UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 24 June, 2021 12:00 pm to 3:00 pm

PAPER 207

STATISTICS IN MEDICINE

Before you begin please read these instructions carefully

Candidates have THREE HOURS to complete the written examination.

*Attempt no more than **FOUR** questions.*
*There are **SIX** 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 Statistics in Medical Practice

Suppose we are following a cohort of patients in order to estimate the risk of stroke and dementia, and how the risk of death changes for people with these conditions. Patients visit a clinic intermittently to be examined for symptoms of dementia, and everyone who has a stroke is immediately admitted to hospital on a known date.

(a) (i) Draw a diagram of the states and the permitted transitions in a five-state multi-state model, with states representing

(1) alive, no disease
(2) alive, with dementia, but never had a stroke
(3) alive, without dementia, and has had a stroke
(4) alive with dementia, and has had a stroke
(5) dead, from any cause

Label the states and the transition intensities, assuming that the intensities are constant through time and the same for all individuals. No recovery from dementia is possible.

(ii) We obtain the following data from three patients. In each case, write down an expression for the contribution of this person's data to the likelihood of the vector of transition intensities $\mathbf{q}$ in the five-state model above, in terms of the transition intensities and transition probabilities. All patients are known to start in state 1. It is not required to know the transition probabilities as a closed-form function of the intensities.

(1) Their first dementia clinic visit occurs 2 years after the start of their follow up. At this visit they are diagnosed with dementia, and they died 1 year later without having a stroke.
(2) Hospitalised with a stroke 0.5 years after the start of their follow-up, but beyond that time it is not known what happened to them.
(3) Hospitalised with a stroke 0.5 years after the start of their follow-up. First examined for dementia, and diagnosed with dementia, 1.5 years after their stroke, and died 1 year after the dementia diagnosis.

(iii) For any two of the assumptions behind the likelihood written in part (ii), explain plausible reasons why they might not hold in this application.

(iv) Explain how we might formally assess the hypothesis that having dementia does not affect a person's risk of death after stroke.

[QUESTION CONTINUES ON THE NEXT PAGE]

(b) (i) Suppose that dementia is diagnosed if $X < 10$, where X is a clinical measurement that can vary non-monotonically through time. We want to estimate the risk of getting dementia and the risk of death after dementia, using a simplified three-state version of the model in part (a):

(1) no dementia
(2) dementia
(3) death after dementia

where dementia is irreversible. What other two quantities might we want to know in order to estimate these risks accurately?

(ii) Suppose we fitted this three-state model, with piecewise-constant transition intensities that depended on age, obtaining estimates of the 1-2 transition intensities of $q_{12} = 1/10$ for people aged 50–60, and $q_{12} = 1/5$ for people aged 60 or more. Show that the expected future time spent alive without dementia, for a person aged 50, is $10 - 5\exp(-1)$.

2 Statistics in Medical Practice

Myopia, or short-sightedness, is one of the leading causes of visual disability worldwide. For more than a century, myopia has been shown to be observationally associated with higher levels of educational attainment.

(a) What is meant by the claim that "education is a causal risk factor for myopia"? Provide and briefly justify two reasons why myopia may be more prevalent amongst highly-educated individuals without education being a causal risk factor for myopia.

(b) The Chancellor of Tomania, a country with a complete absence of ethical regulations for research, asks you to design a research investigation to demonstrate conclusively whether education is a causal risk factor for myopia. Describe how you would design such a research investigation, and justify carefully on what basis the investigation is able to make a causal claim.

To address this research question, a group of researchers in a country with ethical research regulations applied Mendelian randomization, the use of genetic variants as instrumental variables. The researchers selected 69 genetic variants associated with educational attainment at a p-value threshold of $p < 5 \times 10^{-8}$ as instrumental variables for education based on analyses performed by Okbay *et al.* They also selected 44 genetic variants associated with refractive error mean spherical equivalent (MSE), a measure of myopia (negative values suggest increased myopia) at a p-value threshold of $p < 5 \times 10^{-8}$ as instrumental variables for myopia based on analyses performed by Pickrell *et al.*

Genetic associations with education and refractive error were calculated using linear regression, and expressed as beta-coefficients (β). They are plotted in the Figure below. The left panel displays beta-coefficients for the 69 genetic variants previously associated with educational attainment, and their associations with time spent in education in years and refractive error in the UK Biobank cohort. The right panel displays beta-coefficients for the 44 genetic variants previously associated with refractive error.

