Causal inference

Fundamental of Inference - Yan Li

2018 - 2019 Spring

individual level causality

A_i actual treatment

$A_i = 0$ not treated ; $A_i = 1$ treated.

Y_i observed outcome

Y_i^a potential / hypothetical outcome

Y_i^0 outcome if person i is not treated

Y_i^1 outcome if person i is treated.

↓

The difference between these two potential outcomes informs the causal effect for individual i:

individual causal effect = $\tau_i = Y_i^1 - Y_i^0$

Assumption of consistency:

This is an assumption that links the potential outcomes to the observed outcome. The key behind this assumption is that there are no multiple treatments.

∴ the observed outcome is the potential outcome if the corresponding treatment is realized.

If $A_i = a$, $Y_i = Y_i^a$

Alternatively written as

$$Y_i = A_i Y_i^{A_i} + (1 - A_i) Y_i^{1-A_i}$$

population level causality

We are not necessarily interested in individual causal effect, but are interested in the effect of the treatment on the population

⇒ Average causal effect

$$\tau = \frac{1}{N} \sum_{i=1}^{N} \tau_i = \frac{1}{N} \sum (Y_i^1 - Y_i^0) = \frac{1}{N} \sum Y_i^1 - \frac{1}{N} \sum Y_i^0 = E(Y^1) - E(Y^0)$$

Causal effect can be measured in different ways (measures of causality)

- Causal risk difference (RD)
  This is what we have been talking about.
  Compute the absolute number of cases of the disease attributable to the treatment $E(Y^1) - E(Y^0)$

  if $E(Y^1) = E(Y^0)$ ⇒ no average causal effect

- Causal relative risk (RR) $\frac{E(Y^1)}{E(Y^0)}$

- Causal odds ratio $\frac{E(Y^1)}{1-E(Y^1)} \Big/ \frac{E(Y^0)}{1-E(Y^0)}$

The fundamental problem, however, is that potential outcomes can not be observed for individuals because each person is either treated and not treated → impossible to individual causal effect

∴ what we really have in reality is measures of associations.
For example, in the table below, we have a sample of treated and untreated individuals.

Table 1.2

	A	Y
Rheia	0	0
Kronos	0	1
Demeter	0	0
Hades	0	0
Hestia	1	0
Poseidon	1	0
Hera	1	0
Zeus	1	1
Artemis	0	1
Apollo	0	1
Leto	0	0
Ares	1	1
Athena	1	1
Hephaestus	1	1
Aphrodite	1	1
Cyclope	1	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0

Parallel to the measures of causality, we have different measures of association

- Associational risk difference
  $E(Y=1 \mid A=1) - E(Y=1 \mid A=0)$

- Associational relative risk $\frac{E(Y=1 \mid A=1)}{E(Y=1 \mid A=0)}$

- Associational odds ratio $\frac{E(Y=1 \mid A=1)}{1-E(Y=1 \mid A=1)} \Big/ \frac{E(Y=1 \mid A=0)}{1-E(Y=1 \mid A=0)}$

if $E(Y=1 \mid A=1) = E(Y=1 \mid A=0)$
⇒ no association

The difference between association and causality is that

$$E(Y=1 \mid A=1) - E(Y=1 \mid A=0) \neq E(Y^1) - E(Y^0)$$

because $E(Y=1 \mid A=1) \neq E(Y^1)$

Expected value of those who actually received the treatment ≠ Expected value if every one in the population would have been treated.

└ due to selection bias ┘

linking association with causality

those who actually received treatments are very different people.

Assumption of ignorability / unconfoundedness / exchangeability.

To link association with causality, what is needed is exchangeability

$$E(Y^a \mid A=a) = E(Y^a \mid A \neq a)$$

The potential outcomes of those who actually received the treatment ↔ same ↔ The potential outcomes of those who did not receive the treatment

∴ the treated and untreated are exchangeable.

$$\Rightarrow E(Y^a \mid A=a) = E(Y^a \mid A \neq a) = E(Y^a)$$

Association: $E(Y \mid A=1) - E(Y \mid A=0)$

Causal effect: $= E(Y^1) - E(Y^0)$

$E(Y \mid A=1) = E(Y^1 \mid A=1)$

$E(Y \mid A=0) = E(Y^0 \mid A=0)$

Assumption of positivity

Exchangeability basically means that the treated and untreated groups are comparability.

