Causal Inference I: Matching and Inverse Probability of Treatment Weighting (IPTW)

Li Ge

Ph.D. student in Biomedical Data Science

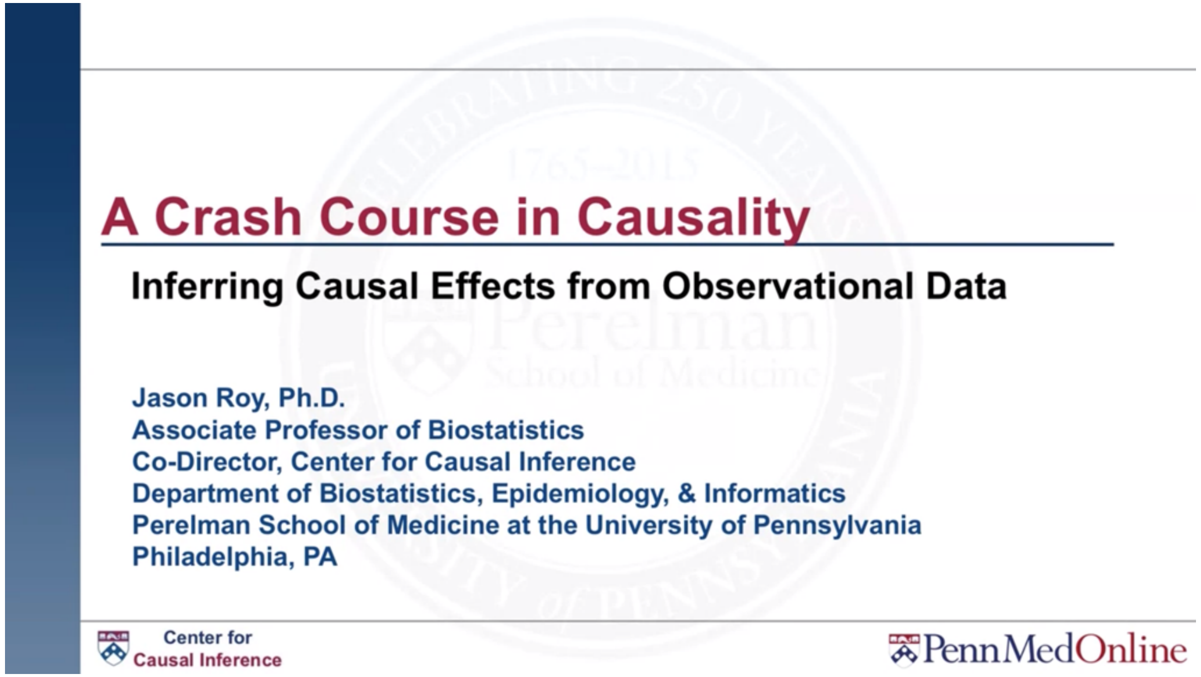
Jul 3, 2020

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

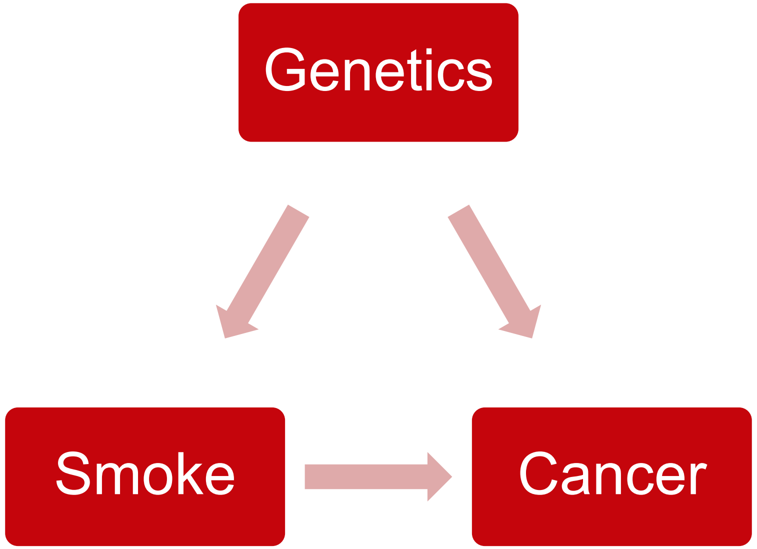
PennMedOnline

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

Part I: Introduction to Causal Effects

What makes a relationship causal?

- Cigarettes, damn cigarettes and statistics



- Bradford Hill criteria (1965)

Cigarettes, damn cigarettes and statistics, Tim Harford: https://timharford.com/2015/04/cigarettes-damn-cigarettes-and-statistics/

Potential Outcomes and Counterfactuals

Suppose we are interested in the causal effects of some treatment A on some outcome Y.

Treatment examples:

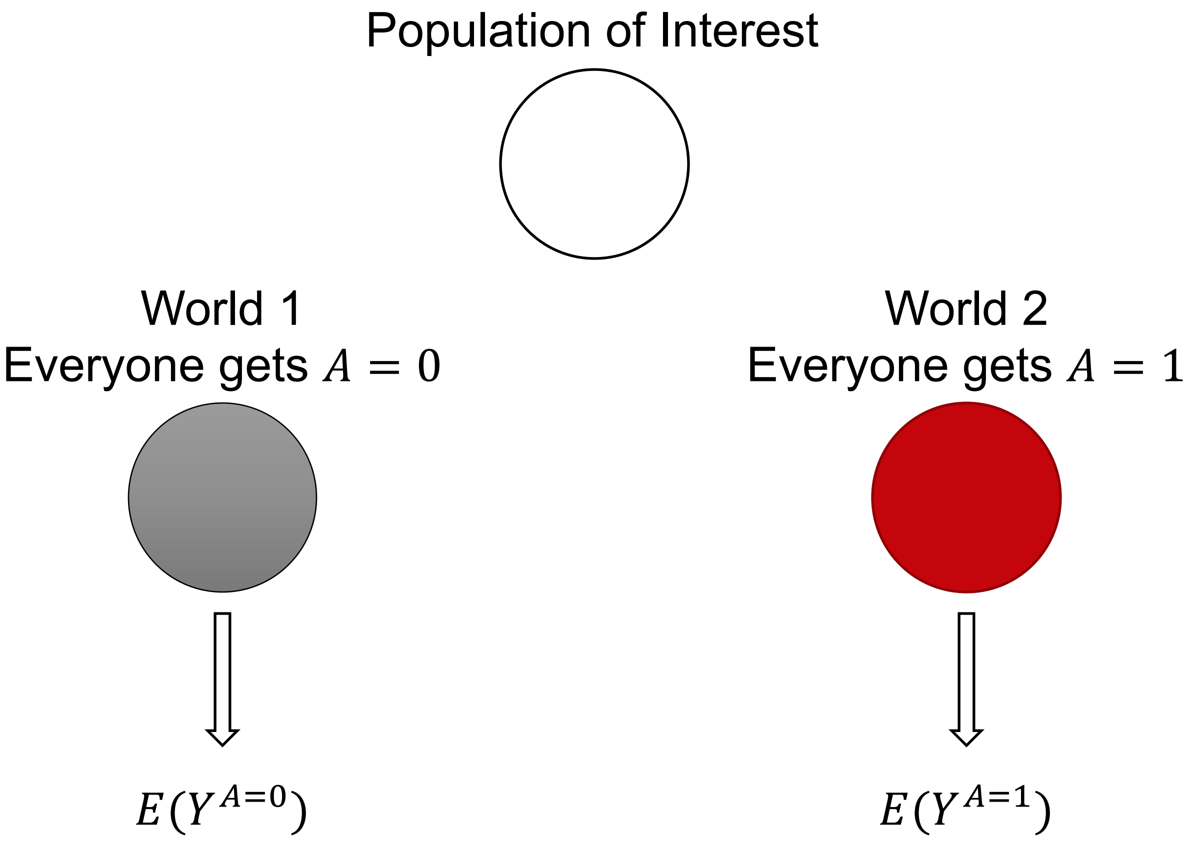
- $A = 1$ if receive active drug; $A = 0$ if receive placebo
- $A = 1$ if smoke; $A = 0$ otherwise

