Causal Inference using Difference-in-Differences

Lecture 2: Classical 2x2 DiD Setup

Pedro H. C. Sant'Anna
Emory University

January 2025

Summary of previous lecture

Summary of Lecture 1

- We have highlighted DiD's popularity and its practical appeal;
- We have discussed Potential Outcomes;
- We have also talked about causal parameters of interest.

The canonical 2 × 2 Difference-in-Differences estimator

- **The canonical 2 × 2 DiD estimator** is given by

$$\widehat{\theta}_n^{DiD} = \left(\overline{Y}_{g=treated,t=post} - \overline{Y}_{g=treated,t=pre}\right) - \left(\overline{Y}_{g=untreated,t=post} - \overline{Y}_{g=untreated,t=pre}\right),$$

where $\overline{Y}_{g=d,t=j}$ is the sample mean of the outcome Y for units in group d in time period j,

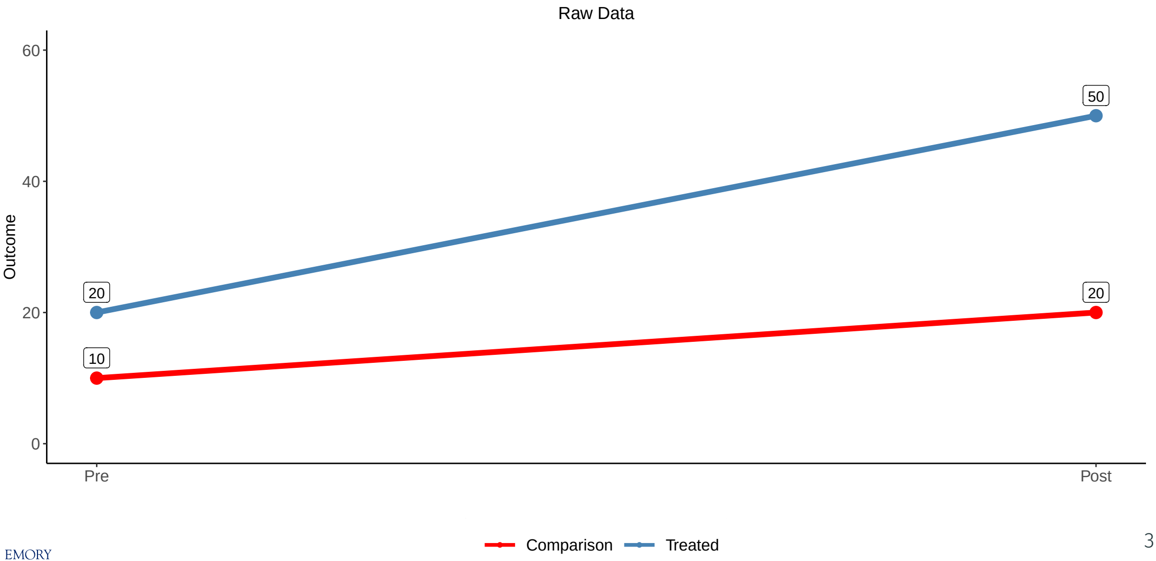
$$\overline{Y}_{g=d,t=j} = \frac{1}{N_{g=d,t=j}} \sum_{i=1}^{N_{all}} Y_i 1\{G_i = d\}1\{T_i = j\},$$

with

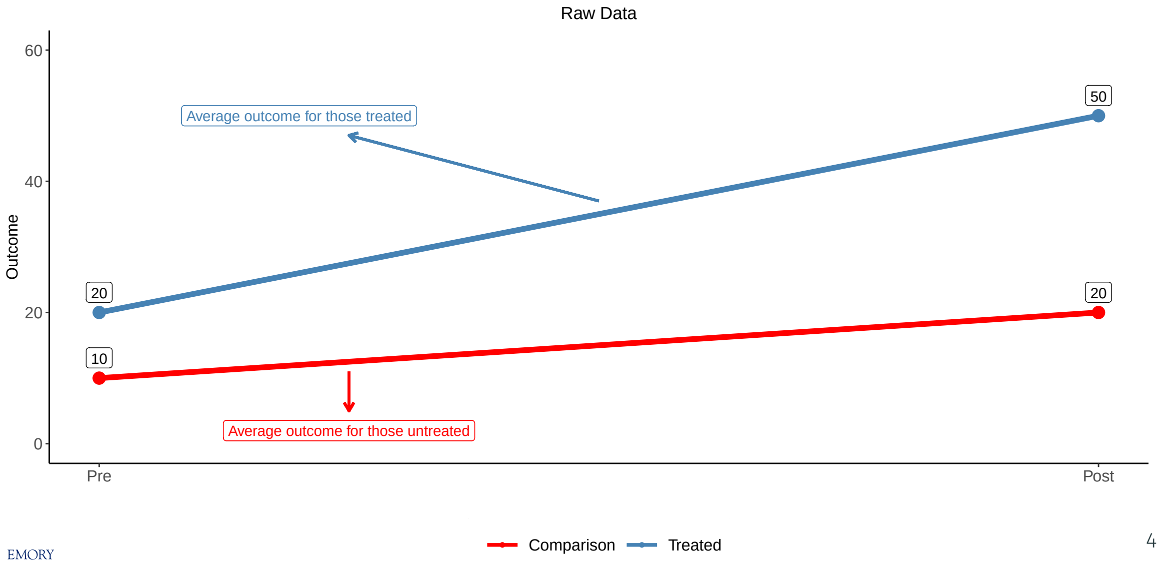
$$N_{g=d,t=j} = \sum_{i=1}^{N_{all}} 1\{G_i = d\}1\{T_i = j\},$$

G_i and T_i are group and time dummy, respectively, and Y_i is the "pooled" outcome data.

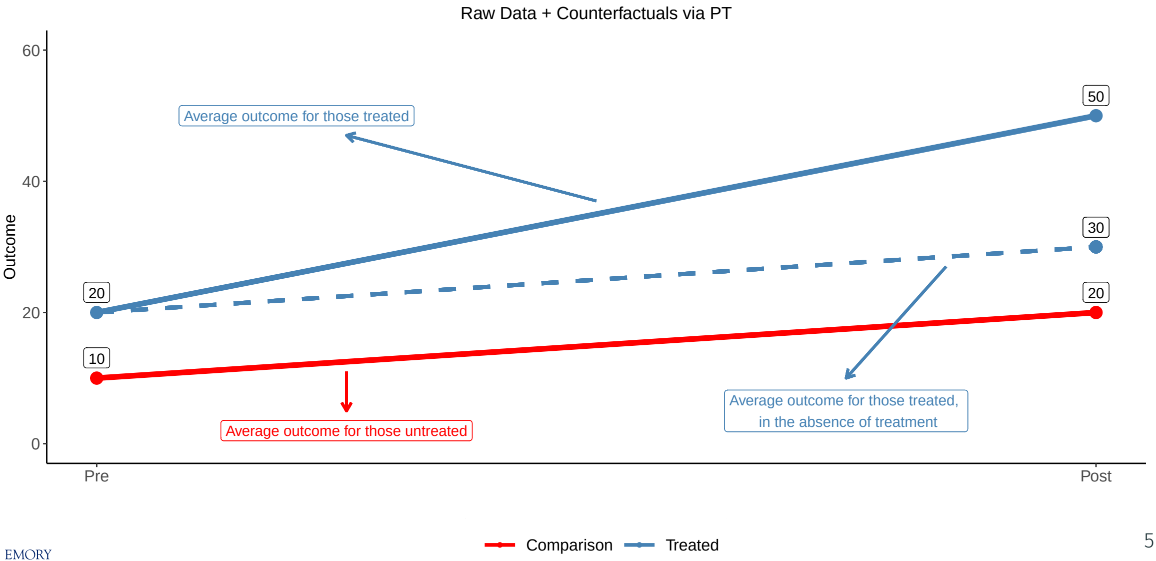
Difference-in-Differences via graphs



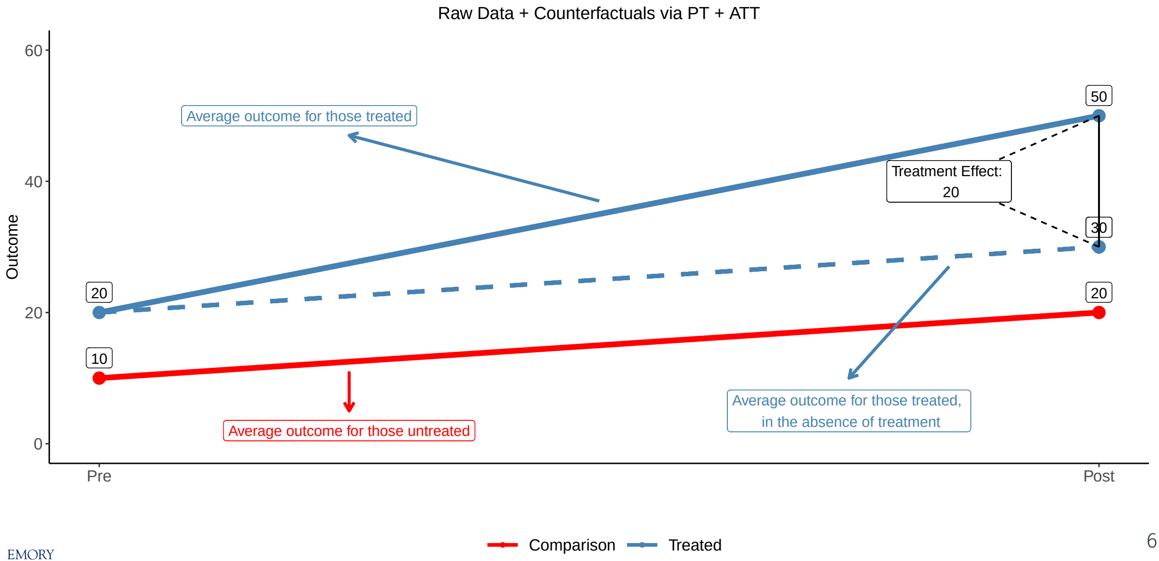
Difference-in-Differences via graphs



Difference-in-Differences via graphs



Difference-in-Differences via graphs



But what kind of treatment effect

parameter $\widehat{\theta}_n^{DiD}$ is actually recovering?

