P-Set III

POLI 784 - Spring 2026

2026-02-19

Complete all of the questions below together with your group members. Please submit both (1) a compiled PDF or HTML file with your responses and (2) the `R Markdown`, `Quarto`, etc. file containing both your responses and code as a group. You are encouraged to use `R Studio` projects and virtual environments to complete this p-set, although it is not required. Similarly, you are encouraged to use consistent style and linters when writing code.

Question 1

a. *(2 points)* Simulate the following data generating process in `R`, save the results to a dataframe called `df`, and create a directed acyclic graph (DAG) representing the network of variables simulated. Make sure to represent $U1$ as an *unobserved* variable. Describe the role (e.g., outcome, treatment, confounder, etc.) that each variable plays in the theory that D causes Y, which you are interested in testing empirically. *(Note: You may use any package / language of your choice to create the DAG. Moreover, using DAGitty may prove helpful in describing the roles of variables and understanding their relationships.)*

1. set seed to `42` to make your results reproducible
2. set sample size to $N = 1000$
3. simulate $U1 \sim Bernoulli(0.5)$
4. simulate $X1 = 1 + 0.5 * U1 + \varepsilon_1$, where $\varepsilon_1 \sim N(0, 1)$
5. simulate $X2 \sim Bernoulli(0.5)$
6. simulate $D = 1.25 + 0.5 * X1 + 0.25 * X2 + \varepsilon_2$, where $\varepsilon_2 \sim N(0, 0.1)$
7. simulate $X3 = 1 + 0.5 * D + \varepsilon_3$, where $\varepsilon_3 \sim N(0, 0.1)$
8. simulate $Y = 5 + 1 * D + 1 * X3 + 0.1 * X2 + 0.1 * U1 + \varepsilon_4$, where $\varepsilon_4 \sim N(0, 0.25)$
9. simulate $X4 = -0.5 + 2 * D + 0.75 * Y + \varepsilon_5$, where $\varepsilon_5 \sim N(0, 1)$

b. *(1 point)* Let's start by estimating the *direct* causal effect of D on Y. Explain which nodes have to be included in the conditioning set C (i.e., which variables do we need to adjust for) to block all non-causal and indirect causal paths between D and Y and, thus, for us to be able

to obtain an unbiased estimate of the *direct* effect. Fit an OLS model using `lm()` based on your explanation. Does the model recover an unbiased estimate of the direct effect of D on Y?

c. *(1 point)* Now let's estimate the *total* causal effect of D on Y. Explain which nodes have to be included in the conditioning set C for us to be able to obtain an unbiased estimate of the *total* effect. Fit an OLS model using `lm()` based on your explanation. Does the model recover an unbiased estimate of the total effect of D on Y?

d. *(1 point)* Fit a new OLS model using `lm()` that is identical to the one you fit in *1c* except that it additionally adjusts for $X4$. What changes about the coefficient on D? Why is that and what is this phenomenon called?

e. *(1 point)* Create a new DAG by adapting the DAG you created in *1a* by (1) removing the direct effect of D on Y, (2) removing $X1$ from the graph, and (3) adding a direct effect of $U1$ on D. Which nodes would the conditioning set C now have to consist of for us to be able to obtain an unbiased estimate of the effect of D on Y? Explain the logic of the approach. *(Note: Do not fit a new OLS model here. Simply provide your response based on your understanding of the graph.)*

f. *(1 point)* Create a new DAG by adding a new unobserved variable $U2$, which directly causes both $X2$ and Y, to the DAG you created in *1e*. How would the conditioning set C change from that in *1e*, and could we still estimate the causal effect of D on Y given the new graph? Assuming that the new DAG accurately captures the true data generating process, would the DAG from *1e* be considered a causal DAG or not and why? Which assumption did you draw on to make this determination?

g. *(1 point)* Based on your results from *1a* to *1f*, which types of variables should we *always* adjust for, which should we *never* adjust for, and which *may* we want to adjust for (and under what circumstances) when estimating causal effects using observational data? Based on what knowledge do we decide to hold a variable constant or not and how can we go about achieving the correct selection of variables in practice?

Question 2

In this exercise, we will work with data from Blattman and Annan (2010). The paper investigates the consequences of being abducted into a rebel organization as a youth in Northern Uganda. The authors argue that conditional on personal characteristics, whether an individual was abducted ($D_i = 1$) was a random event. The outcome of interest is years of education for each individual. You can find the data needed to solve this exercise in `data.csv`.

a. *(0.5 points)* Given the above, is this an experimental or an observational study and why? What is the identification assumption that allows for the estimation of causal estimands (e.g., ATT) based on this design? Why does the identification assumption, if met, allow us to interpret estimates based on this design as causal quantities? What must be true about the

set of covariates $\mathbf{X}$ employed in the matching procedure for the identification assumption to be plausible?

b. *(0.5 points)* Examine covariate balance in the unadjusted data by conducting a balance test and printing the covariate means in the treatment conditions, the standardized differences, and the t-statistic p-values to console. Create a plot of covariate balance in the unadjusted data using `love.plot()`. Explain what the results tell us about the plausibility of the identification assumption before matching.

c. *(0.5 points)* Conduct nearest-neighbor matching with one neighbor per treated unit and report covariate balance after matching using the results of a balance test and a plot. Has matching increased the plausibility of the identification assumption? What issues in assessing covariate balance and, by extension, the plausibility of the identification assumption, do we face when using balance tests and plots based on low-dimensional summaries?

d. *(0.5 points)* Conduct nearest-neighbor matching with five neighbors per treated unit and report covariate balance after matching using the results of a balance test and a plot. Estimate the ATT using the matched data, print the results to console, and interpret the ATT estimate.