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

MAS463

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2010–11**

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 99 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 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. Figure ?? shows a plot of tooth length versus dose by delivery method. Consider the following linear model:

$$y_i = \beta_0 + \beta_1 dose_i + \beta_2 OJ_i + \epsilon_i$$

where y_i is tooth length of guinea pig i, $dose_i$ is the vitamin C dose of guinea pig i, OJ_i is an indicator variable for guinea pig i taking the value 1 if the dose was administered by orange juice and zero otherwise and ϵ_i has a $N(0, \sigma^2)$ distribution. The following `R` output is available:

```
> tooth.lm<-lm(len~dose+OJ)

> summary(tooth.lm)
Call:
lm(formula = len ~ dose + OJ)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2725     1.2824   7.231 1.31e-09
dose          9.7636     0.8768  11.135 6.31e-16
OJ           -3.7000     1.0936  -3.383   0.0013
---
Residual standard error: 4.236 on 57 degrees of freedom
Multiple R-squared: 0.7038,     Adjusted R-squared: 0.6934
F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16

> vcov(tooth.lm)
            (Intercept)   dose        OJ
(Intercept)    1.644    -0.897    -0.598
dose          -0.897     0.769         0
OJ            -0.598         0     1.196

> influence(tooth.lm)$hat
     1     2     3     4     5     6     7     8     9    10    11
0.052 0.052 0.052 0.052 0.052 0.052 0.052 0.052 0.052 0.052 0.035
   12    13    14    15    16    17    18    19    20    21    22
0.035 0.035 0.035 0.035 0.035 0.035 0.035 0.035 0.035 0.063 0.063
   23    24    25    26    27    28    29    30    31    32    33
0.063 0.063 0.063 0.063 0.063 0.063 0.063 0.063 0.052 0.052 0.052
   34    35    36    37    38    39    40    41    42    43    44
0.052 0.052 0.052 0.052 0.052 0.052 0.052 0.035 0.035 0.035 0.035
   45    46    47    48    49    50    51    52    53    54    55
0.035 0.035 0.035 0.035 0.035 0.035 0.063 0.063 0.063 0.063 0.063
   56    57    58    59    60
0.063 0.063 0.063 0.063 0.063
```

 Question 1 continued on next page

1 (continued)

(i) For the `tooth.lm` linear model fitted above, the residual for observation 6 is -0.454; provide an `R` command that would give the residual for observation 6. ***(2 marks)***

(ii) Calculate the 6th scaled residual. ***(5 marks)***

(iii) Calculate the 6th standardized residual. ***(3 marks)***

(iv) Calculate the 6th standardized deletion residual. ***(4 marks)***

(v) Fig **??** shows some diagnostic plots for the model fitted. Comment on whether the plots indicate that the assumptions of the linear model are met. ***(3 marks)***

(vi) Calculate a 95% confidence interval for β_1. ***(6 marks)***

(vii) Let X be the design matrix in which the first column has a one in every position, the second column contains the vitamin C dose and the third column contains the indicator variable values for delivery method as described earlier in the question. Given that the variance-covariance matrix of $\hat{\beta}$ given in the `R` output above is $\hat{\sigma^2}(X^TX)^{-1}$, perform a single hypothesis test to test whether both $\beta_0 = 2\beta_1$ and $\beta_2 = 0$. You may find the following helpful:

$$\left(\begin{pmatrix} 1 & -2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1.64 & -0.90 & -0.60 \\ -0.90 & 0.77 & 0 \\ -0.60 & 0 & 1.20 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -2 & 0 \\ 0 & 1 \end{pmatrix}\right)^{-1} = \begin{pmatrix} 0.12 & 0.06 \\ 0.06 & 0.87 \end{pmatrix}$$