We need to talk about:

1. Potential outcomes

We've talked about this in Lecture 1

2. Assumptions

We will now zoom into this!

We will focus into the 2×2 DiD setup

SUTVA and No-Anticipation Assumption

Stable Unit Treatment Value Assumption (SUTVA)

Assumption (SUTVA)

Observed outcomes at time t are realized as

$$Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g).$$

- In the 2x2 DiD case, observed outcomes at time t are realized as

$$Y_{i,t} = 1\{G_i = 2\} Y_{i,t}(2) + 1\{G_i = \infty\} Y_{i,t}(\infty).$$

Stable Unit Treatment Value Assumption (SUTVA)

Assumption (SUTVA)

Observed outcomes at time t are realized as

$$Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g).$$

- Implicitly implies that potential outcomes for unit *i* are not affected by the treatment of unit *j*.
    - Rules out interference across units
    - Rules out spillover effects
    - Rules out general equilibrium effects

Stable Unit Treatment Value Assumption (SUTVA)

Assumption (SUTVA)

Observed outcomes at time t are realized as

$$Y_{i,t} = \sum_{g \in \mathcal{G}} 1\{G_i = g\} Y_{i,t}(g).$$

- This assumption may be problematic in some applications
- **We should choose the units of analysis to minimize interference across units.**

Are there “causal effects” before treatment takes place?

No-Anticipation Assumption

Assumption (No-Anticipation)

For all units i, $Y_{i,t}(g) = Y_{i,t}(\infty)$ for all groups in their pre-treatment periods, i.e., for all $t < g$.

- Common assumption in duration analysis (Abbring and van den Berg, 2003; Sianesi, 2004).
- This assumption says that unit-specific treatment effects are zero in all **pre-treatment periods**.
- It does not restrict treatment effect heterogeneity in **post-treatment periods**.
- This is plausible in many setups, especially if treatment is not announced in advance.
- But it is not innocuous (Malani and Reif, 2015).

No-Anticipation Assumption

Assumption (No-Anticipation)

For all units i, $Y_{i,t}(g) = Y_{i,t}(\infty)$ for all groups in their pre-treatment periods, i.e., for all $t < g$.

- This assumption also allows us to “simplify” notation.
- Replace all “untreated” (or ”not-yet-treated”) potential outcomes by $Y_{i,t}(\infty)$
- Many times, this assumption is already ”baked” into the potential outcome notation (replace $Y_{i,t}(\infty)$ with $Y_{i,t}(0)$ in all pre-treatment periods).
- I prefer to be explicit about assumptions to enforce transparency.

Fundamental problem of causal inference in the 2×2 DiD setup

	Data				
Unit	$Y_{i,t=1}(2)$	$Y_{i,t=2}(2)$	$Y_{i,t=1}(\infty)$	$Y_{i,t=2}(\infty)$	G_i
1	?	?	✓	✓	∞
2	✓	✓	?	?	2
3	?	?	✓	✓	∞
4	✓	✓	?	?	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	✓	✓	?	?	2

✓: Observed data

?: Missing data (unobserved counterfactuals)

Missing Data + SUTVA + No-Anticipation (1)

	Data				
Unit	$Y_{i,t=1}(2)$	$Y_{i,t=2}(2)$	$Y_{i,t=1}(\infty)$	$Y_{i,t=2}(\infty)$	G_i
1	✓	?	✓	✓	∞
2	✓	✓	✓	?	2
3	✓	?	✓	✓	∞
4	✓	✓	✓	?	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	✓	✓	✓	?	2

✓: Observed data

?: Missing data (unobserved counterfactuals)

Missing Data + SUTVA + No-Anticipation (2)

	Data				
Unit	$Y_{i,t=1}(\infty)$	$Y_{i,t=2}(2)$	$Y_{i,t=1}(\infty)$	$Y_{i,t=2}(\infty)$	G_i
1	✓	?	✓	✓	∞
2	✓	✓	✓	?	2
3	✓	?	✓	✓	∞
4	✓	✓	✓	?	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	✓	✓	✓	?	2

✓: Observed data

?: Missing data (unobserved counterfactuals)

Missing Data + SUTVA + No-Anticipation (3)

	Data			
Unit	$Y_{i,t=1}(\infty)$	$Y_{i,t=2}(2)$	$Y_{i,t=2}(\infty)$	G_i
1	✓	?	✓	∞
2	✓	✓	?	2
3	✓	?	✓	∞
4	✓	✓	?	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	✓	✓	?	2

✓: Observed data

?: Missing data (unobserved counterfactuals)

Missing Data + SUTVA + No-Anticipation (alternative, more classical notation)

			Data		
Unit	$Y_{i,t=1}(0)$	$Y_{i,t=2}(1)$		$Y_{i,t=2}(0)$	D_i
1	✓	?		✓	0
2	✓	✓		?	1
3	✓	?		✓	0
4	✓	✓		?	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
n	✓	✓		?	1

✓: Observed data

?: Missing data (unobserved counterfactuals)

Selection bias

Selection bias

Problem:

Comparison of outcomes at $t = 2$ between the treated and the untreated units do not usually give the right answer.

$$\begin{aligned}\mathbb{E}\left[Y_{i,t=2}|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}|G_i = \infty\right] &= \mathbb{E}\left[Y_{i,t=2}(2)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right] \\ &= \mathbb{E}\left[Y_{i,t=2}(2) - Y_{i,t=2}(\infty)|G_i = 2\right] \\ &\quad + \left(\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right]\right) \\ &= \text{ATT} + \text{Selection bias}\end{aligned}$$

- Selection bias term unlikely to be zero in most applications.
- Selection into treatment is often associated with the potential outcomes.

Selection bias

Example: A job training program for disadvantaged

- Participants are self-selected from a subpopulation of individuals in difficult labor situations

- Post-training period earnings for participants would be lower than those for nonparticipants in the absence of the program:

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right] < 0$$

Selection bias in the stylized example

Unit	$Y_{i,t=1}(\infty)$	$Y_{i,t=2}(2)$	$Y_{i,t=2}(\infty)$	$Y_{i,t=2}$	G_i	$Y_{i,t=2}(2) - Y_{i,t=2}(\infty)$
1	3	3	4	4	∞	-1
2	4	6	6	6	2	0
3	1	5	3	3	∞	2
4	1	7	2	7	2	5
$\mathbb{E}\left[Y_{i,t=2} \mid G_i = 2\right]$				6.5		
$\mathbb{E}\left[Y_{i,t=2} \mid G_i = \infty\right]$				3.5		
$\mathbb{E}\left[Y_{i,t=2}(\infty) \mid G_i = 2\right]$			4			
$\mathbb{E}\left[Y_{i,t=2}(\infty) \mid G_i = \infty\right]$			3.5			
$\mathbb{E}\left[Y_{i,t=2} \mid G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2} \mid G_i = \infty\right]$				3		
$\mathbb{E}\left[Y_{i,t=2}(\infty) \mid G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}(\infty) \mid G_i = \infty\right]$			0.5			
$\mathbb{E}\left[Y_{i,t=2}(2) - Y_{i,t=2}(\infty) \mid G_i = 2\right]$						2.5

✓: Observed data

?: Missing data (unobserved counterfactuals)

Can we exploit the time dimension

to tackle “selection bias”?

Parallel Trends Assumption

Parallel Trends Assumption - 2x2 setup

Since a simple comparison of means at time $t = 2$ does not recover a parameter of interest (ATT), we can take a different route.

Assumption (Parallel Trends Assumption)

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = 2\right] = \mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = \infty\right]$$

The parallel trends (PT) assumption states that, in the absence of treatment, the evolution of the outcome among the treated units is, on average, the same as the evolution among the untreated units.

But how can the parallel trends assumption help us?

Parallel Trends and the ATT

- We will start from the perspective that the ATT at time $t = 2$ is the target parameter.
- From the definition of the ATT and SUTVA, we have

$$\begin{aligned} ATT &\equiv \mathbb{E}\left[Y_{i,t=2}\left(2\right)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i = 2\right] \\ &= \underbrace{\mathbb{E}\left[Y_{i,t=2}|G_i = 2\right]}_{by\ SUTVA} - \mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i = 2\right] \end{aligned}$$

- Green object is estimable from data (under SUTVA).
- Red object still depends on potential outcomes, and we aim to find ways to "impute" it.
- This is where PT comes into play!

