RESTRICTED OPEN BOOK EXAMINATION (Not to be removed from the examination hall)
Data provided: "Statistics Tables" by H.R. Neave

The University Of Sheffield.

MAS463

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2012–13**

Linear Models **2 hours**

Marks will be awarded for your best ***three*** *answers.*

RESTRICTED OPEN BOOK EXAMINATION
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator that conforms to University regulations.
There are 60 marks available on the paper.

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 An experiment is conducted to investigate the effect of vitamin C intake (0.5, 1, and 2 mg) and delivery method (orange juice or vitamin C supplement) on the tooth length (in mm) of guinea pigs. 10 guinea pigs are used at each of the levels of vitamin C and delivery method so that there are 60 guinea pigs in the experiment.

The following R output is available in which 'len' is the tooth length, 'dose' is the vitamin C intake and 'supp' is an indicator variable taking the value 0 if the dose was administered by orange juice and 1 if it was administered by vitamin C supplement:

```
> tooth1.lm<-lm(len~dose+I(dose^2)+supp)
> summary(tooth1.lm)

Call:
lm(formula = len ~ dose + I(dose^2) + supp)

Coefficients:
            Estimate  Std. Error  t-value   Pr(>|t|)
(Intercept)  -0.6400      2.9094   -0.220   0.826690
dose         30.1550      5.5467    5.437   1.23e-06 ***
I(dose^2)    -7.9300      2.1349   -3.714   0.000471 ***
supp         -3.7000      0.9883   -3.744   0.000429 ***
---
Residual standard error: 3.828 on 56 degrees of freedom
Multiple R-squared: 0.7623,     Adjusted R-squared: 0.7496
F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16

> tooth2.lm<-lm(len~dose+supp)
> anova(tooth2.lm)
Analysis of Variance Table

Response: len
          Df  Sum Sq Mean Sq F value    Pr(>F)
dose       1 2224.30 2224.30 123.989 6.314e-16 ***
supp       1  205.35  205.35  11.447  0.001301 **
Residuals 57 1022.56   17.94
```

(i) With reference to the R output, discuss the fit of the model `tooth1.lm` and the need for the parameters in the model. You should include discussion of the F-statistic and the associated p-value, the p-values for the parameters and the multiple R-squared value. State the null hypothesis for any hypothesis tests you refer to. ***(5 marks)***

(ii) Figure 1 shows some diagnostic residual plots for the `tooth1.lm` linear model. State the underlying assumptions for this linear model and comment on whether the plots support these assumptions. ***(3 marks)***

(iii) The plots in Figure 1 are based on the raw residuals ($y_i - \hat{y}_i$). State what other residuals might be more appropriate and why. ***(3 marks)***

 Question 1 continued on next page

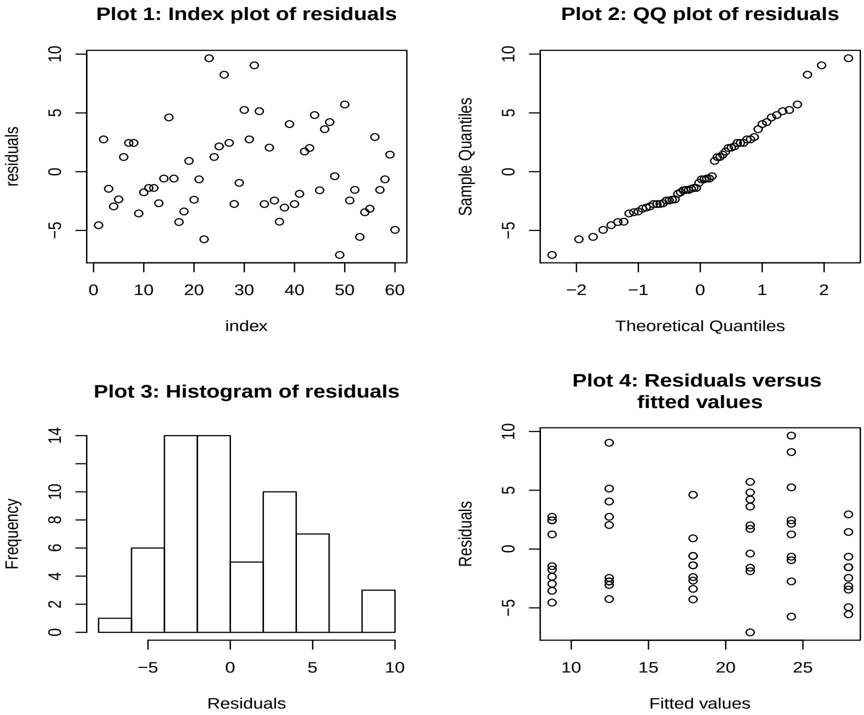

Figure 1: Residual plots for the `tooth1.lm` model

1 (continued)

(iv) Figure 2 shows the log-likelihood for the Box-Cox family of transformations for model `tooth1.lm` for values of λ between -1 and 3. Explain what the parameter λ represents and comment on what Figure 2 tells you about the need for a transformation of the response for the `tooth1.lm` model.

