SISCER Module 13
Lecture 6: Difference-in-Differences and Time-Varying Treatments

Ting Ye & Qingyuan Zhao

University of Washington & University of Cambridge

July 2025

Acknowledgment: this lecture benefits from course materials by Andrea Rotnitzky and Richard Guo.

Plan

Difference-in-Differences

DID extensions

Time-varying treatments

Marginal structural model

Key references for this lecture

- Difference-in-differences: Wing et al. (2018) and Roth et al. (2022) for
- Time-varying treatments: Hernan and Robins book "What if" chapters 19, 20 and 21

Review: casual inference in observational studies

- Methods under the no unmeasured confounders assumption
  1. Matching and entropy balancing weight (Lecture 2)
  2. G-computation, IPW, AIPW (Lecture 4)
- Methods to address unmeasured confounding
  1. Sensitivity analysis (Lecture 3)
  2. Natural experiment: instrumental variable (Lecture 5), regression discontinuity design[1]
  3. Causal exclusion: negative control exposure/outcome (Proximal inference), difference-in-differences (this lecture)

[1] See https://en.wikipedia.org/wiki/Regression_discontinuity_design. Biggs et al. (2017) applied the regression discontinuity design to compare those who received abortions and those were denied abortion in the near-limit group.

Motivations

- We can draw causal inference if controlling for all confounders
- If important confounders are unobserved, we might try to get at causal effects using instrumental variables (IVs) or other methods
- Good IVs are hard to find, however, so we'd like to have other tools to deal with unobserved confounders.
- DID is another strategy that uses data with a time dimension to control for unmeasured but fixed confounding

Example in labor economics: do minimum wage laws affect employment (Card and Krueger, 1993)

- On April 1, 1992, New Jersey raised the state minimum from $4.25 to $5.05.
- Card and Krueger collected data on employment at fast food restaurants (Burger King, Wendy's, and so on) in New Jersey in February 1992 and again in November 1992.
- Card and Krueger collected data from the same type of restaurants in eastern Pennsylvania, just across the Delaware river. The minimum wage in Pennsylvania stayed at $4.25 throughout this period.
- They compared the change in employment in New Jersey to the change in employment in Pennsylvania around the time New Jersey raised its minimum (a DID estimate).

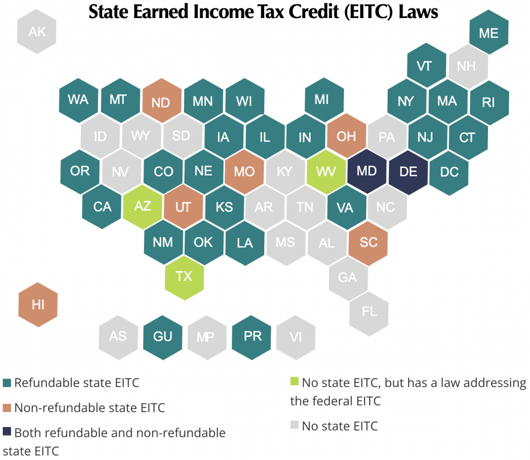
Example in labor economics: do minimum wage laws affect employment (Card and Krueger, 1993)

Table 5.2.1: Average employment per store before and after the New Jersey minimum wage increase

Variable		PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1.	FTE employment before,	23.33	20.44	-2.89
	all available observations	(1.35)	(0.51)	(1.44)
2.	FTE employment after,	21.17	21.03	-0.14
	all available observations	(0.94)	(0.52)	(1.07)
3.	Change in mean FTE	-2.16	0.59	2.76
	employment	(1.25)	(0.54)	(1.36)

Notes: Adapted from Card and Krueger (1994), Table 3. The table reports average full-time equivalent (FTE) employment at restaurants in Pennsylvania and New Jersey before and after a minimum wage increase in New Jersey. The sample consists of all stores with data on employment. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing. Standard errors are reported in parentheses

Example: do earned income tax credits (EITC) reduce deaths of despair?

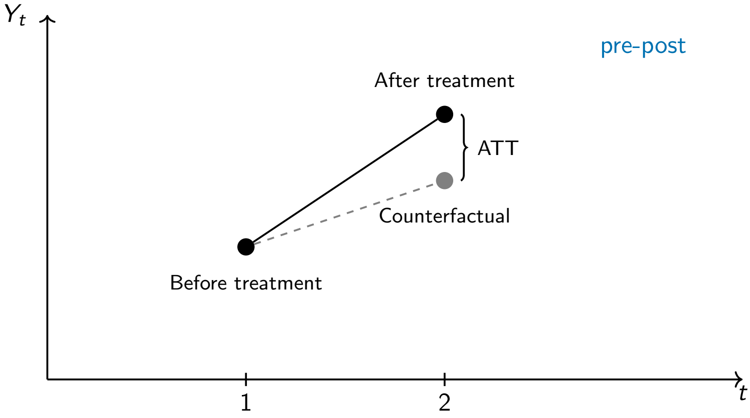


(National Conference of State Legislatures)

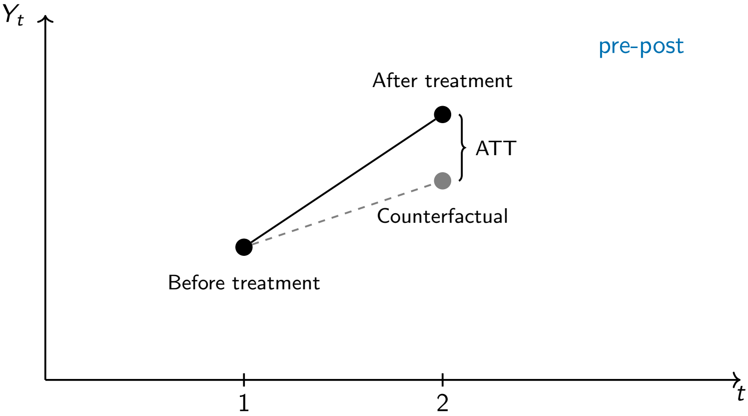
Difference-in-Differences (DID) for causal effect

- Challenges from unmeasured confounding: states with EITC laws differ from states without them in other ways that may be related to deaths of despair
- A before-after comparison of the same units can also be biased due to time trends in the outcome even without the treatment
- DID uses both comparison, and is commonly used for estimating causal effects with panel data
- Prototypical DID application: how do changes in state policies affect individual
  - Did Missouri's handgun purchaser licensing law affects firearm homicide rates?
  - Did minimum wage laws change employment levels?
  - Motivating application: do EITC reduce deaths of despair?