Outcome examples:

- $Y = time\ until\ death$
- $Y = 1$, if lung cancer

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 $A = 1$

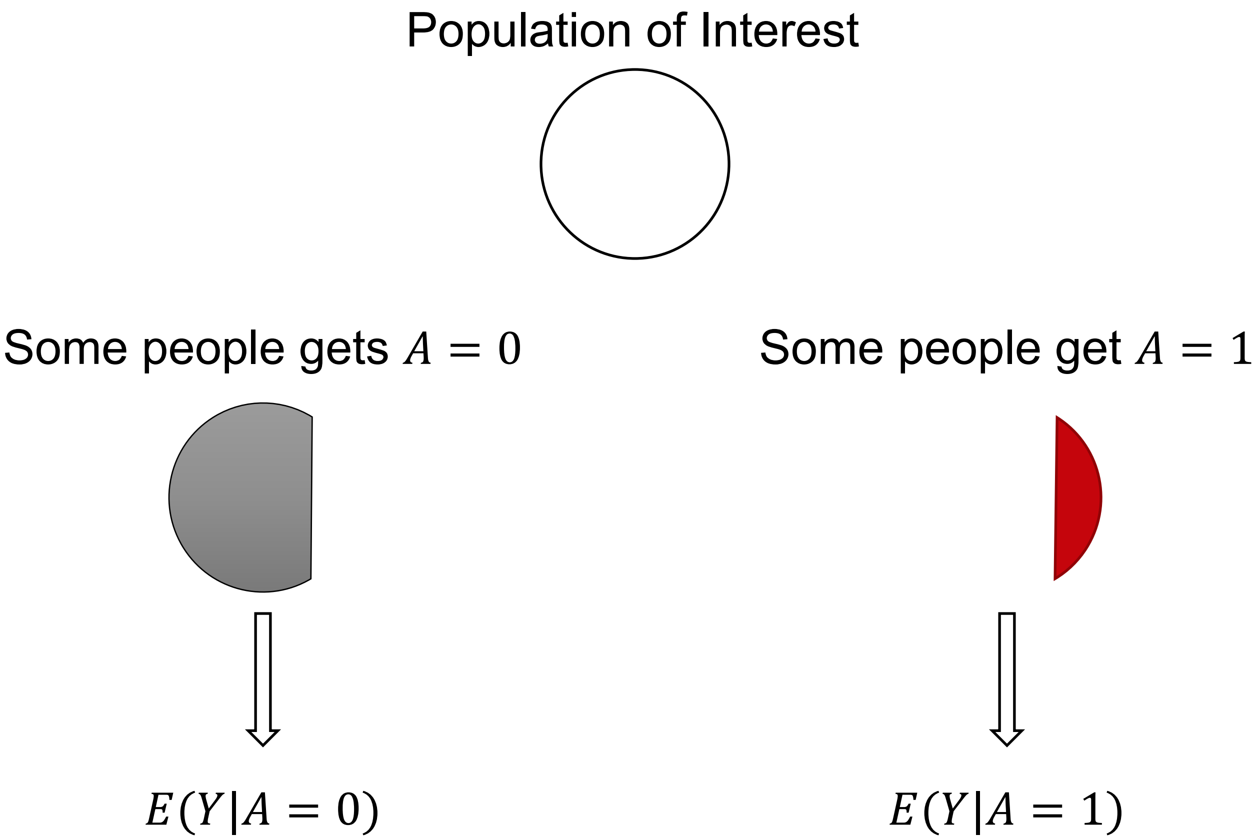
$E(Y^{A=0})$

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

- 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 Effects

- $E(Y^{A=1} - Y^{A=0})$: causal risk difference. ✓
- $E(Y|A=1) - E(Y|A=0)$: treatment effect difference, but it is comparing two difference populations of people. ✗
- $E(Y^{A=1} - Y^{A=0}|A=1)$: causal effect of treatment on the treated. ✓
- $E(Y^{A=1} - Y^{A=0}|V=v)$: causal effect in the subpopulation with covariate $V=v$. ✓
- $E\left(\frac{Y^{A=1}}{Y^{A=0}}\right)$: causal relative risk. ✓

Caveats of Causal Effects

"No causation without manipulation." - Holland (1986)

- Causal effects of (hypothetical) interventions are generally well-defined and actionable, *e.g.*, drug A vs. drug B.
    - Hidden versions of treatment, *e.g.*, body mass index (BMI).
    - Immutable variables, *e.g.*, race, gender, age.

- The Fundamental Problem of Causal Inference is that we can only observe one potential outcome for each person.
    - $Y_i^{A=1} - Y_i^{A=0}$: individual treatment effect (ITE), hopeless.
    - $E(Y^{A=1} - Y^{A=0})$: average treatment effect (ATE), possible with models and assumptions.

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)
  - Consistency
  - Ignorability
  - Positivity

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

SUTVA

The Stable Unit Treatment Value Assumption (SUTVA) really involves two assumptions.

- No interference:
  - Units do not interfere with each other.
  - Treatment assignment of one unit does not affect that outcome of another unit.
  - Spillover or contagion are also terms for interference.
- One version of treatment

SUTVA allows us to write potential outcome for the i^{th} person in terms of only that person’s treatment.

