Which Policy Works and Where? Estimation and inference for state level treatment effects using difference-in-differences

Sunny Karim, Matthew D. Webb, Nichole Austin, Erin Strumpf

May 31, 2025

Overview

- Difference-in-differences is a very common tool in economics. Roughly 20% of all 2024 AER papers use DID.
- Concerns over negative weights and forbidden comparisons have made users shy away from Two Way Fixed Effects (TWFE)
- Callaway & Sant'Anna (CSDID) is one of the most popular alternatives.
- Aggregation is unusual in the method.
- This proposes alternative aggregations, using two recently developed estimators, and demonstrates their reliability for inference.

Unconditional and Conditional $ATT(g,t)$'s

- Under the assumptions of conditional parallel trends (CPT) and no anticipation (NA) assumptions, the $ATT(g,t)$ is defined as:

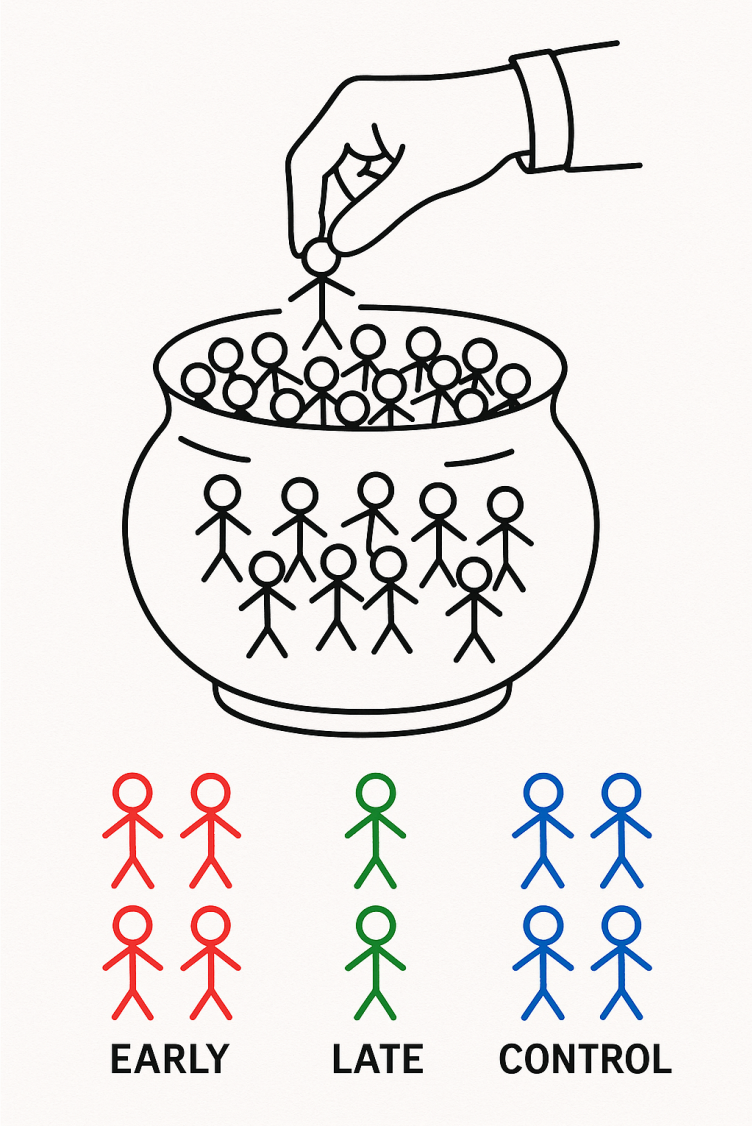
$$ATT(g,t) = \Bigg[E\big[E[Y_{i,g,t}|G_i = g, X_{i,g,t}] - E[Y_{i,g,r-1}|G_i = g, X_{i,g,r-1}]\big] - E\big[E[Y_{i,g',t}|G_i = g', X_{i,g',t}] - E[Y_{i,g',r-1}|G_i = g', X_{i,g',r-1}]\big]\Bigg] \Bigg| G_i = g\Bigg].$$

- The **conditional** $ATT(g,t)$ for a given value of $X_{i,g,t}$ is given by:

$$\begin{aligned} ATT(g,t,x) = &\big[E[Y_{i,g,t}|G_i = g, X_{i,g,t}] - E[Y_{i,g,r-1}|G_i = g, X_{i,g,r-1}]\big] \\ &-\big[E[Y_{i,g',t}|G_i = g', X_{i,g,t}] - E[Y_{i,g',r-1}|G_i = g', X_{i,g,r-1}]\big]. \end{aligned} \tag{1}$$

