Data provided: Neave's Tables

MAS273

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 (i) Concern has been expressed by a college admissions officer over the effectiveness of the present college entrance examination, which is a purely academic test, as an indicator of performance in the final examination. Define x_i and y_i to be the i-th student's entrance exam mark and final exam mark respectively. The following summary statistics have been computed.

$$\sum_{i=1}^{20} x_i = 1285, \quad \sum_{i=1}^{20} x_i^2 = 84139, \quad \sum_{i=1}^{20} y_i = 1296,$$

$$\sum_{i=1}^{20} x_i y_i = 84270, \quad \sum_{i=1}^{20} y_i^2 = 87532.$$

(a) Fit a simple linear regression model to these data, with the final exam mark as the dependent variable. ***(7 marks)***

(b) Give an estimate of the error variance in your model in part (a). ***(3 marks)***

(c) Based on your model in part (a), calculate the proportion of variation in the final examination marks that is explained by the variation in the entrance examination marks, and comment on your result. ***(3 marks)***

(d) Conduct a suitable hypothesis test to address the admissions officer's concern. Give a brief statement to respond to the admissions officer, following the results of your test. ***(8 marks)***

(e) State any assumptions you have made in part (d), and give one procedure to check the validity of your assumptions. ***(2 marks)***

(ii) A new entrance exam based on personality and aptitude as well as academic ability has been tried on a further random sample of twenty students. The data are analysed in R, with `entrance` a vector of the twenty entrance exam marks, and `final` a vector of the corresponding final exam marks. Edited output from an R session is given below.

```
Call: lm(formula = final ~ entrance)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.6480     8.9531   1.413    0.175
entrance      0.9090     0.1478   6.148 8.33e-06

Residual standard error: 9.708 on 18 degrees of freedom
Multiple R-Squared: 0.6774
```

(a) Assess the performance of the new entrance exam in predicting the final exam mark, using all the information in the R output that you judge to be relevant. ***(7 marks)***

2 (i) Each of the following equations is a mathematical model for the relation between variables y_i and x_i, for $i = 1, \ldots, n$. Each model depends on two unknown parameters β_0 and β_1, and the error term ε_i is normally distributed. State which of these models are linear models, and specify the design matrix X for each linear model.

(a) $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$.

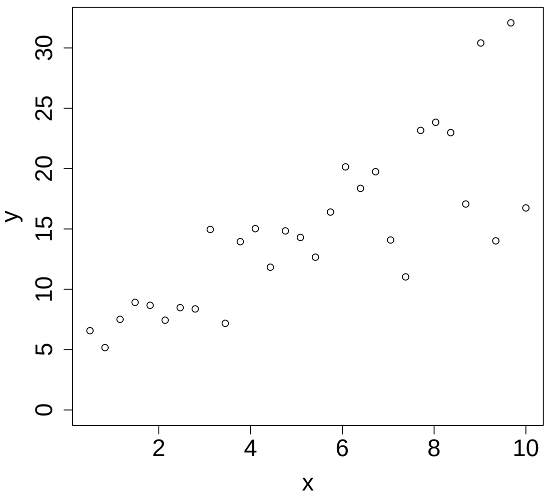
(b) $y_i = \beta_0 + \beta_1 \sin x_i + \varepsilon_i$.

(c) $y_i = \beta_0 + \cos(\beta_1 x_i) + \varepsilon_i$.

(d) $y_i = \Phi(\beta_0 + \beta_1 x_i) + \varepsilon_i$, where Φ is the $N(0,1)$ distribution function.

(6 marks)

(ii) Suppose the simple linear regression model is fitted to the data plotted below



(a) Without detailed calculation, roughly sketch the corresponding plot of standardised residuals against fitted values. ***(4 marks)***

(b) Explain why the usual model assumptions may not be valid in this case, and suggest one possible transformation of the data to deal with this problem. ***(3 marks)***

(c) Why is it not sensible to use a plot of standardized residuals against the observed values? ***(3 marks)***

2 (continued)

(iii) Suppose that the simple linear regression model is written in the form

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i,$$

for $i = 1, \ldots, n$ with $\varepsilon_i \sim N(0, \sigma^2)$ and $\bar{x} = \sum_{i=1}^{n} x_i/n$.