Canonical DID

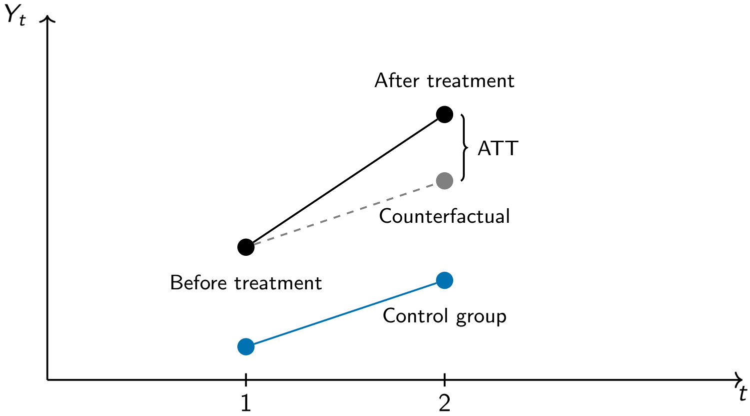


Canonical DID



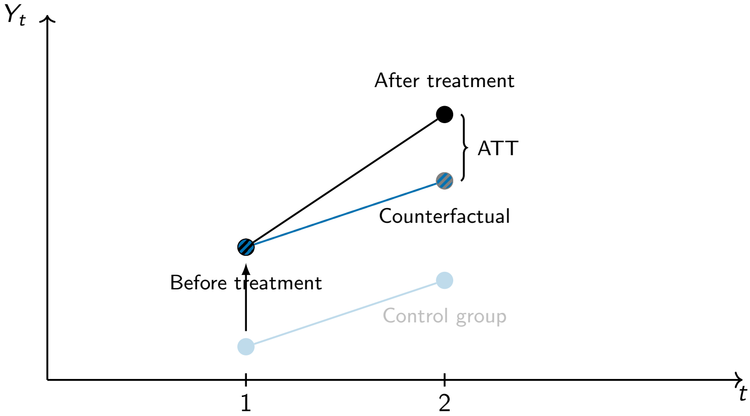
Canonical DID

Parallel Trends: Absent treatment, treated and control would evolve over time in the same way. (functional-form dependent)



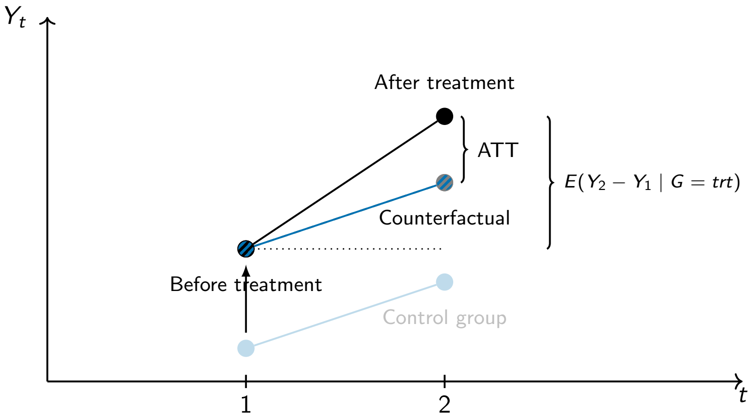
Canonical DID

Parallel Trends: Absent treatment, treated and control would evolve over time in the same way. (functional-form dependent)



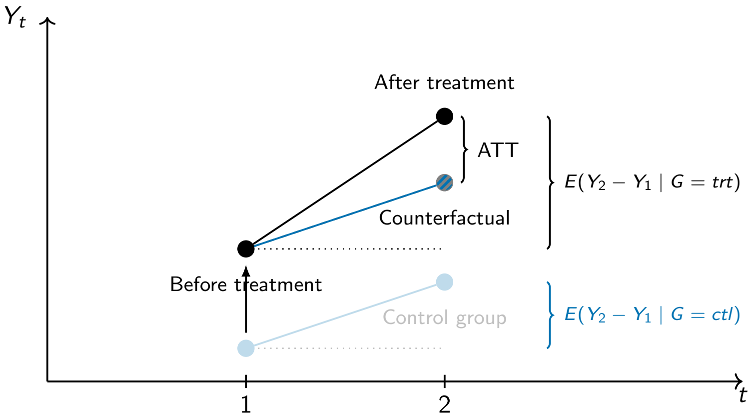
Canonical DID

Parallel Trends: Absent treatment, treated and control would evolve over time in the same way. (functional-form dependent)



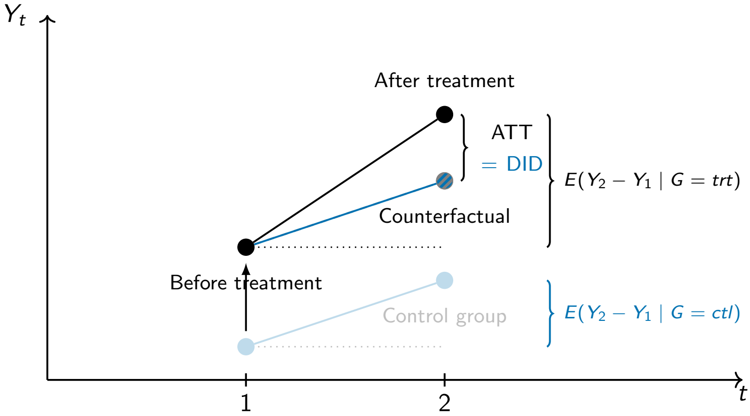
Canonical DID

Parallel Trends: Absent treatment, treated and control would evolve over time in the same way. (functional-form dependent)



