In-Class Exercise 9

For this in-class exercise, work with your group of 2-3 people, to answer the following questions. These questions are not necessarily easy and sometimes they will not have a clear "correct" answer. The goal is to get you thinking about the material we've learned. Some of these questions may require you to discuss and debate with your group members to come up with an answer or can cover topics that we have not yet covered in class.

Be prepared to share your answers with the class and add to the discussion.

After class submit your a do-file with your answers in comments to Moodle for grading. You will be graded as a group on your submission. Only one group member needs to submit the assignment, but make sure add all group member names.

Problem 1

The government has launched a free, voluntary 12-week coding bootcamp for unemployed workers. You have data on two groups of people:

- Participants ($D = 1$): Those who chose to enroll and completed the bootcamp.
- Non-Participants ($D = 0$): Unemployed workers who chose not to enroll.

You observe that Participants have significantly higher post-program wages (Y) than Non-Participants.

1. For a specific worker i, define Y_i^1 and Y_i^0 in plain English.
2. Using the "Switching Equation" $Y_i = D_i Y_i^1 + (1 - D_i)Y_i^0$, explain why we can never calculate the individual treatment effect $\delta_i = Y_i^1 - Y_i^0$ for any specific worker.

! Solution

1. **Definitions:**
   - Y_i^1: The wage worker i would earn **if** they attended the bootcamp.
   - Y_i^0: The wage worker i would earn **if** they did **not** attend the bootcamp.
2. **Fundamental Problem of Causal Inference:**
   - The switching equation says we only observe Y_i^1 when $D_i = 1$ and Y_i^0 when $D_i = 0$.
   - We never observe both potential outcomes for the same person at the same time. Therefore, we cannot calculate the difference $Y_i^1 - Y_i^0$ directly (the missing data problem).

Problem 2

In order for a simple comparison of means ($Mean_{treated} - Mean_{untreated}$) to represent the true causal effect, the "Independence Assumption" what must hold?

Explain what this mathematical statement means in the context of the coding bootcamp. Does it hold here? Why or why not? (Hint: Think about who signs up for a voluntary coding class)

! Solution

- **Independence Assumption:** $\{Y_i^1, Y_i^0\} \perp D_i$ (Potential outcomes are independent of treatment assignment).
- **In English:** This means that the people who *chose* the bootcamp are, on average, comparable to those who didn't. Specifically, their baseline potential wages (Y^0) are the same.
- **Does it hold? No.**
- **Why?** Selection bias. Those who sign up for a voluntary coding bootcamp are likely more motivated, "tech-savvy," or have higher aptitude to begin with. Thus, their Y^0 (wages without the bootcamp) would likely have been higher than the non-participants' Y^0 anyway.

Problem 3

Suppose we have the following "God's Eye View" data (where we can see potential outcomes) for 4 workers:

Worker	Type	Bootcamp (D)	Wage if Bootcamp (Y1)	Wage if No Bootcamp (Y0)
A	"Tech Savvy"	1	\$100k	\$80k
B	"Tech Savvy"	1	\$90k	\$75k
C	"Not Interested"	0	\$60k	\$50k
D	"Not Interested"	0	\$50k	\$40k

1. Calculate the True Average Treatment Effect (ATE) for the population.
2. Calculate the Observed Simple Difference in Means ($E[Y \mid D = 1] - E[Y \mid D = 0]$)
3. Using the decomposition formula from the slides, calculate the Selection Bias term: $E[Y^0 \mid D = 1] - E[Y^0 \mid D = 0]$. Explain what this specific number represents in the real world.