Parallel Trends and the ATT

1) First, recall the PT assumption:

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = 2\right] = \mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}(\infty)|G_i = \infty\right].$$

2) By simple manipulation, we can write it as

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] \quad = \quad \mathbb{E}\left[Y_{i,t=1}\left(\infty\right)|G_i = 2\right] + \left(\mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}\left(\infty\right)|G_i = \infty\right]\right)$$

3) Now, exploiting No-Anticipation and SUTVA:

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] \quad = \quad \underbrace{\mathbb{E}\left[Y_{i,t=1}\left(2\right)|G_i = 2\right]}_{by\ No-Anticipation} + \left(\mathbb{E}\left[Y_{i,t=2}\left(\infty\right)|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}\left(\infty\right)|G_i = \infty\right]\right)$$

$$\mathbb{E}\left[Y_{i,t=2}(\infty)|G_i = 2\right] \quad = \quad \underbrace{\mathbb{E}\left[Y_{i,t=1}|G_i = 2\right] + \left(\mathbb{E}\left[Y_{i,t=2}|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}|G_i = \infty\right]\right)}_{by\ SUTVA}$$

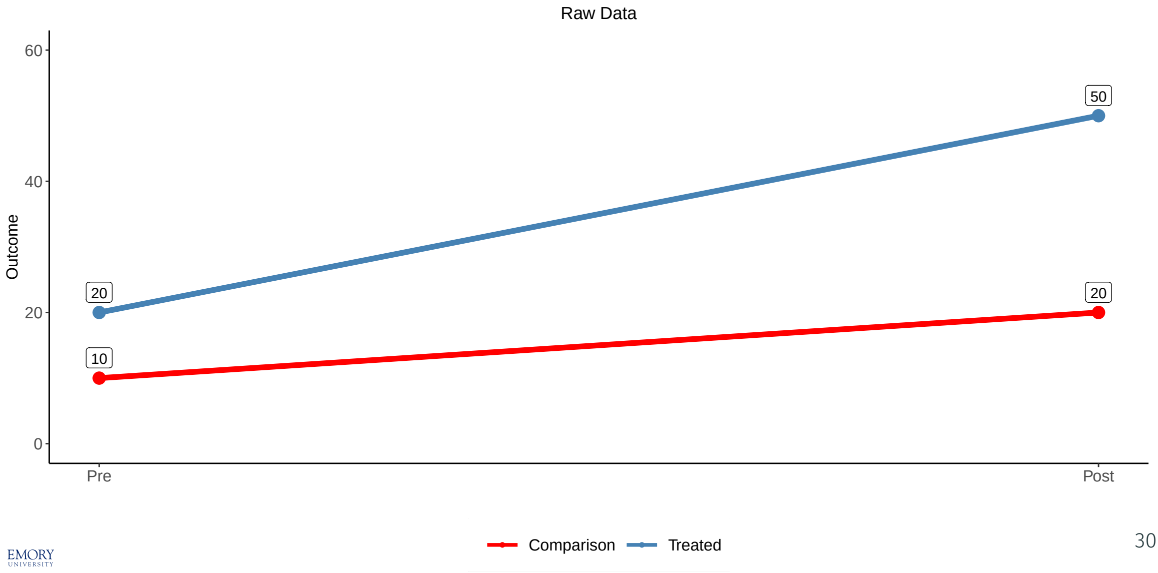
Parallel Trends and the ATT

- Combining these results, we have that, under SUTVA + No-Anticipation + PT assumptions, it follows that

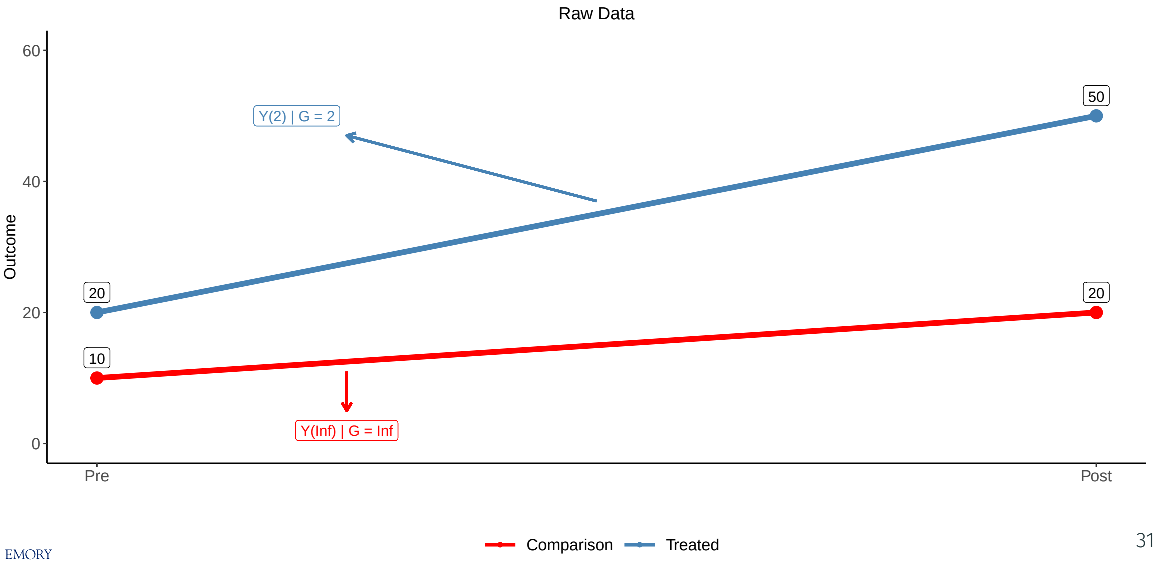
$$\begin{aligned}
\textbf{ATT} &= \mathbb{E}\left[Y_{i,t=2}|G_i = 2\right] - \left(\mathbb{E}\left[Y_{i,t=1}|G_i = 2\right] + \left(\mathbb{E}\left[Y_{i,t=2}|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}|G_i = \infty\right]\right)\right) \\
&= \left(\mathbb{E}\left[Y_{i,t=2}|G_i = 2\right] - \mathbb{E}\left[Y_{i,t=1}|G_i = 2\right]\right) - \left(\mathbb{E}\left[Y_{i,t=2}|G_i = \infty\right] - \mathbb{E}\left[Y_{i,t=1}|G_i = \infty\right]\right)
\end{aligned}$$

- This is “the birth” of the DiD estimand!