For this to be possible, individuals regardless of their characteristics should have a positive change of getting ($A=1$) and not getting the treatment (otherwise one group would consist of substantively different people)

$P(A=a \mid L=l) > 0$ for all l

Let's formally express how association is a function of causality and selection bias

$$E(Y|A=1) - E(Y|A=0)$$
$$= E(Y^1|A=1) - E(Y^0|A=0) \quad \text{bc of consistency}$$
$$= E(Y^1|A=1) - E(Y^0|A=1) + E(Y^0|A=1) - E(Y^0|A=0)$$
$$= \underbrace{E(Y^1 - Y^0|A=1)}_{\text{Average treatment effect on the treated (ATT)}} + \underbrace{E(Y^0|A=1) - E(Y^0|A=0)}_{\text{Selection bias}}$$

To realize exchangeability, randomization in experiments is the way randomization eliminates selection bias

$$SB = E(Y^0|A=1) - E(Y^0|A=0) = 0$$

$$\text{Association} = ATT$$
$$E(Y^1|A=1) - E(Y^0|A=1) = \overbrace{E(Y^1) - E(Y^0)}^{\text{Average causal effect}}$$

With different assumptions realized through randomized experiment, Average causal effect can be identified.

There are different types of randomized experiments.

1) completely randomized experiments

$$A \perp Y^a$$
$$ACE = E(Y^1) - E(Y^0) = E(Y^1|A=1) - E(Y^0|A=0)$$
$$= E(Y|A=1) - E(Y|A=0) = \frac{1}{n_t}\sum^{n_t} Y_i - \frac{1}{n_c}\sum^{n_c} Y_i$$

2) stratified randomized experiments
- completely randomized experiments may not be the most efficient, forming blocks based on covariates increase efficiency (recall SRS sampling vs. stratification)
- complete randomized experiments within blocks

$$A \perp Y^a | L$$

3) Pair randomized experiments
- extreme stratification; 2 cases per block; one receive treatment

Mimicing $Y_i^1 - Y_i^0$

ACE in the case of stratification.

1. Standardization

$$E(Y^a) = E_L[E(Y^a \mid L=l)] = \sum_l E(Y^a \mid L=l) \cdot P(L=l)$$

conditional exchangeability

$$= \sum_l E(Y^a \mid A=a, L=l) P(L=l)$$

consistency

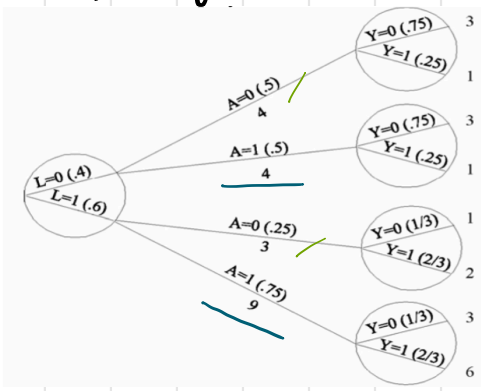
$$= \sum_l E(Y \mid A=a, L=l) P(L=l)$$

$$\text{Causal risk difference} = E(Y^1) - E(Y^0)$$

$$= \sum_l P(L=l)[E(Y \mid A=1, L=l) - E(Y \mid A=0, L=l)]$$

$$\text{causal relative risk} = \frac{E(Y^1)}{E(Y^0)} = \frac{\sum_l P(L=l) E(Y \mid A=1, L=l)}{\sum_l P(L=l) E(Y \mid A=0, L=l)}$$

2. Inverse probability weighting



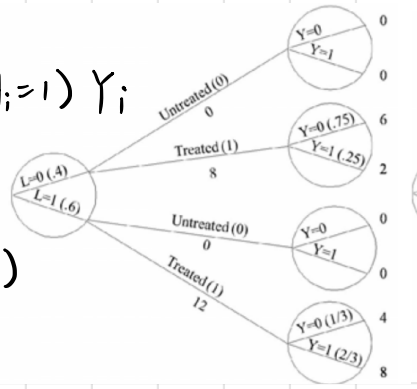
if these untreated people were also assigned to the treatment group