Canonical DID

Parallel Trends: Absent treatment, treated and control would evolve over time in the same way. (functional-form dependent)



Potential outcomes and causal effect

Observed data

- $A_i = 1$ is the treated group and $A_i = 0$ is the control group
- For every unit i, we measure Y_{i1}, Y_{i2} before and after the treated group adopts the treatment

Potential outcomes

- $Y_{it}^{(1)}$ potential outcome for unit i at time t if being treated, $t = 1, 2$
- $Y_{it}^{(0)}$ potential outcome for unit i at time t if being untreated, $t = 1, 2$
- Consistency (SUTVA): $Y_{i1} = Y_{i1}^{(0)}$ and $Y_{i2} = A_i Y_{i2}^{(1)} + (1 - A_i) Y_{i2}^{(0)}$
- Parallel trends assumption (subscript i omitted):

$$E(Y_2^{(0)} - Y_1^{(0)} \mid A = 1) = E(Y_2^{(0)} - Y_1^{(0)} \mid A = 0)$$

- Causal effect: $ATT = E(Y_2^{(1)} - Y_2^{(0)} \mid A = 1)$

Causal identification

Theorem

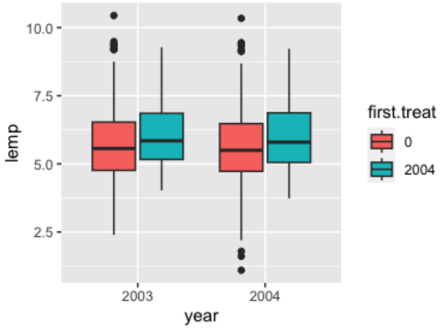
Under the consistency and parallel trends assumptions,
$E(Y_2^{(1)} - Y_2^{(0)} \mid A = 1) = E(Y_2 - Y_1 \mid A = 1) - E(Y_2 - Y_1 \mid A = 0)$

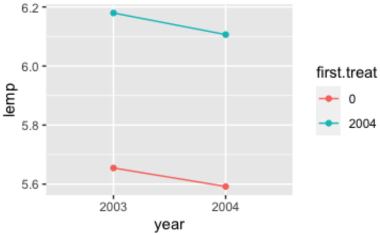
Proof.

$$
\begin{aligned}
& E(Y_2^{(1)} - Y_2^{(0)} \mid A = 1) \\
&= E\left[(Y_2^{(1)} - Y_1^{(0)}) - (Y_2^{(0)} - Y_1^{(0)}) \mid A = 1\right] \\
&= E\left[Y_2^{(1)} - Y_1^{(0)} \mid A = 1\right] - E\left[Y_2^{(0)} - Y_1^{(0)} \mid A = 1\right] \\
&= E\left[Y_2^{(1)} - Y_1^{(0)} \mid A = 1\right] - E\left[Y_2^{(0)} - Y_1^{(0)} \mid A = 0\right] && \text{(parallel trends)} \\
&= E\left[Y_2 - Y_1 \mid A = 1\right] - E\left[Y_2 - Y_1 \mid A = 0\right] && \text{(consistency)}
\end{aligned}
$$

Example in R

- Let's apply the DID to study the effect of the minimum wage on log teen employment.
- The dataset includes county-level data during 2003-2004.
- Treated group: states that increased their minimum wage in 2004
- Control group: states that did not increase their minimum wage during 2003-2004





Example in R

```
> mpdta.sub<-mpdta.sub %>% mutate(after.treat=1*(year>=first.treat))
> # hand-coded DID
> mean(with(mpdta.sub,lemp[first.treat==2004 & year==2004]))-
+     mean(with(mpdta.sub,lemp[first.treat==2004 & year==2003]))-
+     mean(with(mpdta.sub,lemp[first.treat==0 & year==2004]))+
+     mean(with(mpdta.sub,lemp[first.treat==0 & year==2003]))
[1] -0.01050325
> # TWFE version
> twfe_sub<-lm(lemp~year+first.treat+after.treat,data=mpdta.sub)
> # cluster-robust variance estimator with CR2 small-sample correction
> coeftest.twfe <- coef_test(twfe_sub,
+                            vcov = "CR2",
+                            cluster = mpdta.sub$countyreal)
> coeftest.twfe[4,]
       Coef. Estimate     SE t-stat d.f. (Satt) p-val (Satt) Sig.
 after.treat  -0.0105 0.0238 -0.442        21.5        0.663
```

Remarks

- Parallel trends is a strong assumption
  - Testing for pre-trends is a common practice but with caveats (Roth, 2022)
  - Methods to relax parallel trends is an active research area (Rambachan and Roth, 2023; Ye et al., 2023)
  - It is functional form dependent (log or not?)
- All assumptions are on the $Y_{it}^{(0)}$'s, no restrictions on the treatment effect
- DID works under two different settings:
  - Panel data: same units followed over time
  - Repeated cross-sectional: a random (possibly overlap) sample of units at each time
- Estimation:
  - Canonical DID estimator: $(\bar{Y}_{trt,2} - \bar{Y}_{trt,1}) - (\bar{Y}_{ctl,2} - \bar{Y}_{ctl,1})$
  - Static two-way fixed effects (TWFE) model:
    - Panel data: $Y_{it} = \alpha + \delta_t + \gamma A_i + \beta A_i I(t = 2) + \varepsilon_{it}$
    - Repeated cross-sectional data: $Y_{iT_i} = \alpha + \delta_{T_i} + \gamma A_i + \beta A_i I(T_i = 2) + \varepsilon_i$
- Use cluster-robust variance estimator (robust to heteroscedasticity and correlation within county), available from the clubSandwich R package.
  - "CR2" is a type of small sample adjustment (analogous to HC adjustments)

More general set up

Research on DID has been evolving rapidly during the past few years (Roth et al., 2022). But we will cover the main setting and present the key takeaways.

