Problems

Part 1: Causal research questions

Which of the following are causal research questions? Think of what (if any) decision problems underlie the question and how a target trial might look.
When your answer is no, think about if/how you can turn the question into a causal one.

a) What is the probability that a new bank customer (with given age, gender, profession, post code, credit history) will default on a loan?

b) Should Type-2 diabetes patients, who started a first-line treatment but exhibit low glycaemic control after one year, begin with second-line treatment or defer the latter in view of the risk of severe side effects?

Part 2: Causal Graphs

An epidemiologist is interested in the effect of a drug (A) on the risk of a heart attack, Y. The drug works partly directly, and partly by an indirect effect of it acting as a muscle relaxant (M), which in turn affects Y. Its muscle relaxing properties may also have some side effects, S.

The drug may be recommended by a doctor (Z), which increases the chance of it being taken. The propensity of the patient to take the treatment is also a function of their age D, which will in turn affect the muscular composition of their heart (C). This latter aspect will influence the muscle relaxation M.

In addition both the likelihood of taking the treatment and of having a heart attack are dependent on the patient's sex (G), and Y is also influenced by the patient's weight (W).

a) Draw a causal diagram that minimally represents the story told above.

b) Consider the following independences. For each one, say whether or not it holds under the graph you have drawn. If it does hold, state which Markov property you use to deduce the independence; if it does not, give an open path between the variables given those in the conditioning set.

   (i) $D \perp\!\!\!\perp G$;

   (ii) $D \perp\!\!\!\perp Z \mid A$;

   (iii) $Z \perp\!\!\!\perp Y \mid A, G, D$;

   (iv) $C \perp\!\!\!\perp A \mid D$.

c) Suppose that we consider only the subset of patients who are known to have suffered from the side-effects of the muscle relaxation effect (i.e. $S = 1$). How does this change your answers for (b)?

d) In the graph from (a), how many distinct back-door adjustment sets are there for the causal effect of A on Y? Which of these will give the most efficient estimator? *[Hint: Note that you do not need to list every single set explicitly, merely describe how they could be constructed.]*

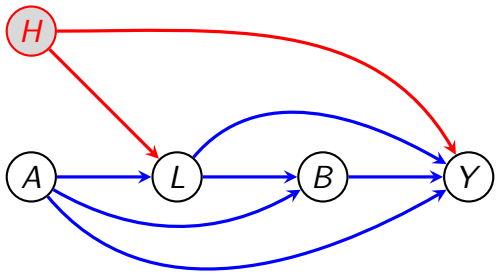
e) Suppose that C is unobserved. Argue that the optimal adjustment set to use now is $\{G, D, W\}$. *[You may use any standard results, provided that you state them clearly.]*

Part 3: Assumptions for estimation

Let $\mathbb{E}[Y(1) - Y(0)|A = 1]$ be the effect of treatment on the treated (ETT).
Assume that X is sufficient to adjust for confounding. Derive the modified adjustment formula for the ETT and comment on how you can weaken the positivity assumption.

Part 4: Sequential Treatment and the g-formula

(Consider the sequential treatment DAG $\mathcal{G}$ shown below.



The variable H is unobserved, A and B represent treatments, and L and Y an intermediate and final outcome respectively.

a) Form the SWIG $\mathcal{G}[a, b]$.

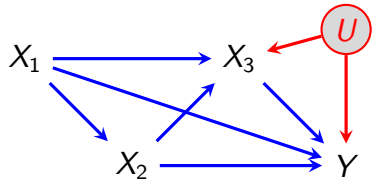
b) Using d-separation, show that $A \perp\!\!\!\perp L(a)$ under distributions Markov with respect to $\mathcal{G}[a, b]$. Then, via consistency and the independence provide a formula to compute $P(L(a) = \ell)$ using only $P(A = a, L = \ell)$.

c) Show that $Y(a, b)$ is d-separated from $\{A, B(a)\}$ given $L(a)$ in $\mathcal{G}[a, b]$.

d) Use the fact proved in c) to find a simple identifying expression for $P(Y(a, b) = y \mid L(a) = \ell)$ in terms of a conditional probability that can be computed from the observed distribution $P(A = a, L = \ell, B = b, Y = y)$.

e) Use your answers to b) and d) to derive an identifying expression for $P(L(a) = \ell, Y(a, b) = y)$, and hence obtain one for $P(Y(a, b) = y)$. *[Hint: don't overthink this!]*

Part 4: Causal DAGs and multiple regression



Assume the simplistic causal DAG above is correctly specified, where Y = (a measure of) infant health, X_3 = birth weight, X_2 = maternal smoking during pregnancy, X_1 = maternal education, U = unmeasured genetic predisposition.
Further consider the following (linear) regressions and assume for simplicity these models are correctly specified. Explain whether the coefficients of X_1, X_2, and/or X_3 have an interpretation as a causal effect, and if so state what type of effect it is.

a) Regress Y on X_1.

b) Regress Y on X_2.

c) Regress Y on X_3.

d) Regress Y on X_1 and X_2 jointly.

e) Regress Y on X_1 and X_3 jointly.

f) Regress Y on X_2 and X_3 jointly.

g) Regress Y on X_1, X_2 and X_3 jointly.