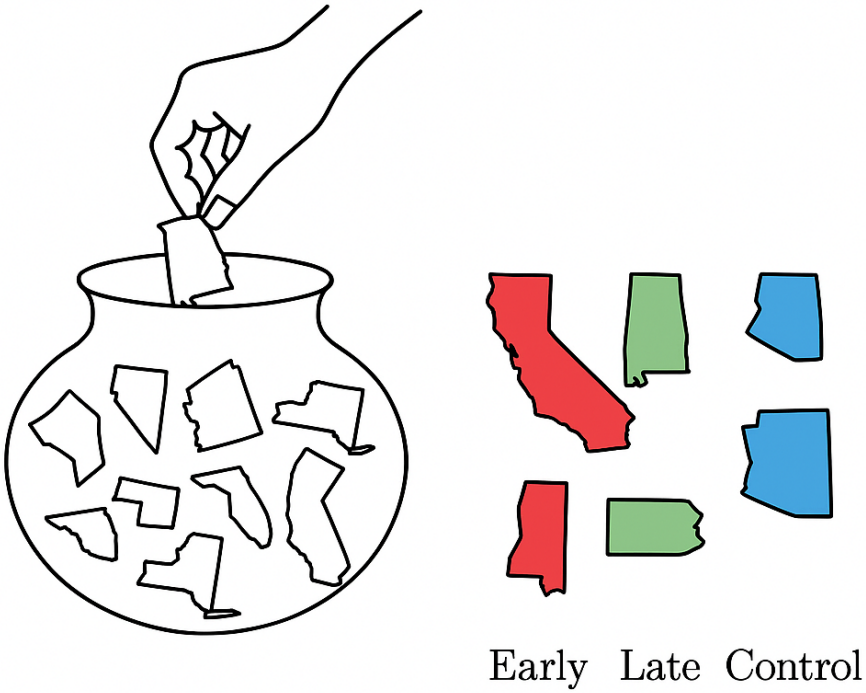
- where, g' is a group that has not been treated at time t
- The ATT is a weighted average of the $ATT(g,t)$'s.

What are samples in CSDID?



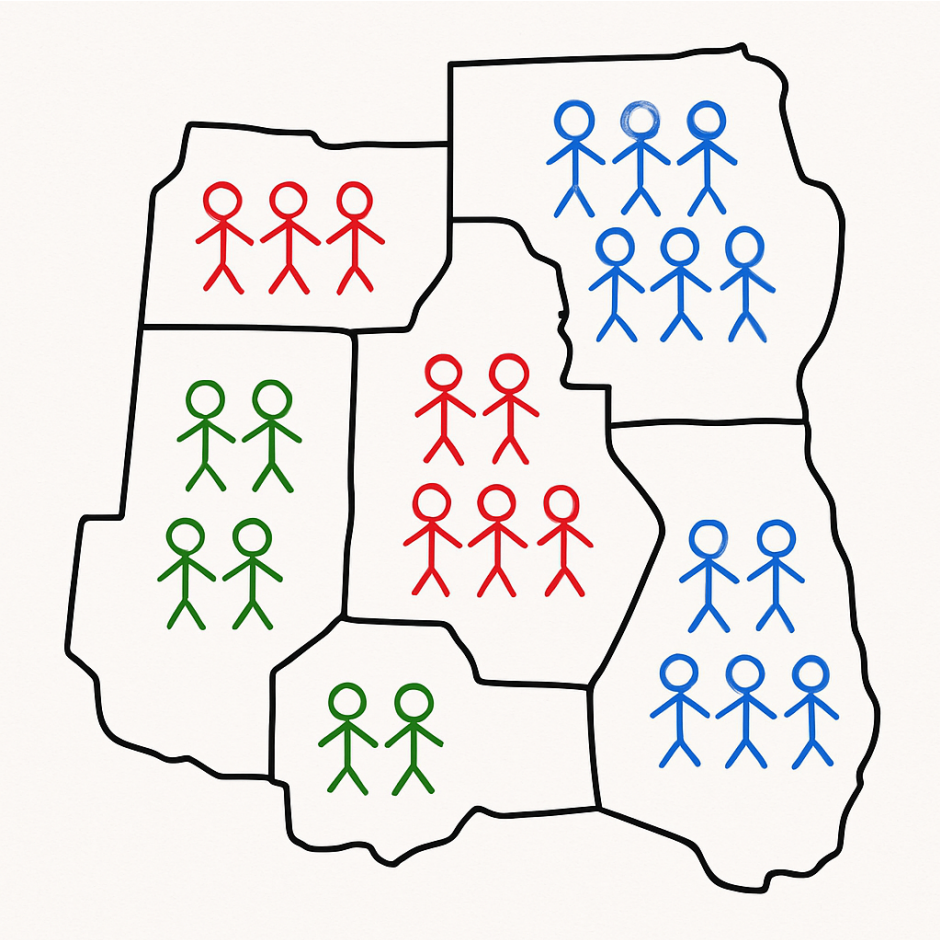
- People are treated randomly
- Timing seems to matter more than place
- Place isn't an argument in their code

What are samples in TWFE DID?



- Places are treated
- Timing (was) secondary
- All people in a treated place are regarded as treated

What does the data look like?



- Data that is typically analyzed "looks" more like the traditional DID data
- We have both treated people and treated groups
- Does it matter that multiple states are treated simultaneously?

IVF Coverage in Canada - Austin and Apold (2023)

Figure: Who is Treated?

