THE UNIVERSITY OF MELBOURNE

Semester 1 Assessment, 2015

Department of Mathematics and Statistics

MAST30025 Linear Statistical Models

Writing time: 3 hours

Reading time: 15 minutes

This is NOT an open book exam

This paper consists of 6 pages (including this page)

Authorised materials:

- Scientific calculators are premitted, but not graphical calculators.
- One A4 double-sided handwritten sheet of notes.

Instructions to Students

- You may remove this question paper at the conclusion of the examination
- You should attempt all questions. Marks for individual questions are shown.
- The total number of marks available is 69.

Instructions to Invigilators

- Students may remove this question paper at the conclusion of the examination

Blank page (ignored in page numbering)

Question 1 (12 marks) Consider the linear model

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where the ϵ_i are i.i.d. $N(0, \sigma^2)$.

(a) Write the model in matrix form

(b) Write down the normal equations and explain two ways they can be obtained (you do not have to derive them).

(c) Show that for this model we have

$$\begin{aligned} \hat{\alpha} &= \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2}\left(\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i\right) \\ \hat{\beta} &= \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2}\left(n\sum_i x_i y_i - \sum_i x_i \sum_i y_i\right) \end{aligned}$$

(d) Show that if $x_i = \sum_i x_i/n$ then the i-th observation has leverage $1/n$.

Question 2 (11 marks)

(a) State the general linear hypothesis for a linear model of full rank.

(b) Give a test statistic for the general linear hypothesis against a general alternative.

What is the distribution of the test statistic under the null?

List the main steps required to prove this, starting from the model assumptions. A complete proof is not required: half a page should prove sufficient.

(c) How does the test statistic behave when the null is not true?

Justify your claim.

Question 3 (6 marks)

(a) Define the conditional inverse of a matrix.

(b) Obtain **all** conditional inverses of

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

(c) For A as above, find a conditional inverse B such that

$$BAB \neq B.$$

Question 4 (8 marks)

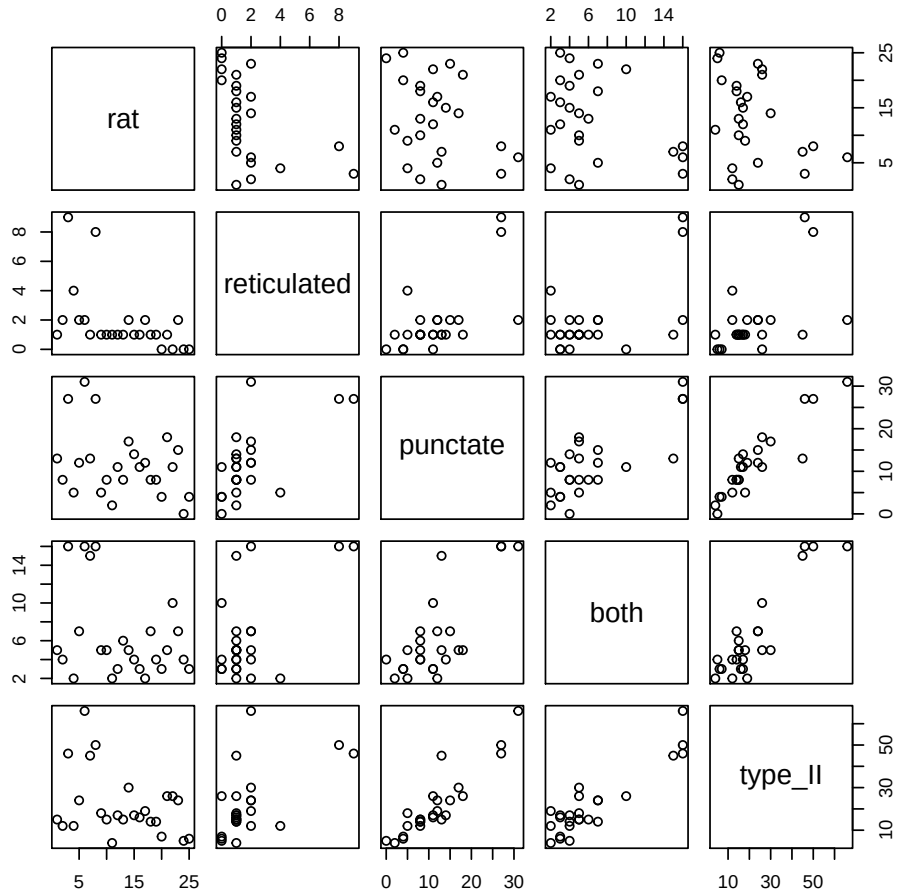
(a) What is a confounding factor, and describe four ways in which an experimental design can (try to) deal with such a thing. Give a brief description of each principle (no more than a quarter of a page on each).

