METHODS FOR CAUSAL INFERENCE: TUTORIAL 2

* = For formative assignment (marked but not evaluated)

1. In Lecture 4, we used a dataframe consisting of 1000 individuals, together with values for their observed treatments and outcomes as well as the corresponding counterfactual values. We showed

$$\mathbb{E}[Y_1 - Y_0] \neq \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] \tag{1}$$

and discussed the reason behind this inequality, *i.e.* , a confounding variable. Considering the usual confounding diagram, perform a simulation (in Python or R), to generate a similar table and show the above inequality.

Hint: You may choose $W \sim \mathcal{N}(\mu, \sigma)$. To obtain T, you may add a gaussian noise to W and then binarise using a cutoff. Then generate Y as a linear function of W and T, together with a guassian noise.

2. For the example above, use regression adjustment (to adjust for the confounder W) and obtain an estimate of the causal effect $\mathbb{E}_W[\mathbb{E}_Y[Y|T=1,W] - \mathbb{E}_Y[Y|T=0,W]]$. Compare your answer to the results obtained in Question 1.

3. * Explain the issue of positivity violation (no overlap) when performing extrapolation with no support in each of the following cases and discuss why positivity violation will lead to incorrect causal inference.

FIG. 1: Three scenarios with positivity violation

4. Recall the adjustment formula:

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}_X\left[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\right]. \tag{2}$$

By expanding out as a sum both the inner expectation over X and the outer expectation over Y on the right hand side, followed by the product rule, show that violation of positivity *i.e.* , $p(T=1|X=x)=0$ leads to an inestimable ATE.

5. Prove the following statements:

   (a) For a binary variable X show that $p[X=1|Y] = \mathbb{E}[X|Y]$.

   (b) Law of iterated expectations, $\mathbb{E}_Z[Z] = \mathbb{E}_W[\mathbb{E}_{Z|W}[Z|W]]$.

   (c) Law of iterated expectations (conditional), $\mathbb{E}[Z|W] = \mathbb{E}_{V|W}\Big[\mathbb{E}[Z|W,V]\Big|W\Big]$.

6. *

   (a) Show that in a multiple regression $Y = \alpha_0 + \beta_1 X_1 + \beta_2 X_2$, where X_1 and X_2 are **independent** of each other, one can obtain a consistent estimate for β_1 without needing to fit for β_2. In other words, fitting $Y = \alpha_0 + \beta_1 X_1$ is sufficient for obtaining β_1. To this by minimising the squared loss function.

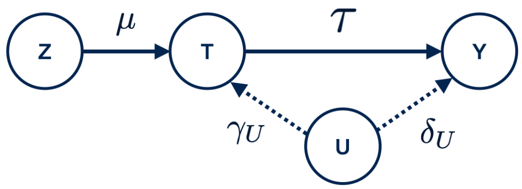

FIG. 2: Instrumental variable graph

(b) Use the above fact, and Structural Causal Equations for the following instrumental variable graph, to show that the least-squares method is justified for obtaining the causal effect τ of T on Y.

(c) Perform a ground truth simulation based on the above SCM, fixing the values μ, τ, γ_U and δ_U. Then use the simulated data (blind to variable U), to show the 2-step least-squares procedure indeed results in the correct estimate of the causal effect τ.