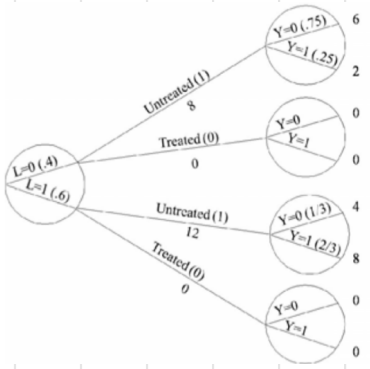
if these treated people were also assigned to the untreated group

$$E(Y^1) = \frac{1}{n} \sum_{i=1}^{n} w_i I(A_i=1) Y_i$$

where

$$w_i = P(A_i=1 \mid L_i)$$





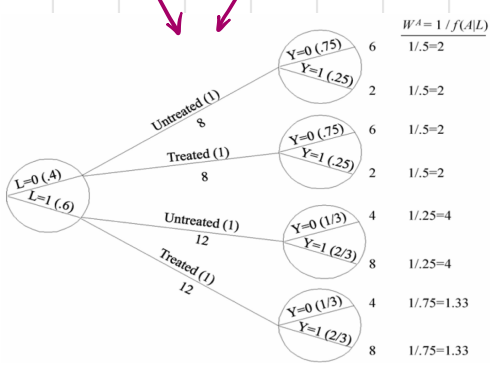
$$E(Y^0) = \frac{1}{n} \sum_{i=1}^{n} w_i' I(A_i=0) Y_i$$

where

$$w_i' = P(A_i=0 \mid L_i) = 1 - P(A_i=1 \mid L_i)$$

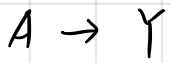
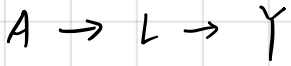
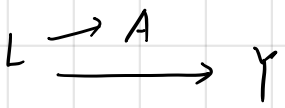
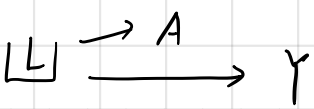
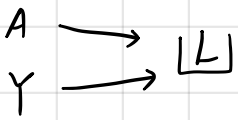
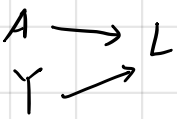
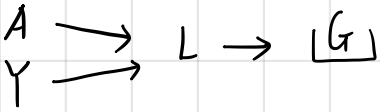
divide by n because the sample is inflated

individual's observation is weighted by his/her probability of being assigned to the treatment or control group.



pseudo-population

Directed Acyclic Graphs (DAGs)

Associated A & Y	Independent A & Y
	
	
	
	
	

d-separation means breaking off all back-door (but not front door) association between A and Y.
so that we know that the observed effect is attributable to the causal effect of A on Y.

conceptual introduction

Observational studies

more feasible, cheaper, better external validity

How to make causal inference in the context of observational studies?

Observational study can be viewed as conditional randomized experiments, in which

- the conditions are not designed but are observed and measured
- the conditional exchangeability is not guaranteed, but only assumed based on substantive knowledge.
  There might be unobserved confounders that are out of researcher's control.

To demonstrate the consequence of uncontrolled confounders.

Suppose $E(Y|A,L) = \alpha_A + \beta_L$ where A is treatment and L is smoking

their effects on Y are additive.

Let $\beta_0 = 0$ for model identification

if smoking is not controlled

$$E(Y|A=1) - E(Y|A=0)$$

$$= p_1 E(Y|A=1, L=1) + (1-p_1) E(Y|A=1, L=0) - p_0 E(Y|A=0, L=1) - (1-p_0) E(Y|A=0, L=0)$$

$$= p_1 (\alpha_1 + \beta_1) + (1-p_1)(\alpha_1 + 0) - p_0 (\alpha_0 + \beta_1) - (1-p_0)(\alpha_0 + 0)$$

$$= \underbrace{\beta_1 (p_1 - p_0)}_{\text{It's contaminated by the effect of smoking as an uncontrolled confounder.}} + \underbrace{\alpha_1 - \alpha_0}_{\text{This is the effect of treatment A that we wish to identify.}}$$

How large the contamination is depends on the effect of smoking and the unbalanceness of the treatment and control groups

How to control for confounders ?