(b) What type of model should be used to analyse the results from a (Randomised) Complete Block Design?

Define the reduced normal equations for this model.

Question 5 (19 marks) Counts of fibres in skeletal tissue were made on 25 rats. There are two types of fibres: Type I and Type II. Type I fibres are further divided into three categories: (i) reticulated; (ii) punctate; and (iii) both reticulated and punctate. The aim was to set up a model which predicts the number of Type II fibres from the numbers of the three different categories of Type I fibre. The data is plotted below.

```
> fibres <- read.csv("rat_fibres.csv")
> pairs(fibres)
```



(a) Comment on the data. Is a linear model appropriate? Do you think any transformations will be required?

A linear model was fitted as follows:

```
> model <- lm(type_II ~ reticulated + punctate + both, data = fibres)
> summary(model)

Call:
lm(formula = type_II ~ reticulated + punctate + both, data = fibres)

Residuals:
   Min     1Q Median     3Q    Max
-6.537 -2.509 -0.756  1.280  7.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4135     1.5146  -0.933    0.361
reticulated  -0.8432     0.4925  -1.712    0.102
punctate      1.1563     0.1787   6.471 2.06e-06 ***
both          1.7525     0.2844   6.163 4.09e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.003 on 21 degrees of freedom
Multiple R-squared:   0.94,        Adjusted R-squared:  0.9314
F-statistic: 109.6 on 3 and 21 DF,  p-value: 5.466e-13
```

(b) The final line of the summary gives the results of an F-test comparing two models. What are those models?

Explain how the degrees of freedom are obtained.

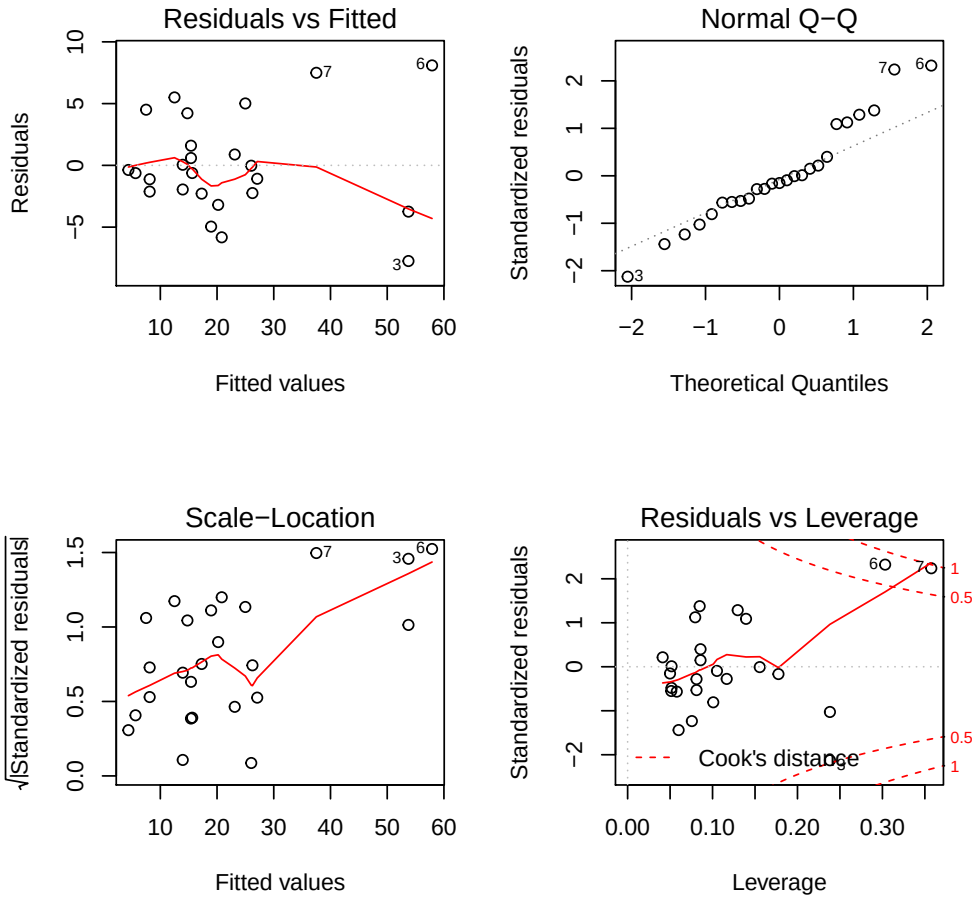
(c) What is the corrected sum of squares, that is $\sum(y_i - \bar{y})^2$, for this model? Show your working.