! Solution

1. **True ATE ($E[Y^1 - Y^0]$):**
   - Individual effects ($Y^1 - Y^0$):
     - A: $100 - 80 = 20$
     - B: $90 - 75 = 15$
     - C: $60 - 50 = 10$
     - D: $50 - 40 = 10$
   - ATE = $(20 + 15 + 10 + 10)/4 = 13.75k$.
2. **Observed Simple Difference in Outcomes (SDO):**
   - Mean Treated ($D = 1$): $(100 + 90)/2 = 95k$
   - Mean Untreated ($D = 0$): $(50 + 40)/2 = 45k$
   - SDO = $95 - 45 = 50k$.
3. **Selection Bias ($E[Y^0 \mid D = 1] - E[Y^0 \mid D = 0]$):**
   - $E[Y^0 \mid D = 1]$ (What participants would have earned without bootcamp): $(80 + 75)/2 = 77.5k$.
   - $E[Y^0 \mid D = 0]$ (What non-participants earned without bootcamp): $(50 + 40)/2 = 45k$.
   - **Selection Bias** = $77.5 - 45 = 32.5k$.
   - **Meaning:** Even without the bootcamp, the "Tech Savvy" group would have earned $32.5k more than the others purely due to their pre-existing characteristics/motivation.

Problem 4

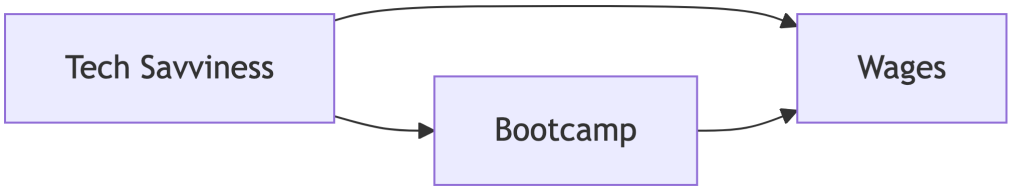
Based on your answer to Part 2 and 3, draw a DAG representing the relationship between:

- D: Bootcamp Attendance
- Y: Future Wages
- U: "Tech Savviness" (Unobserved Confounder)

Identify the Backdoor Path that creates the bias. Is this path open or closed?

! Solution

DAG:



- **Structure:** $U \to D$ (Savvy people join) and $U \to Y$ (Savvy people earn more).
- **Backdoor Path:** $D \leftarrow U \to Y$.
- **Status:** This path is **OPEN**. This creates a spurious correlation between Bootcamp and Wages that isn't causal.

Problem 5

The government is so impressed by the results that they decide to make the bootcamp mandatory for all unemployed workers in the country. Based on the "General Equilibrium" and "SUTVA" slides, why might the treatment effect you calculated in Part 3 ($N = 4$) not apply when the program is scaled up to millions of people? (Hint: What happens to the supply of coders and their wages?)

! Solution

- **SUTVA Violation:** Stable Unit Treatment Value Assumption requires that one person's treatment doesn't affect another's outcome.
- **General Equilibrium Effects:** Scaling the program up increases the labor supply of entry-level coders significantly.
- **Result:** If the supply of coders skyrockets, the equilibrium wage for coders will likely **decrease**. The "partial equilibrium" effect we calculated for 4 people ($13.75k$) assumed market wages stayed constant. The "general equilibrium" effect would likely be much smaller.

Problem 6

A researcher wants to "fix" the selection bias by controlling for more variables. They decide to run a regression controlling for "Job Title at End of Year" (e.g., Software Engineer vs. Barista).

Draw a new DAG including "Job Title" (J). Note that Bootcamp affects Job Title ($D \to J$) and Job Title affects Wages ($J \to Y$).

Why is "Job Title" a "Bad Control"? Explain how controlling for it blocks the very mechanism you are trying to study (mediation) or induces collider bias.

! Solution

DAG:

- **Bad Control (Mechanism/Mediator):**
  - "Job Title" is a **mediator** ($D \to J \to Y$). The whole point of the bootcamp is to help you get a better job title (Software Engineer), which then pays more.
  - If you **control for Job Title** (hold J constant), you are asking: "What is the effect of bootcamp on wages for people who *have the same job*?"
  - If you compare a Bootcamp Software Engineer to a Non-Bootcamp Software Engineer, the effect might be zero. But that misses the fact that the bootcamp *made* them a Software Engineer!
  - By controlling for the mechanism, you "block" the causal path, biasing your estimate of the total effect toward zero.