SISCER Module 2: Causal Inference with Observational Data: Common Designs and Statistical Methods

Ting Ye & Richard Guo

Department of Biostatistics, UW

Day 3, Lecture 6: Time-varying treatments

July 10, 2024

Overview

1. Causal directed acyclic graphs (DAGs)
2. A single treatment
3. Time-varying treatments and confounding
4. Marginal structural model

Causal directed acyclic graphs (DAGs)

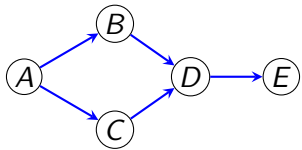
DAG

A graph $\mathcal{G}$ that consists of

- vertices $\boldsymbol{V}$,
- directed edges $\boldsymbol{E}$

such that there is no directed cycle.

DAG



- $\text{Pa}(D) = \{B, C\}$
- $\text{Ch}(A) = \{B, C\}$
- $A \to B \to D \to E$ is a directed path $\qquad A \in \text{An}(E)$ and $E \in \text{De}(A)$
- A and B are adjacent

- Topological ordering: $A \prec B \prec C \prec D \prec E$ (not unique) such that

$$i \text{ and } j \text{ are adjacent with } i \prec j \implies i \to j.$$

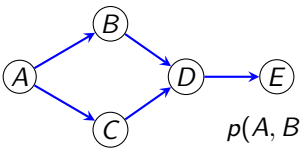
Probability model

☞ Associate every vertex with a random variable. ▶ State space can be $\{0,1\}$, $\mathbb{R}$ or anything

Then a DAG $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ is associated with

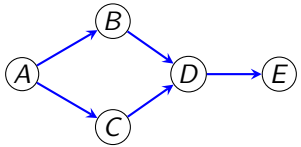
$$\mathcal{M}_{\mathcal{G}} := \{P : p(\boldsymbol{V}) \text{ factorizes according to } \mathcal{G}\}$$
$$= \left\{ P : p(\boldsymbol{V}) = \prod_{v \in \boldsymbol{V}} p(v \mid \mathrm{Pa}(v)) \right\}.$$

▶ Bayesian network. ▶ semiparametric model



$$p(A, B, C, D, E) = p(A)\, p(B \mid A)\, p(C \mid A)\, p(D \mid B, C)\, p(E \mid D)$$

Equivalent description: NPSEM-IE



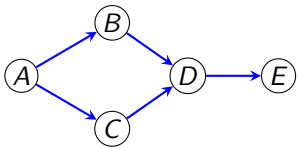
$$p(A, B, C, D, E) = p(A)\, p(B \mid A)\, p(C \mid A)\, p(D \mid B, C)\, p(E \mid D).$$

is equivalent to positing a nonparametric structural equation model with independent errors (NPSEM-IE):

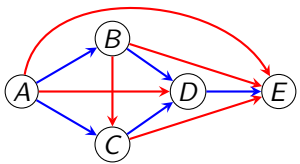
$$\begin{aligned}
\varepsilon_a, \varepsilon_b, \varepsilon_c, \varepsilon_d, \varepsilon_e &\overset{\text{iid}}{\sim} \text{unif}(0,1) \\
A &= f_a(\varepsilon_a) \\
B &= f_b(A, \varepsilon_b) \\
C &= f_c(A, \varepsilon_c) \\
D &= f_d(B, C, \varepsilon_d) \\
E &= f_e(D, \varepsilon_e)
\end{aligned}$$

Constraints: missing edges

Topological ordering: $A \prec B \prec C \prec D \prec E$



$$p(A)\,p(B \mid A)\,p(C \mid A)\,p(D \mid B, C)\,p(E \mid D)$$



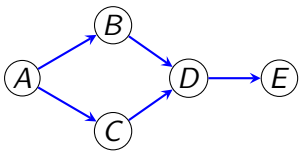
$$p(A)\,p(B \mid A)\,p(C \mid A, B)\,p(D \mid B, C, A)\,p(E \mid D, A, B, C)$$

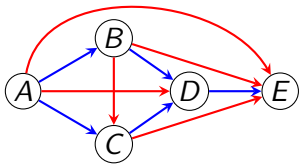
- The full DAG represents any P ▶ the **nonparametric** model.