IVF Coverage in Canada - Austin and Apold (2023)

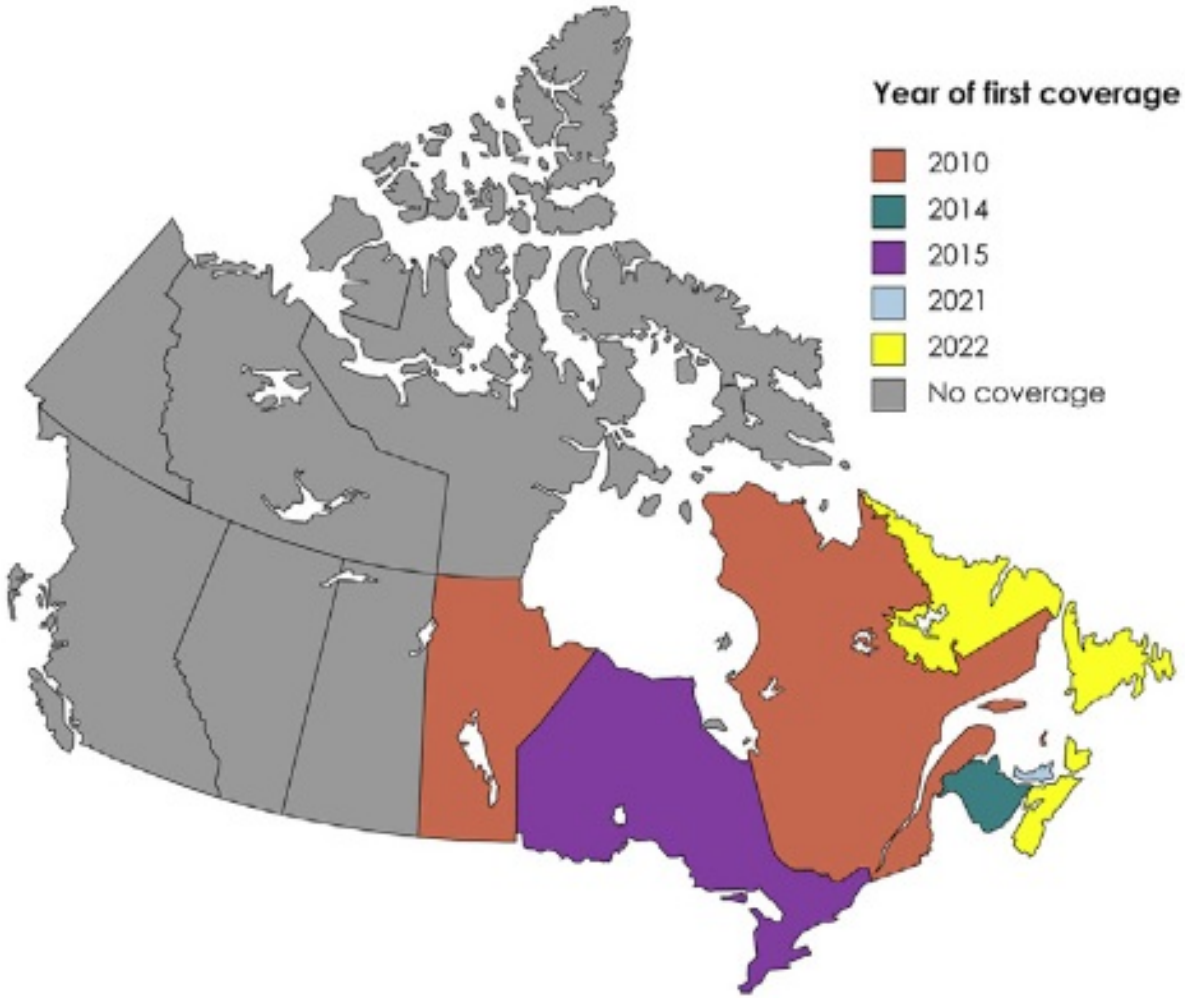


Figure: When Are They Treated?

IVF Coverage in Canada - Austin and Apold (2023)

Figure: How Much is Covered?

IVF Coverage in Canada - Austin and Apold (2023)

