The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS Spring Semester 2019–2020

Sampling, Design, Medical Statistics 1.5 hours

This is an open book exam.

*Answer **all** questions.*

The submission deadline is 10 am (BST), twenty-four hours after it is released. Late submission will not be considered without extenuating circumstances. It is expected that you will be able to complete this exam in approximately one and a half hours and it is recommended that you submit the work within four hours. You will not be penalised for taking longer, however.

Unless it is explicitly stated otherwise, it is intended that calculations are performed by hand (possibly with the aid of a calculator). To gain full marks, you will need to show your working. You will not get full marks if you simply write down output from a computer package.

By uploading your solutions you declare that your submission consists entirely of your own work, that any use of sources or tools other than material provided for this module is cited and acknowledged and that no unfair means have been used.

1 A doctor proposes using a trial of 200 patients (in two parallel groups of 100 patients each) to test whether a new drug improves response rate above the 55% observed on the current drug.

(i) Check that such a trial would give reasonable power if the new drug improved the response rate to 75% or more. ***(5 marks)***

(ii) He proposes using simple randomization to decide which of the subjects will receive the new drug and which the current drug. Explain why this might result in unequally sized groups and suggest an alternative method he could use to ensure equal group sizes. ***(2 marks)***

(iii) After a suitable randomization, the trial is conducted. The doctor notes various items of general information on the patients (age, sex, employment status, prior medical history,...) and records the following data:

	Respond	Do not respond	Total
Current drug	56	44	100
New drug	66	34	100

Test whether the new drug seems to be an improvement on the current drug. ***(4 marks)***

(iv) Give two reasons why analyzing the data using logistic regression might be preferable to your analysis in (iii). ***(2 marks)***

2 Relapsing-remitting multiple sclerosis (RRMS) is a serious disease of the central nervous system. RRMS patients typically present with symptoms, followed by remission, possibly followed by relapse. A clinical trial was conducted on 150 patients to compare two treatments for RRMS, Extavia and Avonex. Patients were randomly allocated to treatments and followed up for 4 years for a potential relapse. Patients still in remission after the end of study were considered right-censored. The data are stored in `RRSM` and coding for the different variables is shown below:

Coding:

Sex:	0 = male; 1 = female
Age:	age of patient centred on 35 years (i.e. 30 year old is -5)
Time:	time until relapse (years)
Treatment:	0 = Extavia; 1 = Avonex
Status:	0 = censored; 1 = relapse

(i) An analysis was implemented in *R* producing the following output:

```
> RRMS.sv <- Surv(Time, Status)
>
> RRMS.fit1<-survreg(RRMS.sv~Sex+Age+Treatment,dist="exponential")
> summary(RRMS.fit1)

Call:
survreg(formula = RRMS.sv ~ Sex + Age + Treatment,dist = "exponential")
               Value Std. Error     z       p
(Intercept)  0.77301    0.16881  4.58 4.7e-06
Sex         -0.12475    0.19577 -0.64 0.52398
Age          0.03466    0.00982  3.53 0.00042
Treatment   -0.61445    0.19586 -3.14 0.00171

Scale fixed at 1

Exponential distribution
Loglik(model)= -147.3   Loglik(intercept only)= -159.4
Chisq= 24.34 on 3 degrees of freedom, p= 2.1e-05
Number of Newton-Raphson Iterations: 5
n= 150
```

(a) Describe the analysis performed and write down the form of the fitted model for the time to relapse using appropriate notation. ***(3 marks)***

(b) Based on the analysis, would you say that Extavia seems to be more effective than Avonex? Justify your answer. ***(1 mark)***

(c) Estimate the expected time to relapse for a 28 year old female assigned to Extavia. ***(1 mark)***

2 (continued)

(ii) An alternative analysis was conducted by another statistician, who presented the following results:

```
> RRMS.fit2 <- coxph(RRMS.sv ~Sex+Age+Treatment)
> summary(RRMS.fit2)
Call:
coxph(formula = RRMS.sv ~ Sex + Age + Treatment)

  n= 150, number of events= 106

              coef exp(coef) se(coef)      z Pr(>|z|)
Sex         0.12999   1.13881  0.19771  0.657 0.510897
Age        -0.03356   0.96700  0.01019 -3.295 0.000986 ***
Treatment  0.58730   1.79913  0.20093  2.923 0.003467 **
---
```