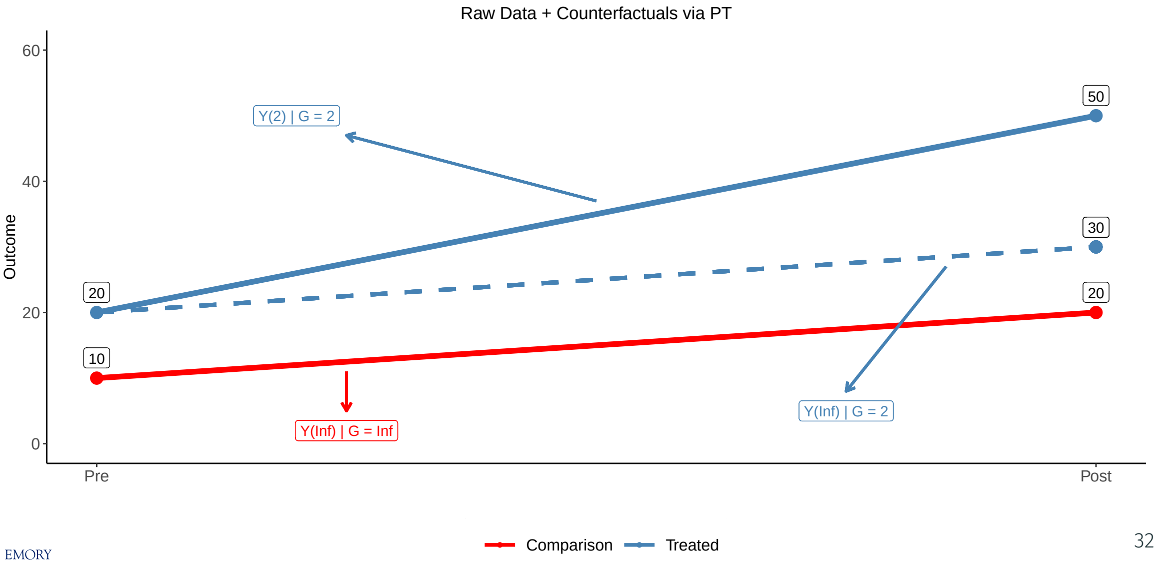
Parallel Trends via graphs



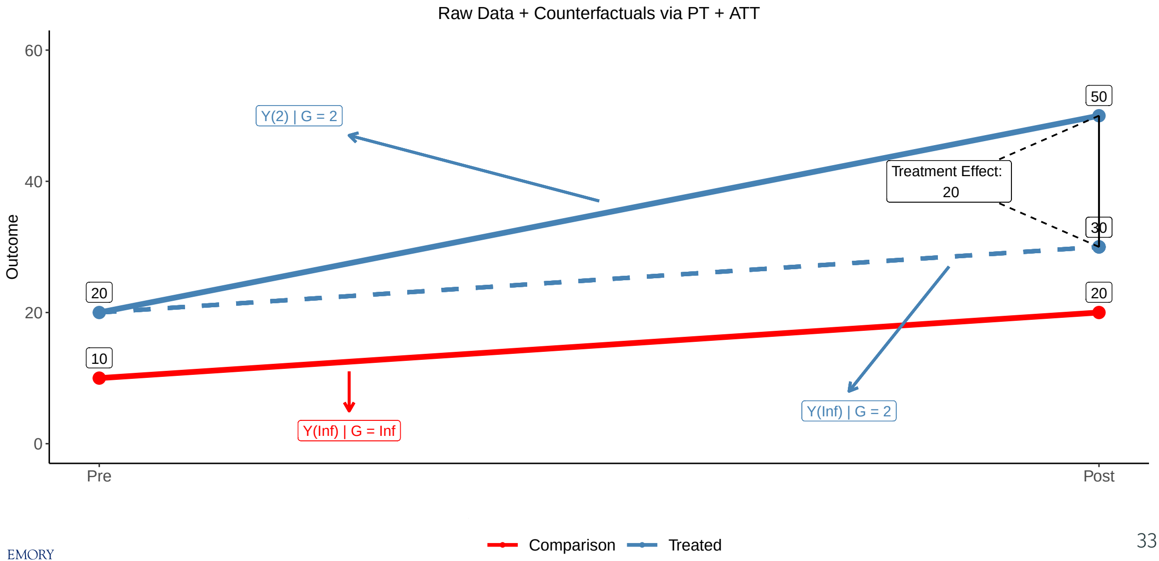
Parallel Trends via graphs



Parallel Trends via graphs



Parallel Trends via graphs



But how can we actually estimate and make inference about the ATT?

Estimating the ATT in the 2x2 DiD Setup

Estimating the ATT in 2x2 DiD setups

- Up to now, we have only shown that the ATT is **identified** under SUTVA + No-anticipation + PT assumptions.
- But our estimand involves **population** expectations, and, in practice, we do not really know the true DGP such that we can compute them.
- However, we can **estimate** them using the "analogy principle": replace population expectations by their sample analogs

“Brute-force” DiD estimator in 2x2 setups

- By using the analogy (or plug-in) principle, we have that our canonical DiD **estimator** for the ATT is given by

$$\widehat{\theta}_n^{DiD} = \left(\overline{Y}_{g=2,t=2} - \overline{Y}_{g=2,t=1}\right) - \left(\overline{Y}_{g=\infty,t=2} - \overline{Y}_{g=\infty,t=1}\right),$$

where $\overline{Y}_{g=d,t=j}$ is the sample mean of the outcome Y for units in group d in time period j,

$$\overline{Y}_{g=d,t=j} = \frac{1}{N_{g=d,t=j}} \sum_{i=1}^{N \cdot T} Y_i 1\{G_i = d\} 1\{T_i = j\},$$

with

$$N_{g=d,t=j} = \sum_{i=1}^{N \cdot T} 1\{G_i = d\} 1\{T_i = j\},$$

G_i and T_i are group and time dummy, respectively, and Y_i is the “poolled” outcome data.

TWFE DiD regression estimator in the 2x2 Setup

- I usually refer to $\widehat{\theta}_n^{DiD}$ as the "brute-force" DiD estimator (or "DiD by hand")/
- The explicit starting point is the estimand (population parameter), and all we do is follow the "plug-in" principle.
- In practice, however, many researchers choose to estimate the ATT in DiD using the following two-way fixed-effects (TWFE) regression specification

$$Y_{i,t} = \alpha_0 + \gamma_0 1\{G_i = 2\} + \lambda_0 1\{T_i = 2\} + \beta_0^{twfe}\left(1\{G_i = 2\} \cdot 1\{T_i = 2\}\right) + \varepsilon_{i,t},$$

where $\mathbb{E}[\varepsilon_{i,t}|G_i, T_i] = 0$ *almost surely*.

- We can show that β_0^{twfe} is equal to the DiD estimand in the canonical 2x2 setup.

TWFE DiD regression estimator in the 2x2 Setup

- The TWFE specification is given by

$$Y_{i,t} = \alpha_0 + \gamma_0 1\{G_i = 2\} + \lambda_0 1\{T_i = 2\} + \beta_0^{twfe}\left(1\{G_i = 2\} \cdot 1\{T_i = 2\}\right) + \varepsilon_{i,t},$$

where $\mathbb{E}[\varepsilon_{i,t}|G, T] = 0$ *almost surely*.

- Now, let's play with its terms:

$$\begin{aligned} \mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 1] &= \alpha_0 \\ \mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 2] &= \alpha_0 + \lambda_0 \\ \mathbb{E}[Y_{i,t}|G_i = 2, T_i = 1] &= \alpha_0 + \gamma_0 \\ \mathbb{E}[Y_{i,t}|G_i = 2, T_i = 2] &= \alpha_0 + \gamma_0 + \lambda_0 + \beta_0^{twfe} \end{aligned}$$

TWFE DiD regression estimator in the 2x2 Setup

- Set of moment restrictions:

$$\begin{aligned}
\mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 1] &= \alpha_0 \\
\mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 2] &= \alpha_0 + \lambda_0 \\
\mathbb{E}[Y_{i,t}|G_i = 2, T_i = 1] &= \alpha_0 + \gamma_0 \\
\mathbb{E}[Y_{i,t}|G_i = 2, T_i = 2] &= \alpha_0 + \gamma_0 + \lambda_0 + \beta_0^{twfe}
\end{aligned}$$

- These imply that

$$\mathbb{E}[Y_{i,t}|G_i = 2, T_i = 2] - \mathbb{E}[Y_{i,t}|G_i = 2, T_i = 1] = \lambda_0 + \beta_0^{twfe}$$

  and that

$$\mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 2] - \mathbb{E}[Y_{i,t}|G_i = \infty, T_i = 1] = \lambda_0 .$$

- Thus, we can clearly see that β_0^{twfe} is equal to the DiD estimand.

TWFE DiD regression estimator in the 2x2 Setup

We can then **estimate** β_0^{twfe} (or the ATT) via ordinary least squares:

$$\left(\widehat{\alpha}, \widehat{\gamma}, \widehat{\lambda}, \widehat{\beta}^{twfe}\right) = \underset{\alpha,\gamma,\lambda,\beta^{twfe}}{\text{argmin}} \sum_{j=1}^{N_{all}} \left(Y_j - \alpha - \gamma 1\left\{G_j = 2\right\} - \lambda 1\left\{T_j = 2\right\} - \beta^{twfe}\left(1\left\{G_j = 2\right\} \cdot 1\left\{T_j = 2\right\}\right)\right)^2$$

- We must stress that we use this regression procedure as a way to get what we are after - the ATT.
- Regression is the estimation tool - it does not fix the target parameter!
- We like this because we have a good understanding of regressions!
- We can leverage it to conduct asymptotically valid inference.

Should we cluster?

Yes, but why?

Importance of clustering

- We should cluster at least at the cross-sectional level.
- We do this because we want to allow for **arbitrary auto-correlation** for the outcomes for the same units across periods.
- Of course, standard inference procedures (without additional strong distributional assumptions) will only be reliable when we have a large number of clusters
  - Without normality assumptions, we need to apply a Central Limit Theorem (CLT) to justify inference.
  - Reliability of CLT depends on the effective sample size (number of clusters) to be large.

Importance of clustering: TWFE regression vs DiD-by-hand

- I also have a more “mechanical” explanation (or anecdote) about why we should cluster at least at the unit level

- We have discussed that the DiD-by-hand estimator, $\widehat{\theta}_n^{DiD}$, is numerically the same as the TWFE estimator, $\widehat{\beta}^{twfe}$ in the 2x2 setup.

- If we were to derive the large sample properties of both estimators (under the same assumptions), they should be the same. (this is obvious, right?!)

- But if we do not cluster, this does not happen:
  - DiD-by-hand is explicit about effective sample size being number of units (or number of cluster)
  - TWFE effective sample size pooled data - 2× number of units in (balanced panel).

It all comes down from sampling!

Panel data sampling schemes

- There are two main leading sampling schemes in the 2x2 DiD setup.

Assumption (Panel Data Sampling Scheme)

The data $\{Y_{i,t=1}, Y_{i,t=2}, G_i\}_{i=i}^{n}$ *is a random sample of the population of interest.*

- We observe data at periods $t = 1$ and $t = 2$ for the same units.

- This, in general, leads to more precise estimators.

Repeated cross-section data sampling schemes

Assumption (Repeated Cross-Section Data Sampling Scheme)

The pooled repeated cross-section data $\{Y_i, G_i, T_i\}_{i=1}^n$ *consist of iid draws from the mixture distribution*

$$\begin{aligned} P\left(Y \leq y, G = g, T = t\right) \quad = \quad & 1\{t = 2\} \cdot \lambda \cdot P\left(Y_{t=2} \leq y, G = g | T = 2\right) \\ & +1\{t = 1\} \cdot (1 - \lambda) \, P\left(Y_{t=1} \leq y, G = g | T = 1\right), \end{aligned}$$

where $(y, g, t) \in \mathbb{R} \times \{2, \infty\} \times \{1, 2\}$, $\lambda = \mathbb{P}\left(T = 2\right) \in (0, 1)$.

