UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Tuesday, 7 June, 2022 1:30 pm to 3:30 pm

PAPER 221

CAUSAL INFERENCE

Before you begin please read these instructions carefully

Candidates have TWO HOURS to complete the written examination.

Attempt no more than **THREE** questions.
There are **FOUR** questions in total.
The questions carry equal weight.

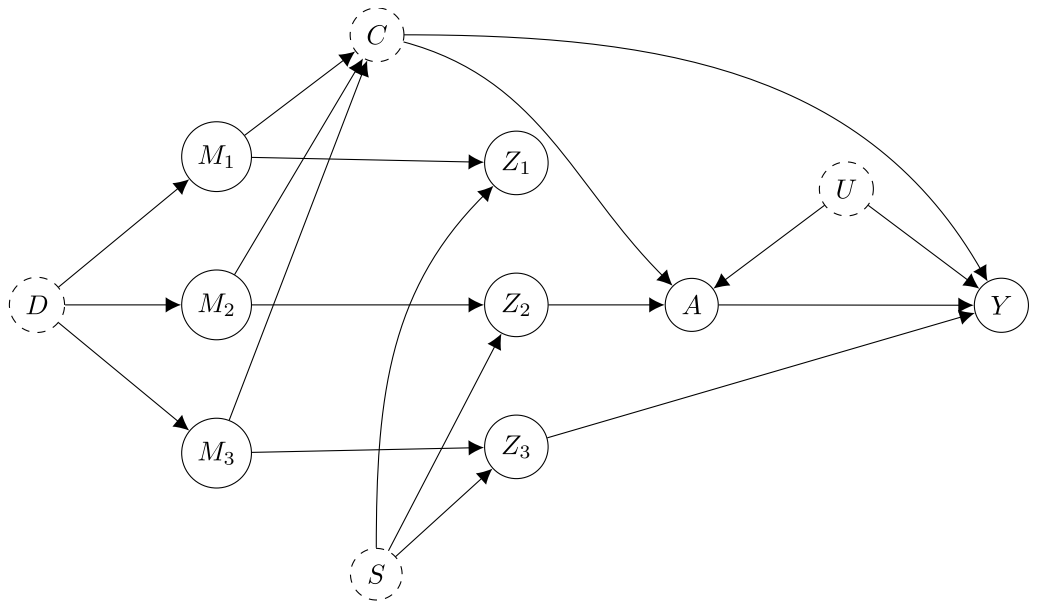
STATIONERY REQUIREMENTS
Cover sheet
Treasury tag
Script paper
Rough paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1

Consider the following causal diagram, in which observed variables are indicated by solid circles and unobserved varibles are indicated by dashed circles. Suppose we are interested in estimating the causal effect of A on Y.



(i) Use the counterfactual language introduced in lectures to state the core assumptions for an instrumental variable. Then show that Z_1 is not a valid instrumental variable for estimating the effect of A on Y.

(ii) Let X be a set of observed variables. We say X is a sufficient adjustment set for the potential instrumental variable Z_1, if the assumptions in (i) are satisfied given X. Find all the sufficient adjustment set(s) for Z_1 and justify your answer.

(iii) Suppose the marginal distribution of S and the conditional distribution of (Z_1, Z_2, Z_3) given (M_1, M_2, M_3, S) are known. Describe, in detail, a randomization test for the sharp null hypothesis that A has no effect on Y and prove that it controls the type I error rate. [You may use any results from the lectures.]

2

Twins studies are commonly used to estimate the genetic heritability of a human trait. Consider a collection of pairs of twins, and let Y_{ij} be a trait (for example, height) of twin j in pair i (so $j \in \{1,2\}$). The popular ACE model assumes that Y_{ij} has the same marginal distribution for all i and j and can be decomposed into

$$Y_{ij} = A_{ij} + C_{ij} + E_{ij},$$

where A_{ij}, C_{ij}, E_{ij} are independent and unobserved terms corresponding to genetic influences, common environmental influences (so $C_{i1} = C_{i2}$), and unique environmental influences (so E_{i1} and E_{i2} are independent and identically distributed), respectively. Genetic heritability of the trait is defined as

$$h^2 = \frac{\mathrm{Var}(A_{ij})}{\mathrm{Var}(Y_{ij})}.$$

For this question, it is useful to know that there are two kinds of twins: monozygotic (MZ) or "identical" twins, who share exactly the same genes, and dizygotic (DZ) or "fraternal" twins, who are no more different than ordinary sibilings besides being born at the same time. A crucial assumption in the ACE model is that $\mathrm{Cor}(A_{i1}, A_{i2}) = 1/2$ if i indexes a pair of DZ twins.

(i) Use the information above to draw a causal diagram and construct a linear structural equation model for the following variables: $Y_{i1}, A_{i1}, C_{i1}, E_{i1}, Y_{i2}, A_{i2}, C_{i2}, E_{i2}$. Then explain how your model is different when i indexes a pair of MZ twins and DZ twins.

(ii) Use your model to prove the so-called Falconer's formula:

$$h^2 = 2(r_{\mathrm{MZ}} - r_{\mathrm{DZ}}),$$

where r_{MZ} is the correlation of the observed trait between MZ twins and r_{DZ} is the correlation of the trait between DZ twins.

