Causal Inference II: Causal Graphs (DAGs) and Instrumental Variables Methods

Li Ge

Ph.D. student in Biomedical Data Science

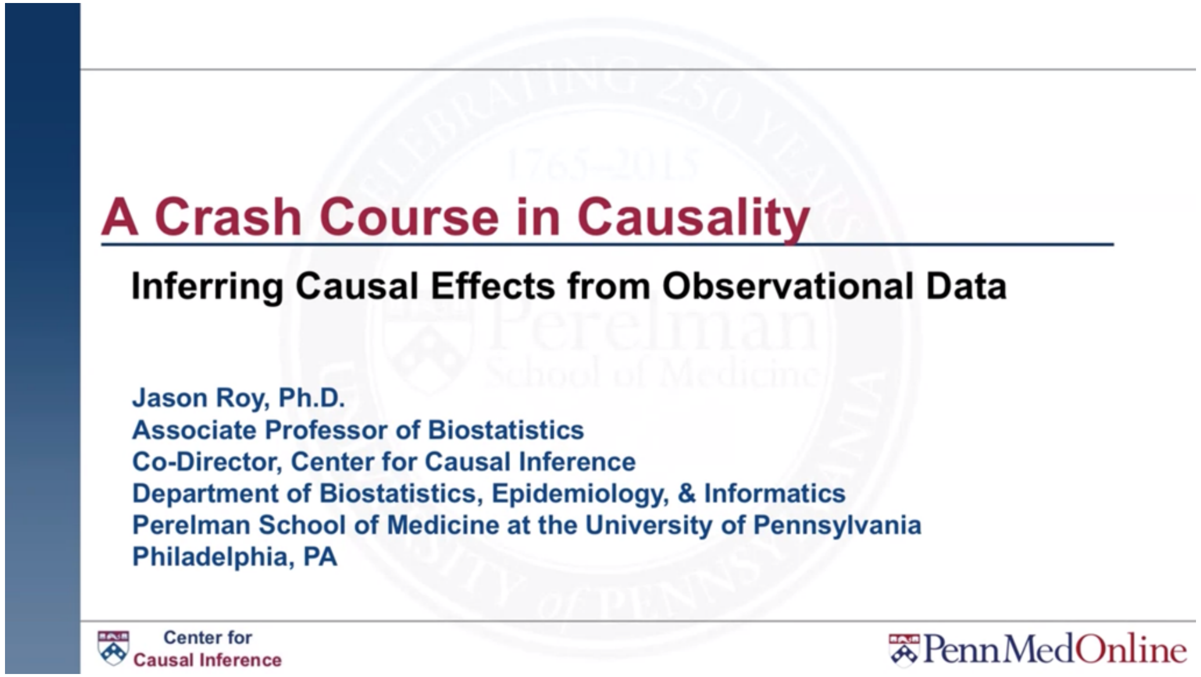
Jul 10, 2020

Department of Biostatistics and Medical Informatics

University of Wisconsin-Madison

Acknowledgement

- I made these slides with my own thoughts but adapted a lot from a course taught by Jason Roy on Coursera.
- I HIGLY recommend this course.



A Crash Course in Causality

Inferring Causal Effects from Observational Data

Jason Roy, Ph.D.
Associate Professor of Biostatistics
Co-Director, Center for Causal Inference
Department of Biostatistics, Epidemiology, & Informatics
Perelman School of Medicine at the University of Pennsylvania
Philadelphia, PA

Center for Causal Inference

PennMedOnline

Course website: https://www.coursera.org/learn/crash-course-in-causality/

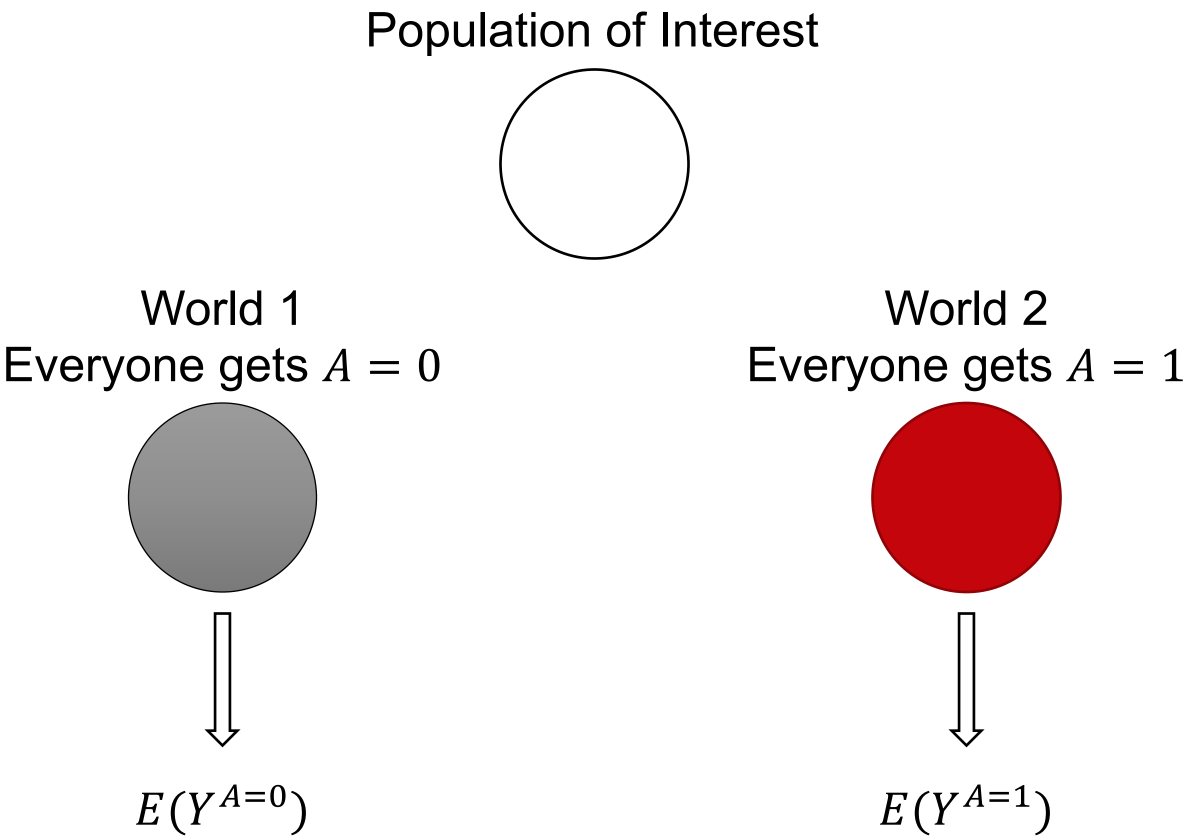
Contents

- Review of last talk:
  - Part I: Introduction to Causal Effects
  - Part II (briefly): Confounding and Directed Acyclic Graphs (DAGs)
  - Part III: Matching and Propensity Scores
  - Part IV: Inverse Probability of Treatment Weighting (IPTW)

- Today's focus:
  - Part II: Confounding and Directed Acyclic Graphs (DAGs)
  - Part V: Instrumental Variables Methods

