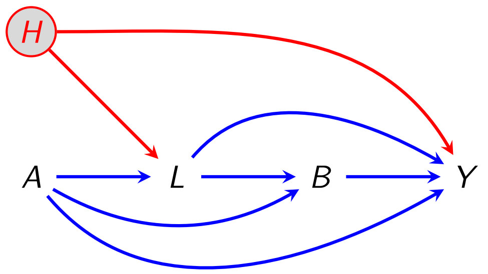
Post-Module Exercises

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 $\mathcal{G}[a, b]$.

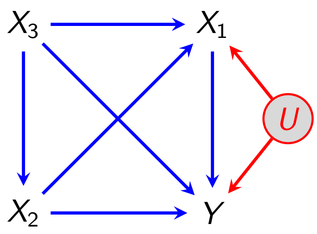
(b) Using d-separation, show that $A \perp\!\!\!\perp L(a)$ under distributions Markov with respect to $\mathcal{G}[a, b]$. Then, via consistency and the independence provide a formula to compute $P(L(a) = \ell)$ using only $P(A = a, L = \ell)$.

(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_1 = birth weight, X_2 = maternal smoking during pregnancy, X_3 = 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.

(h) What other sensible analyses might you suggest?

Part 5: Instrumental Variables

Consider the standard IV set-up with instrument G, exposure X, outcome Y, unobserved confounder U, and assume that the IV conditions are satisfied.

(a) Assume all observable variables G, X, Y are binary.

  (i) Use a SWIG to show that $Y(x) \perp\!\!\!\perp G$.

  (ii) Show that $E(Y(1) - Y(0)|X = 1, G = g) = \psi$ is equivalent to

$$E(Y|X = x, G = g) - E(Y(0)|X = x, G = g) = \psi x.$$

  (iii) Use (i) and (ii) to show that

$$\psi = \frac{E(Y|G = 1) - E(Y|G = 0)}{E(X|G = 1) - E(X|G = 0)}.$$

  Trick: take expectation over X given $G = g$.

(b) Now, for continuous Y, assuming

$$E(Y|X = x, U = u) = \mu_Y + \beta x + h(u),$$

show that

$$\beta = \frac{Cov(Y, G)}{Cov(X, G)}.$$

State clearly what IV assumptions you use.
Trick: define $\tilde{G} = G - E(G)$ and work out $E(Y\tilde{G})$.

(c) Typical data, where IVs might be useful, are obtained from case-control studies: this means that 50% of the observations were sampled from known 'cases' $Y = 1$ and the other 50% from known 'controls' $Y = 0$.

  (i) Draw a DAG that includes a sampling indicator S to represent this situation.

  (ii) Give arguments for or against the validity of IV-based inference regarding (I) testing the null-hypothesis of no $X \to Y$ edge; (II) estimating the causal effect of X on Y using G with a standard IV-method.