(iii) Behind the mathematical assumption $\mathrm{Cor}(A_{i1}, A_{i2}) = 1/2$ for DZ twins is the real world assumption that there is no assortative mating, so DZ twins share 50% of their genes on average. In reality, however, DZ twins share more than 50% of their genes. Comment on the bias of Falconer's formula in this situation.

3

Consider the problem of inferring the causal effect of a binary treatment variable A on a real-valued outcome variable Y. Let X be some observed covariate. Let $Y(a)$ be the counterfactual outcome when A is set to $a = 0, 1$.

(i) State the no unmeasured confounders assumption. Use this to prove that average treatment effect (ATE), $\beta_1 = \mathbb{E}[Y(1) - Y(0)]$ is identified by

$$\beta_1 = \mathbb{E}\{\mathbb{E}[Y \mid A = 1, X]\} - \mathbb{E}\{\mathbb{E}[Y \mid A = 0, X]\}.$$

Carefully state any other assumptions you used when deriving this formula.

For the rest of this question, suppose all the assumptions in (i) are given.

(ii) The *partially linear model* assumes that

$$\mathbb{E}[Y \mid A, X] = \beta_2 A + g_2(X). \tag{1}$$

Show that if this model is correctly specified, then $\beta_1 = \beta_2$.

The rest of this question investigates whether this equality continues to hold when the partially linear model is incorrect.

(iii) Define

$$(\beta_3, g_3(\cdot)) = \arg\min_{\beta, g(\cdot)} \mathbb{E}[\{Y - \beta A - g(X)\}^2].$$

Show that, if (1) is not correctly specified, then $\beta_3 \neq \beta_1$ in general. Find the counterfactual quantity that β_3 identifies and explain how it is related to the ATE.

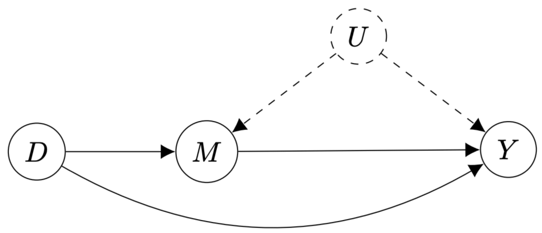
(iv) Let $e(X) = \mathbb{E}[A \mid X]$ and $\mu(X) = \mathbb{E}[Y \mid X]$. Let $\tilde{A} = A - e(X)$ and $\tilde{Y} = Y - \mu(X)$ be the centred treatment and response, respectively. Consider

$$\beta_4 = \arg\min_{\beta} \mathbb{E}[(\tilde{Y} - \beta \tilde{A})^2].$$

Show that $\beta_4 = \beta_3$. Given an independent and identically distributed sample $(X_i, A_i, Y_i), i = 1, \ldots, n$, use this formulation to suggest a semiparametric estimator of β_3.

4

A social scientist would like to investigate racial discrimination in policing. They decide to use the following causal diagram to describe their assumptions about a random police-civilian encounter:



In this diagram, D is a binary variable for race of the civilian ($D = 1$ means the minority race); M is a binary indicator for police stop ($M = 1$ means the civilian was stopped by police); Y is a binary indicator for police violence ($Y = 1$ means the police used force); U is an unmeasured variable that confounds M and Y.

(i) By law, every police officer is required to take a record of every stop. (Of course, they do not need to record a encounter if the civilian is not stopped.) The social scientist uses an administrative dataset of police stops in a major city to compute the following contingency table:

	$Y = 0$	$Y = 1$
$D = 0$	10000	400
$D = 1$	20000	800

The social scientist finds that, according to this table, the civilians had a 4% chance of experiencing police violence regardless of their race. They then conclude that there is no racial discrimination in policing. Using the diagram above, briefly comment on why this reasoning is flawed.

(ii) After consulting a statistician, the social scientist realised their mistake and decide to make a more careful causal analysis. However, they do not know how to formalise the intuition that there should be no police violence if the civilian is never stopped by the police. Help them to state this assumption using counterfactual variables.

[QUESTION CONTINUES ON THE NEXT PAGE]

(iii) The social scientist would like to estimate the causal risk ratio (CRR)

$$\mathrm{CRR} = \frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]},$$

where $Y(d)$ is the counterfactual outcome had the race of the civilian been d (they are willing to assume that race is manipulable in a hypothetical experiment). Using the above causal diagram and the assumption in (ii), show that the CRR can be identified by

$$\mathrm{CRR} = \frac{\mathbb{E}[Y \mid D = 1, M = 1]}{\mathbb{E}[Y \mid D = 0, M = 1]} \cdot \frac{\mathbb{P}(D = 1 \mid M = 1)/\mathbb{P}(D = 0 \mid M = 1)}{\mathbb{P}(D = 1)/\mathbb{P}(D = 0)}.$$

[Hint: Show that $\mathbb{E}[Y(d)] = \mathbb{E}[Y \mid M = 1, D = d] \cdot \mathbb{P}(M = 1 \mid D = d)$ *for* $d = 0, 1$*.]*

(iv) Identify the terms in the last formula that are not estimable from the police administrative dataset which the social scientist has access to. Then suggest how the social scientist can try to estimate them by using some external data sources or conducting a new survey.

END OF PAPER