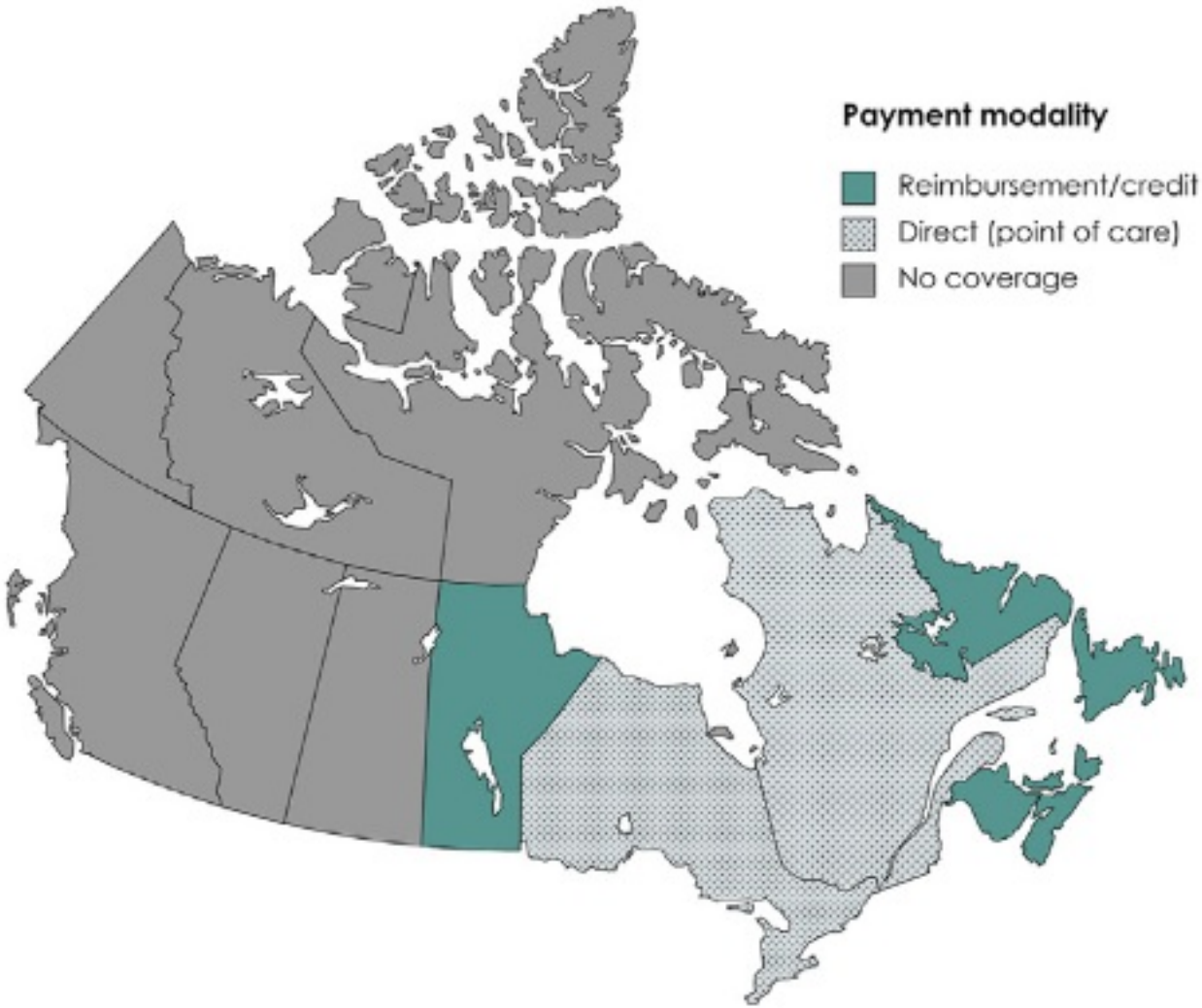


Figure: How are Benefits Paid?

ATT(g,t) or ATT(s,t)?

From Mikola and Webb, *Economics of Education Review*, 2023

	SK	MB	NS	NB
start year	2007	2007	2006	2005
maximum amount	20k	25k	15k	20k
rebate per year	10%,20%	4k, 10%	2.5k	4k
NPV @ 5%	16.9	14.1	13.3	12.6
refundable credit	Y*	N	N	N
rollover credit	N*	Y	N	Y
eligibility duration	7	10	6	20
application req.	N	N	Y	Y
tuition based	Y	Y	N	Y
tuit % refunded	100%	60%	-	50%
program costs	35m	34m	25m	

Unconditional and Conditional $ATT(s,t)$'s

- Under the assumptions of *complete* conditional parallel trends (CPT) and no anticipation (NA) assumptions, the $ATT(s,t)$ is defined as:

$$ATT(s,t) = \Big[E\big[E[Y_{i,s,t}|S_i = g, X_{i,s,t}] - E[Y_{i,s,r-1}|S_i = g, X_{i,s,r-1}]\big] - E\big[E[Y_{i,g',t}|G_i = g', X_{i,g',t}] - E[Y_{i,g',r-1}|G_i = g', X_{i,g',r-1}]\big] \Big| G_i = g \Big].$$

- The **conditional** $ATT(s,t)$ for a given value of $X_{i,s,t}$ is given by:

$$\begin{aligned} ATT(s,t,x) = &\big[E[Y_{i,s,t}|S_i = g, X_{i,s,t}] - E[Y_{i,s,r-1}|S_i = g, X_{i,s,r-1}]\big] \\ &- \big[E[Y_{i,g',t}|G_i = g', X_{i,g,t}] - E[Y_{i,g',r-1}|G_i = g', X_{i,g,r-1}]\big]. \end{aligned} \tag{2}$$

- where, g' is a group that has not been treated at time t

ATT(s,t)

- There is a trade-off of using ATT(s,t) versus ATT(g,t)
- The former are probably more policy relevant
- The latter can offer more reliable inference, but not always
- Reliability can depend on how many S are in a given g
- We compare randomization inference and the jackknife in the simulations
- The parallel trends assumptions are similar, but differ in an interesting way

