








Common Practice: Dynamic Treatment Effect Plots

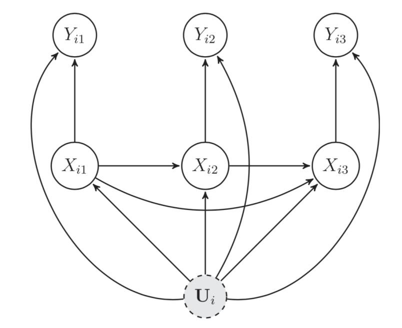


Sun & Abraham (2020)

1. parametric assumptions → biased estimates
2. arbitrarily chosen base category
3. unreliable tests

Reinterpreting Strict Exogeneity (Imai and Kim 2019)

1. No unobserved time-varying confounder exists
2. Past outcomes don't directly affect current outcome (no LDV)
3. Past treatments don't directly affect current outcome (no "carryover effect")
4. Past outcomes don't directly affect current treatment (no "feedback")



Reinterpreting Strict Exogeneity (Imai and Kim 2019)

1. No unobserved time-varying confounder exists
2. Past outcomes don't directly affect current outcome (no LDV)
3. Past treatments don't directly affect current outcome (no "carryover effect")
4. Past outcomes don't directly affect current treatment (no "feedback")

- Violation of 2 alone is fine because past outcomes are not correlated with current treatment; controlling for FEs and LDV simultaneously causes Nickell bias
- To relax 3, "block"/control for past treatments → but how many?
- To relax 4, need instrumental variables (Arellano and Bond 1991) → hard to justify instruments; bad finite sample properties
- Often end up directly controlling for arbitrary number of past treatments and LDVs → Nickel bias

Shortcomings of 2WFE

$$Y_{it} = \tau D_{it} + X'\beta + \alpha_i + \xi_t + \varepsilon_{it}$$

Assumptions

1. Functional form
2. Strict exogeneity

Challenges

1. Treatment effect heterogeneity leads to bias (more to follow)
2. Prevalent parallel trends failure
3. Strict exogeneity means a lot more than what you think
4. A deeper question: what does fixed effects approach imply from a design-based inference perspective?$\rightarrow$ hypothetical experiment?

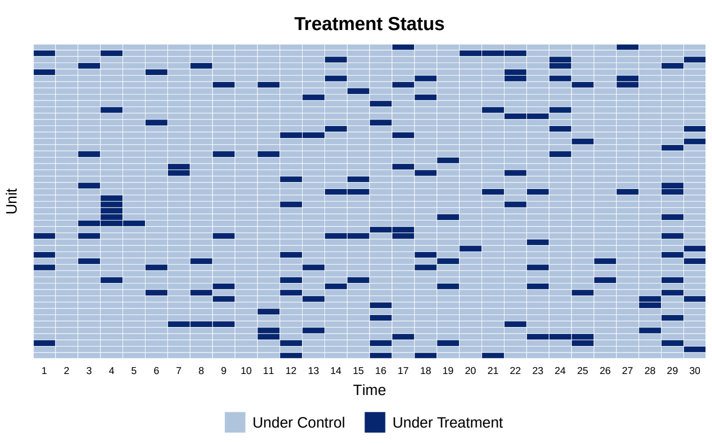
Hypothetical Experiment?

DGPs consistent with strict exogeneity:

$$\alpha_i, \boldsymbol{X_i} \rightarrow \boldsymbol{D_i} \rightarrow \boldsymbol{Y_i}$$

treatment status are assigned randomly or at one shot, not sequentially!

Examples: random assignment within units



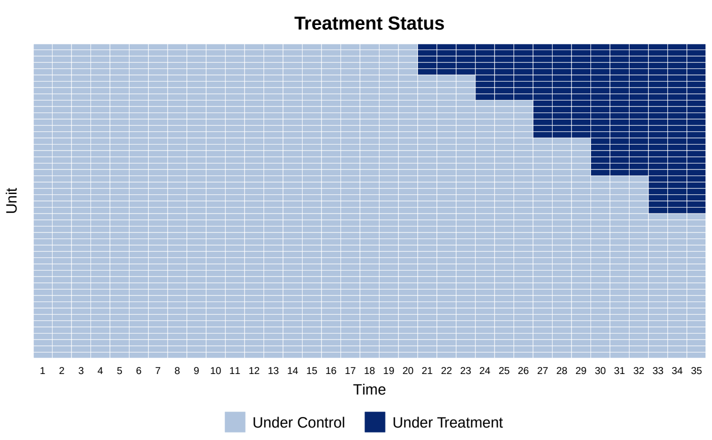
Hypothetical Experiment?

Strict exogeneity implies the following data generating processes:

$$\alpha_i, \boldsymbol{X_i} \rightarrow A_i \rightarrow \boldsymbol{D_i} \rightarrow \boldsymbol{Y_i}$$

treatment status are assigned randomly or at one shot, not sequentially!