Potential Outcomes and Counterfactuals

- A thought experiment: parallel universe, time machine, magic



Population of Interest

World 1
Everyone gets $A = 0$

World 2
Everyone gets $A = 1$

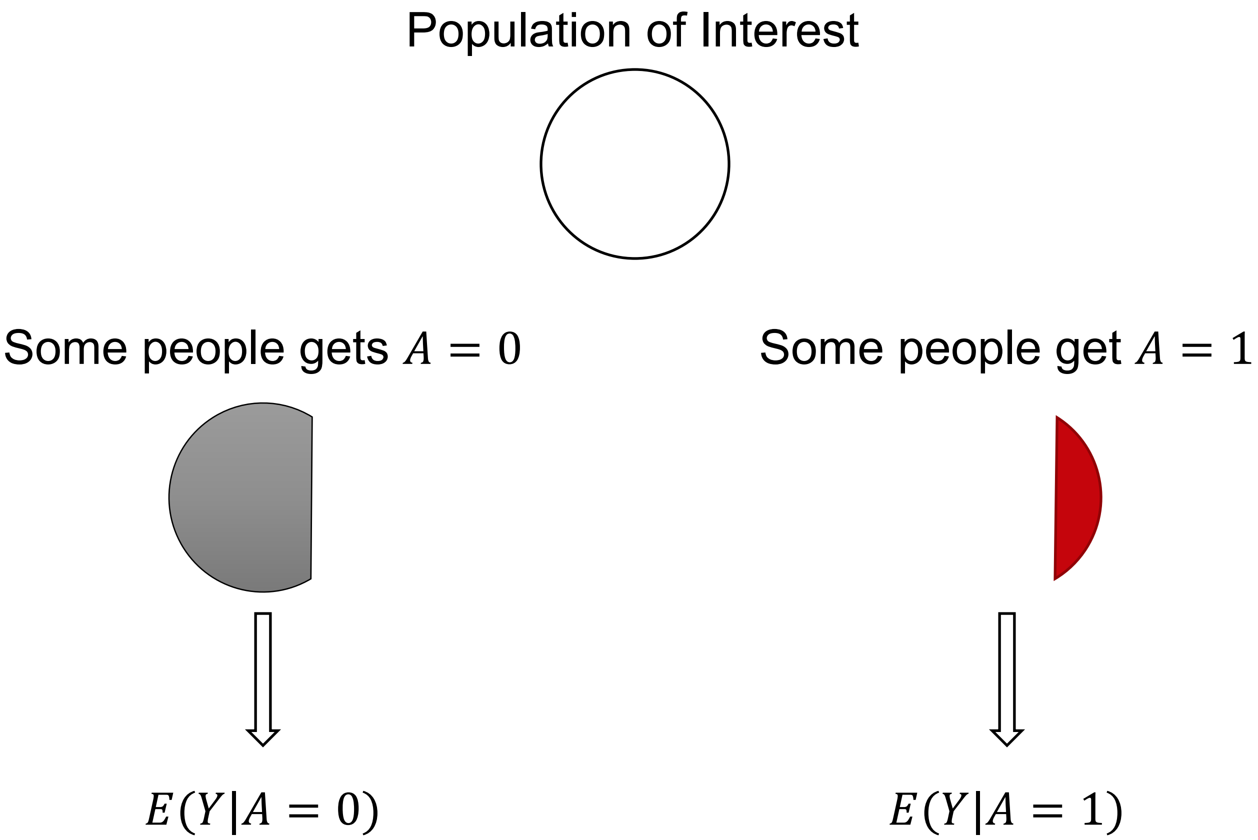
$E(Y^{A=0})$

$E(Y^{A=1})$

- Causal Effect (Estimand): $E(Y^{A=1} - Y^{A=0})$

Real World

- Only can observe treatment effects on subpopulations



- $E(Y|A = 1) - E(Y|A = 0)$ is generally not a causal effect

Causal Assumptions

- Identifiability of causal effects $E(Y^{A=1} - Y^{Y=0})$ requires some untestable assumptions. These are generally called causal assumptions.

- The most common are:
  - Stable Unit Treatment Value Assumption (SUTVA): no interference
  - Consistency: $Y = Y^a, \text{if } A = a, \text{for all } a$
  - Ignorability: $Y^0, Y^1 \perp A|X$
  - Positivity: $P(A = a|X = x) > 0, \text{for all } a \text{ and } x$

- Assumptions will be about the observed data: outcome - Y, treatment - A, and a set of pre-treatment covariates - X.

Causal Estimands

We can put causal assumptions together to identify causal effects.

$E(Y|A=a,X=x)$ involves only the observed data.

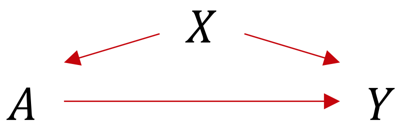
$$\begin{aligned} E(Y|A=a,X=x) &= E(Y^a|A=a,X=x) \text{ by consistency} \\ &= E(Y^a|X=x) \text{ by ignorability} \end{aligned}$$

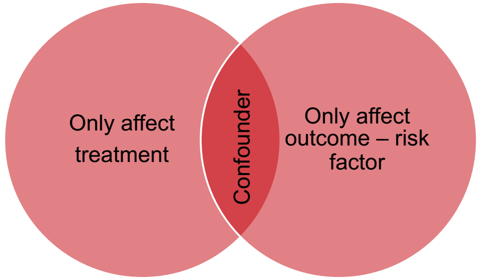
If we want a marginal causal effect, we can average over X.

$$E(Y^a) = E\big(E(Y^a|X)\big) = \sum_X E(Y|A=a,X=x)P(X=x)$$

Confounding

- Confounders are variables that affects both the treatment and the outcome.

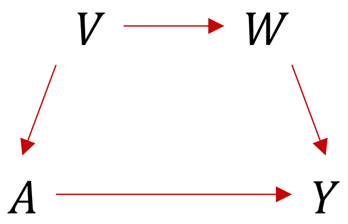




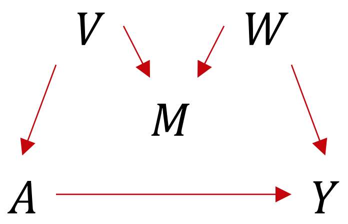
- Controlling confounders means to identify a set of variables X that will make the ignorability assumption $Y^0, Y^1 \perp A|X$ holds.
- What matters is not identifying specific confounders but identifying a set of variables that are sufficient to control for confounding.
  - Backdoor path criterion (Pearl 1995)
  - Disjunctive cause criterion (VanderWeele 2011)

DAG

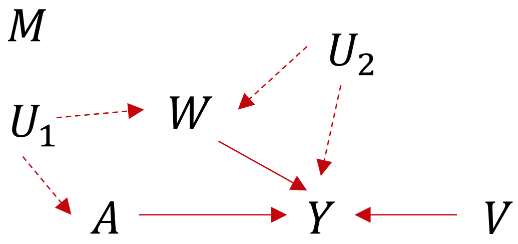
Consider more complex examples:



DAG 1



DAG 2



DAG 3

Sets of variables that are sufficient to control for confounding:

- $\{V\}$
- $\{W\}$
- $\{V, W\}$

Sets of variables that are sufficient to control for confounding:

- $\emptyset$, $\{V\}$, $\{W\}$, $\{M, V\}$, $\{M, W\}$, $\{M, V, W\}$
- But not $\{M\}$

Sets of variables that are sufficient to control for confounding:

- $\{U_1\}$, however, it is unobservable
- Unachievable with observed variables $\{M, W, V\}$

DAG

We will formally introduce the DAG shortly.