If we only have one or a couple of confounders, we can use
- standardization or
- Inverse probability weighting as introduced in stratified randomization experiments

But if we have many confounders and the treatment vs. control groups are highly unbalanced, then some other logically the same but practically slightly different methods can be used.

1. Matching

step 1: Find each treatment unit its counterfactual identical twin

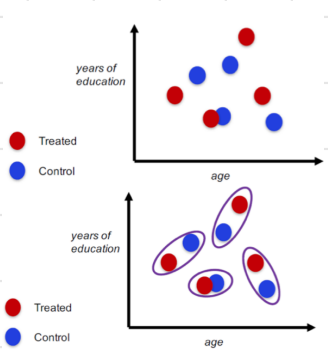
How to find counterfactual identical twin ?

a. distance metric
- > Exact distance
- > Euclidean distance
- > Mahalanobis distance
- > Using propensity score

$P(A_i = 1 \mid X_i) \Rightarrow e(X_i)$

This reduces X dimensions to 1 dimension

difference in propensity scores also indicates distance $d(i,j) = |e(X_i) - e(X_j)|$



b. nearest neighbor (NN) matching using one of distance metrices

I_t denotes the set of treatment units; I_c denotes the set of control units

for each unit in I_t, look for its matching partner in I_c

If this is set as a without replacement process, then order matters

step 2: Check up on the outcome of matched cases

step 3: Estimate individual-level "causal effect" (ICE) and average "causal effect" (ACE)

$$\widehat{ICE}(x_i) = y_i - y_{j(i)}$$

$$\widehat{ACE} = \frac{1}{n} \sum_{i=1}^{n} ICE(x_i)$$

Evaluating the matching

with the technique of matching, the confounders are controlled in the above-mentioned way

How good is the matching? check covariant balance

a. two-sample t-test for the continuous covariates or chi-square independence for discrete covariates

b. standardized difference

$$smd = \frac{\bar{X}_{treatment} - \bar{X}_{control}}{\sqrt{\frac{S^2_{treatment} - S^2_{control}}{2}}} \quad \text{(for each } X \text{ variable)}$$

scale invariant; don't depend on sample size

$smd < 0.1$ adequate

$smd > 0.2$ serious imbalance

Proof that ACE is identifiable under matching:

$$\begin{aligned}
& E(Y^1 - Y^0) \\
&= E(Y^1 \mid A=1) - E(Y^0 \mid A=1) && \text{exchangeability} \\
&= E(Y \mid A=1) - E(Y^0 \mid A=1) && \text{consistency} \\
&= E(Y \mid A=1) - \sum_x E(Y^0 \mid A=1, X=x) P(X=x \mid A=1) \\
&= E(Y \mid A=1) - \sum_x E(Y^0 \mid A=0, X=x) P(X=x \mid A=1) && \text{exchangeability} \\
&= E(Y \mid A=1) - \sum_x E(Y \mid A=0, X=x) P(X=x \mid A=1) && \text{consistency} \\
&= E(Y \mid A=1) - \sum_x E(Y \mid A=0, X=x) P(X=x \mid A=0, I_C) && \text{exact matching}
\end{aligned}$$

☆ why the trouble? where do the exchangeability come from?

2. Propensity score $P(A_i \mid X_i) \Rightarrow e(X_i)$

In matching method, propensity score has debutted as a distance metric. But propensity score can be treated as an independent method of controlling for confounders.

Why controlling for $e(X_i)$ is equivalent to controlling for X_i?

$$P(A_i=1 \mid e(X_i)) = E(A_i=1 \mid e(X_i)) = E_x[E(A_i \mid e(X_i), X_i) \mid e(X_i)]$$
$$= E_x[E(A_i \mid X_i) \mid e(X_i)] = E_x[e(X_i) \mid e(X_i)] = e(X_i)$$
$$= P(A_i=1 \mid X_i)$$

$\therefore$ if $A_i \perp Y^a \mid X_i$ then $A_i \perp Y^a \mid e(X_i)$

Estimating propensity $P(A_i=1 \mid X_i)$ by logistic regression

X should be the set of variables that achieve d-separation and block all the backdoor paths from A to Y

Besides using propensity score as a distance metric in matching, it can also be used as

- stratification
  - identify boundary points $0 = b_0 < b_1 \cdots b_K = 1$ and separate propensity score into strata
  - calculate within strata causal effect → weighted sum across strata
- Matching and stratification can be combined
  - Among the matched sample, stratification further increases the comparability
- Inverse propensity score weighting

(next page ...