We will cover:

- Observed (time-varying) covariates
- More than two time periods
- Staggered adoption: adopting treatment at different times

We won't cover:

- Non-binary treatments (Callaway et al., 2024)

No staggered adoption

If all treated groups adopt the treatment at the same time:

- Static TWFE model,

$$Y_{git} = \beta D_{gt} + \gamma^T X_{git} + \alpha_g + f_t + \varepsilon_{git}$$

  where D_{gt} is the indicator of being treated, X_{git} are the observed time-varying covariates, α_g is the group indicator, f_t is the time indicator.

- Dynamic TWFE model (event study),

$$Y_{git} = \sum_{-\underline{k} \leq \ell \leq -2} \beta_\ell^{\text{lead}} I(t - E_g = \ell) + \sum_{0 \leq \ell \leq \bar{k}} \beta_\ell^{\text{lag}} I(t - E_g = \ell) + \gamma^T X_{git} + \alpha_g + f_t + \varepsilon_{git}$$

  where E_g is when group g initiates the treatment ($E_g = \infty$ if group g is never treated), and for $\ell \notin [-\underline{k}, \bar{k}]$, usually bin them at $-\underline{k}$ and $\bar{k}$.

Staggered adoption

However, estimators from TWFE models can be difficult to interpret under treatment heterogeneity and staggered adoption (Goodman-Bacon, 2021; Sun and Abraham, 2021; Roth et al., 2022).

- TWFE estimator is a weighted average of group-year treatment effects and the weights (especially the treatment effect for early adopters at a late period) can be negative!

Recommendations:

- Use the static TWFE model only if confident in treatment effect homogeneity
- Use the dynamic TWFE model only if confident that there is heterogeneity only in time since treatment
- Otherwise, consider using a "heterogeneity-robust" estimator, e.g., Callaway and Sant'Anna (2021)

Time-varying treatments

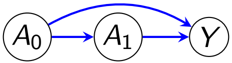
Two treatments, randomized

1. Time 0: randomly assign A_0 (1: treated; 0: control)
2. Time 1: randomly assign A_1 (1: treated; 0: control) depending on A_0.
3. Time 2: measure outcome Y

$$
\begin{tikzcd}
A_0 \arrow[r] \arrow[rr, bend left] & A_1 \arrow[r] & Y
\end{tikzcd}
$$

Two treatments, randomized

1. Time 0: randomly assign A_0 (1: treated; 0: control)
2. Time 1: randomly assign A_1 (1: treated; 0: control) depending on A_0.
3. Time 2: measure outcome Y



(A_0, A_1) as a whole is randomized (why?), so

$$E[Y(a_0, a_1)] = E[Y \mid A_0 = a_0, A_1 = a_1].$$

Two treatments, randomized (more complicated)

Study of the effect of antiretroviral therapy on a health score (Robins and Hernan, 2008): 32,000 HIV infected subjects followed for one year.

Two treatments, randomized (more complicated)

Study of the effect of antiretroviral therapy on a health score (Robins and Hernan, 2008): 32,000 HIV infected subjects followed for one year.

1. Month 0: Assign therapy ($A_0 = 1$: treated; $A_0 = 0$: control) at the start of the follow-up. ☞ Suppose A_0 is randomly assigned.

Two treatments, randomized (more complicated)

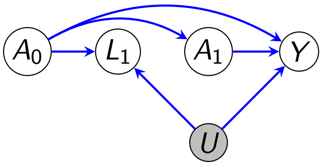
Study of the effect of antiretroviral therapy on a health score (Robins and Hernan, 2008): 32,000 HIV infected subjects followed for one year.

1. Month 0: Assign therapy ($A_0 = 1$: treated; $A_0 = 0$: control) at the start of the follow-up. ☞ Suppose A_0 is randomly assigned.
2. Month 6: Measure blood CD4 counts L_1 and assign therapy ($A_1 = 1$: treated; $A_1 = 0$: control). ☞ Suppose A_1's assignment depends only on A_0 but not L_1

Two treatments, randomized (more complicated)

Study of the effect of antiretroviral therapy on a health score (Robins and Hernan, 2008): 32,000 HIV infected subjects followed for one year.

1. Month 0: Assign therapy ($A_0 = 1$: treated; $A_0 = 0$: control) at the start of the follow-up. ☞ Suppose A_0 is randomly assigned.
2. Month 6: Measure blood CD4 counts L_1 and assign therapy ($A_1 = 1$: treated; $A_1 = 0$: control). ☞ Suppose A_1's assignment depends only on A_0 but not L_1
3. Month 12: Measure the final health score Y.



U represents **unobserved** health status that affects both L_1 and Y.

☞ Can we identify $E[Y(a_0, a_1)]$? Yes, non-causal path is blocked and thus $E[Y(a_0, a_1)] = E[Y \mid A_0 = a_0, A_1 = a_1]$.

Two treatments, with time-varying confounder

Two treatments, with time-varying confounder

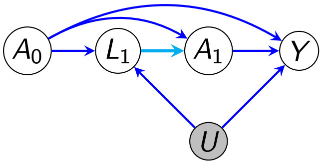
1. Month 0: Assign therapy ($A_0 = 1$: treated; $A_0 = 0$: control) at the start of the follow-up. ☞ Suppose A_0 is randomly assigned.