Consistency

The consistency assumption (no different versions of treatment):

The potential outcome under treatment $A = a$, *i.e.*, Y^a, is equal to the observed outcome if the actual treatment received is $A = a$.

$$Y = Y^a, \text{if } A = a, \text{for all } a$$

Ignorability

The ignorability* assumption:

Given pre-treatment covariates X, treatment assignment is independent from the potential outcomes.

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

- Among people with same values of X, we can think of treatment A as being randomly assigned.
- Note it does not imply that the observed outcome $Y \perp A$. In fact, $Y = Y^1A + Y^0(1-A) = A(Y^1 - Y^0) + Y^0$. Therefore, $Y \perp A \Leftrightarrow Y^1 = Y^0$ (null hypothesis of no treatment effect).

* sometimes referred to as the "no unmeasured confounders" assumption

Positivity

The positivity assumption states that, for every set of values of X, treatment assignment is not deterministic:

$$P(A = a|X = x) > 0, \text{for all } a \text{ and } x$$

Everybody has some chances of getting either treatment. Otherwise, we will have no information for some subpopulations.

Solution if violation: redefine the population of interest.

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.

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

Part II: Confounding and Directed Acyclic Graphs (DAGs)

I'll talk more about this at next week's group meeting.

Confounding

Recall the causal effects, *e.g.*, $E(Y^1 - Y^0)$.

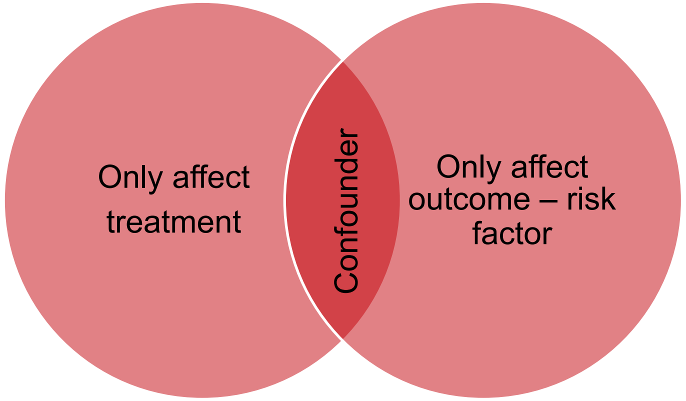
To be able to estimate it from observational data, we make several assumptions, including ignorability: $Y^0, Y^1 \perp A|X$

- Violation means treatment assignment depends on the potential outcomes even within each stratum of the pre-treatment covariates X, *i.e.*, treatment assignment is not randomized within levels of X.
- Assuming the marginal independence between potential outcomes and treatment assignment $Y^0, Y^1 \perp A$ is too strong. It is probably only valid in randomized controlled trials (RCTs).
- The question is how to identify a set of variables X that will make the ignorability assumption hold.

Confounding

Confounders are often defined as variables that affect both the treatment and the outcome.

- If assign treatment based on a coin flip, then that affects treatment but should not affect the outcome (the coin flip is not a confounder).
- If people with a family history of cancer are more likely to develop cancer (the outcome), but family history is not a factor in the treatment decision. Family history is not a confounder (it is a risk factor).
- If older people are at higher risk of cardiovascular disease (the outcome) and are more likely to receive statins (the treatment), then age is a confounder.

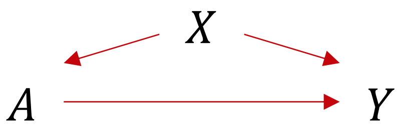


DAG

Recall the informal definition of a confounder:

- A variable that affects both the treatment and the outcome.

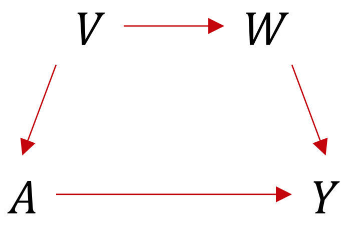
A simple direct acyclic graph (DAG) where X is a confounder between the relationship between treatment A and outcome Y:



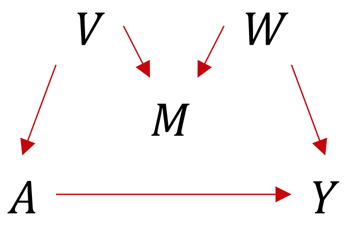
Here X is sufficient to control for confounding, because ignorability holds: $Y^1, Y^0 \perp A|X$.

DAG

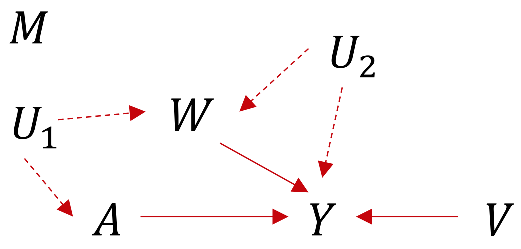
Consider more complex examples:



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 in next week's talk.

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

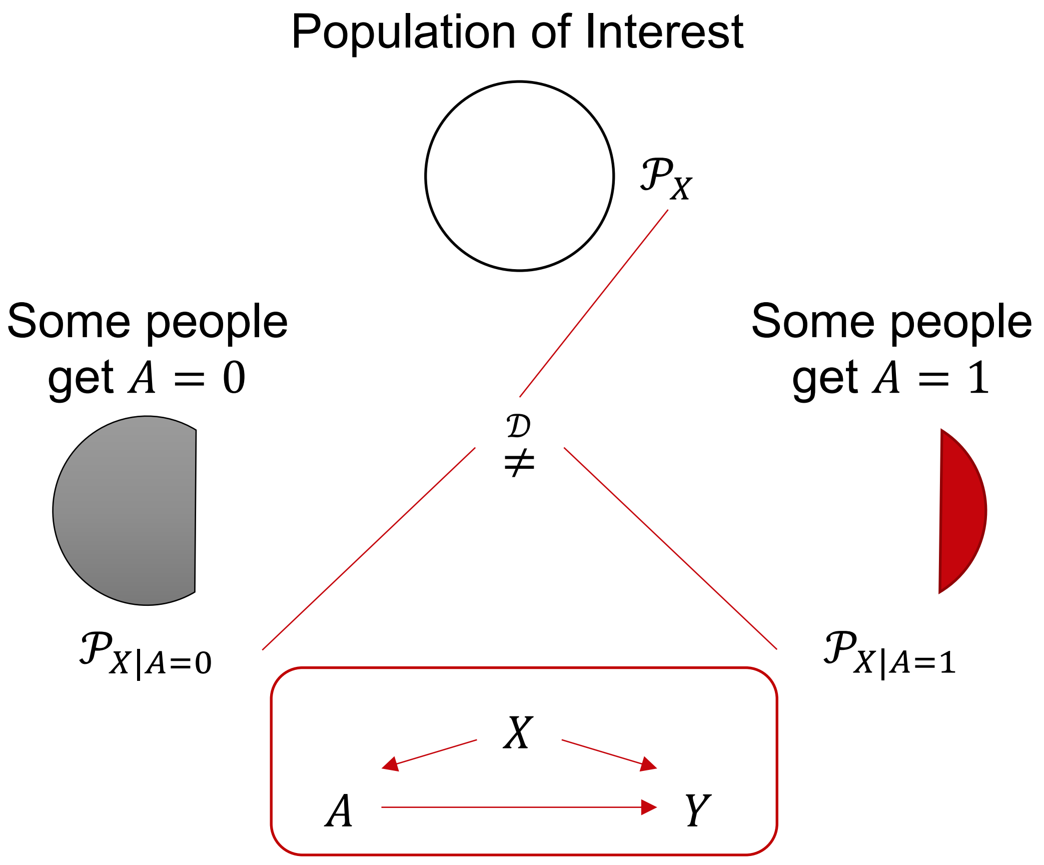
- We'll see that DAGs encode probability distributions.
- We'll be able to recognize different types of paths and understand which of them induce association between nodes.
- We'll see how to block paths to impose conditional independence (d-separation).
- We'll use the backdoor path criterion and the 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.