- Inverse propensity score weighting

$$E(Y^a) = E_x(E(Y^a \mid X_i)) = \sum_x E(Y^a \mid X_i) P(X_i)$$
$$= \sum_X E(Y^a \mid A=a, X_i) P(X_i) \quad \text{conditional exchangeability}$$
$$= \sum_X E(Y \mid A=a, X_i) P(X_i) \quad \text{consistency}.$$

In the case that there are many X covariates $P(X_i)$ is inconvenient to deal with

$\therefore$ wish to work with $P(A_i \mid X_i) = e(X_i)$ instead

$$P(X_i) = \frac{P(A=a, X_i)}{P(A=a \mid X_i)} = \frac{P(X_i \mid A=a) P(A=a)}{P(A=a \mid X_i)}$$

$$= \sum_X E(Y \mid A=a, X_i) \frac{P(X_i \mid A=a) P(A=a)}{P(A=a \mid X_i)}$$

This is the observed outcome

if we weight $P(X_i \mid A=a)$ by $\frac{P(A=a)}{P(A=a \mid X_i)}$ then we get what we want.

- $P(A=a)$ is observed
- $P(A=a \mid X_i)$ is the result of propensity model.

A categorical example: sample distribution

	X=0	X=1
A=0	4	3
A=1	4	9

weight 1 $\frac{1}{P(A \mid X)}$

w_i	X=0	X=1
A=0	$\frac{1}{p(A=0 \mid X=0)} = \frac{1}{0.5} = 2$	$\frac{1}{p(A=0 \mid X=1)} = \frac{1}{0.25} = 4$
A=1	$\frac{1}{p(A=1 \mid X=0)} = \frac{1}{0.5} = 2$	$\frac{1}{p(A=1 \mid X=1)} = \frac{1}{0.75} = 1.33$

weighted frequency 1

	X=0	X=1
A=0	8	12
A=1	8	12

weight 2 $\frac{P(A)}{P(A \mid X)}$ stabilized weight

w_i^s	X=0	X=1
A=0	$\frac{p(A=0)}{p(A=0 \mid X=0)} = \frac{0.35}{0.5} = 0.7$	$\frac{p(A=0)}{p(A=0 \mid X=1)} = \frac{0.35}{0.25} = 1.4$
A=1	$\frac{p(A=1)}{p(A=1 \mid X=0)} = \frac{0.65}{0.5} = 1.3$	$\frac{p(A=1)}{p(A=1 \mid X=1)} = \frac{0.65}{0.75} = 0.86$

weighted frequency 2

	X=0	X=1
A=0	2.8	4.2
A=1	5.2	7.8

The reason why weighting both $\frac{1}{P(A \mid X)}$ and $\frac{P(A)}{P(A \mid X)}$ works is because frequency / count rather than probability is being weighted here.

Proving that weighting by $\frac{1}{P(A=a|X)}$ gives unbiased estimate

- under $A=1$

$$E(w_i A_i Y_i) = E\left(\frac{A_i Y_i}{e(x_i)}\right) = E\left(\frac{A_i Y_i^1}{e(x_i)}\right) = E_x\left[E\left(\frac{A_i Y_i^1}{e(x_i)} \mid X_i\right)\right]$$

bc constrained to $A=1$; i.e.

$$= E_x\left[\frac{1}{e(x_i)} E(A_i Y_i^1 | X_i)\right] = E_x\left[\frac{1}{e(x_i)} E(Y_i^1 | A_i X_i) E(A_i | X_i)\right]$$

$$= E_x[E(Y_i^1 | A_i X_i)] = E_x[E(Y_i^1 | X_i)] = E(Y_i^1)$$

conditional exchangeability

$\therefore \tau = E(Y^1) - E(Y^0)$

$$= \frac{1}{N}\sum^N w_i A_i Y_i - \frac{1}{N}\sum^N w_i (1-A_i) Y_i$$

bc $\sum^N w_i A_i = \sum^N w_i (1-A_i) = N$

$$= \frac{\sum^N w_i A_i Y_i}{\sum^N w_i A_i} - \frac{\sum^N w_i (1-A_i) Y_i}{\sum^N w_i (1-A_i)}$$

where $A_i = 1$ for treatment group and $A_i = 0$ for control group

$w_i = \frac{1}{P(A_i|X_i)}$ if $A_i = 1$ (treated unit)