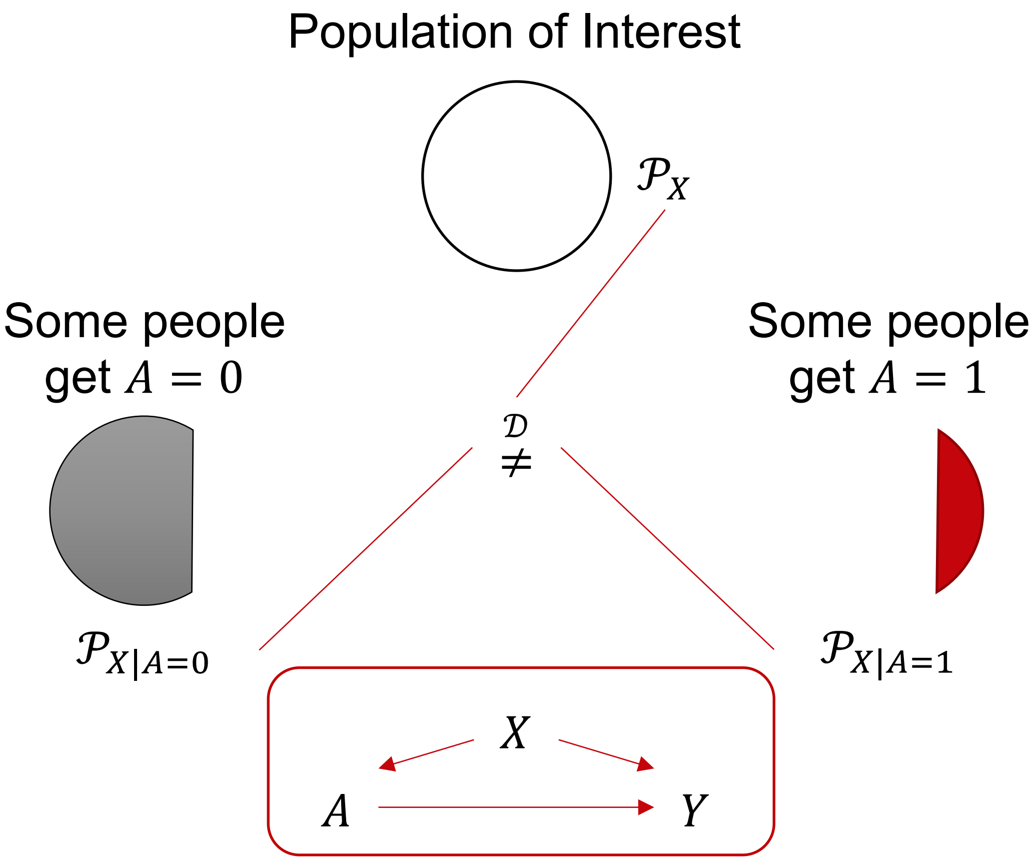
DAGs help us effectively determine the set of variables to control for to achieve ignorability.

- We'll see that DAGs encode probability distributions.
- We'll be able to recognize different types of paths and understand which of them induce association between nodes.
- We'll see how to block paths to impose conditional independence (d-separation).
- We'll use the backdoor path criterion and the disjunctive cause criterion to determine if a set of variables is sufficient to control for confounding.

Once we know which variables to control for them, the question is how to control for them.

General approaches include matching and inverse probability of treatment weighting.

