The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2014–2015**

MAS273 Statistical Modelling (Special Paper) **2 hours**

*Attempt **ALL FOUR** questions. The allocation of marks is shown in brackets. Total marks 85.*

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 In a study of the effect of hormones on the productivity of a certain variety of tomato plant, ten plants were treated at different hormone strengths, and their yields, in kg, noted. A simple linear regression model is fitted in R, with the dependent variable (yield in kg) stored as the vector `y` and the independent variable (hormone strength) stored as the vector `x`. Some R commands and edited output are shown below.

```
> lm1<-lm(y~x)
> summary(lm1)

Coefficients:
            Estimate Std. Error
(Intercept) -3.72335    0.28072
x            0.42999    0.01899

> qt(0.975,8)
[1] 2.306004
```

(i) Write down the equation of the model that has been fitted to the data, defining your notation carefully. State the distribution of any error terms in your model. ***(5 marks)***

(ii) If X is the design matrix for the model fitted in R, then

$$(X^TX)^{-1} = \begin{pmatrix} 2.6485 & -0.1758 \\ -0.1758 & 0.0121 \end{pmatrix}.$$

Also,

$$\sum_{i=1}^{10} y_i = 26.76, \qquad \sum_{i=1}^{10} x_i y_i = 424.33,$$

where y_i is the i-th yield and x_i is the i-th hormone strength. Give suitable calculations that show how the parameter estimates have been obtained in the R output. ***(5 marks)***

(iii) If $x_3 = 11$ and $y_3 = 1.66$, calculate the corresponding fitted value and residual. ***(4 marks)***

(iv) Calculate 95% confidence intervals for the intercept and gradient in the regression model. ***(3 marks)***

(v) Test the hypothesis that there is no relationship between hormone strength and yield, stating your conclusion clearly. State the size of your hypothesis test. ***(4 marks)***

(vi) State the assumptions used to construct the confidence interval and do the hypothesis test in parts (iv) and (v). State one hypothesis test you could use to check these assumptions. You should state precisely what the null hypothesis would be in your hypothesis test. ***(4 marks)***

2 In the simple linear regression model through the origin

$$y_i = \beta x_i + \varepsilon_i,$$

for $i = 1, \ldots, n$, and with the errors being independent $N(0, \sigma^2)$, random variables.

(i) Prove that the least squares estimate of β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Show from first principles that $\hat{\beta}$ is an unbiased estimator of β, obtain its variance (you should consider the mean and variance of the estimator directly; do not quote general properties of least squares estimates). State the distribution of $\hat{\beta}$. ***(10 marks)***

(ii) Show that another estimator of β, denoted by $\tilde{\beta}$ and given by $\bar{Y}/\bar{x}$, is also an unbiased estimator of β , and obtain its variance. ***(6 marks)***

(iii) Which of the estimators $\hat{\beta}$ and $\tilde{\beta}$ is preferred? Give a reason for your choice. Hint:

$$\sum_{i=1}^{n} x_i^2 = n\bar{x}^2 + \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

(5 marks)

3 In a study which compared the effects of a standard treatment A and two new treatments B and C for blood pressure, 36 individuals with high blood pressure were assigned at random to the treatments so that 12 individuals were assigned to each treatment. The reduction in blood pressure (mm/Hg) was recorded for each individual one month after the start of the treatment.

(i) Determine the values $a, b, \ldots, e$ to complete the following analysis of variance table.

Source of variation	Degrees of freedom	Sums of squares	Mean squares	F
Treatment	a	c	e	6.095
Residual	b	d	2475.33	

(5 marks)

(ii) Test the null hypothesis of no differences between the mean effects of the treatments, stating your conclusion clearly. State the two models that you have compared to test this hypothesis, defining your notation clearly. For the hypothesis test you will need one of the following F-quantiles:

$$F_{2,12;0.95} = 3.885, \quad F_{2,33;0.95} = 3.285, \quad F_{3,12;0.95} = 3.490, \quad F_{3,36;0.95} = 2.866.$$

State which F distribution you have used, justifying the choice of the two degrees of freedom parameters.

(7 marks)

(iii) State any assumptions you have made in part (ii). Describe two plots you would use to check these assumptions, stating precisely what you would plot in each case.

(3 marks)

4 School A is known to be progressive whilst school B is traditional. In both schools children aged 8 are assigned at random to be taught by one of three methods to perform a simple task. In each school two children are selected at random from each teaching method, and timed performing the task. The following data, in coded units, were obtained.

		Method		
		1	2	3
School	A	2.2, 2.6	3.9, 3.4	4.5 , 4.0
	B	2.5, 3.2	3.1, 3.5	3.7,4.0

(i) Defining your notation carefully, write down the most complex model that can be fitted to these data. Specify any necessary parameter constraints. State the distribution of any error terms in your model. Give the fitted value for each combination of school and method.

(10 marks)

(ii) The data are stored in R, with the dependent variable stored in the vector `time` and independent variables stored in the variables `school` and `method`. The following R output is obtained.

```
> lm1<-lm(time~school*method)
> deviance(lm1)
[1] 0.7
> lm2<-lm(time~school+method)
> deviance(lm2)
[1] 1.155
```

Using this output, test for an interaction effect between school and method on time taken to perform the task, stating your conclusion clearly. State your null hypothesis clearly, referring to your model defined in part (i). You will need one of the following F-quantiles:

$$F_{2,6;0.95} = 5.143, \quad F_{2,12;0.95} = 3.885, \quad F_{6,6;0.95} = 4.283, \quad F_{6,12;0.95} = 2.996.$$

State which F distribution you have used, justifying the choice of the two degrees of freedom parameters.

(11 marks)

(iii) Suppose that each child's ability is measured before instruction: children are tested with a different task before instruction, and their times for performing this task are recorded. Write down a model for analysing the effectiveness of method and school, that takes into account each child's ability before instruction. Define any extra notation carefully.

(3 marks)

End of Question Paper