(a) Give an interpretation of the parameter β_0. ***(1 mark)***

(b) Write this model in matrix notation $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and hence obtain the least squares estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0,\ \hat{\beta}_1)^T$, using the result that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. ***(8 marks)***

(c) What is the covariance between the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$? ***(2 marks)***

(d) Compare your estimator of $\hat{\beta}_1$ with the least squares estimator of θ_1 in the simple linear regression model

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i.$$ ***(3 marks)***

3 (i) In an analysis of vitamin C concentration in orange juice, 240g servings of four different brands are examined for their vitamin C content. Five separate servings per brand are examined. Define $y_{i,j}$ to be the vitamin C content (in mg) in the j-th serving within brand i, for $i = 1, \dots, 4$ and $j = 1, \dots, 5$. Then

$$\sum_{i=1}^{4}\sum_{j=1}^{5}(y_{i,j} - \bar{y}_{\bullet\bullet})^2 = 181.80 \text{ and } \sum_{i=1}^{4}\sum_{j=1}^{5}(y_{i,j} - \bar{y}_{i\bullet})^2 = 91.60.$$

(a) Complete the following ANOVA table:

Source	Df	Sum of Sq	Mean Sq	F
brand				
error				

(7 marks)

(b) Test the hypothesis that there is no difference between mean vitamin C content for the four brands. Give appropriate bounds for the p-value.

(4 marks)

(c) Suppose each sample is stored in an open container for one hour, either at room temperature or in a refrigerator, before the vitamin C content is measured. Suggest a single model to describe the relationship between vitamin C content and both brand and storage conditions, including an interaction effect, defining all notation carefully. Specify any necessary parameter constraints.

(5 marks)

(ii) Nine patients are divided at random into three groups. Suppose that each group receives a different treatment for a month and the data below indicate the individuals' responses to the treatments.

Group 1: 9, 10, 11;
Group 2: 12, 16, 17;
Group 3: 10, 10, 13.

(a) Is there any evidence that mean responses differ between the three groups? ***(10 marks)***

(b) Give a 90% confidence interval for the mean response in group 3.
(4 marks)

4 (i) Two training methods for novice racing drivers are compared in an experiment. Twenty drivers have their best lap times measured at the start of the experiment. Half the drivers are trained using method A (group $i = 1$), and the remainder are trained using method B (group $i = 2$). The drivers have their best lap times measured again after their training programmes have finished. Define x_{ij} and y_{ij} to be driver i, j's best lap times before and after training respectively, with $i = 1, 2$ and $j = 1, \ldots, 10$.

Below is edited output from an R session, in which two models are fitted to the data. The variables `before`, `after`, and `method` represent the pre-training best lap time, the post-training best lap time, and the training method respectively.

Model 1:

```
Call: lm(formula = after ~ before)

Coefficients:
(Intercept)       before
      0.108        0.945
Residual standard error: 3.017 on 18 degrees of freedom
```

Model 2 (the `method` variable is specified as `A` or `B` in R):

```
Call: lm(formula = after ~ before + method)

Coefficients:
(Intercept)       before      methodB
     0.7885       0.9688      -4.0729
Residual standard error: 2.196 on 17 degrees of freedom
```

(a) Give equations in terms of x_{ij} and y_{ij} of the two models that have been fitted, stating any parameter constraints. In each case, state separately the estimates of the model parameters, including the error variance. ***(9 marks)***

(b) Perform one appropriate hypothesis test given the output. State the hypothesis that you are testing, and interpret the outcome. ***(6 marks)***

4 (continued)

(ii) A trial has been conducted to investigate the effects of both choice of treatment and diet on blood pressure. Each patient is assigned to one of three treatments and one of four diets, and the patient's blood pressure is recorded at the end of the trial. There are 48 patients in total, and a balanced design across treatment and diet has been used.

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

```
               Df  Sum Sq Mean Sq F value
treatment                 406.58
diet                      545.89
treatment:diet             15.76
Residuals                  46.66
```

(a) Complete the above ANOVA table ***(6 marks)***

(b) Perform any tests of significance that you feel are appropriate. Comment on the result of any test, and so provide an assessment of the effect of both treatment and diet on blood pressure. You may use the following R output to help you.

```
> qf(0.95,6,36)
[1] 2.363751
qf(0.99,3,36)
[1] 4.377096
> qf(0.99,2,36)
[1] 5.247894
```

(9 marks)

End of Question Paper