$= \frac{1}{1-P(A_i|X_i)}$ if $A_i = 0$ (untreated unit)

Proving that weighting by $\frac{P(A=a)}{P(A=a|X)}$ gives unbiased estimate

$$\frac{\sum^N w_i' A_i Y_i}{\sum^N w_i' A_i} = \frac{\sum^N P(A=a) w_i A_i Y_i}{\sum^N P(A=a) w_i A_i Y_i}$$

⇒ back to the above situation

$\therefore$ also unbiased.

Caveat

$P(A_i = 1 \mid X_i)$ can be really small

→ introduce extreme weights → truncate and cap the weight

→ violate positivity assumption → matching?

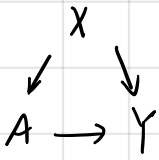
since these people have nothing similiar to the treated group

3. Marginal structure model (MSM)

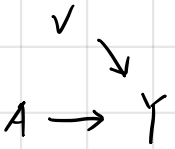
The logic is the same as inverse propensity weighting.
The main idea is to weight to create psedo population and then perform analyses in the psedo-population.
In simple case, MSM = IPSW.
Let's consider two more complex examples to demonstrate MSM as a flexible variant of IPSW.

Example a.

The situation is



The research question is

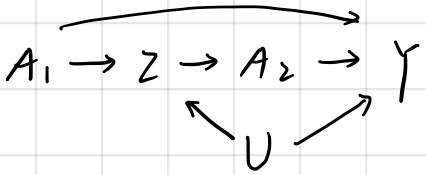


where V is a subset of X.

step 1: Model $P(A|X)$
step 2: Derive weight $w_i = \frac{1}{P(A_i|X_i)}$
step 3: Use the weights to create a psedo population
step 4: In the already balanced psedno-population, we can fit the model $Y \sim A + V$ because this is our research question. logistic regression if Y is binary.
Then the effect of A on odds-ratio of Y is e^{β_A}
(recall how to interpret the coefficient of logistic regression)

Example b.

The situation is

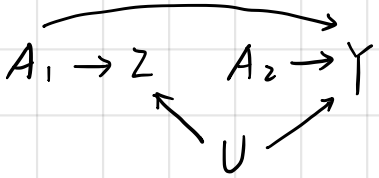


U is unmeasured, but we know it exists

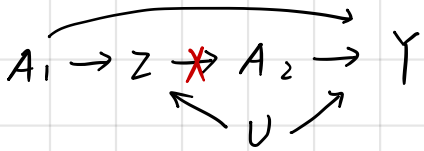
The research question is the causal effect of A_1 and A_2 on Y

Problem is that $A_2 \leftarrow Z \leftarrow U \rightarrow Y$. There is a back door path between A_2 and Y
If we control for Z to try to solve this problem, we run into another problem that $A_1 \rightarrow [Z] \leftarrow U \rightarrow Y$ the back door path between A_1 and Y opens.

∴ The solution is to remove $Z \rightarrow A_2$:
Then no more back door paths



How to remove $Z \to A_2$:



step1 : Propensity modeling

Nothing influence the assignment of A_1 ↘

P_1 : $P(A_1=1|\phi)$ & $P(A_1=0|\phi)$ directly from data

Z influences the assignment of A_2 ↘

P_2 : $P(A_2=1|Z)$ & $P(A_2=0|Z)$ based on propensity.