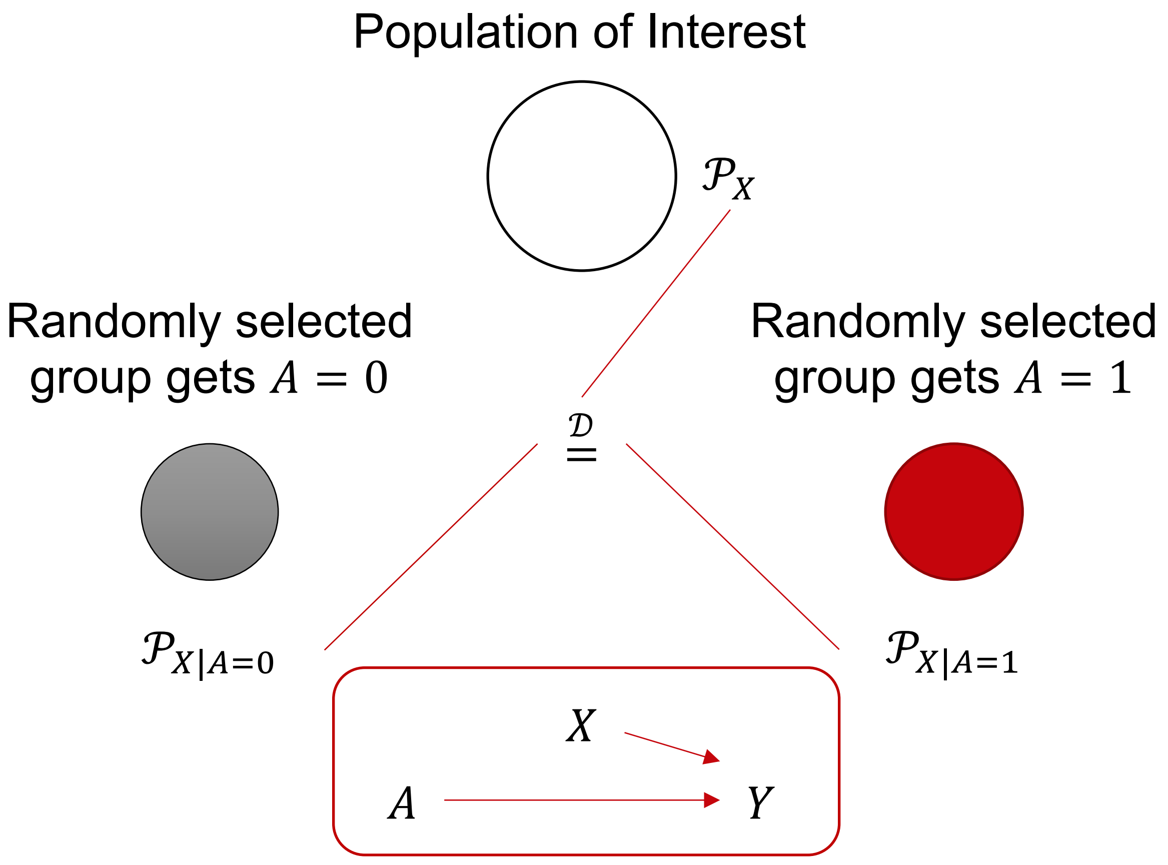
General approaches include matching and inverse probability of treatment weighting.

Part III: Matching and Propensity Scores

Observational Studies



Observational studies

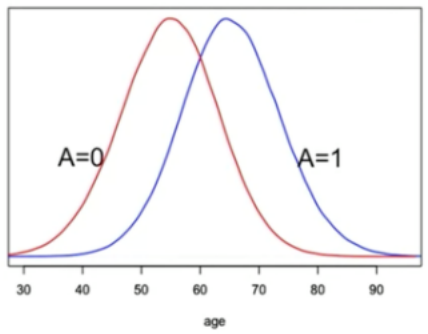

Population of Interest
$\mathcal{P}_X$
Randomly selected group gets $A = 0$
Randomly selected group gets $A = 1$
$\overset{\mathcal{D}}{=}$
$\mathcal{P}_{X|A=0}$
$\mathcal{P}_{X|A=1}$
X
A
Y

Randomized controlled trials

Observational Studies

Observational studies:

- *e.g.*, electric health records, claims, registries.
- Large sample sizes; inexpensive; potential for rapid analysis.
- Data quality typically lower; no uniform standard of collection.



Randomized trials:

- Covariates are balanced by design.
- Expensive; sometimes unethical; people might refuse to participate in trials; time-sensitive, by the time you have outcome data, the question might no longer be relevant.

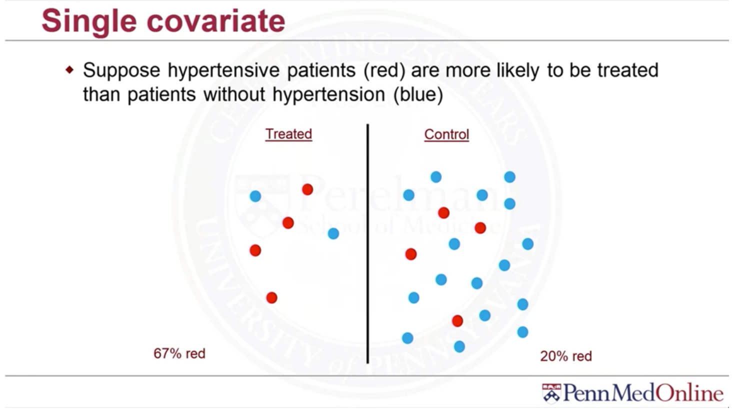
Matching

Matching is a method that attempts to make an observational study more like a randomized trial.