Two treatments, with time-varying confounder

1. Month 0: Assign therapy ($A_0 = 1$: treated; $A_0 = 0$: control) at the start of the follow-up. ☞ Suppose A_0 is randomly assigned.
2. Month 6: Measure blood CD4 counts L_1 and assign therapy ($A_1 = 1$: treated; $A_1 = 0$: control). ☞ Suppose A_1's assignment depends on both A_0 and L_1

Two treatments, with time-varying confounder

1. Month 0: Assign therapy ($A_0 = 1$: treated; $A_0 = 0$: control) at the start of the follow-up. ☞ Suppose A_0 is randomly assigned.
2. Month 6: Measure blood CD4 counts L_1 and assign therapy ($A_1 = 1$: treated; $A_1 = 0$: control). ☞ Suppose A_1's assignment depends on both A_0 and L_1
3. Month 12: Measure the final health score Y.

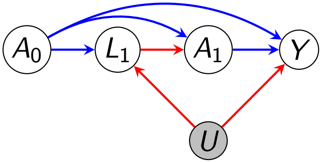


☞ Can we identify $E[Y(a_0, a_1)]$?

Dilemma

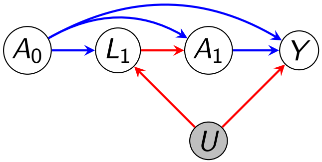
Dilemma

1. Not adjusting for L_1, then A_1 will be confounded

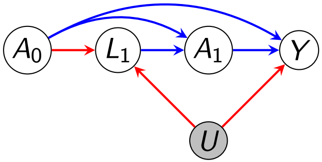


Dilemma

1. Not adjusting for L_1, then A_1 will be confounded

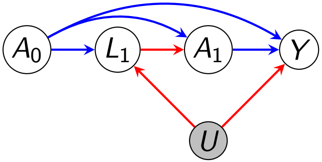


2. Adjusting for L_1, opens a non-causal path from A_0 to Y (collider bias)

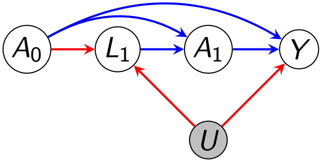


Dilemma

1. Not adjusting for L_1, then A_1 will be confounded

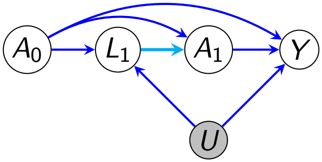


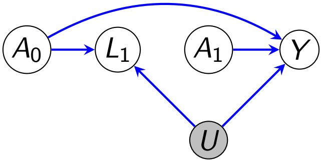
2. Adjusting for L_1, opens a non-causal path from A_0 to Y (collider bias)



- Need something more sophisticated.

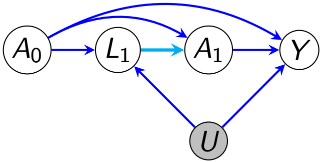
IPW: Removing A_1's dependency on L_1

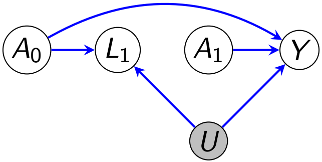




IPW: weight $= 1/p(A_1 \mid A_0, L_1)$

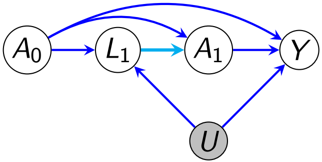
IPW: Removing A_1's dependency on L_1

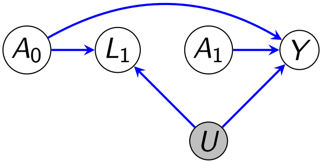




IPW: weight $= 1/p(A_1 \mid A_0, L_1)$

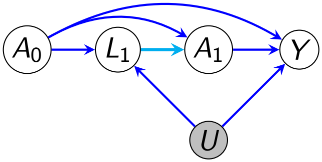
IPW: Removing A_1's dependency on L_1

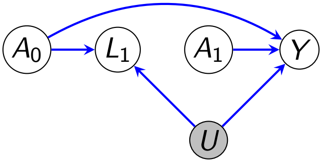




IPW: weight $= 1/p(A_1 \mid A_0, L_1)$

IPW: Removing A_1's dependency on L_1

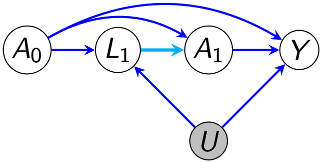


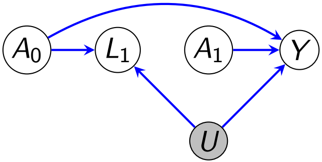


IPW: weight $= 1/p(A_1 \mid A_0, L_1)$

☞ After reweighting, $\mathbb{E}\, Y(a_0, a_1) = \mathbb{E}_w[Y \mid A_0 = a_0, A_1 = a_1]$.

IPW: Removing A_1's dependency on L_1





IPW: weight $= 1/p(A_1 \mid A_0, L_1)$

☞ After reweighting, $\mathbb{E}\, Y(a_0, a_1) = \mathbb{E}_w[Y \mid A_0 = a_0, A_1 = a_1]$.

IPW identification:

$$\boxed{\mathbb{E}\, Y(a_0, a_1) = \mathbb{E}\left\{\frac{Y\, \mathbb{I}_{A_0=a_0, A_1=a_1}}{P(A_1 = a_1 \mid A_0 = a_0, L_1)}\right\} \Big/ \mathbb{E}\left\{\frac{\mathbb{I}_{A_0=a_0, A_1=a_1}}{P(A_1 = a_1 \mid A_0 = a_0, L_1)}\right\}.}$$

Exercise 1: data table

row	n	A_0	L_1	A_1	$E(Y \mid A_0, L_1, A_1)$
1	2000	0	1	0	200
2	6000	0	1	1	220
3	6000	0	0	0	50
4	2000	0	0	1	70
5	3000	1	1	0	130
6	9000	1	1	1	110
7	3000	1	0	0	230
8	1000	1	0	1	250

