The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2020–2021**

Statistical Inference and Modelling **3 hours**

This is an open book exam.

Answer all questions.

You can work on the exam during the 24 hour period starting from 10am (BST), and you must submit your work within 3 hours of accessing the exam paper or by the end of the 24 hour period (whichever is earlier).

Late submission will not be considered without extenuating circumstances. *Calculations should be performed by hand. A university-approved calculator may be used. The use of any other calculational device, software or service is not permitted. To gain full marks, you will need to show your working.*

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 Let $(X, Y)^T$ be a bivariate random variable with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{5}(x+2y) & \text{if } x \in [0,1],\ y \in [0,2] \\ 0 & \text{otherwise} \end{cases}.$$

(i) Calculate $\mathbb{P}(X \geq Y)$. ***(2 marks)***

(ii) Find the marginal probability density function of X, $f_X(x)$. ***(2 marks)***

(iii) Let $x \in [0,1]$. Find $\mathbb{E}(Y|X = x)$. ***(4 marks)***

2 Let $\lambda \in (0, \infty)$. Let X be an $Exp(\lambda)$ random variable, which has probability density function

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases},$$

and let $Y = \sqrt{X}$.

(i) Find the probability density function of Y, $f_Y(y)$. ***(5 marks)***

(ii) What is the name, and associated parameter values, of the distribution of Y? ***(1 mark)***

(iii) Generalise this result to transformations of the form $g(X) = \sqrt[r]{X}$, for $r \in \{2, 3, 4, \ldots\}$. ***(1 mark)***

3 Let $n \in \mathbb{N}$ and let $\mathbf{x} = (x_1, \ldots, x_n)$ be a sample consisting of n independent, identically distributed random variables following the Beta distribution, $Be(1, \theta)$. Here, $\theta \in (0, \infty)$ is an unknown parameter.

(i) Find the likelihood function $L(\theta; \mathbf{x})$ of θ, and hence show that the corresponding log-likelihood function is given by

$$\ell(\theta; \mathbf{x}) = n \log(\theta) + (\theta - 1) \sum_{i=1}^{n} \log(1 - x_i).$$

(3 marks)

(ii) Find the maximum likelihood estimator $\hat{\theta}$ for θ, given the data $\mathbf{x}$.

(3 marks)

(iii) Show that the k-unit likelihood region for θ is

$$\left\{ \theta > 0 \; ; \; \left| \log \frac{\theta}{\hat{\theta}} + \frac{\hat{\theta} - \theta}{\hat{\theta}} \right| \leq \frac{k}{n} \right\}.$$

(2 marks)

(iv) Give the 2-unit likelihood region (to 3 significant figures) if the observed data were

$$\mathbf{x} = (0.13, 0.21, 0.04, 0.41, 0.22, 0.21, 0.67).$$

(2 marks)

4 You are given some data on systolic blood pressure (SBP) for various people, together with each person's body mass index (BMI). You want to test for the effect of BMI on SBP (these are both continuous variables). You also have data on each individual's smoking habit: denoted by $i = 1$ if they smoke regularly and $i = 2$ if they do not. Finally, you have data on gender: whether they are male ($j = 1$) or female ($j = 2$). Let $x_{i,j,k}$ be the BMI of person k within group (i, j). Let $y_{i,j,k}$ be the SBP of person k within group (i, j).

(i) To test the effect of BMI on SBP, whilst controlling for gender and smoking habit, eight linear models are constructed, all assuming (i) there are no interactions between the independent variables and (ii) all relationships between continuous variables are linear (e.g. no quadratic or higher order terms). Below, seven of the models $(M_0, \ldots, M_6)$ are described in terms of the assumptions made, whilst one (M_7) is described in terms of a mathematical expression, where $\epsilon_{i,j,k} \sim N(0, \sigma^2)$ for some σ.

- Model M_0: Assumes there is no relationship between SBP and any of the three independent variables: gender, smoking habit, and BMI.
- Model M_1: Assumes there is a relationship between BMI and SBP, but not between SBP and the other two independent variables
- Model M_2: Assumes there is a relationship between gender and SBP, but not between SBP and the other two independent variables
- Model M_3: Assumes there is a relationship between smoking habit and SBP, but not between SBP and the other two independent variables
- Model M_4: Assumes there is a relationship between SBP and both BMI and gender, but not between SBP and smoking habit
- Model M_5: Assumes there is a relationship between SBP and both BMI and smoking habit, but not not between SBP and gender
- Model M_6: Assumes there is a relationship between SBP and both gender and smoking habit, but not not between SBP and BMI
- Model M_7: $y_{i,j,k} = \beta_0 + \tau_i + \eta_j + \beta_1 x_{i,j,k} + \epsilon_{i,j,k}$

For models M_0, M_1, M_3, M_4 give the corresponding mathematical expression (Note: I have deliberately omitted M_2, M_5, M_6 for brevity). ***(4 marks)***

For Model M_7, describe the assumptions being made and briefly explain the meaning of the symbols τ_i, η_j, and β_1. ***(4 marks)***

(ii) Draw the nesting structure for models M_0 to M_7. ***(2 marks)***

 Question 4 continued on next page

4 (continued)

(iii) To find the best model, models $M_0, ..., M_7$ are coded into `R` using the `lm()` function, and stored as `lm0`, ..., `lm7` respectively. The following commands are put into `R`

```
> anova(lm6,lm7)
> anova(lm5,lm7)
> anova(lm4,lm7)
```

yielding p-values of 0.014, 0.034, and 0.00085, respectively. Assuming a significance level of 5%, which is the best-fit model and why? Make sure your explanation is written in clear, grammatically-correct sentences. ***(2 marks)***

(iv) The following hypothesis tests are also performed

```
> anova(lm0,lm1)
> anova(lm0,lm2)
> anova(lm0,lm3)
```

yielding p-values of 0.000014, 0.063, and 0.0097, respectively. What can you conclude about the effect of gender on SBP (using 5% significance levels)? Account for any apparent contradiction between the results. Make sure your explanation is written in clear, grammatically-correct sentences. ***(3 marks)***

5 Suppose we have some data as follows: $(x_1, y_1) = (1, 42)$, $(x_2, y_2) = (3, 28)$, $(x_3, y_3) = (4, 14)$. We model this as follows:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $i \in \{1, 2, 3\}$ and $\epsilon_i \sim N(0, \sigma^2)$.

(i) Write this model in matrix notation. ***(2 marks)***

(ii) Show that the least-squares estimates for β_0 and β_1 are $\hat{\beta}_0 = 52$ and $\hat{\beta}_1 = -9$ (minus nine), respectively. ***(3 marks)***

(iii) Find the unbiased estimate of σ^2. ***(3 marks)***

(iv) Draw a graph of the data-set with the best-fit line superimposed. ***(2 marks)***

End of Question Paper