Main idea:

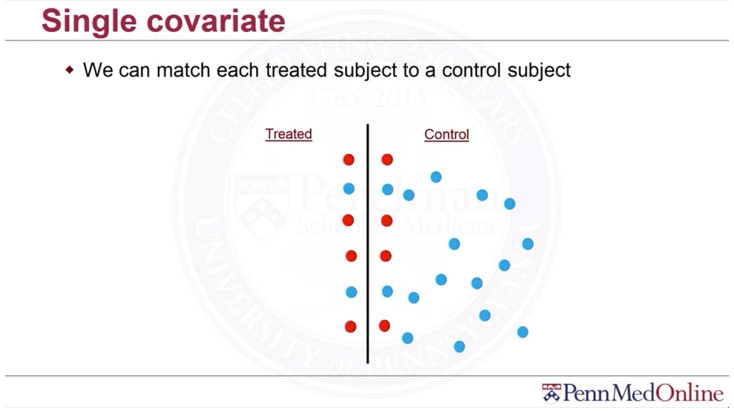
- Match individuals in the treated group $(A = 1)$ to individuals in the control group $(A = 0)$ on the covariates X.
- *e.g.*, in the case where older people are more likely to receive treatment, we can match treated people to control people of the same age, so that there will be about the same number of treated and controls at any age.
- Matching needs to be done at the design phase, *i.e.*, blinded to the outcomes.
- Matching doesn’t always produce perfect balance. Nevertheless, it will reveal lack of overlap in covariate distribution.
- Once data are matched, we can treat them as if from randomized trial. Downstream analysis can be simple.

Matching



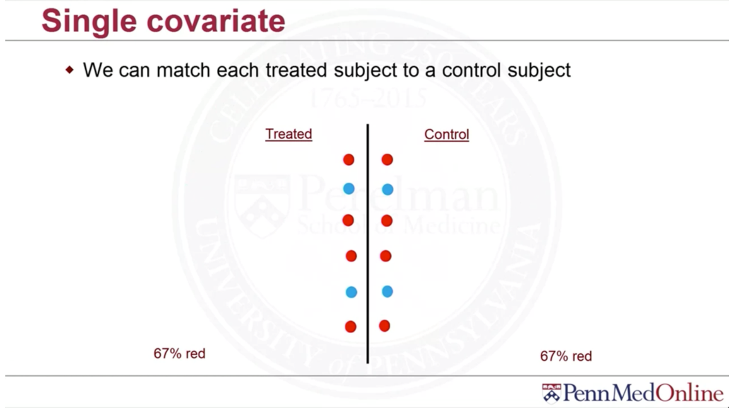
Single covariate

- Suppose hypertensive patients (red) are more likely to be treated than patients without hypertension (blue)



Single covariate

- We can match each treated subject to a control subject



Single covariate

- We can match each treated subject to a control subject

Conceptualize matching using single covariate.

- The goal is to achieve stochastic balance $\mathcal{P}_{X|A=0} \overset{\mathcal{D}}{\sim} \mathcal{P}_{X|A=1}$.
- Notice that we are making the distribution of covariates in the control population look like that in the treated population, thus the causal effect will be on the treated population.
- More details, *e.g.*, target population, fine balance, number of matches.

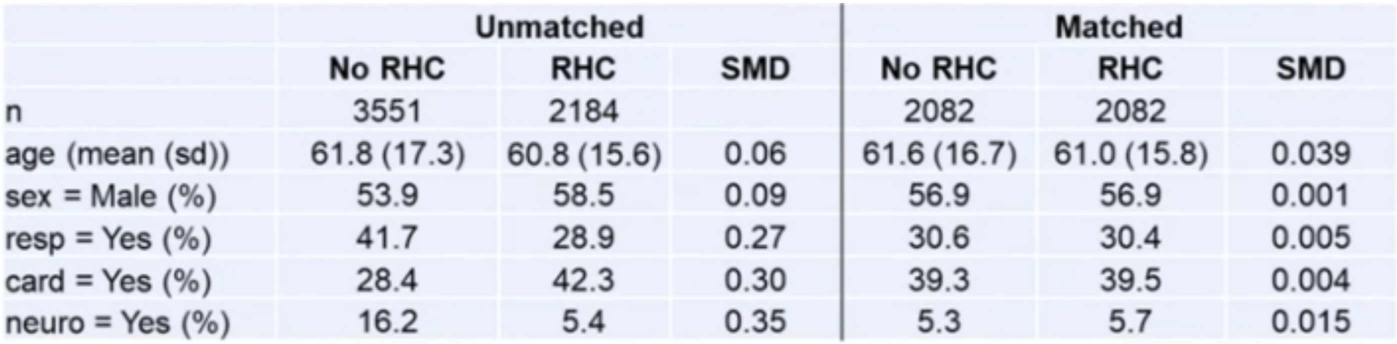
Matching Procedures

1. Select a set of pre-treatment covariates X that (hopefully) satisfy the ignorability assumption.
2. Calculate the distance matrix $D = \left(d_{ij}\right) \in \mathbb{R}_{0+}^{m\times n}$ that contains the pairwise distance $d_{ij} = \mathcal{D}\left(X_i, X_j\right)$ between each treated subject and control subject.
   - *e.g.*, Mahalanobis distance $\mathcal{D}\left(X_i, X_j\right) \coloneqq \sqrt{\left(X_i - X_j\right)^T \Sigma^{-1}\left(X_i - X_j\right)}.$
   - 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.

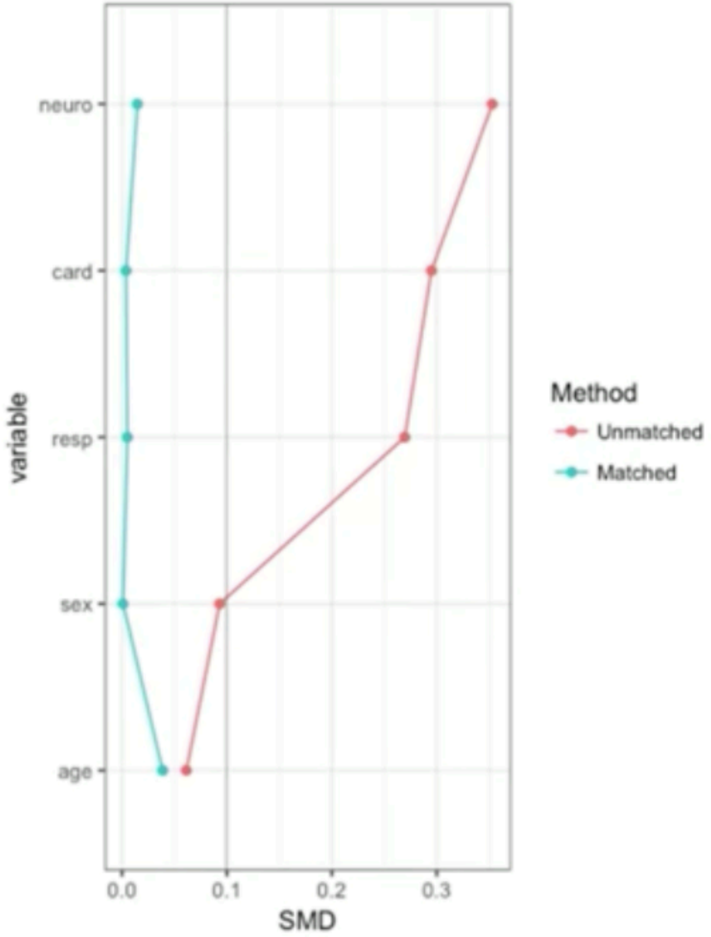
Table 1: Patient baseline characteristics table



	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.