Repeated cross-section data sampling schemes

- It accommodates the binomial sampling scheme where an observation i is randomly drawn from either $(Y_{t=2}, G)$ or $(Y_{t=1}, G)$ with fixed probability λ (here, T is a non-degenerated random variable).

- It also accommodates the "conditional" sampling scheme where $n_{t=2}$ observations are sampled from $(Y_{t=2}, G)$, $n_{t=1}$ observations are sampled from $(Y_{t=1}, G)$ and $\lambda = n_{t=2}/(n_{t=1} + n_{t=2})$ (here, T is treated as fixed).

- In the repeated cross-section setup, we will write $n = n_{t=1} + n_{t=2}$

We can talk with great detail about estimation and inference (details are on the slides).

I will skip details, but we will use them below.

⏩ Large Sample Derivations

How can we concretely conduct inference?

How to make inference?

Inference for the ATT based on the DiD estimator in 2x2 setups with panel data

- We will only cover the Panel data case: repeated cross-sections case is similar.
- We want to conduct inference about the ATT based on our DiD estimator $\widehat{\theta}_n^{DiD}$.
  - We want to do hypothesis testing.
  - We want to construct confidence intervals.
- We know (from the calculations we skipped above) that under our assumptions, as the number of units n grows,

$$\sqrt{n}\left(\widehat{\theta}_n^{DiD} - \theta_0^{DiD}\right) \to N(0, V_p), \quad \text{with } V_p = \mathbb{E}\left[\xi_{G=2}^2\right] + \mathbb{E}\left[\xi_{G=\infty}^2\right],$$

where, for $g \in \{2, \infty\}$,

$$\xi_{G=g} = \frac{1\{G=g\}}{\mathbb{E}\left[1\{G=g\}\right]}\left(\Delta Y - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=g\}\right]}{\mathbb{E}\left[1\{G=g\}\right]}\right).$$

Inference for the ATT in 2x2 DiD setups with panel data

- How can we estimate V_p?
- The analogy principle strikes again!
  - Replace population expectations with sample analogs!
- Estimator for asymptotic variance is

$$\widehat{V}_{n,p} = \mathbb{E}_n\left[\widehat{\xi}^2_{G=2}\right] + \mathbb{E}_n\left[\widehat{\xi}^2_{G=\infty}\right],$$

where, for $g \in \{2, \infty\}$,

$$\widehat{\xi}_{i,G=g} = \frac{1\{G_i = g\}}{\mathbb{E}_n\left[1\{G = g\}\right]}\left(\Delta Y_i - \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G = g\}\right]}{\mathbb{E}_n\left[1\{G_i = g\}\right]}\right).$$

- It is easy to show that, as $n \to \infty$, $\widehat{V}_n \xrightarrow{p} V$.

Inference for the ATT in 2x2 DiD setups with panel data

- Estimated standard error is

$$\widehat{se}_{n,p}(\widehat{\theta}_n^{DiD}) = \sqrt{\frac{\widehat{V}_{n,p}}{n}}.$$

- Std error is clustered at the unit level: it allows for arbitrary time dependence across periods (because $T = 2$).

- 95% confidence interval for ATT based on asymptotic normality:

$$\widehat{\theta}_n^{DiD} \pm 1.96 \cdot \widehat{se}_{n,p}(\widehat{\theta}_n^{DiD}).$$

- Hypotheses tests for the null $H_0 : ATT = c$ for a known $c \in \mathbb{R}$ can also be conducted using the t-statistics:

$$\text{t-stat} = \frac{\widehat{\theta}_n^{DiD} - c}{\widehat{se}_{n,p}(\widehat{\theta}_n^{DiD})}.$$

What if we want to cluster at different level?

How to make inference?

Clustering at a more aggregated level

Clustering at a more aggregate level

- Many times, researchers (or referees) argue that you should cluster standard errors at a more aggregated level than the unit level.

- Sometimes, they recommend clustering at the treatment assignment level, see, e.g., Bertrand, Duflo and Mullainathan (2004); Donald and Lang (2007); Conley and Taber (2010); Ferman and Pinto (2019), among many others.

- These "general" recommendations are tricky, as the choice of the cluster level should depend on the sampling/counterfactual/parameter of interest you are using/considering, see, e.g., Wooldridge (2003), Imbens and Wooldridge (2007), Abadie, Athey, Imbens and Wooldridge (2022) and Rambachan and Roth (2022).

- If we have a large number of clusters (of finite size), (standard) valid inference is feasible; see, e.g., Sherman and Le Cessie (2007); Kline and Santos (2012); Cheng, Yu and Huang (2013); Callaway and Sant'Anna (2021).

Why?

DiD estimator in 2x2 setups with panel data

- Recall that our DiD estimator is

$$\widehat{\theta}_n^{DiD} = \left(\overline{Y}_{g=2,t=2} - \overline{Y}_{g=2,t=1}\right) - \left(\overline{Y}_{g=\infty,t=2} - \overline{Y}_{g=\infty,t=1}\right),$$

- In the panel data case, we can simplify this a bit further:

$$\widehat{\theta}_n^{DiD} = \overline{\Delta Y}_{g=2} - \overline{\Delta Y}_{g=\infty},$$

where $\overline{\Delta Y}_{g=d}$ is the sample mean of $\Delta Y_i \equiv Y_{i,t=2} - Y_{i,t=1}$ for units in group d,

$$\overline{\Delta Y}_{g=d} = \frac{\sum_{i:G_i=d} \Delta Y_i}{n_{G=d}} = \frac{n^{-1}\sum_{i=1}^{n} \Delta Y_i 1\{G_i = d\}}{n^{-1}\sum_{i=1}^{n} 1\{G_i = d\}} = \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G = d\}\right]}{\mathbb{E}_n\left[1\{G = d\}\right]},$$

and $n_{G=d} = \sum_{i=1}^{n} 1\{G = d\}$ is the sample size of group $G = d$.

- Henceforth, for a generic variable A,

$$\mathbb{E}_n\left[A\right] \equiv \frac{\sum_{i=1}^{n} A_i}{n}.$$

DiD estimator in 2x2 setups with panel data

- We then have that

$$\widehat{\theta}_n^{DiD} = \overline{\Delta Y}_{g=2} - \overline{\Delta Y}_{g=\infty} = \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]}.$$

- The above notation emphasizes that the effective sample size is n, the number of units.

- If we are now sampling clusters (of finite size), it is arguably desirable to highlight this in our notation.

- In my view, notation should be used to help with these kinds of things!

DiD estimator in 2x2 setups with panel data and cluster

The key is to rewrite some averages in terms of the number of clusters:

$$\overline{\Delta Y}_{g=d} = \frac{n^{-1}\sum_{i=1}^{n}\Delta Y_i 1\{G_i = d\}}{n^{-1}\sum_{i=1}^{n} 1\{G_i = d\}} = \frac{\frac{1}{N_{cluster}}\sum_{c=1}^{N_{cluster}} \sum_{i=1}^{n}\Delta Y_i 1\{G_i = d\}1\{C_i = c\}}{\frac{1}{N_{cluster}}\sum_{c=1}^{N_{cluster}} \sum_{i=1}^{n} 1\{G_i = d\}1\{C_i = c\}}$$

$$= \frac{\frac{1}{N_{cluster}}\sum_{c=1}^{N_{cluster}} n\,\mathbb{E}_n\left[\Delta Y \cdot 1\{G = d\}1\{C = c\}\right]}{\frac{1}{N_{cluster}}\sum_{c=1}^{N_{cluster}} n\,\mathbb{E}_n\left[1\{G = d\}1\{C = c\}\right]}$$

$$= \frac{\frac{1}{N_{cluster}}\sum_{c=1}^{N_{cluster}} \mathbb{E}_n\left[\Delta Y \cdot 1\{G = d\}1\{C = c\}\right]}{\frac{1}{N_{cluster}}\sum_{c=1}^{N_{cluster}} \mathbb{E}_n\left[1\{G = d\}1\{C = c\}\right]}$$

(This is good but potentially ambiguous)

DiD estimator in 2x2 setups with panel data

We can continue with manipulations:

$$
\begin{aligned}
\overline{\Delta Y}_{g=d} &= \frac{\frac{1}{N_{cluster}} \sum_{c=1}^{N_{cluster}} \mathbb{E}_n \left[\Delta Y \cdot 1\{G=d\} 1\{C=c\} \right]}{\frac{1}{N_{cluster}} \sum_{c=1}^{N_{cluster}} \mathbb{E}_n \left[1\{G=d\} 1\{C=c\} \right]} \\
&= \frac{\frac{1}{N_{cluster}} \sum_{c=1}^{N_{cluster}} \frac{n_c}{n} \cdot \mathbb{E}_{n_c} \left[\Delta Y \cdot 1\{G=d\} | C=c \right]}{\frac{1}{N_{cluster}} \sum_{c=1}^{N_{cluster}} \frac{n_c}{n} \cdot \mathbb{E}_{n_c} \left[1\{G=d\} | C=c \right]} \\
&= \frac{\frac{1}{N_{cluster}} \sum_{c=1}^{N_{cluster}} P_n(C=c) \cdot \mathbb{E}_{n_c} \left[\Delta Y \cdot 1\{G=d\} | C=c \right]}{\frac{1}{N_{cluster}} \sum_{c=1}^{N_{cluster}} P_n(C=c) \mathbb{E}_{n_c} \left[1\{G=d\} 1\{C=c\} \right]}
\end{aligned}
$$

(This is what we want!)