Conditional independence

A DAG $\mathcal{G}$, as a probability model $\mathcal{M}_{\mathcal{G}}$, posits

$$\boxed{\text{missing edges} \implies \text{conditional independence (CI).}}$$



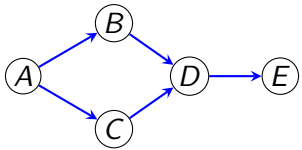


The missing '$B \to C$' posits

$$P(C \mid A, B) = P(C \mid A) \iff \boxed{B \perp\!\!\!\perp C \mid A} \iff P(B, C \mid A) = P(B \mid A)P(C \mid A).$$

Conditional independence

The graph



also implies, e.g.,

$$A, B, C \perp\!\!\!\perp E \mid D, \quad A, C \perp\!\!\!\perp E \mid B, D, \quad \ldots$$

☞ How we read off all the CIs a DAG implies ?

Dependence: mechanisms

Let A, B be the two fair coins.

HH, TT, HT, TH with equal prob. $\iff A \perp\!\!\!\perp B$

$A \perp\!\!\!\perp B$

Mechanisms of inducing dependence

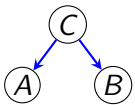
Let A, B be the two fair coins.

$$\text{only HH and TT} \implies A \not\perp\!\!\!\perp B$$

(1) Causal relations

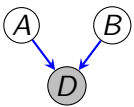
$A \not\perp\!\!\!\perp B$

(2) Common cause (unconditionally)



$A \not\perp\!\!\!\perp B$

(3) Conditioning on a common effect



$A \not\perp\!\!\!\perp B \mid D$

d-connecting path

- A **path** between A and B: a sequence of distinct, adjacent vertices

$$A \to \circ \to \circ \leftarrow \cdots \to B,$$

where every non-endpoint vertex is either a **collider** ($\to \circ \leftarrow$) or a **non-collider** ($\to \circ \to$, $\leftarrow \circ \leftarrow$, $\leftarrow \circ \to$)

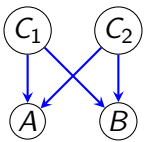
A path is **d-connecting given $\boldsymbol{C}$** if

1. every non-collider $\notin \boldsymbol{C}$, and
2. every collider is $\in \boldsymbol{C}$ or is an ancestor of $\boldsymbol{C}$.

d-separation

Vertex A and vertex B are d-separated by vertex set $\boldsymbol{C}$, written as $A \perp\!\!\!\perp_{\mathcal{G}} B \mid \boldsymbol{C}$, if there is no d-connecting path between A and B given $\boldsymbol{C}$.

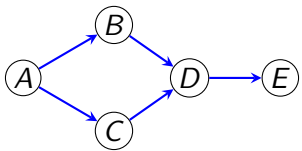
- Extended to $\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid C$ for disjoint vertex sets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$.



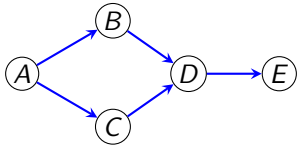
Global Markov property

Global Markov property For disjoint vertex sets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, it holds that

$$\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \; [P], \quad P \in \mathcal{M}_{\mathcal{G}}.$$



Quiz 1: Which CIs hold?



Recall: A **path** between A and B: a sequence of distinct, adjacent vertices

$$A \to \circ \to \circ \leftarrow \cdots \to B,$$

where every non-endpoint vertex is either a **collider** ($\to \circ \leftarrow$) or a **non-collider**. A path is **d-connecting given $\boldsymbol{C}$** if

1. every non-collider $\notin \boldsymbol{C}$, and
2. every collider is $\in \boldsymbol{C}$ or is an ancestor of $\boldsymbol{C}$.

DAG as a CI model

- The global Markov property also holds **reversely**. If P satisfies

$$\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \ [P],$$

then $P \in \mathcal{M}_{\mathcal{G}}$.

