SURV-740_HW1

Namit Shrivastava

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
library(DiagrammeR)
```

Question 1 (20 points)

Supposed that you have a random sample of size n from the population of interest. Answer the following questions that are designed to help you get familiar with potential outcomes. Try to keep your answers brief and your language precise. Throughout the problem, assume that the Stable Unit Treatment Value Assumption (SUTVA) holds.

a) Contrast the meaning of Y^0 with the meaning of Y

- So, Y^0 is the potential outcome which represents what the outcome would have been if the treatment had not been applied.
- And meanwhile Y is the actual outcome that we observe after treatment has been applied (or not).

b) Contrast the meaning of $E(Y^0)$ with the meaning of $E(Y|A = 0)$

- $E(Y^0)$ is the expected value of potential outcome if everyone in the population had not received the treatment.
- On the other hand $E(Y|A = 0)$ is the expected value of observed outcome only among those who did not receive the treatment.

c) Contrast the meaning of $E(Y^0|A = 1)$ with the meaning of $E(Y^0|A = 0)$

Answer: - Ok so $E(Y^0|A = 1)$ is the expected value of potential outcome if the treatment had not been applied, among those in the treated group. - And then $E(Y^0|A = 0)$ is the expected value of potential outcome if the treatment had not been applied, among those who did not receive treatment.

d) Which of the following quantities can be identified from observed data, assuming SUTVA?

a. $E(Y^0|A=1)$
b. $E(Y^0)$
c. $E(Y|A=0)$
d. $E(Y^0|A=0)$

c. $E(Y|A=0)$ **and d.** $E(Y^0|A=0)$. Because both c and d are equal under SUTVA/consistency assumption which according to the definition states that observed outcome for those who didn't receive treatment is equal to their potential outcome if untreated.

e) Now, further assume that the units in this sample are randomly assigned to treatments, which means the assumption of exchangeability holds. Which of the above quantities can be identified from the observed data?

a)-d) Since exchangeability assumption ensures that the potential outcome for a subject would have the same expectation no matter its assignment of treatment and with both assumption of SUTVA and exchangeability, a-d are all identifiable and equal as: a = d = $E(Y|A=0)$, which is equal to (b).

Question 2 (10 points)

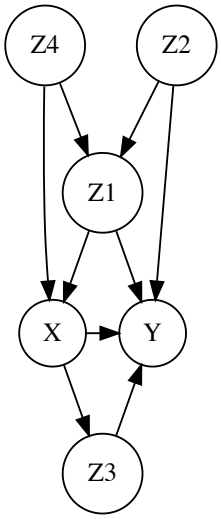
Under the complete randomization of treatment assignment, the average treatment effect is identified because the randomization guarantees statistical independence between the treatment indicator A and the observed outcome Y.

a) True
b) **False** :: Here I think it should be independence between the treatment indicator A and all potential outcomes, not observed outcome Y.

Justification Now, the statement is false because under complete randomization, one needs independence between the treatment indicator A and all potential outcomes (Y^0, Y^1), not just the observed outcome Y. Hence, this independence of potential outcomes from treatment assignment is what allows one to identify causal effects.

Question 3 (20 points)

Consider the following DAG:


Z4
Z2
Z1
X
Y
Z3

a) Enumerate all paths from X to Y

Answer: - X $\rightarrow$ Y - X $\rightarrow$ Z $\rightarrow$ Y - X $\leftarrow$ Z $\rightarrow$ Y - X $\leftarrow$ Z $\rightarrow$ Z $\rightarrow$ Y - X $\leftarrow$ Z $\leftarrow$ Z $\rightarrow$ Y - X $\leftarrow$ Z $\rightarrow$ Z $\leftarrow$ Z $\rightarrow$ Y

b) In the path $X \leftarrow Z_4 \rightarrow Z_1 \leftarrow Z_2 \rightarrow Y$, what type of node is Z_1? Does conditioning on z1 block or unblock this path from X to Y?

So here, Z_1 is a Collider since conditioning on Z_1 will unblock this path from X to Y.

c) In the path $X \leftarrow Z_1 \rightarrow Y$. In this path, what type of node is Z_1? Does conditioning on Z_1 here block or unblock this path?

Here, Z_1 is a Confounder as conditioning on Z_1 blocks the path.

d) Pearl's back-door criterion

i. Is $X \rightarrow Z_3 \rightarrow Y$ a back door path from X to Y? Why?

Umm I think no, because it does not begin with a directed edge which points to the first variable X.

ii. Based on your DAG, enumerate the minimum conditioning sets that satisfy the back door criterion for identifying the effect of X on Y?

Answer: {Z and Z } or {Z and Z }

Question 4

Consider a randomized experiment with four observations, of which two units were randomly assigned to treatment via complete randomization. We use $A_i \in \{0, 1\}$ and Y_i to denote the treatment (1 for treatment and 0 for control) and the observed outcome for unit i, respectively.

i	Y	A	Y^1	Y^0	τ

a) (4 points) Fill in the table

i	Y	A	Y^1	Y^0	τ
1	2	1	2	?	?
2	0	0	?	0	?
3	1	0	?	1	?
4	3	1	3	?	?

So here, the ? values represent **unobservable counterfactual outcomes**. In reality: - For units 1 & 4 (treated): One can observe Y^1 but cannot observe Y^0 - For units 2 & 3 (control): One can observe Y^0 but cannot observe Y^1 - Therefore, the individual treatment effects $\tau_i = Y_i^1 - Y_i^0$ are also unobservable

Hence, this is the fundamental problem of causal inference as one can never observe both potential outcomes for the same unit at the same time.

b) (6 points) Define the population average treatment effect for the treated (ATT)

Definition:

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

So, by consistency: $= E(Y|A = 1) - E(Y^0|A = 1)$

Now by exchangeability due to randomization: $= E(Y|A = 1) - E(Y^0|A = 0)$

And by consistency: $= E(Y|A = 1) - E(Y|A = 0)$

Estimator:

$$\hat{ATT} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i$$

Estimation:

$$\hat{ATT} = \frac{2+3}{2} - \frac{0+1}{2} = 2.5 - 0.5 = 2$$

Question 5 (10 points)

Researchers are studying the effect of education level on income. They believe higher education leads to better job opportunities, which in turn increases income. Additionally, they suspect that family background influences both education and income.

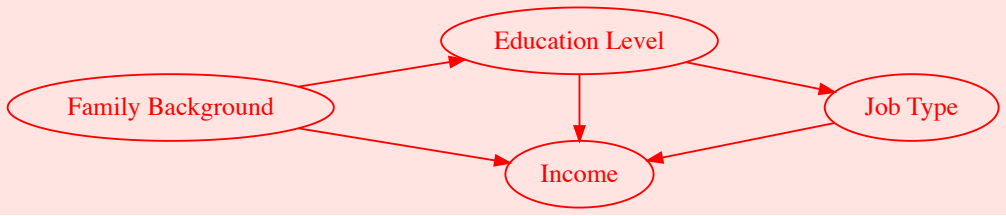
a) Draw a DAG that includes the above variables