Observational Studies



Population of Interest

$\mathcal{P}_X$

Some people get $A = 0$

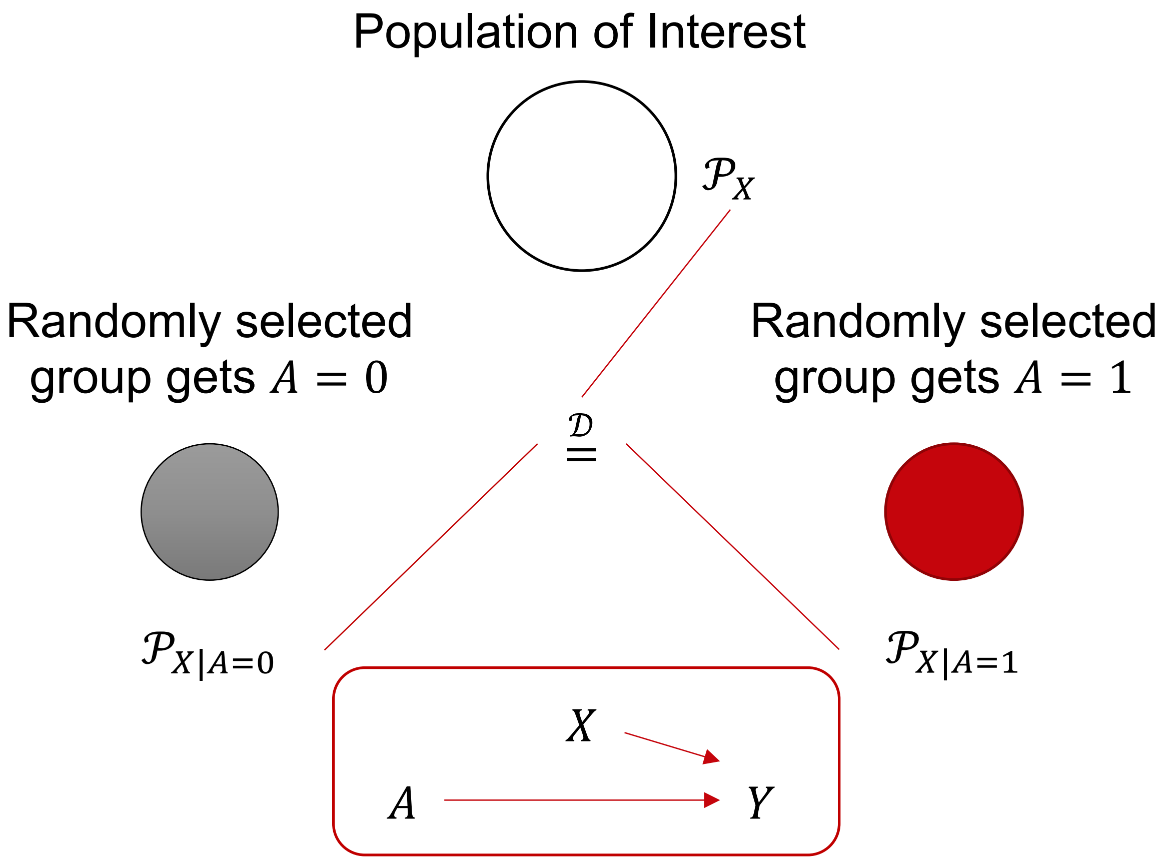
Some people get $A = 1$

$\overset{\mathcal{D}}{\neq}$

$\mathcal{P}_{X|A=0}$

$\mathcal{P}_{X|A=1}$

Observational studies



Population of Interest

$\mathcal{P}_X$

Randomly selected group gets $A = 0$

Randomly selected group gets $A = 1$

$\overset{\mathcal{D}}{=}$

$\mathcal{P}_{X|A=0}$

$\mathcal{P}_{X|A=1}$

Randomized controlled trials

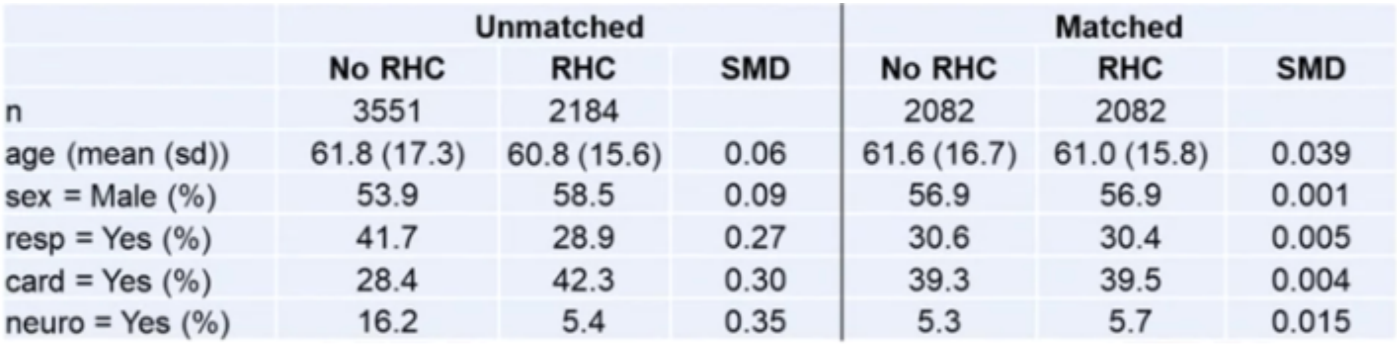
Matching Procedures

1. Select a set of pre-treatment covariates X that (hopefully) satisfy the ignorability assumption.
2. Calculate the distance matrix $D = (d_{ij}) \in \mathbb{R}_{0+}^{m \times n}$ that contains the pairwise distance $d_{ij} = \mathcal{D}(X_i, X_j)$ between each treated subject and control subject.
   - *e.g.*, Mahalanobis distance $\mathcal{D}(X_i, X_j) \coloneqq \sqrt{(X_i - X_j)^T \Sigma^{-1} (X_i - X_j)}$.
   - Replace each covariate value with its rank to get robust distance.
3. Minimize the total distance measure (optimal matching).
   - Can use greedy matching to speed up.
   - Can impose constraints such as caliper (maximum acceptable distance), sparsity (*e.g.*, match within hospitals).

Matching Procedures

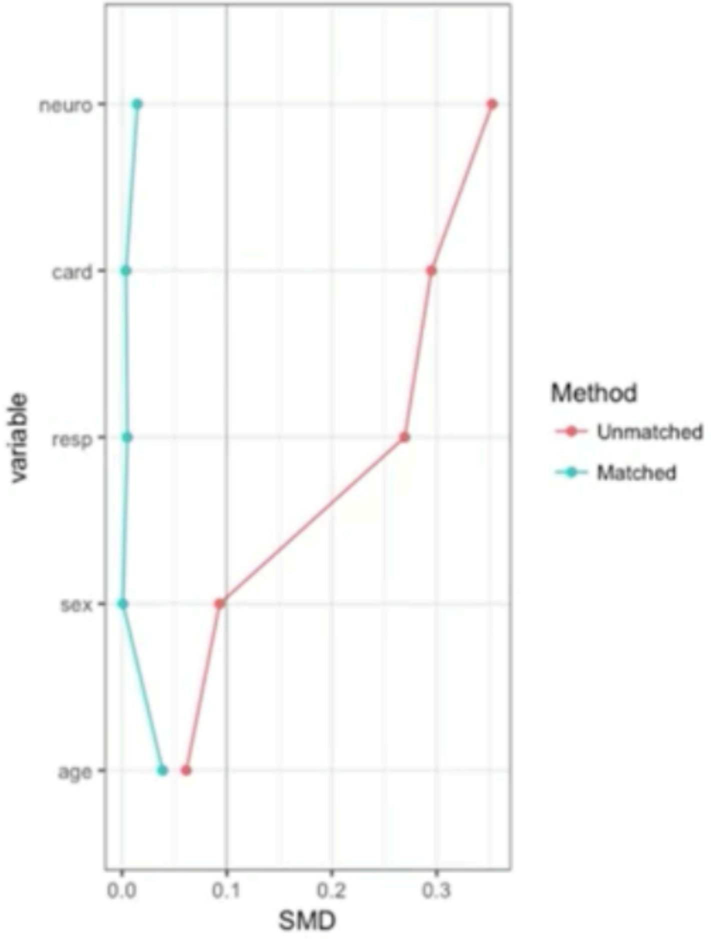
4. Assess covariates balance.

Table 1: Patient baseline characteristics table



	Unmatched			Matched		
	No RHC	RHC	SMD	No RHC	RHC	SMD
n	3551	2184		2082	2082	
age (mean (sd))	61.8 (17.3)	60.8 (15.6)	0.06	61.6 (16.7)	61.0 (15.8)	0.039
sex = Male (%)	53.9	58.5	0.09	56.9	56.9	0.001
resp = Yes (%)	41.7	28.9	0.27	30.6	30.4	0.005
card = Yes (%)	28.4	42.3	0.30	39.3	39.5	0.004
neuro = Yes (%)	16.2	5.4	0.35	5.3	5.7	0.015

5. Analyze post-matching data.
   - Test for treatment effects.
   - Estimate treatment effects and confidence intervals.
   - Methods should take matching into account.
6. Perform sensitivity analysis.
   - Check for hidden bias due to unmeasured confounders.



Standardized Mean Difference (SMD) plot

Propensity Score

The propensity score is the probability of receiving treatment, rather than control, given covariates X.

$$\pi_i = \pi(X_i) = P(A = 1|X_i)$$

Lemma. Assuming ignorability, *i.e.*, $Y^0, Y^1 \perp A|X$, then

$$Y^0, Y^1 \perp A|\pi(X).$$

- Propensity score is a dimension reduction technique.

Definition. $b(X)$ is a balancing score if $A \perp X|b(X)$, *i.e.*,

$$P(X = x|b(X) = p, A = 1) = P(X = x|b(X) = p, A = 0)$$

Remark: $b(X)$ is a balancing score if and only if it is finer than the propensity score, *i.e.*,

$$\pi(X) = h\big(b(X)\big) \text{ for some function } h.$$

If we match on the any balancing score, we should achieve balance,

$$Y^0, Y^1 \perp A|b(X).$$

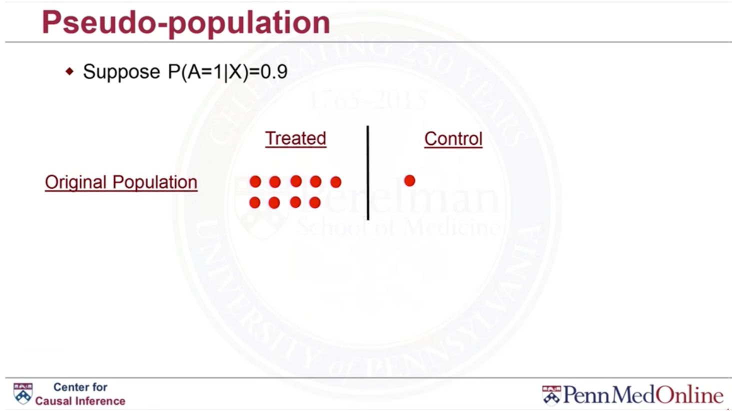
Intuition for IPTW

We can create a pseudo-population by weighting by the inverse of the probability of treatment received.