Examples: staggered adoption (Athey and Imbens 2018)



The Weighting Problem

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

$$Y_{it} = \beta^{2WFE} D_{it} + \alpha_i + \xi_t + \epsilon_{it}$$

- 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:

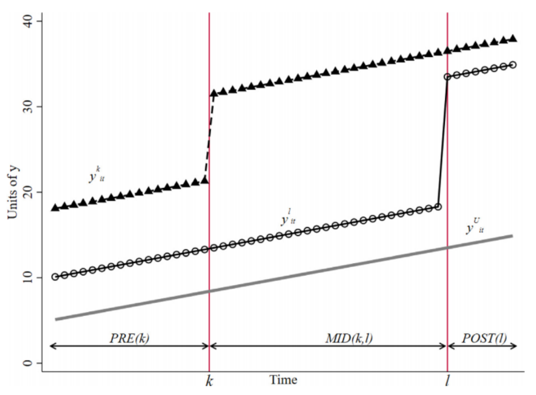
$$\text{plim}_{N\to\infty}\hat{\beta}^{2WFE} = VWATT + VWCT - \Delta ATT$$

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

2WFE Decomposition

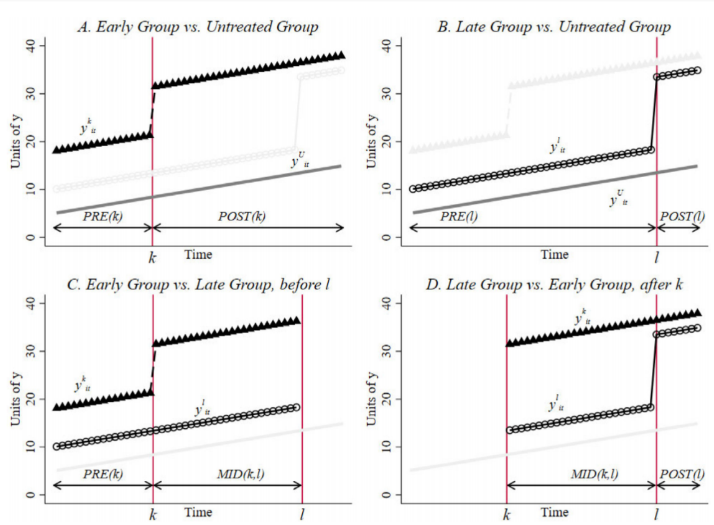
Subgroups under staggered adoption

- 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, $t_i = \infty$.



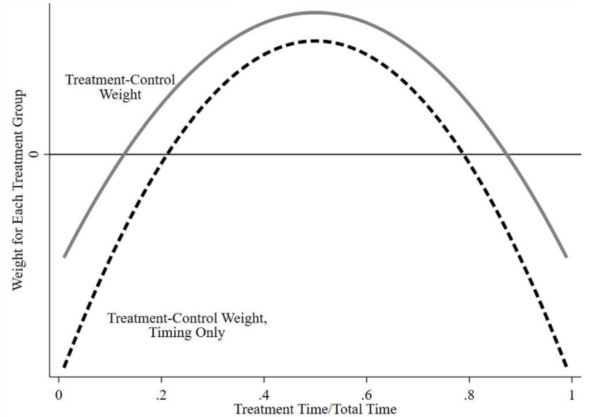
2WFE Decomposition

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



The Intuition for *VWATT* and *VWCT*

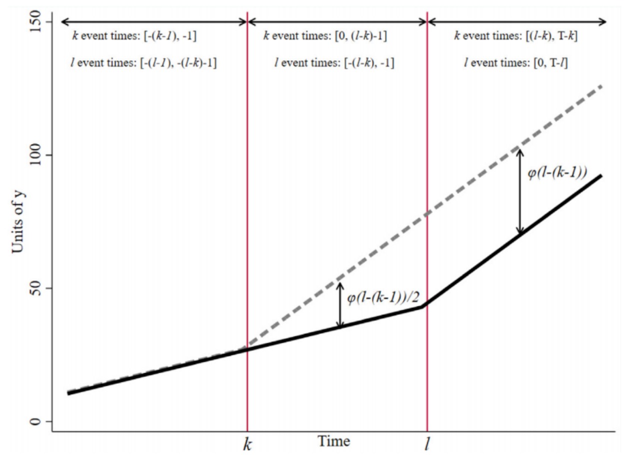
- Plot $w_k^T - w_k^C$ as a function of $\bar{D}$ assuming equal group sizes.