Theorem Factorization $\iff$ Global Markov $\iff$ Local Markov.

- Local Markov: $P \in \mathcal{M}_{\mathcal{G}} \implies A \perp\!\!\!\perp$ non-descendants of $A \mid \mathrm{Pa}(A)$

☞ That is, the model defined as $\mathcal{M}_{\mathcal{G}} := \{P : P \text{ factorizes according to } \mathcal{G}\}$ can be viewed as a **CI model**

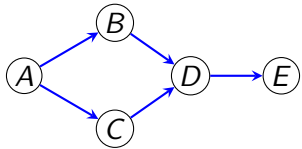
$$\{P : \boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} \mid \boldsymbol{C} \implies \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \ [P]\},$$

i.e.,

$$\{P : P \text{ satisfies CIs that are encoded as d-separations in } \mathcal{G}\}.$$

Sampling from a DAG

- We can simulate (sample) data by sequentially drawing from $p(v \mid \mathrm{Pa}(v))$.



Following the topological ordering $A \prec B \prec C \prec D \prec E$,

1. Draw $A \sim P(A)$
2. Draw $B \sim P(B \mid A)$, $C \sim P(C \mid A)$
3. Draw $D \sim P(D \mid B, C)$
4. Draw $E \sim P(E \mid D)$

DAG as a causal model

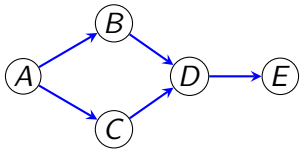
We have already seen that a DAG is a probability model as it defines a set of probability distributions $\mathcal{M}_{\mathcal{G}}$ satisfying the CIs.

- $P \in \mathcal{M}_{\mathcal{G}}$ is an observed distribution over factual random variables.

☞ What makes it a **causal model**?

- It must be augmented with extra semantics that
  1. posits the existence of counterfactuals (i.e., potential outcomes),
  2. makes assumptions about factual (e.g., Y) and counterfactual (e.g., $Y(a)$) variables, and
  3. connects the counterfactual distributions with the observed distribution.

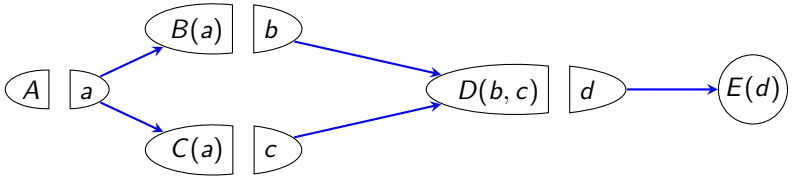
Recall: Sampling from a DAG



Following the topological ordering $A \prec B \prec C \prec D \prec E$,

1. Draw $A \sim P(A)$
2. Draw $B \sim P(B \mid A)$, $C \sim P(C \mid A)$
3. Draw $D \sim P(D \mid B, C)$
4. Draw $E \sim P(E \mid D)$

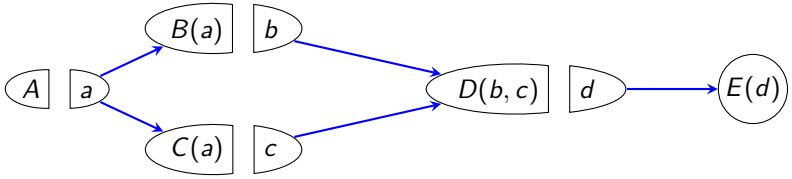
Alternative sampling (I): one-step-ahead counterfactuals



- Single-World Intervention Graph (SWIG) (Richardson & Robins, 2013)

1. Draw $A \sim P(A)$
2. For every potential a, draw $B(a) \sim P(B \mid A = a)$, $C(a) \sim P(A \mid A = a)$ independent of A
3. For every potential (b, c), draw $D(b, c) \sim P(D \mid B = b, C = c)$ independent of previously drawn.
4. For every potential d, draw $E(d) \sim P(E \mid D = d)$ independent of previously drawn.