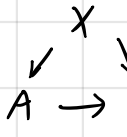
Whether someone gets into the path $Z \to A_2$ in the first place depends on A_1.
(think of P_1 as base weight, weights accumulate)

step2 : derive weights $w_i = \frac{1}{P_1 P_2}$

step3 : create psendo-population

step4 : Estimate model of interest $Y \sim A_1$ & A_2

Points to note :

- Normally, if  , either conditioning on X or removing the relationship between X and A will do. However, this is an example why we must work on removing the relationship.
  ⇒ MSM is flexible in handeling various situations

- To check whether the relationship between Z and A_2 is successfully removed → regress A_2 on Z based on the psendo population
  → if the relationship is non-significant → success

- Can again work with stabalized weights. Recall $w_i = \frac{1}{P(A_i=1|X_i)}$

$$w_i = \underbrace{\frac{P(A_1=1)}{P(A_1=1|\phi)}}_{\text{stabalized } \frac{1}{P_1}} \underbrace{\frac{P(A_2=1)}{P(A_2=1|z_i)}}_{\text{stabalized } \frac{1}{P_2}}$$

$$w_i^s = \frac{P(A_i=1)}{P(A_i=1|X_i)}$$

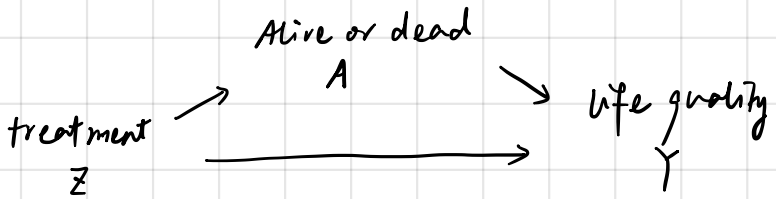
Principle stratification

Before we talked about stratifying on personal characteristics (e.g., smoking) in order to compute causal effect. This is a special type of stratification of a kind of personal characteristics that we know exists, but cannot be directly observed — whether and how people take the assigned treatment

Use 3 examples to demonstrate principle stratification:

Example 1: Truncation due to death

Z — treatment assignment
A — alive or death
Y — quality of life.



Research question is the effect of treatment on life quality.

Problem is that life quality is only defined for the people who are alive. whether someone is alive or dead is, in turn, influenced by treatment.

The key is that there are 4 underlying subpopulations:

[1] $Z=0 \to A=L$ $Z=1 \to A=L$ Always living
[2] $Z=0 \to A=L$ $Z=1 \to A=D$ Induced death
[3] $Z=0 \to A=D$ $Z=1 \to A=L$ Induced living
[4] $Z=0 \to A=D$ $Z=1 \to A=D$ Never living

To make the comparison on life quality, we need the patient to be alive whether he receives the treatment or not: [1] ✓

In scenario [2] for example, the patient is dead if he receives the treatment, Y^1 is not defined → cannot compare Y^0 and Y^1. The logic is the same for scenario [3] and [4].

∴ Research question has to be redefined given this constraint to
— the effect of treatment on life quality among the "always living" group. Average principle causal effect

In reality, we do not know who are the "always living" people and we have to deduce from the data:

What we observe are Z and A

Y_{1L} Y_{0L} Y_{1D} Y_{0D}, each of them is a mixture of 2 subpopulations

$Y_{1L} = \{ Y^1_{DL}, Y^1_{LL} \}$ in the order of A^0 A^1

- Y_{1L}: $Z=1$ Actually live
- Y^1_{DL}: dead if not treated; live if treated
- Y^1_{LL}: live if not treated; live if treated

$Y_{0L} = \{ Y^0_{LD}, Y^0_{LL} \}$ $Y_{1D} = \{ Y^1_{DD}, Y^1_{LD} \}$ $Y_{0D} = \{ Y^0_{DD}, Y^0_{DL} \}$

In a table form

		if Z=0 no treatment	
		A=0 (dead)	A=1 (alive)
		$P_{0\|0} = 0.49$	$P_{1\|0} = 0.51$
if Z=1 treatment	A=0 $P_{0\|1} = 0.35$	DD Never taker $\pi_{DD} = 0.35$	LD defier $\pi_{LD} = 0$
	A=1 $P_{1\|1} = 0.65$	DL complier $\pi_{DL} = 0.49 - 0.35$	LL Always taker $\pi_{LL} = 0.51$

These are what we can observe

These 4 subpopulations are what we wish to infer

To infer π from p, we need assumption

Assumption of monotonicity:

→ individuals are at least as likely to receive treatment when randomized to treatment Z=1, as compared to Z=0

→ no defiers.