. ***(10 marks)***

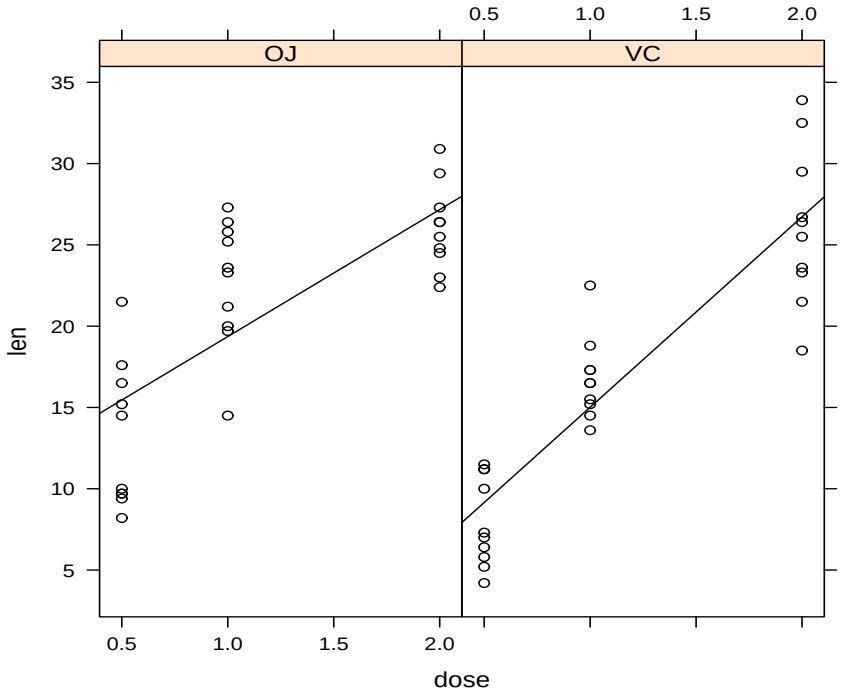

Figure 1: Tooth length against vitamin C dose by delivery method

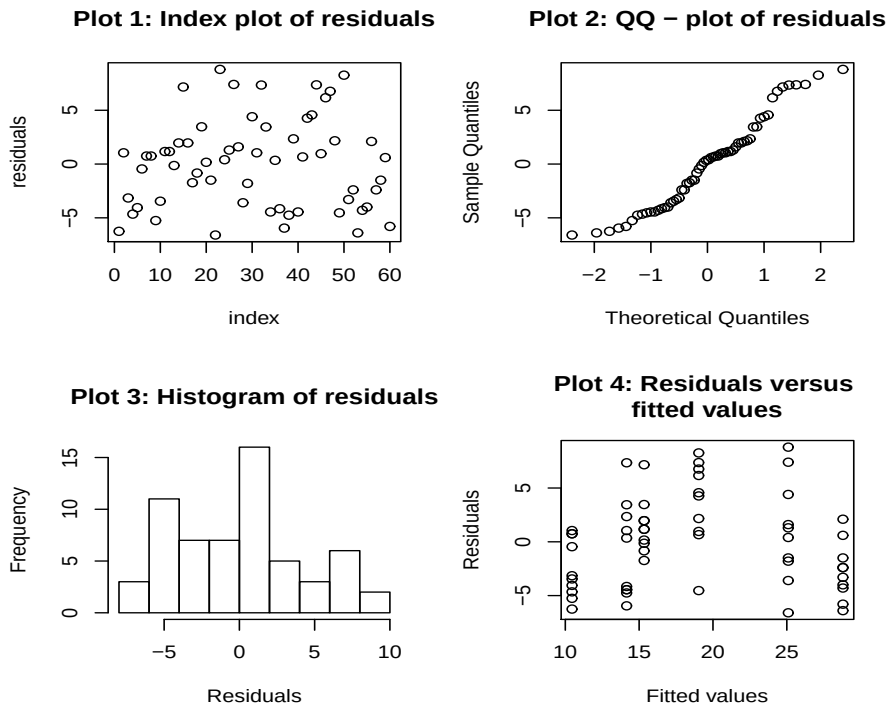

Figure 2: Residual plots for the guinea pig data

2 The data in this question relate to how protein concentration (conc) can be predicted from the optical density of an experimental assay. Output from an R session is given below. Some of the numerical values have been replaced by letters.

```
> DNase
          conc density
1   0.04882812   0.017
2   0.04882812   0.018
3   0.19531250   0.121
4   0.19531250   0.124
5   0.39062500   0.206
6   0.39062500   0.215
7   0.78125000   0.377
8   0.78125000   0.374
9   1.56250000   0.614
10  1.56250000   0.609
11  3.12500000   1.019
12  3.12500000   1.001
13  6.25000000   1.334
14  6.25000000   1.364
15 12.50000000   1.730
16 12.50000000   1.710

> DNase.lm<-lm(conc~density)

> summary(DNase.lm)
Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  -1.3010     0.5878       A      0.044
density       6.5100          B   9.885   1.08e-07
---
Residual standard error: C on D degrees of freedom
Multiple R-squared: E,     Adjusted R-squared: G
F-statistic: 97.71 on 1 and 14 DF,  p-value: 1.078e-07

> anova(DNase.lm)
Analysis of Variance Table
Response: conc
          Df  Sum Sq  Mean Sq  F-value     Pr(>F)
density    1 229.360  229.360   97.712  1.078e-07
Residuals 14  32.862        H
```

2 (continued)

```
> vcov(DNase.lm)
             (Intercept)    density
(Intercept)   0.3455301 -0.2936573
density      -0.2936573  0.4337225

> DNase.lm$resid[1:8]
     1      2    3      4      5      6      7      8
 1.239  1.233    J  0.689  0.351  0.292 -0.372 -0.353

> DNase.lm2<-lm(conc~1)

> anova(DNase.lm2,DNase.lm)
Analysis of Variance Table

Model 1: conc ~ 1
Model 2: conc ~ density
  Res.Df     RSS Df Sum of Sq        F      Pr(>F)
1     15       K
2     14  32.862  1    229.36   97.712   1.078e-07 ***

> influence(DNase.lm)$hat[2]
        2
        L
```

(i) Calculate the value of A. *(2 marks)*

(ii) Calculate the value of B. *(3 marks)*

(iii) Calculate the value of C. *(5 marks)*

(iv) Calculate the value of D. *(1 mark)*

(v) Calculate the value of E. *(3 marks)*

(vi) Calculate the value of G. *(5 marks)*

(vii) Calculate the value of H. *(2 marks)*

(viii) Calculate the value of J. *(4 marks)*

