UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11207 METHODS FOR CAUSAL INFERENCE

Tuesday 24th May 2022

13:00 to 15:00

INSTRUCTIONS TO CANDIDATES

1. **Note that ALL QUESTIONS ARE COMPULSORY.**
2. **DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.**
3. **This is an OPEN BOOK examination.**

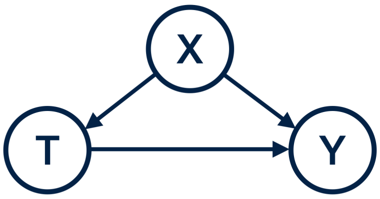
MSc Courses

Convener: A. Pieris
External Examiners: A. Cali, V. Gutierrez Basulto

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. A regional government has set up a financial scheme aimed at incentivising its citizens to have solar panels installed. However, not everyone is equally likely to receive financial aid. If a certain covariate takes value $X = 1$, a citizen has probability q_1 to receive the financial incentive ($T = 1$). If the covariate takes values $X = 0$, the probability of receiving financial aid drops to q_0. The probability to have $X = 1$ is equal to r.

   Similarly, the probability of installing solar panels without financial aid and $X = 0$ is equal to p_0, but increases to p_1 when $X = 1$. The outcome variable Y records if a person installs solar panels ($Y = 1$) or not.



   (a) Give the factorisation of the joint probability distribution $p(y, t, x)$ of (Y, T, X) implied by the causal diagram. Write down the structural causal equations assuming linear relationships between the variables. *[3 marks]*

   (b) (i) Express the propensity score, $e(x) = p(T = 1|X = x)$, in terms of the model parameters. (ii) Use the law of total probability and Bayes' rule to express the probability of a citizen installing solar panels on the roof of their house without financial incentive in terms of the parameters p_0, p_1, q_0, q_1, r. *[4 marks]*

   After running the financial scheme for a year, the government seeks to understand its efficacy. It gathers $n = 1000$ samples (y_i, t_i, x_i) via a questionnaire. (You may assume these samples are independent and identically distributed.)

   (c) Explain how to estimate the average causal effect of T on Y using regression (covariate) adjustment. *[2 marks]*

   (d) It turns out only 45 people have received the financial incentive ($T = 1$). Explain a third method to estimate the average causal effect of T on Y with the aim of accounting for this imbalance. *[2 marks]*

   (e) Describe a method for sensitivity analysis of the average causal effect of T on Y by stating how you would apply the technique to this data. *[2 marks]*

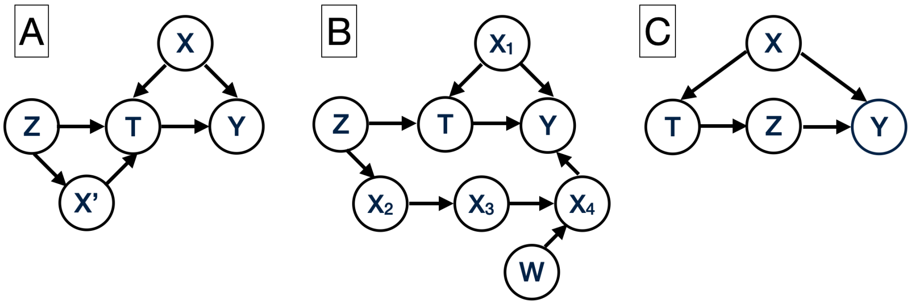
   The results of the questionnaire lead to a policy change in the financial scheme. Besides the covariate X, another covariate A is taken into consideration which measures the angle (from $0 - 90$ degrees, flat to vertical) of the roof, where the optimal angle for efficient energy collection is $A = 30$. The closer A is to 30, the more likely a citizen is to receive financial aid.

(f) Explain one advantage and one disadvantage of including the angle covariate A has on the estimation of the causal effect of the financial scheme on incentivising its citizens to install solar panels. [*2 marks*]

2. Let Y be a continuous outcome variable and let T be a binary treatment variable. In this question, we investigate methods to estimate the causal effect of T and Y as well as the required assumptions for these methods to apply.