(a) Write down the fitted model in terms of the baseline hazard function $h_0(t)$, the covariates and the parameter estimates. Why are such models called semi-parametric? *(3 marks)*

(b) Using the *R* output, describe in detail the effects of sex and treatment on time to relapse. *(2 marks)*

(c) Based on the model fitted, estimate the hazard ratio comparing two males, 33 and 43 years old respectively, both taking Avonex. Would this hazard ratio estimate change if both individuals were females instead? Justify your answer. *(2 marks)*

3 An investigator is studying the dependence of a variable Y on one continuous explanatory variable x which has been scaled to lie between -1 and 1. It is assumed that $E(Y) = 0$ when $x = 0$, and the following model (Model 1) is proposed.

$$\text{Model 1}: \quad E(Y) = \beta_1 x + \beta_{11} x^2.$$

The investigator proposes the following design (design A) using four design points:

Design	Design points (x)
A	$\{-1, -1, 1, 1\}$

(i) Explain whether β_1 and β_{11} in Model 1 are orthogonal to each other for design A. ***(3 marks)***

(ii) Justify whether Design A is G-optimal for Model 1. ***(4 marks)***

(iii) For Model 1, why might an orthogonal design that gives a circular, as opposed to an elliptical, confidence region for $(\beta_1, \beta_{11})^T$ be desirable when designing an experiment? ***(2 marks)***

(iv) The investigator proposes the following model (Model 2)

$$\text{Model 2}: \quad E(Y) = \beta_0 + \beta_1 x + \beta_{11} x^2.$$

Suppose there are n design points, $x_1, \ldots, x_n$, available. Justify whether it is possible to specify a set of design points $\{x_1, \ldots, x_n\}$ for Model 2 that would yield a 95% confidence region for $(\beta_0, \beta_1, \beta_{11})^T$ that is an ellipsoid with all three axes parallel to the co-ordinate axes. ***(5 marks)***

4 (i) An experiment is to be conducted to investigate the effect of four continuous factors, represented by x_1, x_2, x_3 and x_4, on response variable Y. Each factor is restricted to one of two levels labelled -1 and +1. Four design points are to be used.

(a) A fractional factorial design with block generators $x_1x_2 = 1$ and $x_1x_4 = 1$ is proposed. Explain why these generators are a poor choice. ***(2 marks)***

(b) Provide two block generators that would be a better choice than those given in part (a). Explain your answer with reference to the alias structure. ***(4 marks)***

(ii) A further experiment is to be conducted to investigate the effect of two continuous factors, represented by x_5 and x_6 on response variable Y. x_5 is restricted to one of two levels labelled -1 and +1 whilst x_6 is restricted to one of three levels labelled -1, 0 and +1. The investigator has six design points available. They initially propose the following model (Model 3)

$$\text{Model 3}: \quad E(Y) = \beta_0 + \beta_5 x_5 + \beta_6 x_6 + \beta_{56} x_5 x_6 + \beta_{55} x_5^2 + \beta_{66} x_6^2$$

(a) Explain why the investigator cannot include all the parameters in Model 3. ***(2 marks)***

(b) Following advice from a statistician, they propose the following model (Model 4)

$$\text{Model 4}: \quad E(Y) = \beta_0 + \beta_5 x_5 + \beta_6 x_6 + \beta_{56} x_5 x_6 + \beta_{66} x_6^2 + \beta_{566} x_5 x_6^2$$

Based on Model 4, they propose a design that leads to the following design matrix (matrix X)

$$X = \begin{pmatrix} 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

in which the order of the columns matches the order of the terms in Model 4. What is name of the design used to generate the columns for x_5 and x_6 (columns 2 and 3 of X)? ***(1 mark)***

(c) The investigator checks to see whether $|X^TX|$ is zero. What does the result of this check tell them? ***(2 marks)***

End of Question Paper