(ix) Calculate the value of K. *(2 marks)*

(x) Calculate the value of L. *(6 marks)*

3 The linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is to be fitted by the least squares method using n observations; $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters and it is assumed that $\boldsymbol{X}$ is of full rank p. The errors $\boldsymbol{\varepsilon}$ are assumed to be independent and normally distributed and to have zero mean vector and covariance matrix $\sigma^2 I_n$, where $\sigma^2 > 0$. Let $\widehat{\boldsymbol{\beta}}$ be the LS estimate of $\boldsymbol{\beta}$.

(i) Starting with the definition of the residual sum of squares: $S_r = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$; show that S_r can be written as $S_r = \boldsymbol{y}^T\boldsymbol{M}\boldsymbol{y}$ and specify the matrix $\boldsymbol{M}$ in terms of $\boldsymbol{X}$. ***(5 marks)***

(ii) Show that $\boldsymbol{M}$ satisfies $\boldsymbol{M}^2 = \boldsymbol{M}$. ***(5 marks)***

(iii) Explain whether $\boldsymbol{M}$ is invertible. ***(3 marks)***

(iv) Prove that $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$. ***(3 marks)***

(v) Consider $Q = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. By noting that

$$Q = \{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^T\{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}$$

show that

$$Q = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

(5 marks)

(vi) Use Q to derive the distributions of Q/σ^2, $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/\sigma^2$ and $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{X}^T\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2$, stating any general results to which you appeal. ***(8 marks)***

(vii) Derive the distribution of the statistic

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/p}{\boldsymbol{y}^T\boldsymbol{M}\boldsymbol{y}/(n-p)}.$$

(4 marks)

4 An investigator collects data on 200 crabs. The following variables are measured:

- sex - the gender of the crab
- sp - the species (blue or orange)
- FL - frontal lobe size (mm)
- RW - rear width (mm)
- CL - shell length (mm)
- CW - carapace width (mm)
- BD - body depth (mm)

The interest is in how the frontal lobe size depends on the other variables collected. The investigator uses `R` to help choose a statistical model.