All trends parallel

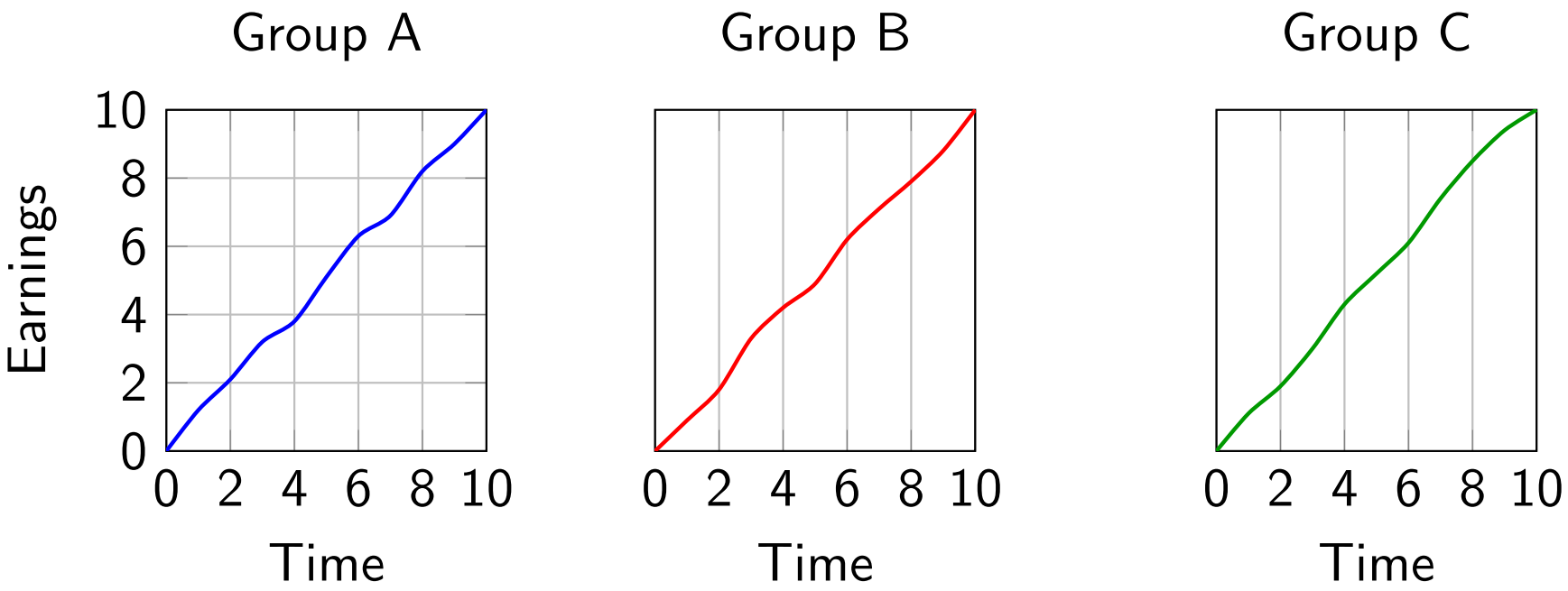


Figure: Earnings trajectories for Groups A, B, and C over time

Other Trend

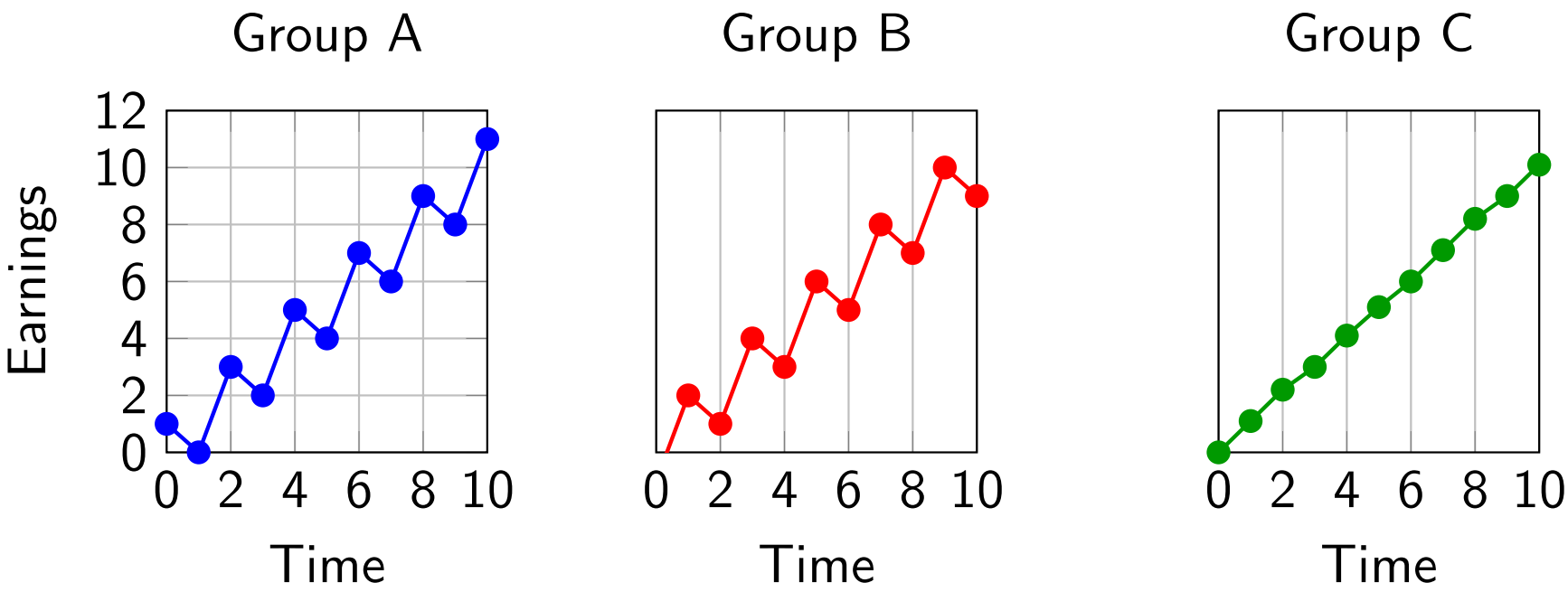


Figure: Earnings with alternating shocks above/below the 45-degree line for A and B

