Causal Inference - Homework 1

Introduction to Causal Inference 097400
Winter 2023

Submission date: 7 December, 2022

Question 1

Using potential outcomes notation, provide the following:

I. give an example of a data generating process (DGP, i.e. a joint distribution) which include a binary covariate X, a binary treatment T and two potential outcomes Y_0 and Y_1.

II. For the DGP you created, calculate the expected potential outcomes $\mathbb{E}[Y_t]$ for $t \in \{0, 1\}$ and the resulting ATE.

III. Let $Y = TY_1 + (1 - T)Y_0$. Give an example of another DGP with the same variables as in I., which has the same distribution on the observables $p(X, T, Y)$ as in I., but has a different ATE.

IV. Explain in your own words the meaning of this difference.

For better understanding, we encourage you to simulate data according to your suggested DGPs. There is no need to submit the simulation results.

Question 2

Let $(t_1, y_1), \ldots, (t_n, y_n)$ be a sample from a randomized controlled trial (RCT), where for each $i = 1, \ldots, n$, $t_i \in \{0, 1\}$ is a binary treatment, and $y_i \in \mathbb{R}$ is an outcome measured after the treatment. We define $ATT := \mathbb{E}[Y_1 - Y_0 \mid T = 1]$ to be the Average Treatment Effect on the Treated population (ATT). Using potential outcome notation, prove that ATT can be estimated from data gathered as the sample above. Please be clear about which assumptions you

have used.

Hint: Think about the relationship between ATT and ATE in this settings.

Question 3

Multiple treatments: Let $(X_1, t_1, y_1), \ldots, (X_n, t_n, y_n)$ be a sample from an observational study, where for each $i = 1, \ldots, n$, $t_i \in \{0, \ldots, T\}$ is a treatment, $y_i \in \mathbb{R}$ is an outcome measured after the treatment, and $X_i \in \mathbb{R}^d$ is a vector of d confounders. Assume X represents all the confounders, i.e. $\{Y_0, \ldots, Y_T\} \perp\!\!\!\perp T|X$.

Given there are T possible treatments (where $|T|$ is finite and greater than 2) suggest an estimand which can replace the ATE. For the estimand you have suggested give:

I. An IPW estimator

II. A matching estimator

Question 4

Give an example of a real-world dataset with features X and one or more observed outcome variables Y. For this dataset give:

1. Two examples of interesting causal questions relating one of the features and one of the outcomes. Explain what would be the treatment and what would be the *potential* outcomes in this case.

2. Two examples of interesting prediction questions which do not require causal reasoning.

Examples can come from the fields of politics, biology, sports, economics, entertainment, medicine, transportation and so on - use your imagination. You might find the article " *A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks*" by Miguel Hernán (available on the course Moodle) to be helpful here, though please do not use the exact same examples.