Exercise 1: data table

row	n	A_0	L_1	A_1	$E(Y \mid A_0, L_1, A_1)$	weight $(1/p(A_1 \mid A_0, L_1))$	n-pseudo
1	2000	0	1	0	200	4	8000
2	6000	0	1	1	220	4/3	8000
3	6000	0	0	0	50	4/3	8000
4	2000	0	0	1	70	4	8000
5	3000	1	1	0	130	4	12000
6	9000	1	1	1	110	4/3	12000
7	3000	1	0	0	230	4/3	4000
8	1000	1	0	1	250	4	4000

Crude means in pseudo-study $\mathbb{E}_w[Y \mid A_0 = a_0, A_1 = a_1] = \mathbb{E}\, Y(a_0, a_1)$

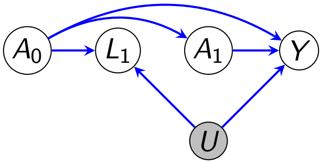
- $\mathbb{E}\, Y(0,0) = \mathbb{E}_w[Y \mid A_0 = 0, A_1 = 0] = (200 * 8000 + 50 * 8000)/(8000 + 8000) = 125$
- Finish calculating $\mathbb{E}\, Y(0,1), \mathbb{E}\, Y(1,0), \mathbb{E}\, Y(1,1)$. Calculate and interpret their contrast.
- How is the above results compared to crude means in the actual study $\mathbb{E}(Y \mid A_0 = a_0, A_1 = a_1)$?

Rationale of the IPW procedure: summary

- We can pretend that the pseudo study is formed by two copies ("clones") of each person, one clone receives $A_1 = 0$ and the other receives $A_1 = 1$. So we can pretend that to assign A_1 we have flipped one same coin for everyone.
- Since we have also flipped one same coin for everyone to assign A_0 (might be different from the imaginary coin for assigning A_1)
- Then, in the pseudo study, the subjects assigned to each of the four treatment arms $(A_0, A_1) = \{(0,0),(0,1),(1,0),(1,1)\}$ are exchangeable, so we can estimate the counterfactual means $\mathbb{E}\, Y(a_0, a_1)$ with the crude means in the pseudo study $\mathbb{E}_w[Y \mid A_0 = a_0, A_1 = a_1]$.

Bonus: stabilized IPW

- The IPW procedure we have just seen creates a pseudo-study in which
  - has size equal to the double of the actual study size
  - the crude mean in the pseudo-study $\mathbb{E}_w[Y \mid A_0 = a_0, A_1 = a_1]$ is the counterfactual mean $\mathbb{E}\, Y(a_0, a_1)$
- There is a modification to IPW (called stablized IPW)
  - has size equal to the actual study size
  - the crude mean in the pseudo-study $\mathbb{E}_{sw}[Y \mid A_0 = a_0, A_1 = a_1]$ is the counterfactual mean $\mathbb{E}\, Y(a_0, a_1)$



stabilized weight $= p(A_1 \mid A_0)/p(A_1 \mid A_0, L_1)$

IPW: Identification

$$\boxed{\mathbb{E}\, Y(a_0, a_1) = \mathbb{E}\left\{\frac{Y\, \mathbb{I}_{A_0=a_0, A_1=a_1}}{P(A_1 = a_1 \mid A_0 = a_0, L_1)}\right\} / \,\mathbb{E}\left\{\frac{\mathbb{I}_{A_0=a_0, A_1=a_1}}{P(A_1 = a_1 \mid A_0 = a_0, L_1)}\right\}.}$$

☞ It makes no difference to use the stabilized weight $P(A_1 = a_1 \mid A_0 = a_0)/P(A_1 = a_1 \mid A_0 = a_0, L_1)$.

☞ For some other estimands (like parameters of marginal structural models), there will be differences.

Standardization / g-formula

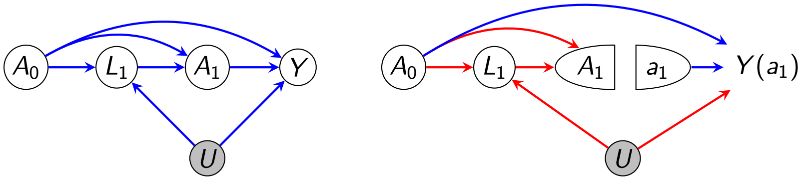
- With a bit more algebra, the IPW formula can be rewritten as

$$
\begin{aligned}
\mathbb{E}\, Y(a_0, a_1) &= \mathbb{E}\left\{\frac{Y\, \mathbb{I}_{A_0=a_0, A_1=a_1}}{P(A_1 = a_1 \mid A_0 = a_0, L_1)}\right\} \Big/ \mathbb{E}\left\{\frac{\mathbb{I}_{A_0=a_0, A_1=a_1}}{P(A_1 = a_1 \mid A_0 = a_0, L_1)}\right\} \\
&= \mathbb{E}\left\{\frac{Y\, \mathbb{I}_{A_0=a_0, A_1=a_1}}{P(A_0 = a_0)\, P(A_1 = a_1 \mid A_0 = a_0, L_1)}\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{E}[Y\, \mathbb{I}_{A_0=a_0, A_1=a_1} \mid L_1]}{P(A_0 = a_0)\, P(A_1 = a_1 \mid A_0 = a_0, L_1)}\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1, L_1] P(A_1 = a_1, A_0 = a_0 \mid L_1)}{P(A_0 = a_0)\, P(A_1 = a_1 \mid A_0 = a_0, L_1)}\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1, L_1] P(A_0 = a_0 \mid L_1)}{P(A_0 = a_0)}\right\} \\
&= \boxed{\sum_{l_1} \mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1, L_1 = l_1] P(L_1 = l_1 \mid A_0 = a_0).}
\end{aligned}
$$

