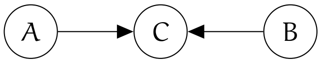
COMP3224: Homework 1

1 Exercises

1. Implement a collider 

to demonstrate how selection bias can induce correlations between uncorrelated variables. There are a number of people auditioning to be actors. Let A be the acting talent of a candidate and B be their perceived attractiveness. The choice of cast C is taken to be an aggregation of scores for A and B and only those above a threshold are selected.

   Working with centred (zero mean) variables, choose A and B to be independent Gaussian variables $\eta_A, \eta_B \sim \mathcal{N}(0, 1)$ and C their sum:

$$\begin{aligned} a &= \eta_A, \\ b &= \eta_B, \\ c &= a + b + \eta_C. \end{aligned}$$

   Draw scatterplots of $(A, B) = (a_n, b_n)$, $n = 1, \ldots, N$ and fit a linear regression line for A as a function of B. What does the regression tell you about how A and B are related?

   Set a positive threshold θ for $C = c_n$ and select the points (a_n, b_n) that lie above the threshold θ: $\mathcal{D} := \{(a_n, b_n) | c_n > \theta, n = 1, \ldots, N\}$. Display the scatterplot for $\mathcal{D}$ and do a linear regression for B against A for the selected data $\mathcal{D}$. What does this regression line say about the relation between acting talent and attractiveness?

2. Generate samples from a conditional probability distribution for binary variables A and B. Let $P(A = 0)$, $P(B = 0|A = 0)$ and $P(B = 0|A = 1)$ to be any numbers between 0 and 1 of your choice. For some choice of $P(A = 0)$, say pA, `numpy.random.binomial(1, pA, n)` will generate a string of n binary digits. (*Hint:* Experiment with the probability pA and n to figure out the conventions used.) Let $pB = P(B = 0|A)$ be set as:

$$pB = (1 - A)P(B = 0|A = 0) + (A)P(B = 0|A = 1).$$

   To generate n samples of B, call `np.random.binomial(1, pB, n)`. Verify that the n samples of A and B are consistent with the defining probabilities in the simulation.

3. Looking ahead: think about how you would generate samples from the network in Study question 1.5.2 of *Causal Inference in Statistics: A Primer* by Pearl, Glymour and Jewell? You will need to extend the method in the previous exercise to define $\mathsf{P}(\mathsf{C} = \mathsf{0}|\mathsf{A}, \mathsf{B})$. How would you check that the simulation was correct?

2 Submission

Submit a 2-page document. For each exercise provide a brief motivation for the task, in the context of questions you might encounter in causal inference from data. Comment on what general lessons you can take away from the evidence you present. You get marks for clarity of presentation using the evidence you have gathered. **Total marks** 15.

Note: For exercise 3, you have to clearly explain, but do not have to implement, your procedure to generate the data for this coursework. That will be required in a later coursework.