- Units treated in the middle get more weight as treated
- Units treated at the beginning or toward the ends get more weight as controls

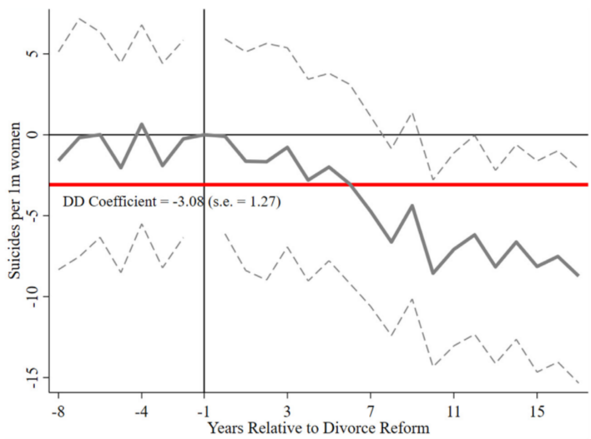
The Consequence of Time-Varying Treatment Effects

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



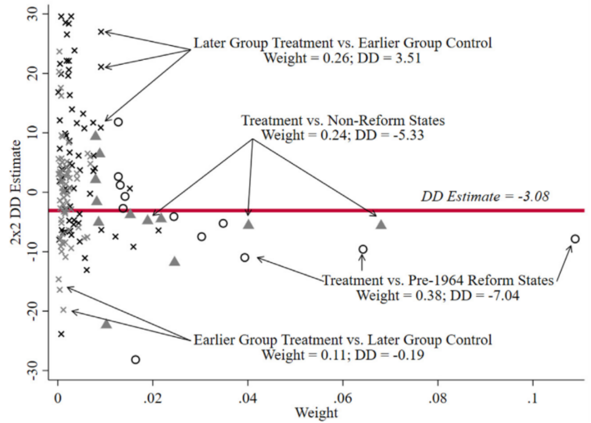
Example: No-Fault Divorce Reforms and Female Suicide

- 37 states from 1969-1985
- Event-study and 2WFE estimates:



2WFE Decomposition

- Plot each 2x2 DiD against its weight and calculate the average effect and total weight for each type of 2x2 comparison:



- The two-way fixed effects estimate, -3.08, is an average of the *y*-axis values weighted by their *x*-axis values.

Summary

- 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'Haultfoeuille (2020)

- Denote τ_{it} the treatment effect for unit i at time t
- 2WFE converges to a weighted average of τ_{it}
$$\mathbb{E}[\hat{\beta}^{2WFE}] = \mathbb{E}\left[\sum_{D_{it}=1} w_{it}\tau_{it}\right]$$
in which $w_{it} = \frac{\hat{\epsilon}_{it}}{\sum_{D_{it}=1}\hat{\epsilon}_{it}}$ and $\hat{\epsilon}_{it}$ 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 τ_{it} are positive, $\hat{\beta}^{2WFE}$ can be negative.

The Negative Weighting Problem

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

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

- 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

DiD_M

- Focus on "joiner" ($D_{i,t-1} = 0, D_{it} = 1$) and "leavers" ($D_{i,t-1} = 1, D_{it} = 0$), assuming their comparison groups (0 0 and 1 1) exist
- Local estimators

$$DiD_{+,t} = \sum_{g:D_{g,t}=1,D_{g,t-1}=0} \frac{N_{g,t}}{N_{1,0,t}}(Y_{g,t} - Y_{g,t-1}) - \sum_{g:D_{g,t}=D_{g,t-1}=0} \frac{N_{g,t}}{N_{0,0,t}}(Y_{g,t} - Y_{g,t-1})$$

$$DiD_{-,t} = \sum_{g:D_{g,t}=D_{g,t-1}=1} \frac{N_{g,t}}{N_{1,1,t}}(Y_{g,t} - Y_{g,t-1}) - \sum_{g:D_{g,t}=0,D_{g,t-1}=1} \frac{N_{g,t}}{N_{0,1,t}}(Y_{g,t} - Y_{g,t-1})$$

- Their weighted sum

$$DiD_M = \sum_{2}^{T} \left(\frac{N_{1,0,t}}{N_S} DiD_{+,t} + \frac{N_{0,1,t}}{N_S} DiD_{-,t} \right)$$

- Placebo test: whether Y changes from $t-k-1$ to $t-k$ in groups that switch or do not switch from $t-1$ to t

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

References

- Haber, Noah A., Emma Clarke-Deelder, Joshua A. Salomon, Avi Feller, and Elizabeth A. Stuart. 2021. "COVID-19 Policy Impact Evaluation: A Guide to Common Design Issues." *A*merican Journal of Epidemiology, June. https://doi.org/10.1093/aje/kwab185.
- Athey, Susan, and Guido W. Imbens. 2018. "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption." Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w24963.
- Roth, Jonathan, and Pedro H. C. Sant'Anna. n.d. "When Is Parallel Trends Sensitive to Functional Form?" arXiv:2010.04814.
- Abadie, Alberto (2005). "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies* (2005) 72, 1–19.
- Imai, Kosuke and In Song Kim (2019). "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science*, Vol. 62, Iss. 2, April 2019, pp. 467–490.
- Sun, Liyang, and Sarah Abraham. 2018. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." arXiv [econ.EM]. arXiv. http://arxiv.org/abs/1804.05785.
- Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics*, June.
- Chaisemartin, Clément de, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *The American Economic Review* 110 (9): 2964–96.