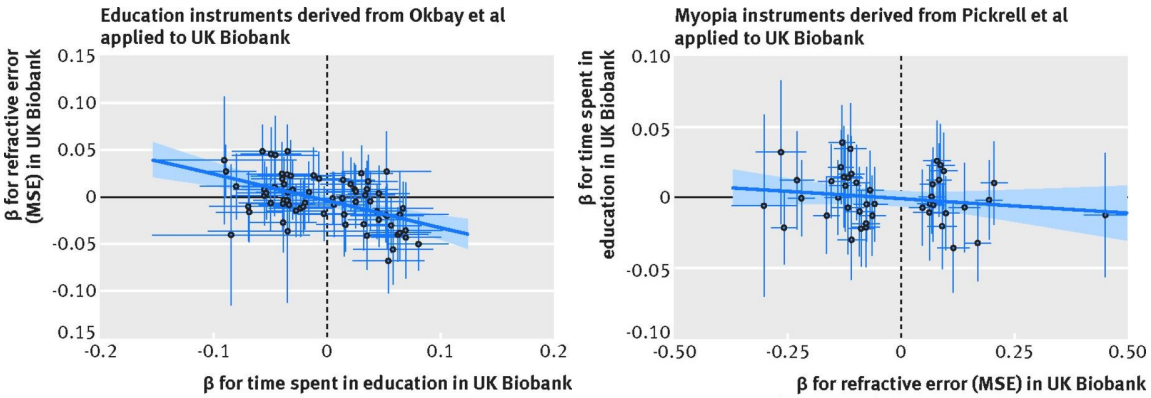

Figure: Genetic associations with time in education and with refractive error mean spherical equivalent (MSE). Error bars represent 95% confidence intervals for the genetic associations. Regression lines and 95% confidence intervals (shaded areas) are fitted using weighted linear regression. Graphs taken from Mountjoy *et al*, British Medical Journal 2018.

[QUESTION CONTINUES ON THE NEXT PAGE]

(c) List the instrumental variable assumptions, and explain briefly how they enable causal inferences to be made without perfect knowledge of all confounding variables.

(d) Assuming that the instrumental variable assumptions are satisfied for all of the genetic variants, what conclusions can be drawn from the Figure?

(e) Provide and justify three ways that this investigation could be extended to increase the strength of evidence for a causal relationship between education and myopia.

3 Statistics in Medical Practice

Consider a clinical trial in which $k \geqslant 1$ new treatments are to be tested against only one standard treatment (a situation known as Dunnett's test). When treatment i is given to a patient (where $i = 0$ denotes the standard treatment, and $i = 1, \ldots, k$ the new treatments), a continuous outcome is observed.

Let n_i denote the sample size for each treatment, and assume there is to be equal allocation amongst the new treatments (i.e. $n_1 = n_2 = \ldots = n_k$). We also assume that it is appropriate to compare between treatments using a normally-distributed test statistic, and that the outcomes for all treatment groups have a common known variance σ^2.

For testing one new treatment i against the standard treatment, we use a Z-test based on the statistic Δ_i / SE_{Δ_i}, where $\Delta_i = \bar{X}_i - \bar{X}_0$ is the difference in means between group i and the standard group, and the standard error of this comparison is given by:

$$SE_{\Delta_i} = \sigma\sqrt{V}, \text{ where } V = \tfrac{1}{n_0} + \tfrac{1}{n_i}.$$

(a) Define q as some positive real constant, which is not necessarily an integer, such that $n_0 = qn_k$. Also let n_{tot} denote the total trial sample size, which is fixed.

(i) Express the sample size n_k and variance V in terms of only n_{tot}, k and q. Hence find the optimal choice of q, and prove that this value of q minimises rather than maximises the variance V.

(ii) Suppose $k = 4$ and we use the optimal choice of q. We want to calculate the sample size for this trial to detect a clinically relevant difference δ^* between the standard treatment and each new treatment, with 80% power and 5% type I error rate when using a two-sided Z-test.