Parallel Trends Assumption: Average Treated vs. Control

Setting:

- Treated groups: A and B
- Control group: C
- Define $G = \frac{1}{2}(Y_A + Y_B)$

Parallel Trends Assumption:

- In the absence of treatment, the average of A and B would have followed the same trend as C:

$$\mathbb{E}[G_t - G_{t-1} \mid \text{no treatment}] = \mathbb{E}[Y_{C,t} - Y_{C,t-1}]$$

- Explicitly, with $G_t = \frac{1}{2}(Y_{A,t} + Y_{B,t})$:

$$\mathbb{E}\left[\frac{1}{2}(Y_{A,t} - Y_{A,t-1} + Y_{B,t} - Y_{B,t-1}) \mid \text{no treatment}\right] = \mathbb{E}[Y_{C,t} - Y_{C,t-1}$$

Parallel Trends Assumptions for $ATT_{s,t}$

Setting:

- Groups: A (treated), B (treated), C (control)
- Outcome: Y_{gt} for group $g \in \{A, B, C\}$ at time t

Parallel Trends Assumptions:

- In the absence of treatment, Group A would have followed the same trend as Group C:

$$\mathbb{E}[Y_{A,t} - Y_{A,t-1} \mid \text{no treatment}] = \mathbb{E}[Y_{C,t} - Y_{C,t-1}]$$

- Similarly, for Group B:

$$\mathbb{E}[Y_{B,t} - Y_{B,t-1} \mid \text{no treatment}] = \mathbb{E}[Y_{C,t} - Y_{C,t-1}]$$

Parallel Trends Assumptions for $ATT_{g,t}$

Setting:

- Treated groups: A and B
- Control group: C
- Define $G = \frac{1}{2}(Y_A + Y_B)$

Parallel Trends Assumption:

- In the absence of treatment, the average of A and B would have followed the same trend as C:

$$\mathbb{E}[G_t - G_{t-1} \mid \text{no treatment}] = \mathbb{E}[Y_{C,t} - Y_{C,t-1}]$$

- Explicitly, with $G_t = \frac{1}{2}(Y_{A,t} + Y_{B,t})$:

$$\mathbb{E}\left[\frac{1}{2}(Y_{A,t} - Y_{A,t-1} + Y_{B,t} - Y_{B,t-1}) \mid \text{no treatment}\right] = \mathbb{E}[Y_{C,t} - Y_{C,t-1}$$

Two Estimators for $ATT_{s,t}$

- We can estimate the $ATT_{s,t}$ terms using two new estimators:
  - UN-DID
  - DID-INT
- Both of these involve a multi-step process for estimating $ATT_{s,t}$ or $ATT_{g,t}$
- Fortunately, the final steps are fast, which allows for both:
  - Fast estimation of different aggregations
  - Fast jackknife or randomization inference

UN-DID - Difference-in-differences With Unpoolable Data

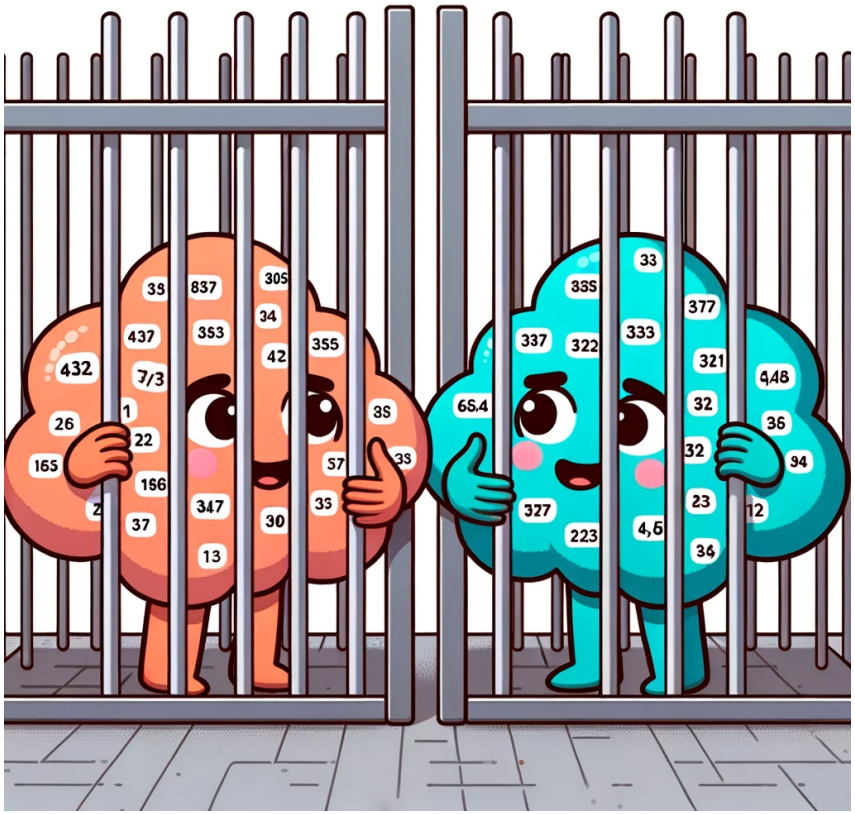


Figure: Unpooled Data

- This research agenda started with the problem of unpoolable data
- Treatment and control observations in different datasets
- This is very common with health data
- Most(?) administrative datasets cannot be merged across regions
- Difference-in-Differences with Unpoolable Data
- Karim, Webb, Austin, Strumpf - Arxiv (2024)