DiD estimator in 2x2 setups with panel data and clusters

We then have that

$$\widehat{\theta}_n^{DiD} = \frac{\frac{1}{N_{cluster}} \sum_{c=1}^{N_{cluster}} P_n(C=c) \cdot \mathbb{E}_{n_c}\left[\Delta Y \cdot 1\{G=2\} | C=c\right]}{\frac{1}{N_{cluster}} \sum_{c=1}^{N_{cluster}} P_n(C=c) \mathbb{E}_{n_c}\left[1\{G=2\} 1\{C=c\}\right]}$$

$$- \frac{\frac{1}{N_{cluster}} \sum_{c=1}^{N_{cluster}} P_n(C=c) \cdot \mathbb{E}_{n_c}\left[\Delta Y \cdot 1\{G=\infty\} | C=c\right]}{\frac{1}{N_{cluster}} \sum_{c=1}^{N_{cluster}} P_n(C=c) \mathbb{E}_{n_c}\left[1\{G=\infty\} 1\{C=c\}\right]}$$

Now, it is only a matter of applying a law of large numbers + a central limit theorem + continuous mapping theorem.

How to make inference?

Cluster-robust inference via multiplier bootstrap

Bootstrap procedure for clustering

- Let's illustrate how we can leverage the influence functions to conduct cluster-robust inference.

- We will use a multiplier-bootstrap procedure, see, e.g., van der Vaart and Wellner (1996); Kline and Santos (2012); Callaway and Sant'Anna (2021).

- A big advantage of this bootstrap procedure is that we do not have to re-estimate all parameters in every bootstrap draw.
  - Much faster.
  - No problem with "small-ish groups" in some bootstrap draws.
  - Easy to prove its validity.

Multiplier bootstrap procedure

- Recall that the estimated influence function for the panel data case is

$$\widehat{\xi}_{i,p} = \widehat{\xi}_{i,G=2} - \widehat{\xi}_{i,G=\infty},$$

where, for $g \in \{2, \infty\}$,

$$\widehat{\xi}_{i,G=g} = \frac{1\{G_i = g\}}{\mathbb{E}_n\left[1\{G = g\}\right]}\left(\Delta Y_i - \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G = g\}\right]}{\mathbb{E}_n\left[1\{G_i = g\}\right]}\right).$$

- Let $\{V_i\}_{i=1}^{n}$ be a sequence of *iid* random variables with zero mean, unit variance, and bounded third moment, independent of the original sample.
  - Example: Rademacher Distribution, $\mathbb{P}(V = 1) = \mathbb{P}(V = -1) = 0.5$.
- $\widehat{\theta}_n^{*,DiD}$, a bootstrap draw of $\widehat{\theta}_n^{DiD}$, is given by

$$\widehat{\theta}_n^{*,DiD} = \widehat{\theta}_n^{DiD} + \mathbb{E}_n\left[V \cdot \widehat{\xi}_p\right]. \tag{1}$$

Multiplier bootstrap algorithm

1. Draw a realization of $\{V_i\}_{i=1}^{n}$.
2. Compute $\widehat{\theta}_n^{*,DiD}$ as in (1), and form a bootstrap draw of its limiting distribution as
$$\hat{R}^* = \sqrt{n}\left(\widehat{\theta}_n^{*,DiD} - \widehat{\theta}_n^{DiD}\right)$$
3. Repeat steps 1-2 B times (say, 999).
4. Estimate $V_p^{1/2}$ by
$$\widehat{V}_{n,p}^{1/2,boot} = \left(q_{0.75} - q_{0.25}\right) / \left(z_{0.75} - z_{0.25}\right),$$
where q_τ is the τ-th sample quantile of $\hat{R}^*$, and z_τ is the τ-th quantile of the standard normal distribution.
5. For each bootstrap draw, compute t-test$^* = \left|\hat{R}^*\right| / \widehat{V}_{n,p}^{1/2,boot}$.
6. Construct $\widehat{c}_{1-\alpha}$ as the empirical $(1-a)$-quantile of the B bootstrap draws of $t - test^*$.
7. Construct the bootstrapped confidence intervals for the ATT as
$$\widehat{C} = [\widehat{\theta}_n^{DiD} \pm \widehat{c}_{1-\alpha} \cdot \widehat{V}_{n,p}^{1/2,boot} / \sqrt{n}].$$

How to cluster at a more “aggregated level”?

- How can we cluster at a level more aggregated than the unit?
- This is straightforward to implement with the multiplier bootstrap described above.
- Example: allow for clustering at the state level
  - draw a scalar U_s S times – where S is the number of states
  - set $V_i = U_s$ for all observations i in state s
- This procedure is justified, provided that the number of clusters is “large”, and cluster size is “fixed”.
- This is the case when we sample (entire) clusters from a super-population.

What if we have small number of clusters?

Empirical exercise

Empirical exercise

- Let’s switch to R/Stata so we can see how to do all these things!

Large sample properties of the 2x2 DiD estimator

Let's derive the large sample properties of the DiD estimator in the panel data case

⏩ Skip all derivations to justify asymptotic normality

Large sample properties of the 2x2 DiD estimator

Panel Data Case

DiD estimator in 2x2 setups with panel data

- Recall that our DiD estimator is

$$\widehat{\theta}_n^{DiD} = \left(\overline{Y}_{g=2,t=2} - \overline{Y}_{g=2,t=1}\right) - \left(\overline{Y}_{g=\infty,t=2} - \overline{Y}_{g=\infty,t=1}\right),$$

- In the panel data case, we can simplify this a bit further:

$$\widehat{\theta}_n^{DiD} = \overline{\Delta Y}_{g=2} - \overline{\Delta Y}_{g=\infty},$$

  where $\overline{\Delta Y}_{g=d}$ is the sample mean of $\Delta Y_i \equiv Y_{i,t=2} - Y_{i,t=1}$ for units in group d,

$$\overline{\Delta Y}_{g=d} = \frac{\sum_{i:G_i=d} \Delta Y_i}{n_{G=d}} = \frac{n^{-1}\sum_{i=1}^{n} \Delta Y_i 1\{G_i = d\}}{n^{-1}\sum_{i=1}^{n} 1\{G_i = d\}} = \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G = d\}\right]}{\mathbb{E}_n\left[1\{G = d\}\right]},$$

  and $n_{G=d} = \sum_{i=1}^{n} 1\{G = d\}$ is the sample size of group $G = d$.

- Henceforth, for a generic variable A,

$$\mathbb{E}_n\left[A\right] \equiv \frac{\sum_{i=1}^{n} A_i}{n}.$$

DiD estimator in 2x2 setups with panel data

- We then have that

$$\widehat{\theta}_n^{DiD} = \overline{\Delta Y}_{g=2} - \overline{\Delta Y}_{g=\infty} = \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]}.$$

- We want to know if this estimator is “reliable”.
  - As the number of units increases, does it converge in probability the true ATT, under our assumptions?
  - How can we conduct reliable ATT inferences without invoking distributional assumptions?
- We will rely on large sample approximation results.
- All those stats classes you took (or teach), can be very handy now!
- We will use law of large numbers (LLN) + continuous mapping theorem (CMT) + CLT.

Consistency of the DiD estimator in 2x2 setups with panel data

- Since

$$\widehat{\theta}_n^{DiD} = \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]},$$

  consistency follows directly from the law of large numbers and the continuous mapping theorem.

- **LLN**: with iid + bounded moments (which we implicitly assume), sample means converge in probability to population means.

- **Continuous mapping theorem**: continuous functionals preserve limits.

- As a result, we have, as $n \to \infty$,

$$\widehat{\theta}_n^{DiD} \xrightarrow{p} \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]} = \mathbb{E}\left[\Delta Y | G=2\right] - \mathbb{E}\left[\Delta Y | G=\infty\right] \equiv \theta^{DiD},$$

  and $\theta^{DiD} = ATT$ under SUTVA + No-Anticipation + PT assumptions.

Asy. Normality of the DiD estimator in 2x2 setups with panel data

- Now, we want to derive the asymptotic distribution of

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_n^{DiD} - \theta^{DiD}\right) &= \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right) \\ &\quad -\sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]}\right).\end{aligned}$$

- To get there, we can use CLT and Delta Method (iid + finite asymptotic variance + denominator bounded away from zero.)

- We will do this slightly differently because I want to get the **influence function**.

- Express $\sqrt{n}\left(\widehat{\theta}_n^{DiD} - \theta^{DiD}\right)$ as as average of *iid* terms + negligible terms.