We know that $\delta^* = 2$ and $\sigma = 2$, and also need to take into account an overall 10% attrition rate.

Find the total trial sample size n_{tot}, as well as the total sample size required before attrition (i.e. the total to be recruited).

Note: The sample size for a two-sided Z-test between two groups, one of size n_k and the other of size qn_k, is given by

$$n_k = \frac{1+q}{q}\left(\frac{\sigma}{\delta^*}\right)^2 \left(z_{1-\beta} + z_{1-\alpha/2}\right)^2.$$

For 80% power ($\beta = 20\%$) and type I error rate $\alpha = 5\%$, $z_{1-\beta} \approx 0.84$ and $z_{1-\alpha/2} \approx 1.96$. Therefore $\left(z_{1-\beta} + z_{1-\alpha/2}\right)^2 \approx 7.84$.

[QUESTION CONTINUES ON THE NEXT PAGE]

(b) Setting $k = 1$, suppose we now want to redesign the clinical trial as a two-stage group sequential design, with equal allocation to the new and standard treatment. The null hypothesis tested in the trial is $H_0 : \delta \leqslant 0$, where δ is the difference between the mean effectiveness of the new treatment and the mean effectiveness of the standard treatment.

At the jth analysis ($j = 1, 2$), a Wald test statistic Z_j is calculated using the patients assessed so far. The group sequential design is defined by pre-specified constants (f_j, e_j), where $f_1 < e_1$ and $f_2 = e_2$.

At the 1st analysis, if $Z_1 \geqslant e_1$ we stop and reject H_0, if $Z_1 \leqslant f_1$ we stop and do not reject H_0, and if $Z_1 \in (f_1, e_1)$ we continue to the 2nd stage.

If the trial continues to the 2nd stage, at the 2nd (and final) analysis, we reject H_0 if $Z_2 \geqslant e_2$, otherwise we do not reject H_0.

(i) Describe an advantage and a disadvantage of using a group sequential trial compared to a trial with a fixed sample size.

(ii) Write down the joint distribution of (Z_1, Z_2) in terms of δ and the (Fisher) information $\mathcal{I}_j$ at the jth analysis.

(iii) Using the joint distribution (Z_1, Z_2), write down an expression for the probability of continuing to the 2nd stage and not rejecting H_0, in terms of a multidimensional integral of a function you should specify.

(iv) Derive an expression for the expected sample size (per treatment arm) as a function of $\delta, f_1, e_1, \mathcal{I}_1$ and N_j, where N_j denotes the cumulative number of observations on each treatment arm if the trial continues to the jth analysis.

4 Analysis of Survival Data

(a) Let T represent a continuous time-to-event random variable and C represent the corresponding time-to-censoring variable. The functions $N(t)$ and $Y(t)$ are defined by:

$$N(t) = \mathbb{I}\left[T \leqslant \min(t, C)\right]$$

and

$$Y(t) = \mathbb{I}\left[t \leqslant \min(T, C)\right]$$

where $\mathbb{I}$ is the indicator function. Explain why these functions are so useful in the analysis of time-to-event data. When is $Y(t) + N(t)$ not equal to unity?

(b) By considering the conditional expectation $\mathbb{E}\left[\mathrm{d}N(t)|Y(t) = y(t)\right]$, write down an expression which defines the integrated hazard $\mathrm{H}(t)$.

(c) A time-to-event dataset comprises n individuals, the ith either having an event ($v_i = 1$) or being censored ($v_i = 0$) at time x_i, $i = 1, \ldots, n$. All the individuals experience the same integrated hazard. Assuming that there are no ties in the dataset (that is: $x_i \neq x_{i'}$ for $i \neq i'$), derive an estimator $\hat{\mathrm{H}}(t)$ for the integrated hazard using the definition in part (b).

(d) Why is it desirable that an estimator $\tilde{\mathrm{H}}(x_i)$ of $\mathrm{H}(t)$ should satisfy $\sum_i \left(v_i - \tilde{\mathrm{H}}(x_i)\right) = 0$?

(e) Now assume further that, without losing generality, the individuals have been ordered such that $x_i < x_{i'}$ for $i < i'$.