Denote the propensity score for subject i by $\pi_i = \pi(X_i) = P(A = 1|X_i)$.

- Suppose age is the only X variable, and older people are more likely to get treatment.
- That is, $\pi_i > \pi_j$ if $X_i > X_j$. Then, $P(A = 1|\text{age} = 60) > P(A = 1|\text{age} = 30)$.
- $\pi_i = 0.3$ means that if a person i has a propensity score value of 0.3, given that person's covariate value X_i, there is a 30% chance the person will be treated.

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.

Balancing Score

Suppose two subjects have the same value of the propensity score, but they possibly have different covariate values X.

Despite the different covariate values, they are both equally likely to be treated.

- This means that both subjects' X is just as likely to be found in the treatment group.
- If you restrict to a subpopulation of subjects who have the same value of the propensity score, there should be balance in two treatment groups.
- Thus, the propensity score is a balancing score.

Balancing Score

More formally, $b(X)$ 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,

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

Estimated Propensity Score

In a randomized trial, the propensity score is generally known, *e.g.*, $P(A = 1|X) = P(A = 1) = 0.5$.

In an observational study, it will be unknown.

- We therefore need to estimate it from the observed data.
- Typically when people talk about a propensity score, they are referring to the estimated propensity score $\hat{\pi}_i$.
- Nonparametric:

$$\hat{\pi}(x) = \frac{\hat{P}(A = 1, X = x)}{\hat{P}(X = x)}$$

- Parametric:

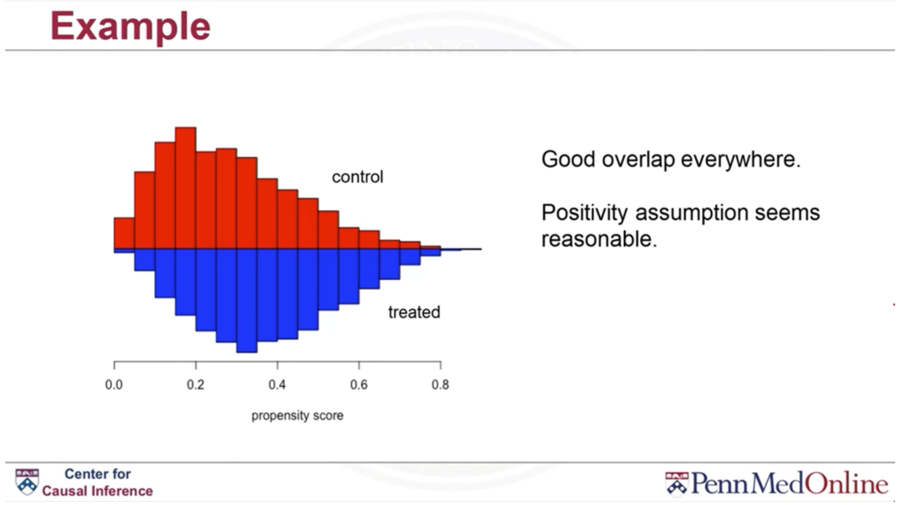
$$\text{logit}\big(\pi(X;\gamma)\big) = X^T\gamma$$

Propensity Score Matching

Follows the general matching procedure.

In practice, $\mathcal{D}\left(X_i, X_j\right) := \left|\text{logit}\left(\pi(X_i)\right) - \text{logit}\left(\pi(X_j)\right)\right|$.

- The propensity score is bounded between 0 and 1, making many values seem similar.
- Use logit to transform the propensity score to $\mathbb{R}$, while preserving ranks.



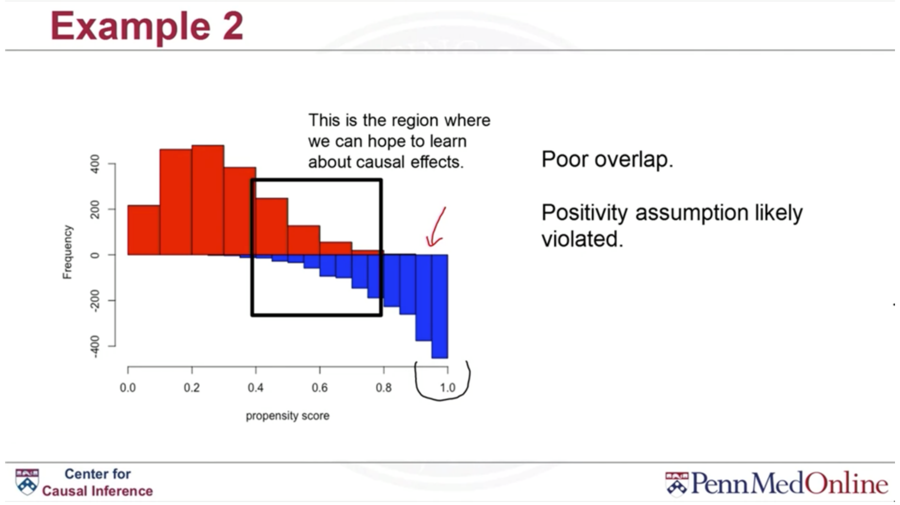
Example

Good overlap everywhere.

Positivity assumption seems reasonable.

Center for Causal Inference

PennMedOnline



Example 2

This is the region where we can hope to learn about causal effects.

Poor overlap.

Positivity assumption likely violated.