Asy. Normality of the DiD estimator in 2x2 setups with panel data

- Let's first analyze

$$\sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y\cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]}-\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right).$$

- With some manipulation, we can rewrite this as

$$\begin{aligned}
&\frac{1}{\mathbb{E}\left[1\{G=2\}\right]}\sqrt{n}\left(\mathbb{E}_n\left[\Delta Y\cdot 1\{G=2\}\right]-\mathbb{E}\left[\Delta Y\cdot 1\{G=2\}\right]\right)\\
&-\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]^2}\sqrt{n}\left(\mathbb{E}_n\left[1\{G=2\}\right]-\mathbb{E}\left[1\{G=2\}\right]\right)\\
&+\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=2\}\right]\cdot\left(\mathbb{E}_n\left[1\{G=2\}\right]-\mathbb{E}\left[1\{G=2\}\right]\right)}{\mathbb{E}\left[1\{G=2\}\right]^2\cdot\mathbb{E}_n\left[1\{G=2\}\right]}\sqrt{n}\left(\mathbb{E}_n\left[1\{G=2\}\right]-\mathbb{E}\left[1\{G=2\}\right]\right)\\
&-\frac{\left(\mathbb{E}_n\left[1\{G=2\}\right]-\mathbb{E}\left[1\{G=2\}\right]\right)}{\mathbb{E}\left[1\{G=2\}\right]\cdot\mathbb{E}_n\left[1\{G=2\}\right]}\sqrt{n}\left(\mathbb{E}_n\left[\Delta Y\cdot 1\{G=2\}\right]-\mathbb{E}\left[\Delta Y\cdot 1\{G=2\}\right]\right).
\end{aligned}$$

Asy. Normality of the DiD estimator in 2x2 setups with panel data

- Red terms converge in probability to zero by LLN
- Blue terms converge in distribution to Normal with finite variance by CLT.
- Then, by Slutsky's Theorem

$$
\begin{aligned}
& \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]}-\frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right) \\
= \quad & \frac{1}{\mathbb{E}\left[1\{G=2\}\right]}\sqrt{n}\left(\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]-\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]\right) \\
& -\frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]^2}\sqrt{n}\left(\mathbb{E}_n\left[1\{G=2\}\right]-\mathbb{E}\left[1\{G=2\}\right]\right) \\
& +o_p(1).
\end{aligned}
$$

Asy. Normality of the DiD estimator in 2x2 setups with panel data

- Rearranging some terms (and with some abuse of notation), we have

$$
\begin{aligned}
&\sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y\cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]}-\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right)\\
= \quad &\sqrt{n}\mathbb{E}_n\left[\frac{\Delta Y\cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right]-\sqrt{n}\mathbb{E}\left[\frac{\Delta Y\cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right]\\
&-\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\sqrt{n}\left(\mathbb{E}_n\left[\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right]-1\right)+o_p(1)\\
= \quad &\sqrt{n}\mathbb{E}_n\left[\frac{\Delta Y\cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right]-\sqrt{n}\mathbb{E}\left[\frac{\Delta Y\cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right]\\
&-\sqrt{n}\mathbb{E}_n\left[\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right]+\sqrt{n}\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}+o_p(1).
\end{aligned}
$$

Asy. Normality of the DiD estimator in 2x2 setups with panel data

- Continuing the manipulations...

$$
\begin{aligned}
&\sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right) \\
= \quad & \sqrt{n}\mathbb{E}_n\left[\frac{\Delta Y \cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] - \sqrt{n}\mathbb{E}_n\left[\frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] + o_p(1) \\
= \quad & \sqrt{n}\mathbb{E}_n\left[\frac{\Delta Y \cdot 1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]} - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\right] + o_p(1) \\
= \quad & \sqrt{n}\mathbb{E}_n\left[\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\left(\Delta Y - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right)\right] + o_p(1).
\end{aligned}
$$

Asy. Normality of the DiD estimator in 2x2 setups with panel data

- Thus, we have that

$$
\begin{aligned}
& \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]}-\frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right) \\
= \quad & \sqrt{n}\mathbb{E}_n\left[\frac{1\{G=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\left(\Delta Y-\frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right)\right]+o_p(1) \\
= \quad & \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(\frac{1\{G_i=2\}}{\mathbb{E}\left[1\{G=2\}\right]}\left(\Delta Y-\frac{\mathbb{E}\left[\Delta Y_i \cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right)\right)}_{=\xi_{i,G=2} \quad \text{: influence function}}+o_p(1) \\
= \quad & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_{i,G=2}+o_p(1),
\end{aligned}
$$

- The $\xi_{i,G=2}$ is the influence function we were after: it is mean zero, has finite variance, and is *iid*.

Asy. Normality of the DiD estimator in 2x2 setups with panel data

- Now, following the same steps as we did, we have that

$$
\begin{aligned}
& \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y\cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]}-\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]}\right) \\
= \quad & \sqrt{n}\mathbb{E}_n\left[\frac{1\{G=\infty\}}{\mathbb{E}\left[1\{G=\infty\}\right]}\left(\Delta Y-\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]}\right)\right]+o_p(1) \\
= \quad & \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(\frac{1\{G_i=\infty\}}{\mathbb{E}\left[1\{G=\infty\}\right]}\left(\Delta Y-\frac{\mathbb{E}\left[\Delta Y_i\cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]}\right)\right)}_{=\xi_{i,G=\infty}\quad \text{: influence function}}+o_p(1) \\
= \quad & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_{i,G=\infty}+o_p(1),
\end{aligned}
$$

Asy. Normality of the DiD estimator in 2x2 setups with panel data

- Putting these pieces together, it follows that

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_n^{DiD}-\theta^{DiD}\right) &= \sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y\cdot 1\{G=2\}\right]}{\mathbb{E}_n\left[1\{G=2\}\right]}-\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=2\}\right]}{\mathbb{E}\left[1\{G=2\}\right]}\right)\\ &\quad -\sqrt{n}\left(\frac{\mathbb{E}_n\left[\Delta Y\cdot 1\{G=\infty\}\right]}{\mathbb{E}_n\left[1\{G=\infty\}\right]}-\frac{\mathbb{E}\left[\Delta Y\cdot 1\{G=\infty\}\right]}{\mathbb{E}\left[1\{G=\infty\}\right]}\right)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\xi_{i,G=2}-\xi_{i,G=\infty}\right)+o_p(1)\end{aligned}$$

- Now, it follows from the CLT that

$$\sqrt{n}\left(\widehat{\theta}_n^{DiD}-\theta^{DiD}\right)\xrightarrow{d} N(0,V_p),$$

where

$$V_p=\mathbb{E}\left[\left(\xi_{G=2}-\xi_{G=\infty}\right)^2\right]=\mathbb{E}\left[\xi_{G=2}^2\right]+\mathbb{E}\left[\xi_{G=\infty}^2\right]$$

What about the repeated cross-section data case?

Large sample properties of the 2x2 DiD estimator

Repeated Cross-Section Data Case

DiD estimator in 2x2 setups with RCS data

- Recall that our “general” DiD estimator is

$$\widehat{\theta}_n^{DiD} = \left(\overline{Y}_{g=2,t=2} - \overline{Y}_{g=2,t=1}\right) - \left(\overline{Y}_{g=\infty,t=2} - \overline{Y}_{g=\infty,t=1}\right),$$

- In the RCS data case, we can’t use the first difference of the outcome as we do not observe outcomes for the same units in both periods.

- So, we will need to work with four sample means instead of two.

- Not very different, though!

- Data are in the ”long” structure (each row is a different cross-section unit)

Consistency of the DiD estimator in 2x2 setups with RCS data

- Note that we can write the DiD estimator as

$$\begin{aligned}\widehat{\theta}_n^{DiD} &= \left(\frac{\mathbb{E}_n\left[Y\cdot 1\left\{T=2,G=2\right\}\right]}{\mathbb{E}_n\left[1\left\{T=2,G=2\right\}\right]}-\frac{\mathbb{E}_n\left[Y\cdot 1\left\{T=1,G=2\right\}\right]}{\mathbb{E}_n\left[1\left\{T=1,G=2\right\}\right]}\right)\\ &\quad -\left(\frac{\mathbb{E}_n\left[Y\cdot 1\left\{T=2,G=\infty\right\}\right]}{\mathbb{E}_n\left[1\left\{T=2,G=\infty\right\}\right]}-\frac{\mathbb{E}_n\left[Y\cdot 1\left\{T=1,G=\infty\right\}\right]}{\mathbb{E}_n\left[1\left\{T=1,G=\infty\right\}\right]}\right).\end{aligned}$$

- Just like in the panel data case, consistency follows directly from the law of large numbers and the continuous mapping theorem.