Standardization / g-formula: Intuition

Standardization / g-formula: Intuition

1. Consider $Y(a_1) := Y(A_0, a_1)$.

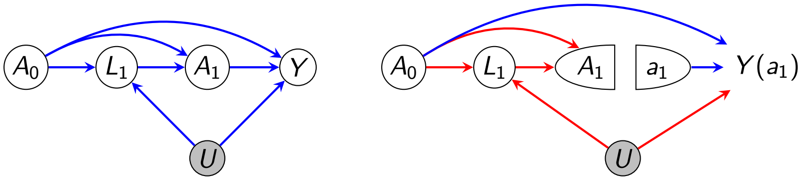


Within the stratum of (A_0, L_1), A_1 is independent of $Y(a_1)$, so (why?)

$$\mathbb{E}[Y(a_1) \mid A_0 = a_0, L_1 = l_1] = \mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1, L_1 = l_1].$$

Standardization / g-formula: Intuition

1. Consider $Y(a_1) := Y(A_0, a_1)$.



Within the stratum of (A_0, L_1), A_1 is independent of $Y(a_1)$, so (why?)

$$\mathbb{E}[Y(a_1) \mid A_0 = a_0, L_1 = l_1] = \mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1, L_1 = l_1].$$

2. Because A_0 is randomly assigned,

$$\mathbb{E}[Y(a_0, a_1)] = \mathbb{E}[Y(a_1) \mid A_0 = a_0] = \sum_{l_1} \mathbb{E}[Y(a_1) \mid A_0 = a_0, L_1 = l_1] P(L_1 = l_1 \mid A_0 = a_0).$$

Positivity

From the standardization / g-formula

$$\boxed{\mathbb{E}\, Y(a_0, a_1) = \sum_{l} \mathbb{E}[Y \mid A_1 = a_1, A_0 = a_0, L_1 = l_1] P(L_1 = l_1 \mid A_0 = a_0),}$$

to identify $\mathbb{E}\, Y(a_0, a_1)$, we must have

$$\forall l_1 : P(L_1 = l_1 \mid A_0 = a_0) > 0 \implies \text{data within } (a_0, a_1, l_1),$$

i.e.,

$$\boxed{\forall l_1 : P(L_1 = l_1 \mid A_0 = a_0) > 0 \implies P(A_1 = a_1 \mid A_0 = a_0, L_1 = l_1) > 0.}$$

☞ This can also be seen in the weighting identification formula.

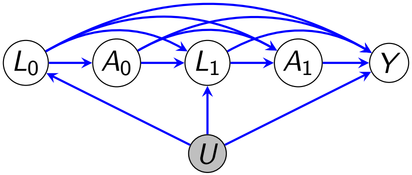
Exercise 2: $\mathbb{E}\, Y(a_0, a_1) = \sum_{l_1} \mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1, L_1 = l_1] P(L_1 = l_1 \mid A_0 = a_0)$

row	n	A_0	L_1	A_1	$E(Y \mid A_0, L_1, A_1)$
1	2000	0	1	0	200
2	6000	0	1	1	220
3	6000	0	0	0	50
4	2000	0	0	1	70
5	3000	1	1	0	130
6	9000	1	1	1	110
7	3000	1	0	0	230
8	1000	1	0	1	250

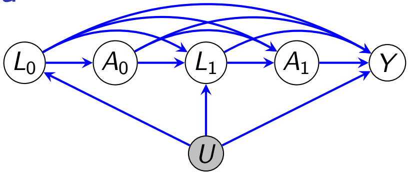
- $P(L_1 = 1 \mid A_0 = 0) = (2000 + 6000)/(2000 + 6000 + 6000 + 2000) = 0.5$, $P(L_1 = 1 \mid A_0 = 1) = (3000 + 9000)/(3000 + 9000 + 3000 + 1000) = 0.75$
- Finish calculating $\mathbb{E}\, Y(0,0), \mathbb{E}\, Y(0,1), \mathbb{E}\, Y(1,0), \mathbb{E}\, Y(1,1)$. Calculate and interpret their contrast.
- How is the above results compared to the IPW results?

Generalization

1. Month 0: Assign therapy ($A_0 = 1$: treated; $A_0 = 0$: control) at the start of the follow-up.
   - Suppose $Y(a_0, a_1) \perp\!\!\!\perp A_0 \mid L_0$ **for baseline covariates** L_0.
2. Month 6: Measure blood CD4 counts L_1 and assign therapy ($A_1 = 1$: treated; $A_1 = 0$: control).
   - Suppose $Y(a_0, a_1) \perp\!\!\!\perp A_1 \mid L_0, A_0, L_1$.
3. Month 12: Measure the final health score Y.



Generalization: g-formula



Under positivity and sequential randomization

$$Y(a_0, a_1) \perp\!\!\!\perp A_0 \mid L_0,$$
$$Y(a_0, a_1) \perp\!\!\!\perp A_1 \mid L_0, A_0, L_1,$$

$$\mathbb{E}\, Y(a_0, a_1) = \sum_{l_0} \sum_{l_1} \mathbb{E}[Y \mid A_1 = a_1, A_0 = a_0, L_1 = l_1, L_0 = l_0]$$
$$\times P(L_1 = l_1 \mid A_0 = a_0, L_0 = l_0) P(L_0 = l_0).$$

▶ Extends to more time points.

Generalization: IPW

Under positivity and sequential randomization

$$Y(a_0, a_1) \perp\!\!\!\perp A_0 \mid L_0,$$
$$Y(a_0, a_1) \perp\!\!\!\perp A_1 \mid L_0, A_0, L_1,$$

$$\mathbb{E}\, Y(a_0, a_1) = \mathbb{E}\left\{ \frac{Y\, \mathbb{I}_{A_0=a_0, A_1=a_1}}{P(A_1 = a_1 \mid A_0 = a_0, L_1, L_0)P(A_0 = a_0 \mid L_0)} \right\} \Big/ \mathbb{E}\left\{ \frac{\mathbb{I}_{A_0=a_0, A_1=a_1}}{P(A_1 = a_1 \mid A_0 = a_0, L_1, L_0)P(A_0 = a_0 \mid L_0)} \right\}.$$

- We can also use stablized weights: $\frac{P(A_1|A_0)P(A_0)}{P(A_1|A_0, L_1, L_0)P(A_0|L_0)}$
- Extends to more time points

Marginal structural model

Marginal structural (mean) model

Consider two treatments A_0, A_1.

