Data provided: Neave's Tables

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS Spring Semester 2010–2011

MAS273 Statistical Modelling 2 hours

Restricted Open Book Examination.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator which conforms to University regulations.
*Marks will be awarded for your best **three** answers. Total marks 90.*

1 A small study has been carried out to investigate the effect of dose of a particular drug on pulse rates of patients. In the study, each patient has his or her pulse measured, and is then given a particular dose of the drug. A second pulse measurement is taken one hour later, and the reduction in pulse rate is recorded.

Define d_i and r_i to be the i-th patient's dose received (in mg) and reduction in pulse (in beats per minute) respectively. The observed data are given in the following table.

Patient (i)	1	2	3	4	5	6	7	8	9	10
Dose (d_i)	10	10	20	20	30	30	40	40	50	50
Reduction (r_i)	4	5	8	9	8	10	9	10	9	11

Some summary statistics are as follows.

$$\sum_{i=1}^{10} r_i = 83, \quad \sum_{i=1}^{10} d_i r_i = 2730,$$

$$\sum_{i=1}^{10} d_i^2 = 11000, \quad \sum_{i=1}^{10} r_i^2 = 733.$$

(i) Fit a simple linear regression model to these data, with reduction in pulse rate as the dependent variable, and calculate the estimated error variance. ***(7 marks)***

(ii) Test the hypothesis that there is no relationship between dose and reduction in pulse. ***(4 marks)***

(iii) Calculate one suitable statistic to measure how well your model in (i) describes the variation in the observed pulse rate reductions, and comment briefly on your result. ***(3 marks)***

(iv) A simple linear regression model with the log of the reduction as the dependent variable is fitted to these data. The estimated error variance is 0.005 (to 3 d.p.). Does this imply a better model compared to your model fitted in part (a)? Justify your answer. ***(2 marks)***

(v) A doctor is concerned about the size of the reduction that could result from a dose of 50mg.

(a) Which would be more appropriate to report to the doctor: a 95% confidence interval for the mean reduction given a 50mg dose, or a 95% prediction interval for a reduction given a 50mg dose? Justify your answer. ***(2 marks)***

(b) The doctor has plotted a histogram of the observed pulse rate reductions, notes that they may not be normally distributed, and questions the validity of the confidence and prediction intervals. How would you respond to the doctor's concern? ***(2 marks)***

1 (continued)

(vi) Below is some output from an R session, with `dose` defined as the vector $(x_1, \ldots, x_{10})$ and `reduction` defined as the vector $(y_1, \ldots, y_{10})$

```
> lm.1<-lm(reduction~dose+I(dose^2))
> deviance(lm.1)
 [1] 8.3
```

Investigate whether there is evidence of a quadratic relationship between pulse reduction and dose. ***(4 marks)***

(vii) Suppose the sex of each patient is recorded, and it is desired to fit a model in which the relationship between pulse reduction and dose is quadratic, and where the relationship may be different for males and females.

(a) Write down suitable notation to describe such a model, defining your notation carefully. ***(3 marks)***

(b) Suppose, in the table given above, patients 1 to 4 were female, and patients 5 to 10 were male. Are there any difficulties in fitting the model in part (a)? Justify your answer. ***(3 marks)***

2 (i) An agricultural trial is to be conducted to measure crop yields in two different fields, using different concentrations of fertilizer. Each field is split into two plots, with the possibility of a different concentration of fertilizer being used in each plot. Overall, two different concentrations are used only: a 'low' concentration and a 'high' concentration. The following model is proposed for the data.

$$y_i = \mu + \beta_1 f_i + \beta_2 c_i + \varepsilon_i,$$

where $i = 1, 2, 3, 4$ corresponding to the four plots, $f_i = -1$ or 1 corresponding to the two fields, $c_i = -1$ for a low concentration of fertilizer, $c_i = 1$ for a high concentration, and $\varepsilon_i \sim N(0, \sigma^2)$.

(a) Give an interpretation of the parameter β_2. ***(2 marks)***

(b) The following combinations of field and fertilizer concentration are proposed.

i	f_i	c_i
1	-1	-1
2	-1	-1
3	1	1
4	1	1

Demonstrate that it will not be possible to obtain least squares estimates of μ, β_1 and β_2 given $y_1, \ldots, y_4$ in this case. Give an intuitive explanation for this result. ***(9 marks)***

(c) An alternative design for the experiment is proposed:

i	f_i	c_i
1	-1	-1
2	-1	1
3	1	-1
4	1	1

Derive expressions for the least squares estimators of μ, β_1 and β_2 in terms of $y_1, \ldots, y_4$. State the variance of each estimator, and state the covariance between any two estimators. Verify that the least squares estimator of β_1 is unbiased. ***(11 marks)***

(ii) In an experiment, the concentration of a particular substance decays exponentially with rate λ. Starting with a concentration of N units at time 0, the concentration at time t is $N \exp(-\lambda t)$.