(2 marks)

(v) For the `anova(tooth2.lm)` command, state the null hypothesis for the two tests performed and describe the conclusion of each hypothesis test.

(3 marks)

(vi) Briefly describe how the partition sum of squares property is used in the hypothesis tests performed in the `anova(tooth2.lm)` command.

(4 marks)

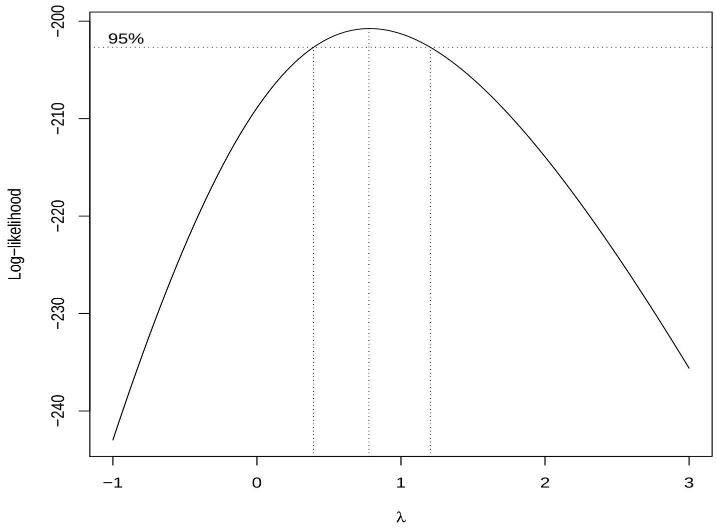

Figure 2: Log-likelihood function for the Box-Cox family of transformations for the `tooth1.lm` model.

2 The effect of the amount of a fertilizer ($F, grammes/m^2$) and the level of watering (W) on the yield ($Y, grammes/m^2$) of a tomato plant is studied. The exact level of watering is not known and is classified as low, medium and high. The data are listed in the first 3 columns of Table 1.

Table 1

Y	W	F	X	$Z1$	$Z2$	$Z3$
210	High	2.5	2	1	0	0
353	High	4.8	2	1	0	0
448	High	7.5	2	1	0	0
590	High	5.8	2	1	0	0
555	High	9.6	2	1	0	0
210	Med	5.2	1	0	1	0
352	Med	6.6	1	0	1	0
440	Med	10.1	1	0	1	0
621	Med	12.2	1	0	1	0
695	Med	14.8	1	0	1	0
247	Low	6.8	0	0	0	1
350	Low	9.8	0	0	0	1
346	Low	13.8	0	0	0	1
450	Low	15.3	0	0	0	1
560	Low	19.8	0	0	0	1

(i) A researcher analyses the data after first creating a new variable (X), as shown in Table 1, corresponding to the level of watering. She then uses the `R` command `lm1<-lm(Y~F+X)` to fit a linear model. Write down the statistical model for the ith observation being fitted by this `R` command.

(3 marks)

2 (continued)

(ii) A statistician recommends that the data should be reanalyzed. However, in her analysis she represents the different levels of watering by three new variables: $Z1$, $Z2$ and $Z3$ shown in Table 1. She uses the `R` command `lm2<-lm(Y~F+Z2+Z3)` to fit a linear model. Write down the statistical model for the ith observation being fitted by this `R` command and interpret the model parameters in terms of the expected yield of tomatoes.

(7 marks)

(iii) The researcher then uses the `R` command `lm3<-lm(Y~F+Z1+Z2+Z3)` to fit a linear model. Discuss what is wrong with this model. ***(3 marks)***

(iv) State, with justification, how the researcher could modify the linear predictor of model `lm3` to allow an additive combination of F, $Z1$, $Z2$ and $Z3$ to be included explicitly. Give an `R` command to do this. ***(2 marks)***

(v) Suppose another researcher records n observations at each of the three levels of watering (high, medium and low). The statistician fits a model with the `R` command `lm3<-lm(Y~Z1)`. Let $\bar{y_1}, \bar{y_2}$ and $\bar{y_3}$ represent the sample mean of the observations at high, medium and low levels of watering respectively. By first specifying the $3n$ by 2 design matrix X, derive an expression for the least squares estimate of the parameter for Z_1 in this model in terms of $\bar{y_1}, \bar{y_2}, \bar{y_3}$ and n. ***(5 marks)***

3 A statistician is asked to analyse data from a chemical-making company. Each day for 21 days, the following covariates are recorded:

- air - air flow
- temp - water temperature
- conc - acid concentration
- yield - amount of ammonia produced

(i) Some R output generated by the statistician is given below. Describe what is being done and what the conclusions are in each part of the R output. What does the output say about the relationship between the amount of ammonia produced and the air flow, water temperature and acid concentration?