UNDID - Unpooled Estimator with covariates, 2X2

$$\begin{aligned} \text{ATT}_X &= (\mathbb{E}[Y \mid T, post, X] - \mathbb{E}[Y \mid T, pre, X]) \\ &\quad - (\mathbb{E}[Y \mid C, post, X] - \mathbb{E}[Y \mid C, pre, X]) \end{aligned}$$

- For j = {T, C}

$$\text{For treated: } Y_{i,t}^{T} = \lambda_1^{T} pre_t^{T} + \lambda_2^{T} post_t^{T} + \lambda_3^{T} X_{i,t}^{T} + \nu_{i,t}^{T} \tag{3}$$

$$\text{For untreated: } Y_{i,t}^{C} = \lambda_1^{C} pre_t^{C} + \lambda_2^{C} post_t^{C} + \lambda_3^{C} X_{i,t}^{C} + \nu_{i,t}^{C} \tag{4}$$

- $\widehat{ATT} = (\widehat{\lambda_2^{T}} - \widehat{\lambda_1^{T}}) - (\widehat{\lambda_2^{C}} - \widehat{\lambda_1^{C}})$

The Intersection Difference-in-Difference Estimator

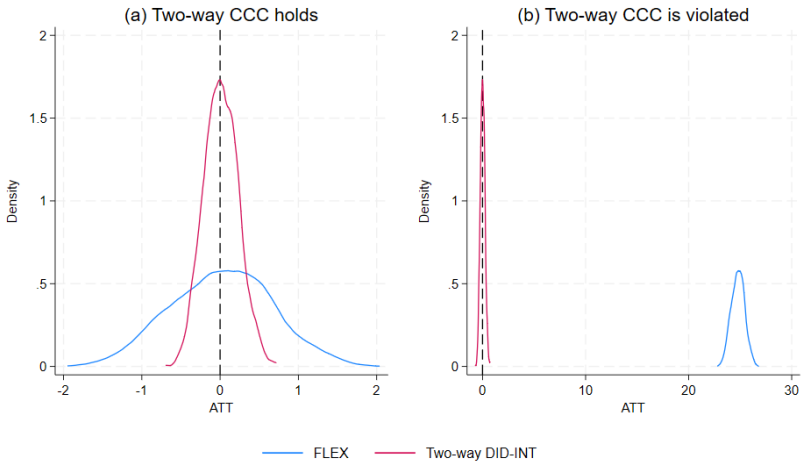


Figure: CCC Violations

- The DID-INT estimator is designed for pooled data
- Is robust to time varying covariates
- Is robust to violations of the *common causal covariate* assumption implicit in most DID estimators
- *Good Controls Gone Bad...* Karim and Webb, Arxiv (2024)

Intersection Difference-in-differences

- Generate three types of dummy variables:
  1. $I(g,t)$ is a dummy which takes a value of 1 if the observation is in group g in period t
  2. $I(g)$ is a dummy variable which takes a value of 1 if the observation is in group g
  3. $I(t)$ is a dummy variable which takes a value of 1 if the observation is from year t
- Here, g is an index for group, t is an index for time and k is an index for control variables (there are K covariates)
- We propose running the following regression:

$$Y_{i,g,t} = \sum_{g} \sum_{t} \lambda_{g,t} I(g,t) + f(X^{k}_{i,g,t}) + \epsilon_{i,g,t} \tag{5}$$

- Note: The regression is done without a constant

Estimate of the ATT from DID-INT

- Let g' be a relevant control group for group g.
- t' is the year group g is **first treated**
- The $\widehat{ATT}(g,t)$ from DID-INT is as follows: Estimand of ATT(g,t)

$$\widehat{ATT}(g,t) = (\widehat{\lambda_{g,t}} - \widehat{\lambda_{g,t'-1}}) - (\widehat{\lambda_{g',t}} - \widehat{\lambda_{g',t'-1}}). \tag{6}$$

- $\widehat{ATT}(g,t)$ is an unbiased estimator of the estimand (proofs in paper).
- The overall estimate of the ATT is given by:

$$\widehat{ATT} = \sum_{g=2}^{G} \sum_{t=2}^{\mathcal{T}} 1\{t' \leq t\} w_{g,t} \widehat{ATT}(g,t). \tag{7}$$

- This does not include "forbidden comparisons"
- Cluster robust inference can be done using Jackknife/randomization inference.

Estimating an ATT from a Second Stage Regression

- Assume a simple staggered adoption setting with 3 groups and 3 periods.
- A is first treated in Period 2, B is first treated in Period 3, and C is never treated.