- For treated subjects, weight by the inverse of $P(A = 1|X) = \pi(X)$.
- For control subjects, weight by the inverse of $P(A = 0|X) = 1 - \pi(X)$.

Hence, it is called the inverse probability of treatment weighting (IPTW).

- In the pseudo-population, treatment assignment doesn't depend on X.



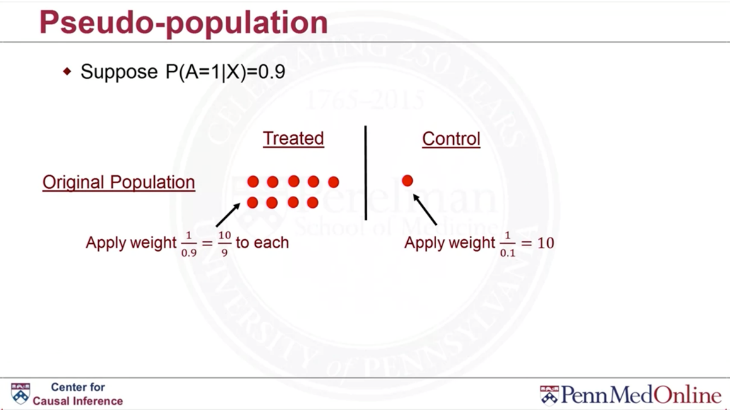
Pseudo-population

- Suppose P(A=1|X)=0.9

	Treated	Control
Original Population	●●●●● ●●●●	●

Center for Causal Inference

PennMedOnline



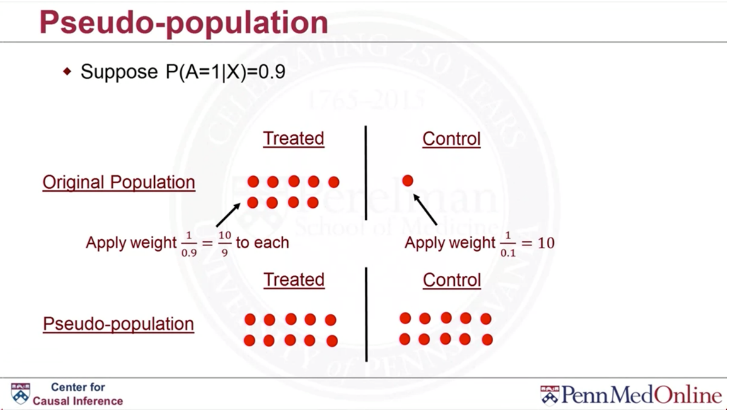
Pseudo-population

- Suppose P(A=1|X)=0.9

	Treated	Control
Original Population	●●●●● ●●●●	●
	Apply weight $\frac{1}{0.9} = \frac{10}{9}$ to each	Apply weight $\frac{1}{0.1} = 10$

Center for Causal Inference

PennMedOnline



Pseudo-population

- Suppose P(A=1|X)=0.9

	Treated	Control
Original Population	●●●●● ●●●●	●
	Apply weight $\frac{1}{0.9} = \frac{10}{9}$ to each	Apply weight $\frac{1}{0.1} = 10$
	Treated	Control
Pseudo-population	●●●●● ●●●●●	●●●●● ●●●●●

Center for Causal Inference

PennMedOnline

Marginal Structural Models

General MSM:

$$g\big(E(Y^a|V)\big) = h(a, V; \psi)$$

- where $g()$ is a link function.
- $h()$ is a function specifying parametric form of a and V (typically additive, linear).

IPTW Estimation

- Recall that the pseudo-population (obtained from IPTW) is free from confounding (assuming ignorability and positivity).
- We can therefore estimate MSM parameters by solving the weighted estimating equation

$$\sum_{i=1}^{n} \frac{\partial \mu_i^T}{\partial \psi} V_i^{-1} W_i \big(Y_i - \mu_i(\psi)\big) = 0$$

- where $W_i = \frac{1}{A_i \hat{\pi}_i + (1 - A_i)(1 - \hat{\pi}_i)}$.

IPTW in Practice

Steps:

1. Estimate propensity score (*e.g.*, LR: $A \sim X$).
2. Create weights ($w_i = \frac{1}{A_i\hat{\pi}_i + (1 - A_i)(1 - \hat{\pi}_i)}$).
3. Specify the MSM of interest.
4. Use software to fit a weighted generalized linear model.

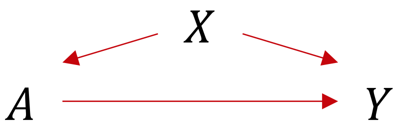
   (*e.g.*, `glm.obj <- glm(y ~ trt, weights = w, family = binomial(link = log))`)
5. Use asymptotic (sandwich) estimator (or bootstrapping) to get standard error.

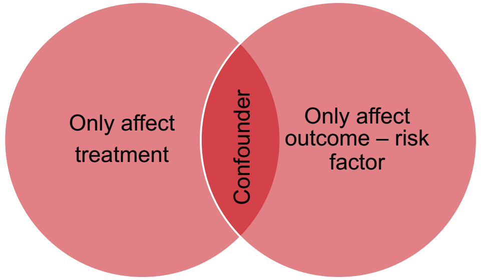
   (*e.g.*, `SE <- sqrt(diag(vcovHC(glm.obj, type = “HC0”)))`)

Part II: Confounding and Directed Acyclic Graphs (DAGs)

Confounding

- Confounders are variables that affects both the treatment and the outcome.





- Controlling confounders means to identify a set of variables X that will make the ignorability assumption $Y^0, Y^1 \perp A|X$ holds.
- What matters is not identifying specific confounders but identifying a set of variables that are sufficient to control for confounding.
  - Backdoor path criterion (Pearl 1995)
  - Disjunctive cause criterion (VanderWeele 2011)

The Basics of Graphical Models

- Graphical model represents a family of distributions

$$P(X_{1:n}) \coloneqq \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i))$$

  - Factorize and simply the joint distribution.

- Inference by enumeration

$$P(\mathcal{A}|\mathcal{B}) = \frac{P(\mathcal{A},\mathcal{B})}{P(\mathcal{B})} = \frac{\sum_{X_{1:n}\setminus\mathcal{A},\mathcal{B}} P(X_{1:n})}{\sum_{X_{1:n}\setminus\mathcal{B}} P(X_{1:n})}$$

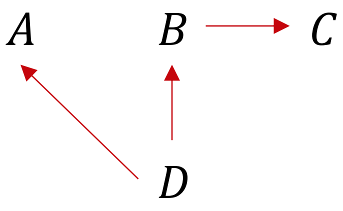
  - Marginalization is answered by summing over joint.