Alternative sampling (I): one-step-ahead counterfactuals



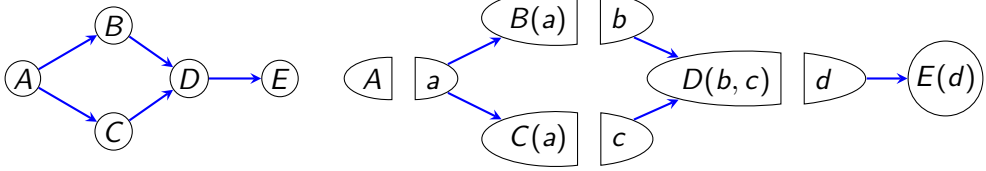
▶ Single-World Intervention Graph (SWIG) (Richardson & Robins, 2013)

- A: factual variable; naturally occurring value of A
- a: imagine that upon observing A, immediately we intervene on A and set its value to a
- $B(a)$, $C(a)$: the potential outcomes (counterfactual) under such an intervention

☞ From this SWIG, we can see that

$$\boxed{A \perp\!\!\!\perp B(a), C(a) \quad \text{for every } a.}$$

Alternative sampling (II): recursive substitution



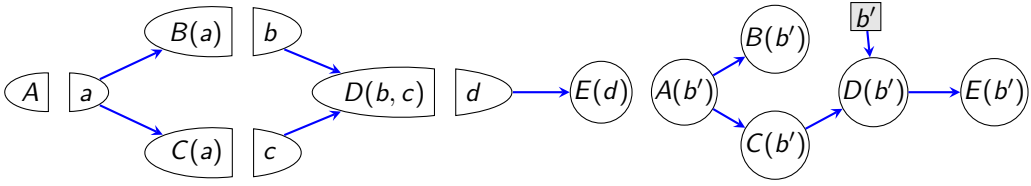
To generate the observed, factual variables,

1. $A = A$,
2. $B = B(A)$, $C = C(A)$,
3. $D = D(B, C)$,
4. $E = E(D)$.

- Apparently, $(A, B, C, D, E) \sim P$

Alternative sampling (III): intervention

☞ Suppose that we **intervene on B and set it to b'** — imposes input to B's children.



1. $A(b') = A$,
2. $B(b') = B(A(b')) = B(A)$, $C(b') = C(A(b')) = C(A)$
   - $B(b')$ is the naturally occurring value of B immediately before it is intervened on
3. $D(b') = D(b', C(b'))$,
4. $E(b') = E(D(b'))$.

- This defines the distribution of $P((A, B, C, D, E)(b'))$, or $P(A, B, C, D, E \mid \text{do}(B = b'))$.

Alternative sampling: the causal model

☞ This set of semantics defines the **FFRCISTG / SWIG causal model** associated with a DAG $\mathcal{G}$.

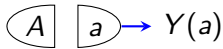
- 'Finest Fully Randomized Causally Interpreted Structured Tree Graph' (Robins, 1986)

☞ It makes **weaker** assumptions than Pearl's NPSEM-IE (nonparametric structural equation model with independent errors) causal model.

Single treatment

Single treatment, randomized

- $A = 1$: treated; $A = 0$: control.



☞ From the SWIG, we can read off

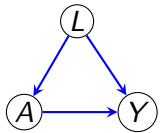
$$A \perp\!\!\!\perp Y(a),$$

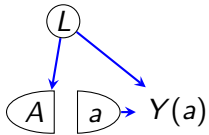
so

$$\boxed{\mathbb{E}\, Y(a) = \mathbb{E}[Y \mid A = a], \quad a = 0, 1.}$$

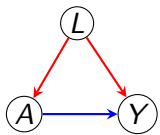
- association = causation

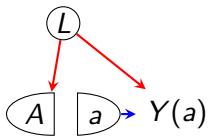
Single treatment, conditionally randomized





Single treatment, conditionally randomized



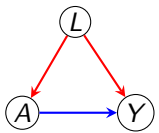


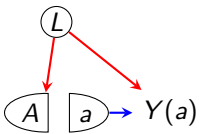
☞ L is an observed confounder between A and Y, so $\mathbb{E}\,Y(a) \neq \mathbb{E}[Y \mid A = a]$.