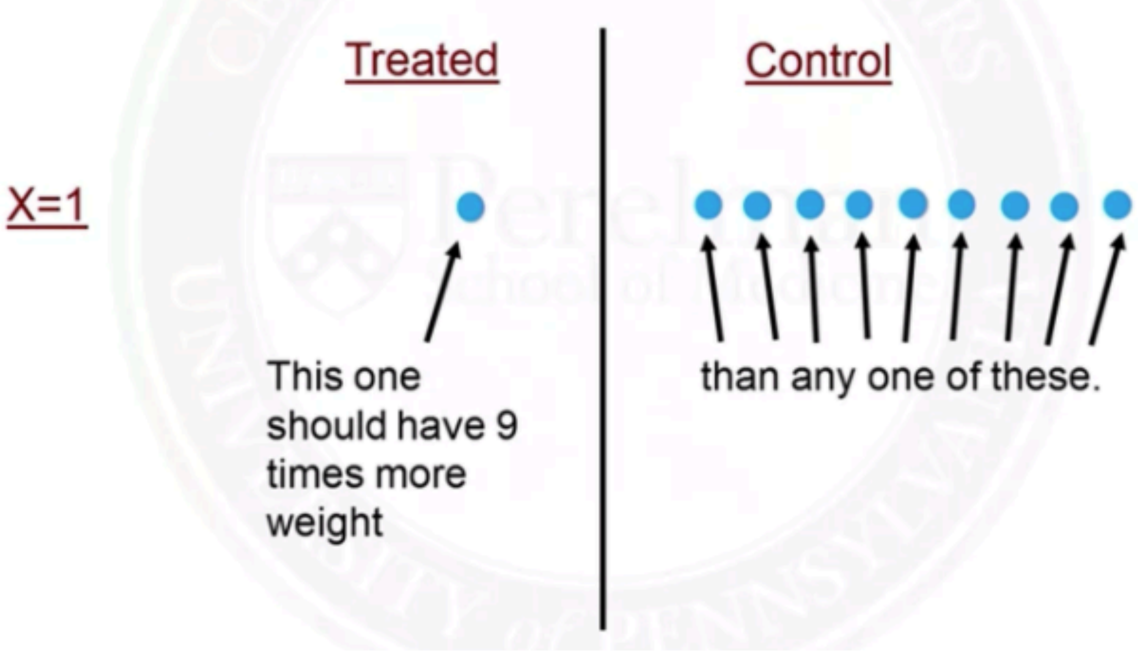
Center for Causal Inference

PennMedOnline

Part IV: Inverse Probability of Treatment Weighting (IPTW)

Intuition for IPTW

Rather than match, we could use all the data, but down-weight over-representative ones and up-weight under-representative ones.



Weight: $\frac{1}{P(A=1|X=1)} = \frac{1}{0.1} = 10$

Weight: $\frac{1}{P(A=0|X=1)} = \frac{1}{0.9} = \frac{10}{9}$

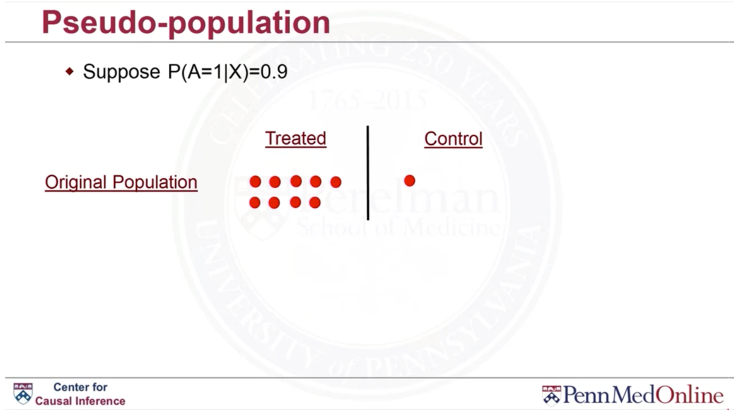
Intuition for IPTW

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 $P(A=0|X)=1-\pi(X)$.

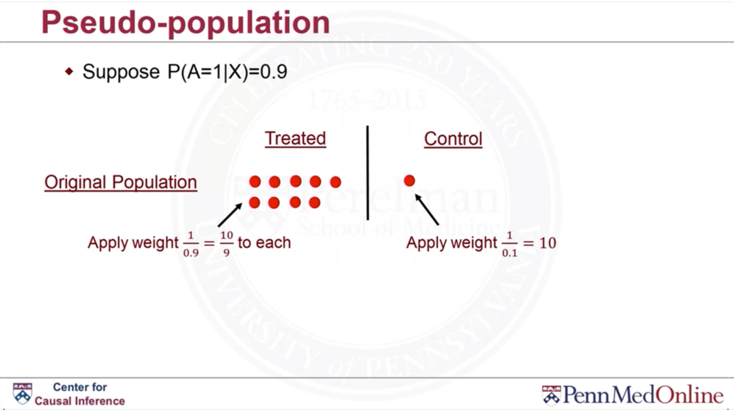
Hence, it is called the inverse probability of treatment weighting (IPTW).

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



Pseudo-population

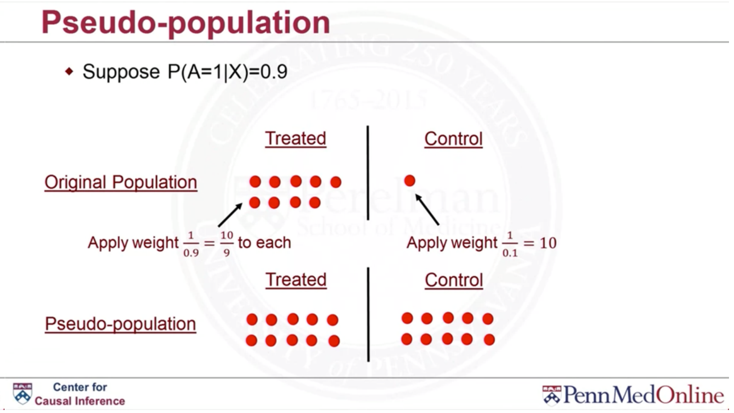
- Suppose P(A=1|X)=0.9



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

IPTW Estimator

Under the assumption of ignorability and positivity, we can estimate $E(Y^1)$ as

$$\frac{\sum_{i=1}^{n} I(A_i=1)\frac{Y_i}{\hat{\pi}_i}}{\sum_{i=1}^{n}\frac{I(A_i=1)}{\hat{\pi}_i}},$$

Sum of the Y’s in treated pseudo-population

Number of subjects in treated pseudo-population

where $\hat{\pi}_i = \hat{P}(A = 1|X_i)$ is the estimated propensity score.

Marginal Structural Models

Motivation

- Previously we discussed IPTW estimation for simple causal effects, such as an average causal effect.
- However, IPTW estimation methods can be used more generally to estimate causal effect parameters from models.