(6 marks)

```
> int.lm<-lm(yield~1)
> step(int.lm,scope=list(upper=yield~air*temp+air*acid+acid*temp),
+ direction="forward")
Start:  AIC=98.4

yield ~ 1
       Df Sum of Sq     RSS    AIC
+ air   1    1750.1  319.12 61.142
+ temp  1    1586.1  483.15 69.852
+ acid  1     330.8 1738.44 96.741
<none>              2069.24 98.399
Step:  AIC=61.14

yield ~ air
       Df Sum of Sq    RSS    AIC
+ temp  1   130.321 188.80 52.119
<none>              319.12 61.142
+ acid  1     9.979 309.14 62.475
Step:  AIC=52.12

yield ~ air + temp
           Df Sum of Sq    RSS    AIC
+ air:temp  1    38.563 150.23 49.321
<none>                  188.79 52.119
+ acid      1     9.965 178.83 52.980
Step:  AIC=49.32
```

 Question 3 continued on next page

3 (continued)

```
yield ~ air + temp + air:temp
       Df Sum of Sq    RSS    AIC
<none>              150.23 49.321
+ acid  1   0.93534 149.30 51.190

Call:
lm(formula = yield ~ air + temp + air:temp)

Coefficients:
(Intercept)          air         temp     air:temp
   22.29030     -0.51551     -1.93006      0.05176
```

(ii) Further partial R output generated by the statistician is given below. Describe what this output says about the relationship between the amount of ammonia produced and the air flow, water temperature and acid concentration. ***(4 marks)***

```
> yield.amm<-regsubsets(yield~air*temp+air*acid+acid*temp)
> summary(yield.amm)
Subset selection object

1 subsets of each size up to 6
Selection Algorithm: exhaustive
         air temp acid air:temp air:acid temp:acid
1  ( 1 ) " " " "  " "  "*"      " "      " "
2  ( 1 ) " " " "  " "  "*"      " "      "*"
3  ( 1 ) " " "*"  " "  "*"      "*"      " "
4  ( 1 ) " " " "  "*"  "*"      "*"      "*"
5  ( 1 ) "*" "*"  "*"  "*"      "*"      " "
6  ( 1 ) "*" "*"  "*"  "*"      "*"      "*"

> summary(yield.amm)$rsq
[1] 0.9193685 0.9257151 0.9276716 0.9300364 0.9336477 0.9337217
> summary(yield.amm)$cp
[1] 0.031833 0.691239 2.277976 3.778440 5.015631 7.000000
> summary(yield.amm)$bic
[1] -46.78614 -45.46323 -42.97921 -40.63279 -38.70118 -35.68009
```

(iii) Other than using the `step` and `regsubsets` commands, what other statistical method(s) could the statistician use to assess the relationship between the amount of ammonia produced and the air flow, water temperature and acid concentration? Briefly outline the advantages and disadvantages of the `step` and `regsubsets` methods as well as the other method(s) you suggest. ***(5 marks)***

 Question 3 continued on next page

3 (continued)

(iv) The statistician decides that the data support the linear model $yield_i = \beta_0 + \beta_1(air_i) + \beta_2(temp_i) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$. Based on this model, the statistician forms a 95% confidence interval for β_1 and a 95% confidence interval for β_2. Let $(b_1 - a, b_1 + a)$ and $(b_2 - c, b_2 + c)$ represent the 95% confidence intervals for β_1 and β_2 respectively. Derive the 95% confidence interval for $\beta_1 + 2\beta_2$ in terms of b_1, b_2, a and c and any other terms you need. Explain how you could obtain the value of any other terms you need. ***(5 marks)***

4 Consider the linear model

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{0}_{n\times p} \\ \boldsymbol{0}_{n\times p} & \boldsymbol{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon}$$

in which $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are both $n \times p$ matrices; $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $p \times 1$ vectors; $\boldsymbol{0}_{n\times p}$ is an n by p matrix of zeroes; $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{2n})$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

(i) Let $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$. Suppose we want to test the null hypothesis that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ and that we want to write this hypothesis in the form $\boldsymbol{C\beta} = \boldsymbol{c}$. Specify the matrix C (in terms of $\boldsymbol{I}_p$) and specify the vector $\boldsymbol{c}$ for this hypothesis. ***(3 marks)***

(ii) Let $X = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{0}_{n\times p} \\ \boldsymbol{0}_{n\times p} & \boldsymbol{X}_2 \end{pmatrix}$. Show that $C(X^TX)^{-1}C^T = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}$. ***(5 marks)***

(iii) Suppose that σ^2 is an **unknown** constant. Show, by using an expression given in the course notes or otherwise, that a test statistic for the null hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ is given by $(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)^T[(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}]^{-1}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)/p\hat{\sigma}^2$ where you should give an expression for $\hat{\sigma}^2$ in terms of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. ***(6 marks)***

(iv) What is the distribution of the test statistic in (iii) under the null hypothesis? Explain how to determine the p-value using this test statistic. ***(3 marks)***

(v) Suppose that σ^2 is a **known** constant. Without further calculation, state the test statistic for testing the null hypothesis that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ and state its distribution under the null hypothesis. ***(3 marks)***

End of Question Paper