- **Causal Markov condition**: Let ℓ be a *topological ordering* of the nodes, which ensures that $\mathcal{p}_i$ occurs in the ordering before i. Let $\mathcal{v}_i$ be the set of indices that before i, not including $\mathcal{p}_i$, then

$$X_i \perp X_{\mathcal{v}_i} | X_{\mathcal{p}_i}$$

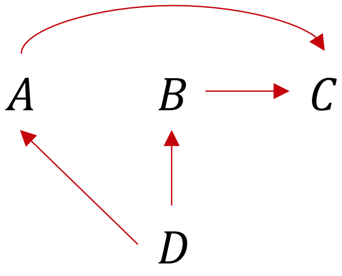
  - Every node is conditionally independent of its nondescendents, given its parents.
  - We can read off conditional independencies by looking at the graph.

DAG examples



This DAG implies:

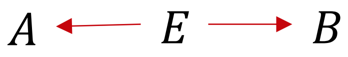
- $P(A,B,C,D) = P(D)P(A|D)P(B|D)P(C|B)$
- $P(A|B,C,D) = P(A|D)$
- $P(D|A,B,C) = P(D|A,B)$
- $P(D|B,C) = P(D|B)$
- $A \perp B,C|D$
- $D \perp C|A,B$
- $D \perp C|B$



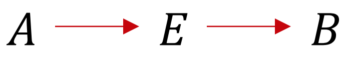
This DAG implies:

- $P(A,B,C,D) = P(D)P(A|D)P(B|D)P(C|A,B)$
- $P(A|B,C,D) = P(A|C,D)$
- $P(D|A,B,C) = P(D|A,B)$
- $P(D|B,C) =$ ...
- $A \perp B|C,D$
- $D \perp C|A,B$
- ~~$D \perp C|B$~~

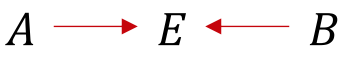
Paths & Associations



Fork (Tree)

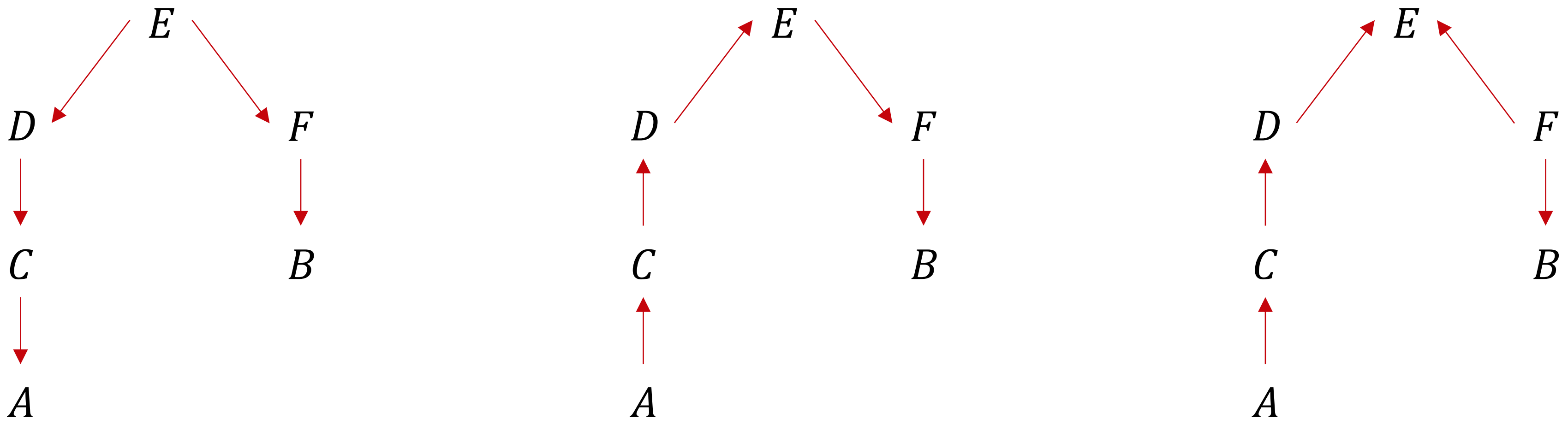


Chain (Sequence)



Inverted fork
(Inverse tree)

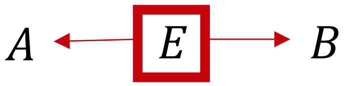
- If nodes A and B are on the ends of a path, they are associated (via this path) if:
  - Some information flows to both
  - Information from one makes it to the other



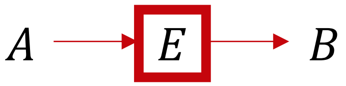
- Paths on the right panel won't introduce association. Information from A and B collide at E. We call E as a collider.

D-Separation

- Paths can be blocked by conditioning on nodes in the path.

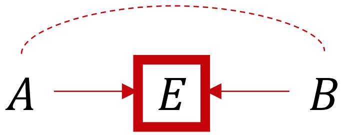


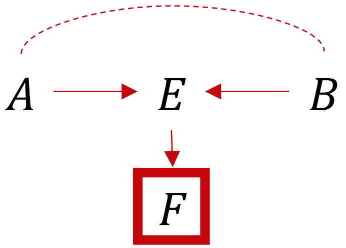
Fork (Tree)



Chain (Sequence)

- The opposite situation occurs if a collider is conditioned on.





Inverted fork
(Inverse tree)

- A path is d-separated by a set of node C if:
  - It contains a chain and the middle part is in C.

    OR
  - It contains a fork and the middle part is in C.

    OR
  - It contains an inverted fork and the middle part is not in C, nor are any descendants of it.

D-Separation

Two nodes, A and B, are d-separated by a set of nodes C if it blocks every path from A to B.

- Then:

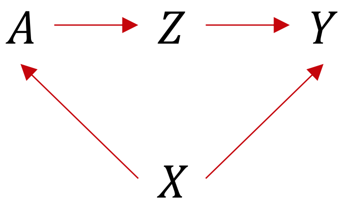
$$A \perp B|C$$

Recall the ignorability assumption:

$$Y^0, Y^1 \perp A|X$$

Backdoor Path Criterion

- A frontdoor path from A to Y is one that begins with an arrow emanating out of A, *e.g.*, $A \to Z \to Y$.
- A backdoor path from A to Y is one that travel through arrows going into A, *e.g.*, $A \leftarrow X \to Y$.