```
DiagrammeR::grViz("
digraph {
  graph [layout = dot, rankdir = LR, bgcolor = 'mistyrose', nodesep = 0.5, ranksep = 0.8]
  node  [shape = ellipse, color = 'red', fontcolor = 'red']
  edge  [color = 'red', arrowsize = 0.8]

  // nodes
  F [label = 'Family Background']
  E [label = 'Education Level']
  J [label = 'Job Type']
  I [label = 'Income']

  // rank constraints
  { rank = min;  F }          // leftmost tier
  { rank = same; E; I }       // align Education and Income

  // edges
  F -> E
  F -> I
  E -> J
  J -> I
  E -> I
}
")
```


Education Level
Family Background
Job Type
Income

b) Is Job Type a mediator or confounder in this DAG?

Here Job Type is a mediator since it explains the relationship between an independent variable (education level) and a dependent variable (income). Simply if I see, Job Type is a mediator because the independent variable affects the mediator, which in turn affects the dependent variable.

c) How would you use the DAG to assess the causal effect of Education Level on Income?

Now, Family Background is identified as a confounder as to assess the causal effect accurately, one needs to control for this confounder in the analysis.

Question 6 (10 points)

You are tasked with investigating the relationship between air pollution and asthma in children. Consider the following: - Parental Smoking increases the likelihood of both air pollution exposure and asthma in children. - Living in an urban area increases exposure to air pollution but is not directly related to asthma.

a) Draw a DAG including the above variables

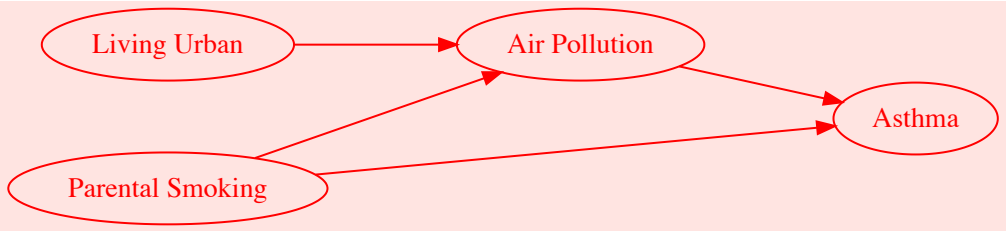
```
DiagrammeR::grViz("
digraph {
  graph [layout = dot, rankdir = LR, bgcolor = 'mistyrose', nodesep = 0.5, ranksep = 0.9]
  node  [shape = ellipse, color = 'red', fontcolor = 'red']
  edge  [color = 'red', arrowsize = 0.9]

  // nodes
  UL [label = 'Living Urban']
  PS [label = 'Parental Smoking']
  AP [label = 'Air Pollution']
  A  [label = 'Asthma']

  // rank constraints (tiers)
  { rank = min;  UL; PS }  // leftmost sources
  { rank = max;  A }       // rightmost outcome

  // edges
  UL -> AP
```

```
  PS -> AP
  AP -> A
  PS -> A
}
")
```


Living Urban
Air Pollution
Asthma
Parental Smoking

b) What are the confounders you should adjust for?

Here, Parental smoking is a confounder because it is associated with both: - An increased likelihood of air pollution exposure (e.g., secondhand smoke contributing to indoor air pollution). - And a higher incidence of asthma in children (due to direct exposure to tobacco smoke).

c) Should you adjust for Urban Living? Explain why or why not based on your DAG.

Well I would not adjust for Urban Living, as it increases exposure to air pollution but does not have a direct causal relationship with asthma. Thus, it is not a confounder in the context of the DAG I am considering.