- As a result, we have that as $n = n_{t=1} + n_{t=2} \rightarrow \infty$, with $n_{t=2}/n \rightarrow \lambda \in (0,1)$,

$$\begin{aligned}\widehat{\theta}_n^{DiD} &\xrightarrow{p} \left(\frac{\mathbb{E}\left[Y\cdot 1\left\{T=2,G=2\right\}\right]}{\mathbb{E}\left[1\left\{T=2,G=2\right\}\right]}-\frac{\mathbb{E}\left[Y\cdot 1\left\{T=1,G=2\right\}\right]}{\mathbb{E}\left[1\left\{T=1,G=2\right\}\right]}\right)\\ &\quad -\left(\frac{\mathbb{E}\left[Y\cdot 1\left\{T=2,G=\infty\right\}\right]}{\mathbb{E}\left[1\left\{T=2,G=\infty\right\}\right]}-\frac{\mathbb{E}\left[Y\cdot 1\left\{T=1,G=\infty\right\}\right]}{\mathbb{E}\left[1\left\{T=1,G=\infty\right\}\right]}\right)\\ &\equiv \theta^{DiD}\end{aligned}$$

Asy. Normality of the DiD estimator in 2x2 setups with RCS data

- The asymptotic normality will follow from similar steps as in the panel data case, as the problems have very similar structures.

$$
\begin{aligned}
\sqrt{n}\left(\widehat{\theta}_n^{DiD} - \theta^{DiD}\right) \quad = \quad & \sqrt{n}\left(\frac{\mathbb{E}_n\left[Y \cdot 1\{T=2, G=2\}\right]}{\mathbb{E}_n\left[1\{T=2, G=2\}\right]} - \frac{\mathbb{E}\left[Y \cdot 1\{T=2, G=2\}\right]}{\mathbb{E}\left[1\{T=2, G=2\}\right]}\right) \\
& -\sqrt{n}\left(\frac{\mathbb{E}_n\left[Y \cdot 1\{T=1, G=2\}\right]}{\mathbb{E}_n\left[1\{T=1, G=2\}\right]} - \frac{\mathbb{E}\left[Y \cdot 1\{T=1, G=2\}\right]}{\mathbb{E}\left[1\{T=1, G=2\}\right]}\right) \\
& -\sqrt{n}\left(\frac{\mathbb{E}_n\left[Y \cdot 1\{T=2, G=\infty\}\right]}{\mathbb{E}_n\left[1\{T=2, G=\infty\}\right]} - \frac{\mathbb{E}\left[Y \cdot 1\{T=2, G=\infty\}\right]}{\mathbb{E}\left[1\{T=2, G=\infty\}\right]}\right) \\
& +\sqrt{n}\left(\frac{\mathbb{E}_n\left[Y \cdot 1\{T=1, G=\infty\}\right]}{\mathbb{E}_n\left[1\{T=1, G=\infty\}\right]} - \frac{\mathbb{E}\left[Y \cdot 1\{T=1, G=\infty\}\right]}{\mathbb{E}\left[1\{T=1, G=\infty\}\right]}\right)
\end{aligned}
$$

Asy. Normality of the DiD estimator in 2x2 setups with RCS data

- Analyzing each of these terms separately, we get that

$$\sqrt{n}\left(\widehat{\theta}_n^{DiD} - \theta^{DiD}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\xi_{i,T=2,G=2} - \xi_{i,T=1,G=2} - \xi_{i,T=2,G=\infty} + \xi_{i,T=1,G=\infty}\right) + o_p(1),$$

  where, for $g \in \{2, \infty\}, t \in \{1, 2\}$,

$$\xi_{T=t,G=g} = \frac{1\{T=t, G=g\}}{\mathbb{E}\left[1\{T=t, G=g\}\right]}\left(Y - \frac{\mathbb{E}\left[Y \cdot 1\{T=t, G=g\}\right]}{\mathbb{E}\left[1\{T=t, G=g\}\right]}\right).$$

- Now, it follows from the CLT that

$$\sqrt{n}\left(\widehat{\theta}_n^{DiD} - \theta^{DiD}\right) \xrightarrow{d} N(0, V_{rcs}),$$

  where

$$V_{rcs} = \mathbb{E}\left[\xi^2_{T=2,G=2}\right] + \mathbb{E}\left[\xi^2_{T=1,G=2}\right] + \mathbb{E}\left[\xi^2_{T=2,G=\infty}\right] + \mathbb{E}\left[\xi^2_{T=1,G=\infty}\right].$$

How can we use these results to conduct inference?

Large sample properties of the 2x2 DiD estimator

How to do inference?

Inference for the ATT based on the DiD estimator in 2x2 setups with panel data

- We will cover the Panel data only: RCS is very similar.
- We want to conduct inference about the ATT based on our DiD estimator $\widehat{\theta}_n^{DiD}$.
  - We want to do hypothesis testing
  - We want to construct confidence intervals
- We then need to find a way to estimate the asymptotic variance

$$V_p = \mathbb{E}\left[\xi_{G=2}^2\right] + \mathbb{E}\left[\xi_{G=\infty}^2\right],$$

where, for $g \in \{2, \infty\}$,

$$\xi_{G=g} = \frac{1\{G=g\}}{\mathbb{E}\left[1\{G=g\}\right]}\left(\Delta Y - \frac{\mathbb{E}\left[\Delta Y \cdot 1\{G=g\}\right]}{\mathbb{E}\left[1\{G=g\}\right]}\right).$$

- How can we do that?

Inference for the ATT in 2x2 DiD setups with panel data

- The analogy principle strikes again!
  - Replace population expectations with sample analogs!

- Estimator for asymptotic variance is

$$\widehat{V}_{n,p} = \mathbb{E}_n\left[\widehat{\xi}^2_{G=2}\right] + \mathbb{E}_n\left[\widehat{\xi}^2_{G=\infty}\right],$$

where, for $g \in \{2, \infty\}$,

$$\widehat{\xi}_{i,G=g} = \frac{1\{G_i = g\}}{\mathbb{E}_n\left[1\{G = g\}\right]}\left(\Delta Y_i - \frac{\mathbb{E}_n\left[\Delta Y \cdot 1\{G = g\}\right]}{\mathbb{E}_n\left[1\{G_i = g\}\right]}\right).$$

- It is easy to show that, as $n \to \infty$, $\widehat{V}_n \xrightarrow{p} V$.

Inference for the ATT in 2x2 DiD setups with panel data

- Estimated standard error is

$$\widehat{se}_{n,p}(\widehat{\theta}_n^{DiD}) = \sqrt{\frac{\widehat{V}_{n,p}}{n}}.$$

- Std error is clustered at the unit level: it allows for arbitrary time dependence across periods.
- 95% confidence interval for *ATT* based on asymptotic normality:

$$\widehat{\theta}_n^{DiD} \pm 1.96 \cdot \widehat{se}_{n,p}(\widehat{\theta}_n^{DiD}).$$

- Hypotheses tests for the null $H_0 : ATT = c$ for a known $c \in \mathbb{R}$ can also be conducted using the t-statistics:

$$\text{t-stat} = \frac{\widehat{\theta}_n^{DiD} - c}{\widehat{se}_{n,p}(\widehat{\theta}_n^{DiD})}.$$

References

Abadie, Alberto, Susan Athey, Guido Imbens, and Jeffrey Wooldridge, “When should you adjust standard errors for clustering?,” Technical Report 2022.

Abbring, Jaap H. and Gerard J. van den Berg, “The nonparametric identification of treatment effects in duration models,” *Econometrica*, 2003, *71* (5), 1491–1517.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, “How Much Should We Trust Differences-In-Differences Estimates?,” *The Quarterly Journal of Economics*, February 2004, *119* (1), 249–275.

Callaway, Brantly and Pedro H. C. Sant’Anna, “Difference-in-Differences with Multiple Time Periods,” *Journal of Econometrics*, 2021, *225* (2), 200–230.

Cheng, Guang, Zhuqing Yu, and Jianhua Z. Huang, “The cluster bootstrap consistency in generalized estimating equations,” *Journal of Multivariate Analysis*, 2013, *115*, 33–47.

Conley, Timothy G. and Christopher R. Taber, “Inference with “Difference in Differences” with a Small Number of Policy Changes,” *The Review of Economics and Statistics*, 2010, *93* (1), 113–125.

Donald, Stephen G. and Kevin Lang, "Inference with Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, 2007, *89* (2), 221–233.

Ferman, Bruno and Cristine Pinto, "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," *The Review of Economics and Statistics*, 2019, *101* (3), 452–467.

Imbens, Guido W. and Jeffrey M. Wooldridge, "What is New in Econometrics? Difference-in-Difference Estimation (Lecture Notes 10)," *NBER Summer Institute*, 2007. Available at https://users.nber.org/confer/2007/si2007/WNE/lect_10_diffindiffs.pdf.

Kline, Patrick and Andres Santos, "A score based approach to wild bootstrap inference," *Journal of Econometric Methods*, 2012, *1* (1), 1–40.

Malani, Anup and Julian Reif, "Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform," *Journal of Public Economics*, 2015, *124*, 1–17.

Rambachan, Ashesh and Jonathan Roth, "Design-Based Uncertainty for Quasi-Experiments," *arXiv:2008.00602*, 2022.

Sherman, Michael and Saskia Le Cessie, "A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models," *Communications in Statistics - Simulation and Computation*, 2007, *26* (3), 901–925.

Sianesi, Barbara, "An evaluation of the Swedish system of active labor market programs in the 1990s," *The Review of Economics and Statistics*, 2004, *86* (1), 133–155.

van der Vaart, Aad W and Jon A Wellner, "Weak Convergence and Empirical Processes: With Applications to Statistics," 1996.

Wooldridge, Jeffrey M, "Cluster-Sample Methods in Applied Econometrics," *American Economic Review P&P*, 2003, *93* (2), 133–138.