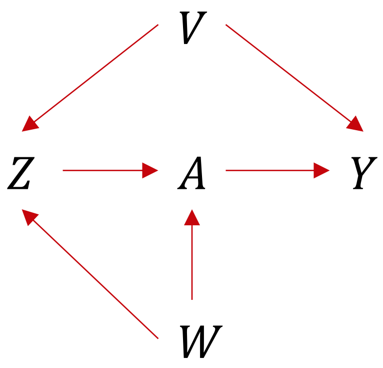


Backdoor path criterion: A set of variables X is sufficient to control for confounding if

- It blocks all backdoor paths from treatment to the outcome
- It does not include any descendants of treatment

Note: not necessarily unique, NP-hard to find all paths

Backdoor Path Criterion



There are 2 back door paths from A to Y:

- $A \leftarrow Z \leftarrow V \rightarrow Y$ is a fork. $\{Z\}, \{V\}, \{Z, V\}$ can block this path.
- $A \leftarrow W \rightarrow Z \leftarrow V \rightarrow Y$ is an inverted fork. $\{V\}, \{W\}, \{Z, V\}, \{Z, W\}, \{V, Z, W\}$ but not $\{Z\}$ alone can block this path.

Together, we say the following set are sufficient to control for confounding:

- $\{V\}, \{V, Z\}, \{Z, W\}, \{V, Z, W\}$, but not $\{Z\}$ or $\{W\}$.

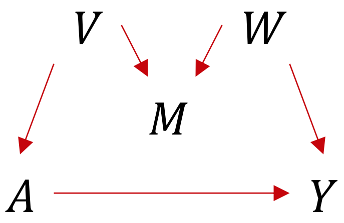
Disjunctive Cause Criterion

Disjunctive Cause Criterion: Control for all (observed) causes of the exposure, the outcome, or both.

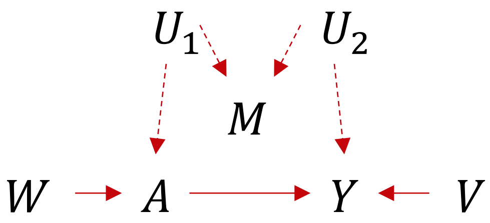
- It is conceptually simper than the backdoor path criterion.
- It guarantees to select a set of variables that are sufficient to control for confounding, if:
  - Such a set exists
  - We correctly identify all the observed causes of A and Y.

Note: it’s not to say we block all pre-treatment covariates.

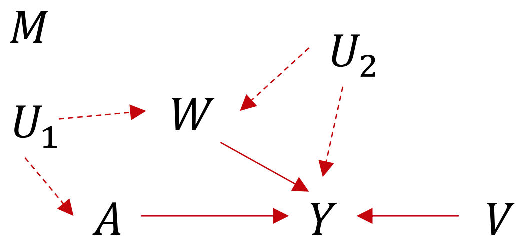
Disjunctive Cause Criterion



1. Using all pre-treatment covariates: $\{M, W, V\}$, satisfies the backdoor path criterion. ✓
2. Using disjunctive cause criterion: $\{W, V\}$, satisfies the backdoor path criterion. ✓



1. Using all pre-treatment covariates: $\{M, W, V\}$, does not satisfy the backdoor path criterion. ✗
2. Using disjunctive cause criterion $\{W, V\}$, satisfies the backdoor path criterion. ✓



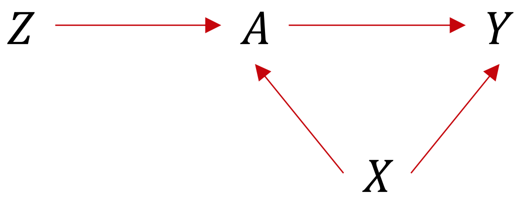
1. Using all pre-treatment covariates $\{M, W, V\}$, does not satisfy the backdoor path criterion. ✗
2. Using disjunctive cause criterion $\{W, V\}$, does not satisfy the backdoor path criterion. ✗

Part V: Instrumental Variables Methods

Instrumental Variables

Instrumental variables (IV) method is an alternative causal inference method that does not rely on the ignorability assumption.

Here, Z is an IV:



- It affects treatment but does not (directly) affect the outcome.
- Think of Z as encouragement.

Randomized Trials with Noncompliance

Setup:

- Z: randomization to treatment (1 if randomized to treatment, 0 otherwise), *e.g.*, encouragement to stop smoking $Z = 1$.
- A: treatment received (1 if received treatment, 0 otherwise), *e.g.*, smoking during pregnancy $A = 1$.
- Y: outcome, *e.g.*, birthweight.
- X: parity, mother's age, weight, etc.
- Noncompliance means not everyone assigned to treatment will receive treatment.

Compliance classes (principle strata):

$A^{Z=0}$	$A^{Z=1}$	**Label**
0	0	Never-takers
0	1	Compliers
1	0	Defiers
1	1	Always-takers

Angrist, Imbens, & Rubin 1996. "Identification of Causal Effects Using Instrumental Variables"

Local Average Treatment Effect

The target of inference is:

$$
\begin{aligned}
&E(Y^{Z=1} - Y^{Z=0} | A^{Z=0} = 0, A^{Z=1} = 1) \\
&= E(Y^{Z=1} - Y^{Z=0} | \text{compliers}) \\
&= E(Y^{a=1} - Y^{a=0} | \text{compliers})
\end{aligned}
$$

- This is causal because it contrasts counterfactuals in a common population.
- Known as complier average causal effect (CACE)
  - It is a causal effect in a subpopulation.
  - It is a causal effect of treatment received.
  - No inference about defiers, always-takers, or never-takers.

Observed Data

For each person we observe an A and Z, not (A^0, A^1).

Z	A	A^0	A^1	Class
0	0	0	?	Never-takers or compliers
0	1	1	?	Always-takers or defiers
1	0	?	0	Never-takers or defiers
1	1	?	1	Always-takers or compliers

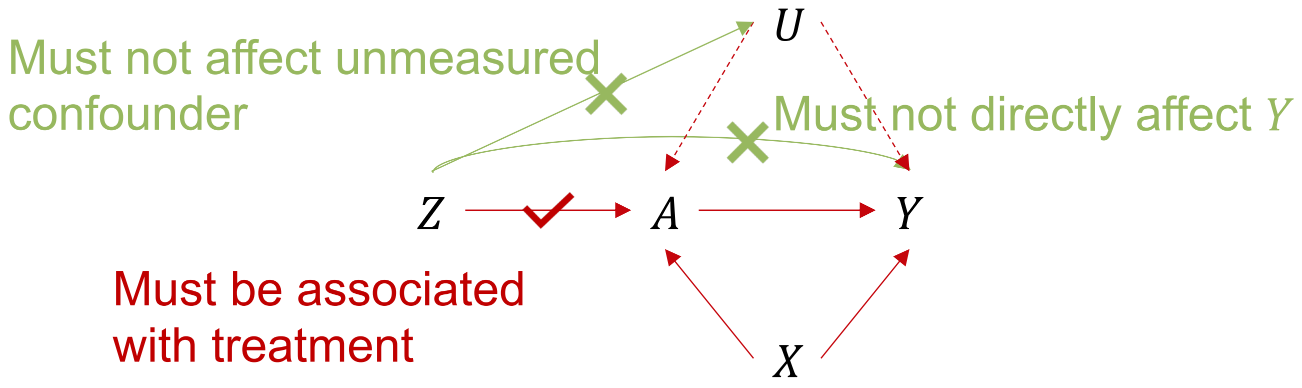
To estimate CACE, we need to make the following assumptions:

1. It is associated with the treatment;
2. (Exclusion restriction) It affects the outcome only through its effect on treatment;
3. (Monotonicity) There are no defiers, *i.e.*, the probability should increase with more encouragement.