Then $\pi_{LD} = 0$; $\pi_{DD} = P_{0|1}$ (not taking the treatment A=0 even if Z=1)

$\pi_{LL} = P_{1|0}$ (taking the treatment A=1 even if Z=0)

$\pi_{DL} = P_{1|1} - P_{1|0} = P_{0|0} - P_{0|1}$

Now we pin down the percentage of the four subpopulations, we can estimate the average principle causal effect

We again make use of the fact that what we can observe is a mixture of different subpopulations.

		if $Z=0$ no treatment	
		$A=D$ $P_{D\|0} = 0.49$	$A=L$ $P_{L\|0} = 0.51$ $Y_{L\|0} = 0.86$
if $Z=1$ treatment	$A=D$ $P_{D\|1} = 0.35$	DD Never taker $\pi_{DD} = 0.35$ ~~Y^1_{DD}~~ ~~Y^0_{DD}~~	LD defier $\pi_{LD} = 0$ ~~Y^1_{LD}~~ ~~Y^0_{LD}~~
	$A=L$ $P_{L\|1} = 0.65$ $Y_{L\|1} = 0.89$	DL complier $\pi_{DL} = 0.49 - 0.35$ Y^1_{DL} ~~Y^0_{DL}~~	LL Always taker $\pi_{LL} = 0.51$ Y^1_{LL} Y^0_{LL}

These are what we can observe

These 4 subpopulations are what we wish to infer

In the specific current example, Y can only be observed for $Y_{L|1}$ and $Y_{L|0}$ – the alive group.

In the current example $Y^1_{DD}, Y^0_{DD}, Y^0_{DL}, Y^1_{LD}$ are not defined.

$Y_{L|0} = \{ Y^0_{LL}, Y^0_{LD} \}$ ∴ $Y_{L|0} = Y^0_{LL}$

↑ do not exist

$$Y_{L|1} = \{ Y^1_{LL}, Y^1_{DL} \} = \frac{0.14}{0.14+0.51} Y^1_{DL} + \frac{0.51}{0.14+0.51} Y^1_{LL} = 0.86$$

with $0 < Y^1_{DL} < 1$ $0 < Y^1_{LL} < 1$

or even $0 < Y^1_{DL} < Y_{L|1} < Y^1_{LL} < 1$ ⇒ we can derive a range for Y^1_{DL} and Y^1_{LL}

Example 2: Principle stratification's principle strata can be used to address specific research question. In the example before, we stratify because the outcome variable Y is only defined for one stratum. But this doesn't have to be the case. Consider this example:

students are randomly assigned to a program (Z)

At the end of first year, some students are on track, others off track (A)

In two years, they graduate (Y)

special program Z → A on track or not → graduation Y

In this example, the outcome variable Y (graduation) is defined for all group. The reason we stratify could be because we are interested in the effect only for compliers.

Example 3: Instrumental variable method.

A special application of principle stratification. It relies on an extra assumption

exclusion restriction: Z must not indirectly influence Y.

All influence of Z on Y go through A

$Z \to A \to Y$

consider this setting:

Z = letter to encourage quit smoking for pregnant women

A = actually quit smoking or not

Y = new born babies' health

Z's assignment is a random experiment: Z's causal effect on A and Y are both legitimate

But the research question is what's A's effect on Y?

Let's break down $A^{Z=1 \text{ or } 0}$:

		if Z = 0	
		A = smoke	A = quit smoking
if Z = 1	A = smoke	Never taker	defier (don't exist)
	A = quit smoking	complier	Always taker

For never taker and always taker, the treatment did not change their behavior, and should have no effect on Y $\quad Z \not\to A \to Y$

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

$$= E_L[E(Y|Z=1, L=l)] - E_L[E(Y|Z=0, L=l)]$$

where L refers to NT, AT or C

$$= \sum_L E(Y|Z=1, L=l) \cdot P(L=l) - \sum_L E(Y|Z=0, L=l) P(L=l)$$

$$= \sum_L [E(Y|Z=1, L=l) - E(Y|Z=0, L=l)] P(L=l)$$

key: For NT and AT, their $E(Y|Z=1, L=\text{NT or AT}) = E(Y|Z=0, L=\text{NT or AT})$ bc Z should have no effect on Y

$$\therefore = P(L=C) \underbrace{[E(Y|Z=1, L=C) - E(Y|Z=0, L=C)]}$$

↑ percentage of complier

complier average causal effect ✓