▶ association $\neq$ causation

▶ There is no unobserved confounding. Recall that we can use L to identify $\mathbb{E}\,Y(a)$ through **standardization** or **IPW** (inverse probability weighting).

Single treatment, conditionally randomized: Standardization





- **Standardization**: From the SWIG, we see that

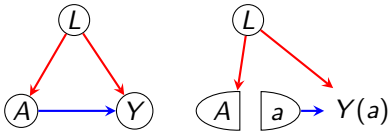
$$A \perp\!\!\!\perp Y(a) \mid L,$$

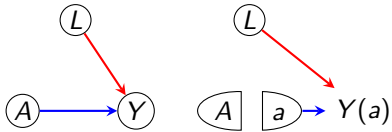
i.e., A is randomized within every stratum of L.

1. Within stratum $L = l$, we have $\mathbb{E}[Y(a) \mid L = l] = \mathbb{E}[Y \mid A = a, L = l]$.
2. Averaging over l to get the whole population:

$$\mathbb{E}\, Y(a) = \sum_l \mathbb{E}[Y(a) \mid L = l] P(L = l) = \boxed{\sum_l \mathbb{E}[Y \mid A = a, L = l] P(L = l).}$$

Single treatment, conditionally randomized: IPW





$p(L = l)p(A = a \mid L = l)p(Y \mid A = a, L = l)$
$p(L = l)p(A = a \mid L = l)p(Y \mid A = a, L = l)$

$p(L = l)\cancel{p(A = a \mid L = l)}p(Y \mid A = a, L = l)$
$\times q(A = a)/\cancel{p(A = a \mid L = l)}$

- **IPW**: Weighting by

$$1/p(A \mid L) \qquad \blacktriangleright\ q(A = 0) = q(A = 1) = 1/2$$

or

$$p(A)/p(A \mid L) \qquad \blacktriangleright\ q(A) = p(A), \text{ stabilized weight}$$

gives a trial where A is randomized (does not depend on L). ▶ association = causation

Single treatment, conditionally randomized

IPW:

$$\mathbb{E}[Y(a)] = \mathbb{E}\left\{\frac{\mathbb{I}_{A=a}\,Y}{P(A=a\mid L)}\right\} / \mathbb{E}\left\{\frac{\mathbb{I}_{A=a}}{P(A=a\mid L)}\right\} = \mathbb{E}\left\{\frac{\mathbb{I}_{A=a}\,Y}{P(A=a\mid L)}\right\}.$$

Stabilized IPW has the same target

$$\mathbb{E}\left\{\frac{\mathbb{I}_{A=a}\,\cancel{P(A=a)}\,Y}{P(A=a\mid L)}\right\} / \mathbb{E}\left\{\frac{\mathbb{I}_{A=a}\,\cancel{P(A=a)}}{P(A=a\mid L)}\right\} = \mathbb{E}\left\{\frac{\mathbb{I}_{A=a}\,Y}{P(A=a\mid L)}\right\}.$$

☞ But it makes a difference when fitting marginal structural models with estimated weights...

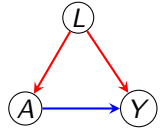
Single treatment, conditionally randomized: unified view

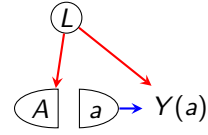
- **Standardization** and **IPW** target the same population quantity:

$$\underbrace{\mathbb{E}\left\{\frac{\mathbb{I}_{A=a}\,Y}{P(A=a\mid L)}\right\}}_{\text{IPW}}=\mathbb{E}\left\{\frac{\mathbb{E}[\mathbb{I}_{A=a}\,Y\mid L]}{P(A=a\mid L)}\right\}=\mathbb{E}\left\{\frac{P(A=a\mid L)\,\mathbb{E}[Y\mid A=a,L]}{P(A=a\mid L)}\right\}=\underbrace{\mathbb{E}\{\mathbb{E}[Y\mid A=a,L]\}}_{\text{standardization}}$$

☞ Again, it makes a difference when replaced by estimates...