(d) What would you do to improve the predictive power of the model, and why?

(e) Suppose I used an F-test to compare the fitted model to the model `type_II ~ punctate + both`, what would be the value of the test statistic? What distribution would you compare it to (include the degree(s) of freedom)? What would be the p-value of the test?

We fitted a reduced model and then produced at some diagnostic plots.

```
> model2 <- lm(type_II ~ punctate + both, data = fibres)
> par(mfrow = c(2,2))
> plot(model2, which = 1)
> plot(model2, which = 2)
> plot(model2, which = 3)
> plot(model2, which = 5)
```



(f) Two observations are of concern. Identify these points, and explain why.

(g) Using the information below, given a 95% prediction interval for the number of type II fibres you would get from a rat with 10 punctate fibres and 10 fibres that are both reticulated and punctate.

```
> X <- cbind(1, fibres$punctate, fibres$both)
> round(solve(t(X) %*% X), 4)

        [,1]    [,2]    [,3]
[1,]  0.1403 -0.0058 -0.0051
[2,] -0.0058  0.0017 -0.0022
[3,] -0.0051 -0.0022  0.0049

> round(qt(0.975, 20:25), 4)

[1] 2.0860 2.0796 2.0739 2.0687 2.0639 2.0595

> deviance(model2)

[1] 383.4728
```

```
> round(model2$coefficients, 4)

(Intercept)    punctate        both
    -1.0452      1.0407      1.6681
```

Question 6 (13 marks) An experiment is to be set up to test the effectiveness of a new teat disinfectant in controlling mastitis in dairy cows. The disinfectant is applied to the cow's four teats immediately after milking. There are only two treatments in the experiment: disinfectant, or no disinfectant. The disinfectant is applied as a spray by the experimenter. There are 24 cows available, and there are 4 sections of the milking shed where the treatment can be applied; 6 cows can fit into each section. Each section is managed by a different farm worker. Following a week of milking, each of the four teats on each cow is given an infection rating on a 7-point scale. Here are six possible experimental designs:

(a) Twelve cows are randomly chosen to get the disinfectant.

(b) The two left or the two right teats on each cow are randomly chosen to get the disinfectant. (Assume that the teats respond to any treatment independently of each other.)

(c) Three cows in each section are randomly chosen to get the disinfectant.

(d) The first three cows to be milked in each section get the disinfectant.

(e) Two sections are randomly chosen, and the cows in those sections get the disinfectant.

(f) All 24 cows get the disinfectant, and the results are compared with measurements taken before the experiment.

For each of the six designs, state the following:

- What the experimental unit is;
- What type of design is used (completely randomised, randomised block, or neither). If it is a randomised block design, state the blocking factor;
- Any flaws in the experiment (statistically unsound aspects).

(g) Rank the designs from best to worst, with a full explanation of your ranking.

(h) For design (a), what is the response variable?

End of Exam—Total Available Marks = 69.