Obtain a condition on $\tilde{\mathrm{H}}(t)$ such that $\sum_i (v_i - \tilde{\mathrm{H}}(x_i)) = 0$. Choose the most natural form of $\tilde{\mathrm{H}}(t)$ which satisfies that condition and show the resulting estimator is the same as the one you derived in part (c).

[Hint: you will find it easier in part (e) to work with the increment in the estimate of the integrated hazard from one observation to the next: $D_i = \tilde{\mathrm{H}}(x_i) - \tilde{\mathrm{H}}(x_{i-1})$.]

5 Analysis of Survival Data

What is meant by a *proportional hazards* model? Explain how the likelihood function for the model parameters can be constructed in the case when the baseline hazard is unspecified (you may assume no tied event times). Why is this likelihood often referred to as a 'partial likelihood'?

A survival dataset comprises four observations (x_i, v_i, g_i) $i = 1, \ldots, 4$ where x_i is the time of event ($v_i = 1$) or censoring ($v_i = 0$) and $g_i \in \{0, 1\}$ indicates which of two treatment groups the ith indvidual belongs to. The table shows the values of x_i, v_i, g_i for the four individuals:

i	x_i	v_i	g_i
1	t_1	1	0
2	t_2	1	1
3	c	0	1
4	t_4	k	0

where $t_1 < t_2 \leqslant c \leqslant t_4$ and $k \in \{0, 1\}$.

(a) Assuming that the hazard function for group 1 is a constant multiple θ of the unspecified hazard function for group 0, write down the partial likelihood function for θ in the cases (i) $t_2 < c < t_4$ and (ii) $c = t_4$. Obtain an estimate $\hat{\theta}$ for θ when $t_2 < c < t_4$ and explain why this estimate does not depend on k. Describe fully the dependence of $\hat{\theta}$ on c for $t_2 \leqslant c \leqslant t_4$ (you may use the result that $\hat{\theta} \simeq 0.39$ when $c = t_4$ and $k = 1$).

(b) What is the maximum likelihood estimator of the rate parameter of an exponential distribution? Explain why fitting an exponential distribution to both treatment groups can be considered equivalent to using a proportional hazards model. Obtain an expression for the maximum likelihood estimate $\hat{\lambda}$ of the hazard ratio λ in terms of t_1, t_2, c, t_4 and k. How, in general terms, does this estimate vary with c for $t_2 \leqslant c \leqslant t_4$? Why does the hazard ratio in this model depend on the actual event and censoring times? Explain why, for $t_2 < c < t_4$, $\hat{\lambda}$ varies with t_4 but $\hat{\theta}$ does not.

6 Analysis of Survival Data

(a) Describe how to construct a Kaplan-Meier estimate of the survivor function for a group of individuals who either have an event or are right-censored.

What is meant by *left-truncation*? Give a practical example of left-truncation. How can the Kaplan-Meier estimate be adapted to cope with left-truncation?

When would you consider using a period survival analysis? Explain how to construct a Kaplan-Meier estimate of the period survivor function for a given calendar year.

(b) The table below lists survival data for 10 patients diagnosed with a particular disease. The date of diagnosis of the ith patient is y_i, the number of months since December 2016 (so that, as examples, $y_i = 0$ if the patient was diagnosed in December 2016, $y_i = 3$ if the patient was diagnosed in March 2017, $y_i = -2$ if the patient was diagnosed in October 2016). The date of death ($u_i = 1$) or censoring ($u_i = 0$) is represented in the same way by z_i.

Calculate the period survival estimate of the survivor function for time from diagnosis to death for the calendar year 2017.

i	y_i	z_i	u_i
1	-9	15	0
2	-8	-2	1
3	-7	8	1
4	-3	4	0
5	-1	18	1
6	2	11	0
7	4	10	1
8	7	21	0
9	9	24	1
10	15	20	1

[You may assume (i) that the year is divided into 12 equal calendar months, (ii) that diagnoses, deaths and censorings occur on the first day of the month and (iii) that a patient is at risk of death from the disease on and after the day of diagnosis. You may ignore any difficulties associated with the imprecise recording of time.]

END OF PAPER