☞ Both standardization and IPW are **adjusting for** L: We use L to block all the **non-causal paths** (paths not in the shape $A\to\cdots\to Y$).

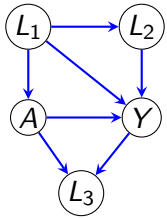




☞ Non-causal ('backdoor') path $A\leftarrow L\rightarrow Y$ is blocked by L (i.e., not d-connected given L).

Quiz 2

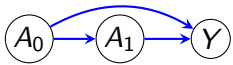
In the setting below, which strategies can correctly identify $\mathbb{E}\, Y(a)$?

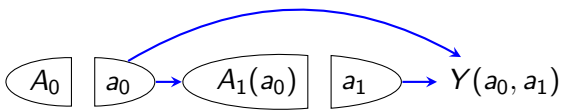


Time-varying treatments

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





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

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

- What is the meaning of $\mathbb{E}\, Y(0, 0)$?
- association = causation

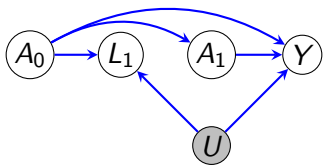
Two treatments: Example

Study of the effect of antiretroviral therapy on a health score (Robins & 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.
2. Month 6: Measure blood CD4 counts L_1 and assign therapy ($A_1 = 1$: treated; $A_1 = 0$: control).
3. Month 12: Measure the final health score Y.

Quiz 3

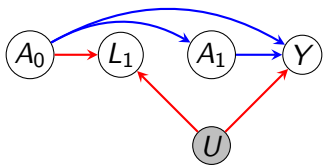
1. Month 0: Assign therapy ($A_0 = 1$: treated; $A_0 = 0$: control) at the start of the follow-up.
   - Suppose A_0 is randomly assigned.
2. Month 6: Measure blood CD4 counts L_1 and assign therapy ($A_1 = 1$: treated; $A_1 = 0$: control).
   - Suppose A_1's assignment depends only on A_0 but not L_1.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



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

Quiz 3

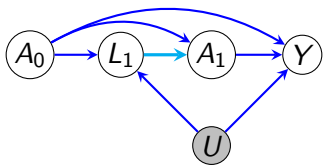
1. Month 0: Assign therapy ($A_0 = 1$: treated; $A_0 = 0$: control) at the start of the follow-up.
   - Suppose A_0 is randomly assigned.
2. Month 6: Measure blood CD4 counts L_1 and assign therapy ($A_1 = 1$: treated; $A_1 = 0$: control).
   - Suppose A_1's assignment depends only on A_0 but not L_1.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



☞ Non-causal path is not d-connected (unless conditioning on L_1).

Two treatments, with time-varying confounder

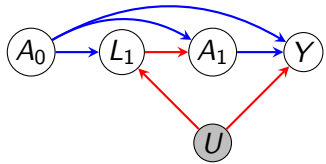
1. Month 0: Assign therapy ($A_0 = 1$: treated; $A_0 = 0$: control) at the start of the follow-up.
   - Suppose A_0 is randomly assigned.
2. Month 6: Measure blood CD4 counts L_1 and assign therapy ($A_1 = 1$: treated; $A_1 = 0$: control).
   - Suppose A_1's assignment depends on **both** A_0 **and**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



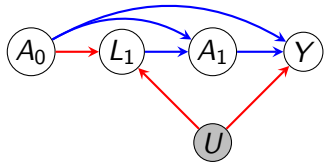
☞ Can we identify $\mathbb{E}\, Y(a_0, a_1)$?

Dilemma

1. Not adjusting for L_1.

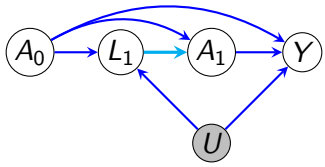


2. Adjusting for L_1.

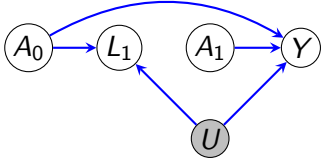


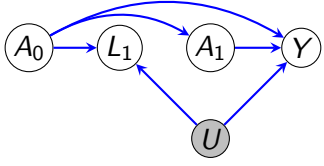
- Need something more sophisticated.