Staggered Example

Contrast	Silo	D	G	T	Difference (diff)
A22	A	1	2	2	$\bar{Y}_{A2} - \bar{Y}_{A1}$
A23	A	1	2	3	$\bar{Y}_{A3} - \bar{Y}_{A1}$
B33	B	1	3	3	$\bar{Y}_{B3} - \bar{Y}_{B2}$
C22	C	0	2	2	$\bar{Y}_{C2} - \bar{Y}_{C1}$
C23	C	0	2	3	$\bar{Y}_{C3} - \bar{Y}_{C1}$
C33	C	0	3	3	$\bar{Y}_{C3} - \bar{Y}_{C2}$

The ATT(g,t)s are estimated using the following regression:

$$\texttt{diff}_{s,g,t} = \alpha + \beta d_{s,g,t} + \epsilon_{s,g,t} \text{ if } G = g \text{ and } T = t. \tag{8}$$

ATT(s,t)

- We can instead estimate a series of ATT(s,t) term using the following equations:

$$\text{diff}_{s,g,t} = \alpha + \beta d_{s,g,t} + \epsilon_{s,g,t} \text{ for each } s \text{ in } g \text{ if } T = t. \quad (9)$$

- Essentially, we loop over treated states, rather than treatment groups
- Randomization Inference involves repeating this regression multiple time, permuting the timing of treatment/control

Jackknife Solution

- If we wish to estimate a simple (weighted) ATT, we can instead run a single regression:

$$\text{diff}_{s,g,t} = \beta d_{s,g,t} + \sum_{g=1}^{G} \sum_{t=1}^{T} \alpha_{g,t} I(g) I(t) + \epsilon_{s,g,t} \tag{10}$$

- So long as there are at least two control states we can estimate cluster standard errors using the cluster jackknife
- The interpretation of the jackknife is conceptually somewhat different if some of the ATT(g,t) groups contain only one state
- Weighting the ATT by states rather than cohorts requires additional steps

Jackknife Example

Contrast	Silo	D	G	T	Difference (diff)
A22	A	1	2	2	$\bar{Y}_{A2} - \bar{Y}_{A1}$
A23	A	1	2	3	$\bar{Y}_{A3} - \bar{Y}_{A1}$
B33	B	1	3	3	$\bar{Y}_{B3} - \bar{Y}_{B2}$
C22	C	0	2	2	$\bar{Y}_{C2} - \bar{Y}_{C1}$
C23	C	0	2	3	$\bar{Y}_{C3} - \bar{Y}_{C1}$
C33	C	0	3	3	$\bar{Y}_{C3} - \bar{Y}_{C2}$
D22	D	1	2	2	$\bar{Y}_{D2} - \bar{Y}_{D1}$
D23	D	1	2	3	$\bar{Y}_{D3} - \bar{Y}_{D1}$
E33	E	1	3	3	$\bar{Y}_{E3} - \bar{Y}_{E2}$
F22	F	0	2	2	$\bar{Y}_{F2} - \bar{Y}_{F1}$
F23	F	0	2	3	$\bar{Y}_{F3} - \bar{Y}_{F1}$
F33	F	0	3	3	$\bar{Y}_{F3} - \bar{Y}_{F2}$

Note, A,D are $g = 2$, B, E are $g = 3$, C,F are $g = \infty/0$. We can estimate each ATT(g,t) in each of the 6 jackknife replications.

Jackknife Variance

$$CV_3(\beta) = \frac{G-1}{G} \sum_{g=1}^{G} (\widehat{\beta}^{(g)} - \widehat{\beta})^2, \tag{11}$$

where $\widehat{\beta}$ is the original (full sample)

estimate of the ATT, and $\widehat{\beta}^{(g)}$ is the ATT estimate from omitting cluster g.

`jackknife: reg diff treat i.g#i.t, cluster(silo)`

Monte Carlo

- We test size at 5% for the aggregates: ATT, or ATT_S
- We use a design similar to Bertrand, Duflo, and Mullainathan (2004)
- This uses CPS data, and women's wages from the ages of 24 to 55
- Covering the period from 2000 to 2019
- These simulations choose random subsets of states each time replication
- Covariates: race, married, educational status

Rejection Rates: Randomization Inference ATT_S

	G8	G16	G32
S1	0.0570	0.0595	0.0524
S2	0.0434	0.0481	0.0478
S3	0.0454	0.0458	0.0434
S4		0.0443	0.0495
S6			0.0464
S8			0.0462
S10			0.0451
S12			0.0514

Table: ATT_s Early Adopters

Rejection Rates: Randomization Inference ATT_S

	G8	G16	G32
S1	0.0496	0.0656	0.0464
S2	0.0453	0.0527	0.0464
S3	0.0444	0.0422	0.0452
S4		0.0383	0.0433
S6			0.0426
S8			0.0388
S10			0.0410
S12			0.0418

Table: ATT_s Late Adopters

Conclusions

- CS-DID aggregate treated groups by the time of adoption
- The details of the policies often matter to researchers, rather than the timing
- UN-DID and DID-INT allow for state$\times$year level treatment effect estimation, $ATT_{(s,t)}$
- These estimates can be aggregated into state level treatment effects, ATT_S
- Inference for the latter seems quite good using Randomization Inference
- Inference for the former needs some work

References

Karim, Sunny, and Matthew D Webb (2024) 'Good controls gone bad: Difference-in-differences with covariates.' *arXiv* p. 2412.14447

Karim, Sunny, Matthew D. Webb, Nichole Austin, and Erin Strumpf (2024) 'Difference-in-differences with unpoolable data.' *arXiv* p. 2403.15910