Marginal Structural Models

Linear MSM:

$$E(Y^a) = \psi_0 + \psi_1 a, \qquad a = 0, 1$$

- $E(Y^0) = \psi_0$
- $E(Y^1) = \psi_1 - \psi_0$
- ψ_1 is the average causal effect $E(Y^1 - Y^0)$

Marginal Structural Models

Logistic MSM for binary outcome:

$$\operatorname{logit}\big(E(Y^a)\big) = \psi_0 + \psi_1 a, \qquad a = 0,1$$

- $\exp(\psi_1)$ is the causal odds ratio

$$\frac{\dfrac{P(Y^1 = 1)}{1 - P(Y^1 = 1)}}{\dfrac{P(Y^0 = 1)}{1 - P(Y^0 = 1)}}$$

Odds that $Y^1 = 1$

Odds that $Y^0 = 1$

Marginal Structural Models

- MSMs can also include effect modifiers.
- Suppose V is a variable that modifies the effect of A.
- A linear MSM with effect modification:

$$E(Y^a|V) = \psi_0 + \psi_1 a + \psi_3 V + \psi_4 aV, \qquad a = 0{,}1$$

- $E(Y^1 - Y^0|V) = \psi_1 + \psi_4 V$

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 the generalized estimation equation (GEE) of a generalized linear model (GLM):

$$E(Y_i|X_i) = \mu_i = g^{-1}\left(X_i^T\beta\right)$$

- Estimation involves solving

$$\sum_{i=1}^{n} \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1}\left(Y_i - \mu_i(\beta)\right) = 0$$

for β.

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 Estimation

- We can estimate $E(Y^1)$ using IPTW:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i}{\hat{\pi}_i}$$

- If the propensity score is correctly specified, this estimator is unbiased:

$$\begin{aligned}
E\left(\frac{AY}{\pi(X)}\right) &= E\left(\frac{AY^1}{\pi(X)}\right) \text{by } AY = A(AY^1 + (1-A)Y^0) = AY^1 \\
&= E\left(E\left(\frac{AY^1}{\pi(X)}|Y^1, X\right)\right) \text{ by law of total expectation} \\
&= E\left(E(A|Y^1, X)\frac{Y^1}{\pi(X)}\right) \\
&= E\left(E(A|X)\frac{Y^1}{\pi(X)}\right) \text{ by ignorability} \\
&= E\left(\pi(X)\frac{Y^1}{\pi(X)}\right) \\
&= E(Y^1)
\end{aligned}$$

Regression-Based Estimation

- Alternatively, we could estimate $E(Y^1)$ by specifying an outcome model $m_1(X) = E(Y|A = 1, X)$ and then average over the distribution of X:

$$\frac{1}{n}\sum_{i=1}^{n}(A_iY_i + (1 - A_i)m_1(X_i))$$

For subjects with $A = 1$, use observed Y

For other subjects, use predicted value of Y given their X, if their A had been 1

- If outcome model is correctly specified, then this estimator is unbiased.

Doubly Robust Estimators

- A doubly robust estimator is an estimator that is unbiased if either the propensity score model or the outcome regression model are correctly specified.
- Example:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\underbrace{\frac{A_iY_i}{\hat{\pi}_i}}_{\text{IPTW}} - \underbrace{\frac{A_i-\hat{\pi}_i}{\hat{\pi}_i}m_1(X_i)}_{\text{Augmentation}}\right\}$$

Doubly Robust Estimators

- If propensity score is correct, but outcome model is not:

Expectation of this is equal to propensity score

$$\frac{1}{n}\sum_{i=1}^{n}\Bigg\{\frac{A_iY_i}{\hat{\pi}_i} - \underbrace{\frac{A_i-\hat{\pi}_i}{\hat{\pi}_i}m_1(X_i)\Bigg\}}_{\text{This part has expectation 0}}$$

Doubly Robust Estimators

- If propensity score is wrong, but outcome model is correct:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_iY_i}{\hat{\pi}_i}-\frac{A_i-\hat{\pi}_i}{\hat{\pi}_i}m_1(X_i)\right\}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\{\underbrace{\frac{A_i\left(Y_i-m_1(X_i)\right)}{\hat{\pi}_i}}_{\text{This part has expectation 0}}+\underbrace{m_1(X_i)}_{\text{This part goes to } E(Y^1)}\right\}$$

Doubly Robust Estimators

- These estimators are also known as augmented IPTW (AIPTW) estimators.
  - Can use semiparametric theory to identify best estimators.
  - In general, AIPTW estimators should be more efficient than regular IPTW estimators.

IPTW in Practice

Steps:

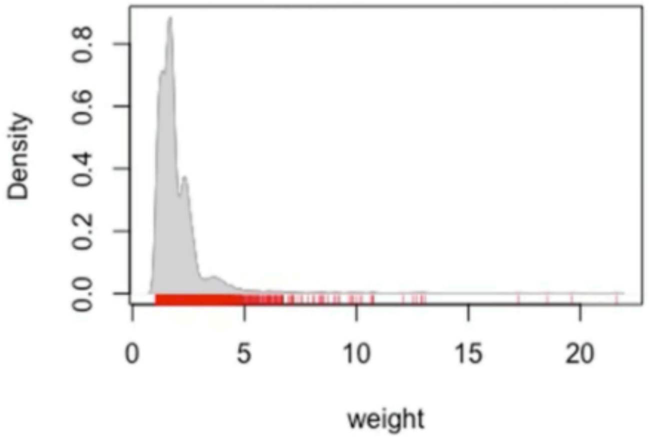
1. Estimate propensity score (*e.g.*, LR: $A \sim X$).
2. Create weights ($w_i = \frac{1}{A_i\hat{\pi}_i+(1-A_i)(1-\hat{\pi}_i)}$).
3. Specify the MSM of interest.
4. Use software to fit a weighted generalized linear model.

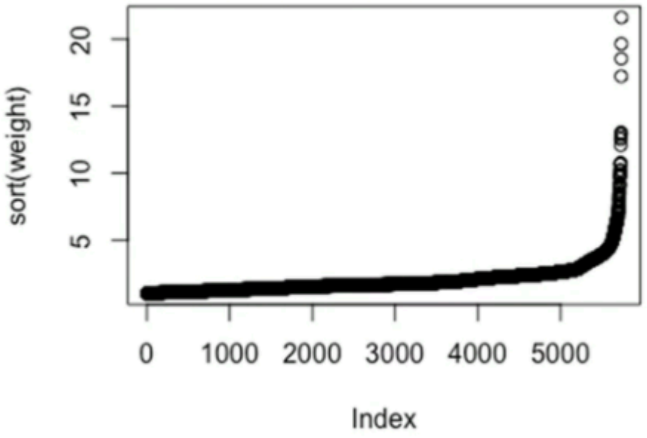
   (*e.g.*, `glm.obj <- glm(y ~ trt, weights = w, family = binomial(link = log))`)
5. Use asymptotic (sandwich) estimator (or bootstrapping) to get standard error.

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

IPTW in Practice

- Also need to assess balance for pseudo-population.
- Might have large weights lead to large standard errors.
  - Check distribution of weights.
  - Trim tail weights or perform weight truncation: biased estimator, smaller variance. Both methods can lead to overall better estimators (lower MSE).


Density
weight


sort(weight)
Index