IPW: Removing A_1's dependency on L_1



- **IPW**





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

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

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

IPW: Identification

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

☞ Does it make a difference to use the stabilized weight $P(A_1 = a_1 \mid A_0 = a_0)/P(A_1 = a_1 \mid A_0 = a_0, L_1)$?

Quiz 4

Suppose A_0, L_1, A_1 are all binary. What is the estimate of $\mathbb{E}\, Y(0,0)$ based on IPW?

n	A_0	L_1	A_1	$\mathbb{E}[Y \mid A_0, L_1, A_1]$
50	0	0	0	4
100	0	0	1	-1
100	0	1	0	2
50	0	1	1	-3
50	1	0	0	4
50	1	0	1	-2
100	1	1	0	7
200	1	1	1	11

☞ The first row means there are 50 subjects with $A_0 = 0, L_1 = 0, A_1 = 0$ and their average outcome is 4.

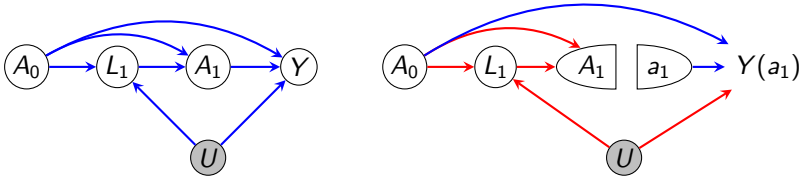
Standardization / g-formula

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

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

Standardization / g-formula: Intuition

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.,

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

Quiz 5

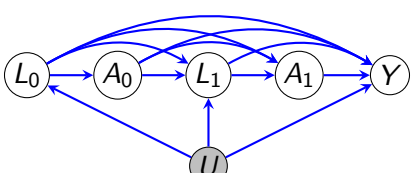
Suppose A_0, L_1, A_1 are all binary. What is the estimate of $\mathbb{E}\, Y(0,0)$ based on standardization / g-formula?

n	A_0	L_1	A_1	$\mathbb{E}[Y \mid A_0, L_1, A_1]$
50	0	0	0	4
100	0	0	1	-1
100	0	1	0	2
50	0	1	1	-3
50	1	0	0	4
50	1	0	1	-2
100	1	1	0	7
200	1	1	1	11

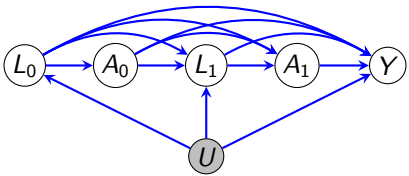
☞ The first row means there are 50 subjects with $A_0 = 0, L_1 = 0, A_1 = 0$ and their average outcome is 4.

Quiz 6: 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



Under positivity and sequential randomization

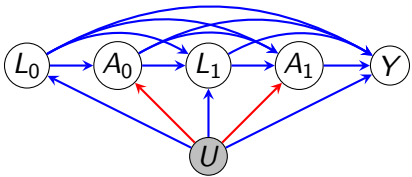
$$Y(a_0, a_1) \perp\!\!\!\perp A_0 \mid L_0,$$
$$Y(a_0, a_1) \perp\!\!\!\perp A_1 \mid L_0, A_0, L_1,$$

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

- Extends to more time points.

Out of luck

- If either red edge is present, then $\mathbb{E}\, Y(a_0, a_1)$ cannot be identified.



Marginal structural model

Marginal structural (mean) model

Consider two treatments A_0, A_1.

- Marginal structural mean model is to postulate and fit

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

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.

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.

Further reading

See Chapters 19, 20 and 21 of Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

References I

- Richardson, T., & Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, *128*(30), 2013.
- Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, *7*(9-12), 1393–1512.
- Robins, J., & Hernan, M. (2008). Estimation of the causal effects of time-varying exposures. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*, 553–599.