Suppose N is known at time 0, and three measurements of the concentration are taken at times t_1, t_2 and t_3. If the measurements are subject to measurement error, by considering a log transformation, derive an estimate of λ. ***(8 marks)***

3 (i) Waiting times (in number of days since referral by a GP) for hip replacement surgery have been recorded for 10 patients at each of five hospitals.

Define $y_{i,j}$ to be the waiting time for the j-th patient within hospital i, for $i = 1, \ldots, 5$ and $j = 1, \ldots, 10$. Then $\sum_{i=1}^{5}\sum_{j=1}^{10}(y_{i,j} - \bar{y}_{\bullet\bullet})^2 = 14141$ and $10\sum_{i=1}^{5}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 = 2180.9$.

(a) If $\bar{y}_{3\bullet} = 41.1$, calculate a 99% confidence interval for the mean waiting time at hospital $i = 3$. ***(6 marks)***

(b) Test the hypothesis that there is no difference between mean waiting times at the five hospitals. ***(6 marks)***

(c) Boxplots of the waiting times for each of the five hospitals are shown below.

What problem with your analysis in part (b) does this suggest? Defining $z_{ij} = \log y_{ij}$ and given $\sum_{i=1}^{5}\sum_{j=1}^{10}(z_{i,j} - \bar{z}_{\bullet\bullet})^2 = 16.6032$ and $10\sum_{i=1}^{5}(\bar{z}_{i\bullet} - \bar{z}_{\bullet\bullet})^2 = 3.1611$, perform any further test that you consider to be appropriate. ***(6 marks)***

 Question 3 continued on next page

3 (continued)

(ii) Consider the one-way ANOVA model

$$M_1 : \ y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

for $i = 1, 2$ and $j = 1, 2$, and with $\varepsilon_{ij} \sim N(0, \sigma^2)$. The constraint $\tau_1 + \tau_2 = 0$ is applied.

(a) Find the least squares estimator of τ_1. ***(6 marks)***

(b) In comparison with the model

$$M_2 : \ y_{ij} = \mu_i + \varepsilon_{ij},$$

for the same data, will the model M_1 have a different value for the residual sum of squares? Justify your answer. ***(6 marks)***

4 (i) A study by Heavenrich et al (1991) for the US Environmental Protection Agency investigated the relationship between average miles per gallon and variables such as weight and engine horsepower in various cars (data obtained from the Data and Story Library).

In an R session, the miles per gallon, weight (in kilograms) and horsepower for each car are stored under the variable names `mpg`, `wt` and `hp` respectively. There are 82 cars in the dataset.

Some edited output is given below.

```
> lm1<-lm(mpg~wt+hp)
> e<-stdres(lm1)
> fi<-fitted(lm1)
> plot(fi,e)
```

```
> lm2<-lm(log(mpg)~wt+hp)
```

(a) State what model has been fitted as `lm1`, giving suitable notation, explain what has been plotted, and explain the motivation for the definition of `lm2`. ***(7 marks)***

4 (continued)

(b) Some further edited R output is given below.

```
> lm2
Coefficients:
(Intercept)           wt           hp
   4.499302    -0.0006321    -0.001171

> deviance(lm2)
[1] 0.6360848

> lm3<-lm(log(mpg)~hp)
> lm3
Coefficients:
(Intercept)           hp
   4.013229    -0.004589

> deviance(lm3)
[1] 1.993003

> lm4<-lm(log(mpg)~wt)
> lm4
Coefficients:
(Intercept)           wt
    4.57248     -0.0007821

> deviance(lm4)
[1] 0.7464087
```

Conduct any suitable hypothesis tests, stating clearly what is being tested in each instance, and summarise the relationship between fuel economy, horsepower and weight. ***(10 marks)***

4 (continued)

(ii) A trial has been conducted to investigate the effect of both diet and exercise on weight loss. In a balanced design, individuals are assigned to one of three diet plans and one of two exercise regimes. There are five people per combination of diet and exercise regime. The weight lost by each person is recorded at the end of the trial.

Some (edited) R output from the analysis of the data is given below.

```
               Df  Sum Sq Mean Sq F value
diet                0.266
exercise           38.073
diet:exercise       6.842
Residuals         119.406
```

Use the output to investigate the effect of both diet and exercise on weight loss, and summarise your findings. The trial investigator asks you if it is true that the diet plans do not result in any weight loss. What would your response be? ***(13 marks)***

End of Question Paper