   (a) What should one be wary of in case unobserved variables are known or suspected to influence the data (T, Y)? State the name of the corresponding assumption for causal identifiability. [*1 mark*]

   (b) Consider the three causal diagrams in the figure below. For each diagram, state if Z has the role of an instrumental variable for the effect size of T on Y. If not, explain which IV condition is violated. [*3 marks*]



   (c) Weak correlation between the instrumental variable and treatment variable may lead to unstable estimates of the causal effect of T on Y. By writing down an IV estimator for continuous variables, explain why *weak instruments* may lead to unstable estimates. [*2 marks*]

Next, we investigate how the front-door criterion may allow the identification of the causal effect of treatment T on outcome Y. Consider the following table of summary information on the variables (T, Z, Y), where Y denotes health outcome:

	Drug $T = t_0$		Drug $T = t_1$		All subjects	
Total individuals	400		200		600	
Mechanism Z	High	Low	High	Low	High	Low
Column total	230	170	126	74	356	244
High health $Y = y_1$	156	58	72	25	228	83
Low health $Y = y_0$	74	112	54	49	128	161

We seek to understand if Y is affected differently by a drug containing iron only $(T = t_0)$, or a more expensive drug containing both iron and zinc $(T = t_1)$. It is known the causal diagram is of type $\boxed{\text{C}}$ above and the variable X is unobserved. However, the intermediate variable Z tells us if the mechanism is highly effective $(Z = z_1)$ or lowly effective $(Z = z_0)$.

(d) Compute the causal effect of $T = t_1$ on $Y = y_1$ and the causal effect of $T = t_0$ on $Y = y_1$ using the above table. Present your calculations clearly so the examiners can follow your reasoning. [*5 marks*]

(e) Suppose the standard deviation on the average causal effect of T on $Y = y_1$ has been estimated, *e.g.*, with a bootstrap procedure, and equals $\sigma = 0.02$. What is the conclusion regarding drug $T = t_1$? [*1 mark*]

We are told by a subject expert that we should incorporate a further variable V in order to obtain the correct causal effect of T on Y. The expert knows that the variable V affects both Z and the health outcome Y, but not T or X.

(f) Draw the new causal graph. By reasoning with d-separation, explain which variables should be conditioned on to identify the causal effect of T on Y. (You will prove the identification in questions (g) and (h).) [*2 marks*]

Recall the derivation of the front-door criterion in the lectures to identify the causal effect of T on Y.

(g) Use the total law of probability and Bayes' rule to express the probability $p(Y = y \mid \mathrm{do}(T = t))$ in terms of

$$p\big(Y = y \mid V = v, \mathrm{do}(Z = z(t, v))\big) \qquad \text{and} \qquad p\big(Z = z(t, v) \mid V = v, \mathrm{do}(T = t)\big). \tag{1}$$

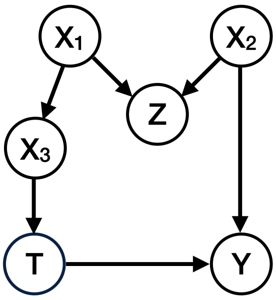
Here $z(t, v)$ denotes the right-hand side of the structural equation expressing z as a function of t and v. [*4 marks*]

(h) Use your result in (g) to complete the identification of the causal effect of T on Y by expressing the hypothetical probability $p(Y = y \mid \mathrm{do}(T = t))$ in terms of (conditional) probabilities of Y, T, Z, V. [*4 marks*]

Hint: identify both probabilities in Eq. 1 separately, and combine these results as per (f).

(i) How does this formula change if V also affects T, in addition to Z, and Y? Explain your answer. [*2 marks*]

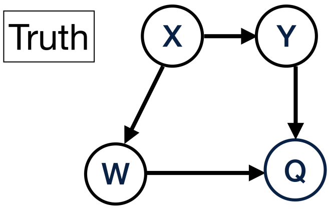
3. Let Y be a continuous outcome variable, T be a binary treatment variable, X_1, X_2, X_3 and Z be variables that influence T and/or Y as depicted in the causal graph below.



(a) The adjustment formula can be used to identify the causal effect of T on Y. Given the causal graph above, write down four adjustment sets that can be used to do so. [*2 marks*]

(b) Now suppose that the variables X_1, X_2 and X_3 are *not* observed. Can you still give an adjustment set to estimate the causal effect of T on Y? Explain why. [*2 marks*]

Suppose we are working with the variables (X, Y, W, Q) in the causal graph shown below. However, suppose we did not known the causal graph and we wish to discover it using a causal discovery algorithm.



(c) Starting with a completely connected graph, run the PC algorithm by hand. More precisely, at each step, draw the causal graph and explain why edges are removed and/or why directionality can be determined. [*7 marks*]