Exclusion Restriction

A variable is an instrumental variable (IV) if:

1. It is associated with the treatment;
2. (Exclusion restriction) It affects the outcome only through its effect on treatment;


Must not affect unmeasured confounder
U
Must not directly affect Y
Z
A
Y
Must be associated with treatment
X

Monotonicity

Observed data with monotonicity:

Z	A	A^0	A^1	Class
0	0	0	?	Never-takers or compliers
0	1	1	1	Always-takers ~~or defiers~~
1	0	0	0	Never-takers ~~or defiers~~
1	1	?	1	Always-takers or compliers

- Intention-to-treat (ITT) effect:

$$E(Y^{Z=1} - Y^{Z=0}) = E(Y|Z=1) - E(Y|Z=0)$$

- Treatment assignment on treatment received effect:

$$\begin{aligned} E(A^{Z=1} - A^{Z=0}) &= E(A|Z=1) - E(A|Z=0) \\ &= P(A|Z=1) - P(A|Z=0) \\ \text{by monotonicity} &= P(\text{compliers}) \end{aligned}$$

Complier Average Causal Effect (CACE)

ITT effect:

$$\begin{aligned} E(Y|Z=1) = {} & E(Y|Z=1,\text{always−takers})P(\text{always−takers}) \\ & + E(Y|Z=1,\text{never−takers})P(\text{never−takers}) \\ & + E(Y|Z=1,\text{compliers})P(\text{compliers}) \end{aligned}$$

$$\begin{aligned} E(Y|Z=0) = {} & E(Y|Z=0,\text{always−takers})P(\text{always−takers}) \\ & + E(Y|Z=0,\text{never−takers})P(\text{never−takers}) \\ & + E(Y|Z=0,\text{compliers})P(\text{compliers}) \end{aligned}$$

Notice that:

$$\begin{aligned} E(Y|Z=1,\text{always−takers}) &= E(Y|\text{always−takers}) \\ E(Y|Z=1,\text{never−takers}) &= E(Y|\text{never−takers}) \\ E(Y|Z=0,\text{always−takers}) &= E(Y|\text{always−takers}) \\ E(Y|Z=0,\text{never−takers}) &= E(Y|\text{never−takers}) \end{aligned}$$

Because treatment assignment has no impact on always-takers or never-takers.

Complier Average Causal Effect (CACE)

Therefore,

$$\begin{aligned} E(Y|Z=1)-E(Y|Z=0) = {} & E(Y|Z=1,\text{compliers})P(\text{compliers}) \\ & - E(Y|Z=0,\text{compliers})P(\text{compliers}) \end{aligned}$$

Which implies:

$$\begin{aligned} & \frac{E(Y|Z=1)-E(Y|Z=0)}{P(\text{compliers})} \\ & = E(Y|Z=1,\text{compliers}) - E(Y|Z=0,\text{compliers}) \\ & = E(Y^{a=1}|\text{compliers}) - E(Y^{a=0}|\text{compliers}) \\ & = \text{CACE} \end{aligned}$$

Complier Average Causal Effect (CACE)

$$\text{CACE} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(A|Z=1) - E(A|Z=0)}$$

ITT: causal effect of treatment assignment on the outcome

Causal effect of treatment assignment on the treatment received

Note:

- If perfect compliance, $\text{CACE} = \text{ITT}$.
- Denominator always between 0 and 1. Thus, CACE will be at least as large as ITT.
- The denominator is the proportion of compliers, which also measures the strength of an IV. A weak instrument leads to large variance estimates.

IVs in Observational Studies

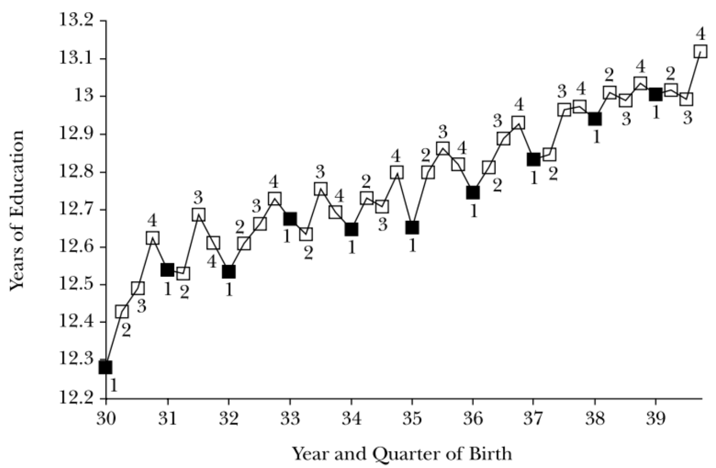
IVs can be thought of as randomizers in natural experiments.

Examples:

- Mendelian randomization: some genetic variant is associated with some behavior (*e.g.*, alcohol use) but it is assumed to not be associated with outcome of interest.
- Provider preference: use treatment prescribed to previous patients as an IV for current patients. Idea: previous decision should not be associated with current decision, but previous decision should not directly affect outcome.
- Quarter of birth → years in school → income.

Figure 1

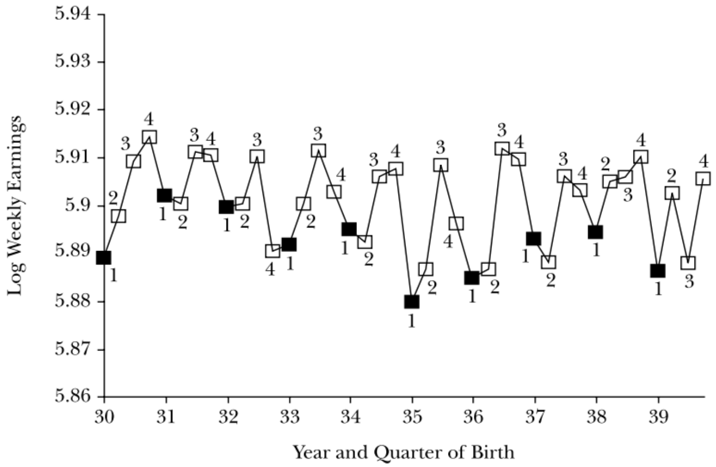
Mean Years of Completed Education, by Quarter of Birth



Source: Authors' calculations from the 1980 Census.

Figure 2

Mean Log Weekly Earnings, by Quarter of Birth



Source: Authors' calculations from the 1980 Census.

Two Stage Least Squares

1. Regress $A \sim Z$, to estimate $\hat{A}_i$

$$A_i = \alpha_0 + Z\alpha_1 + \epsilon_i$$

- $\hat{A}_i$ is projection of A onto space spanned by Z.
- $\alpha_1 = E(A|Z=1) - E(A|Z=0) = P(\text{compliers})$

2. Regress $Y \sim \hat{A}$

$$Y_i = \beta_0 + \hat{A}_i\beta_1 + \varepsilon_i$$

- $\beta_1 = \frac{E(Y|\hat{A}=\widehat{\alpha}_0+\widehat{\alpha}_1) - E(Y|\hat{A}=\widehat{\alpha}_0)}{\widehat{\alpha_1}} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(A|Z=1) - E(A|Z=0)} = \text{CACE}$