- Marginal structural mean model is to postulate and fit

$$\mathbb{E}[Y(a_0, a_1)] = f(a_0, a_1; \theta).$$

Marginal structural (mean) model

Consider two treatments A_0, A_1.

- Marginal structural mean model is to postulate and fit

$$\mathbb{E}[Y(a_0, a_1)] = f(a_0, a_1; \theta).$$

For example, when A_0, A_1 are both **binary:**

- **Saturated model**

$$\mathbb{E}[Y(a_0, a_1)] = \alpha + \beta_0 a_0 + \beta_1 a_1 + \gamma a_0 a_1$$

- **Main effect only**

$$\mathbb{E}[Y(a_0, a_1)] = \alpha + \beta_0 a_0 + \beta_1 a_1$$

Fitting model with IPW

If (A_0, A_1) is randomized, we have $\mathbb{E}[Y(a_0, a_1)] = \mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1]$, so the model can be simply fitted with least squares.

Fitting model with IPW

If (A_0, A_1) is randomized, we have $\mathbb{E}[Y(a_0, a_1)] = \mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1]$, so the model can be simply fitted with least squares.

Now under time-varying confounding, we can use IPW to reweigh data such that we can treat the data as if it comes from a randomized experiment.

Fitting model with IPW

If (A_0, A_1) is randomized, we have $\mathbb{E}[Y(a_0, a_1)] = \mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1]$, so the model can be simply fitted with least squares.

Now under time-varying confounding, we can use IPW to reweigh data such that we can treat the data as if it comes from a randomized experiment.

- To fit marginal structural mean model,

Fitting model with IPW

If (A_0, A_1) is randomized, we have $\mathbb{E}[Y(a_0, a_1)] = \mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1]$, so the model can be simply fitted with least squares.

Now under time-varying confounding, we can use IPW to reweigh data such that we can treat the data as if it comes from a randomized experiment.

- To fit marginal structural mean model,
  1. Estimate the propensity score $\widehat{P}(a_1 \mid a_0, l_1)$ (e.g., with logistic regression)

Fitting model with IPW

If (A_0, A_1) is randomized, we have $\mathbb{E}[Y(a_0, a_1)] = \mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1]$, so the model can be simply fitted with least squares.

Now under time-varying confounding, we can use IPW to reweigh data such that we can treat the data as if it comes from a randomized experiment.

- To fit marginal structural mean model,
  1. Estimate the propensity score $\widehat{P}(a_1 \mid a_0, l_1)$ (e.g., with logistic regression)
  2. Compute weights $\widehat{w} = 1/\widehat{P}(A_1 \mid A_0, L_1)$ or the stabilized weights

$$\widehat{w}_s = \left(\sum_{l_1} \widehat{P}(A_1 \mid A_0, l_1) \widehat{P}(l_1 \mid A_0) \right) / \widehat{P}(A_1 \mid A_0, l_1).$$

Fitting model with IPW

If (A_0, A_1) is randomized, we have $\mathbb{E}[Y(a_0, a_1)] = \mathbb{E}[Y \mid A_0 = a_0, A_1 = a_1]$, so the model can be simply fitted with least squares.

Now under time-varying confounding, we can use IPW to reweigh data such that we can treat the data as if it comes from a randomized experiment.

- To fit marginal structural mean model,
  1. Estimate the propensity score $\widehat{P}(a_1 \mid a_0, l_1)$ (e.g., with logistic regression)
  2. Compute weights $\widehat{w} = 1/\widehat{P}(A_1 \mid A_0, L_1)$ or the stabilized weights

$$\widehat{w}_s = \left(\sum_{l_1} \widehat{P}(A_1 \mid A_0, l_1)\widehat{P}(l_1 \mid A_0) \right) / \widehat{P}(A_1 \mid A_0, l_1).$$

  3. Fit least squares using $\widehat{w}$ or $\widehat{w}_s$ as weights.

- It makes a difference here.
- Statistical inference: bootstrap.

Discussions: Static vs dynamic treatment regimes

- **Static regime**: everybody receives $A_0 = a_0$ and $A_1 = a_1$ regardless of the patient characteristics,
  - e.g. everybody receives ART the second time but not the first
- **Dynamic regime**: subject receives ART depending on the values of recorded covariates
  - E.g. nobody receives ART the first time and only those whose CD4 count are below 200 receive ART the second time.
- Today we have focused on the effects of static. The same idea applies to dynamic regime, with some delicate differences.

Biggs, M. A., Upadhyay, U. D., McCulloch, C. E., and Foster, D. G. (2017). Women's mental health and well-being 5 years after receiving or being denied an abortion: A prospective, longitudinal cohort study. *JAMA psychiatry*, 74(2):169–178.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. (2024). Difference-in-differences with a continuous treatment. Technical report, National Bureau of Economic Research.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Card, D. and Krueger, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591.

Robins, J. and Hernan, M. (2008). Estimation of the causal effects of time-varying exposures. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, pages 553–599.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–322.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2022). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *arXiv preprint arXiv:2201.01194*.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Wing, C., Simon, K., and Bello-Gomez, R. A. (2018). Designing difference in difference studies: best practices for public health policy research. *Annu Rev Public Health*, 39(1):453–469.

Ye, T., Keele, L., Hasegawa, R., and Small, D. S. (2023). A negative correlation strategy for bracketing in difference-in-differences. *Journal of the American Statistical Association*, pages 1–13.