(i) In the following output, explain what the '`regsubsets`' command does and which of the various sized subsets you would select based on the outputs of the three '`summary`' commands.

```
> a<-regsubsets(FL~factor(sex)+factor(sp)+RW+CL+CW+BD)
> summary(a)
Subset selection object
Call:regsubsets.formula(FL~factor(sex)+factor(sp)+RW+CL+CW+BD)
6 Variables  (and intercept)

1 subsets of each size up to 6
Selection Algorithm: exhaustive
         factor(sex)M factor(sp)O RW  CL  CW  BD
1  ( 1 ) " "          " "         " " " " " " "*"
2  ( 1 ) " "          "*"         " " " " "*" " "
3  ( 1 ) "*"          "*"         " " "*" " " " "
4  ( 1 ) "*"          "*"         " " "*" "*" " "
5  ( 1 ) "*"          "*"         " " "*" "*" "*"
6  ( 1 ) "*"          "*"         "*" "*" "*" "*"
> summary(a)$rsq
[1] 0.975407 0.986325 0.987568 0.988003 0.988074 0.988087
> summary(a)$cp
[1] 202.433  27.548   9.404   4.354   5.213   7.000
> summary(a)$bic
[1] -730.466 -842.548 -856.314 -858.142 -854.021 -848.944
```

(8 marks)

4 (continued)

(ii) Describe what is being done at each stage of the R output below.

```
> crabs.lm<-lm(FL~1,data=crabs)
> step(crabs.lm,scope=list(upper=FL~factor(sex)+CW+BD+RW,
+ data=crabs),direction="both")
Start:  AIC=501.57
FL ~ 1

               Df Sum of Sq     RSS     AIC
+ BD            1   2371.45   59.79 -237.49
+ CW            1   2263.83  167.42  -31.57
+ RW            1   2000.00  431.24  157.67
<none>                      2431.24  501.57
+ factor(sex)   1      4.56 2426.68  503.19

Step:  AIC=-237.49
FL ~ BD

               Df Sum of Sq     RSS     AIC
+ RW            1      9.62   50.17 -270.59
+ factor(sex)   1      5.03   54.76 -253.07
+ CW            1      3.19   56.60 -246.47
<none>                      59.79 -237.49
- BD            1   2371.45 2431.24  501.57

Step:  AIC=-270.59
FL ~ BD + RW

               Df Sum of Sq    RSS     AIC
+ CW            1      0.58  49.59 -270.91
<none>                      50.17 -270.59
+ factor(sex)   1      0.24  49.92 -269.56
- RW            1      9.62  59.79 -237.50
- BD            1    381.07 431.24  157.67

Step:  AIC=-270.91
```

4 (continued)

```
FL ~ BD + RW + CW

               Df Sum of Sq     RSS      AIC
<none>                      49.588 -270.913
- CW            1     0.579  50.167 -270.593
+ factor(sex)  1     0.034  49.554 -269.052
- RW            1     7.011  56.599 -246.466
- BD            1    99.143 148.732  -53.235

Call:
lm(formula = FL ~ BD + RW + CW, data = crabs)

Coefficients:
(Intercept)           BD           RW           CW
    0.71812      0.82993      0.16991      0.02901
```

(9 marks)

(iii) The AIC for the chosen model using the '`steps`' command above is -270.91; show how to calculate this value based on the 'RSS' values in the output.

(3 marks)

(iv) The investigator looks at a plot for a particular statistical model and concludes that it shows evidence of heteroscedasticity.

(a) Explain what heteroscedasticity means in a linear model.

(2 marks)

(b) Justify which plot would allow heteroscedasticity to be observed and what you would expect to see. ***(3 marks)***

(c) Explain why heteroscedasticity is a problem in linear models.

(2 marks)

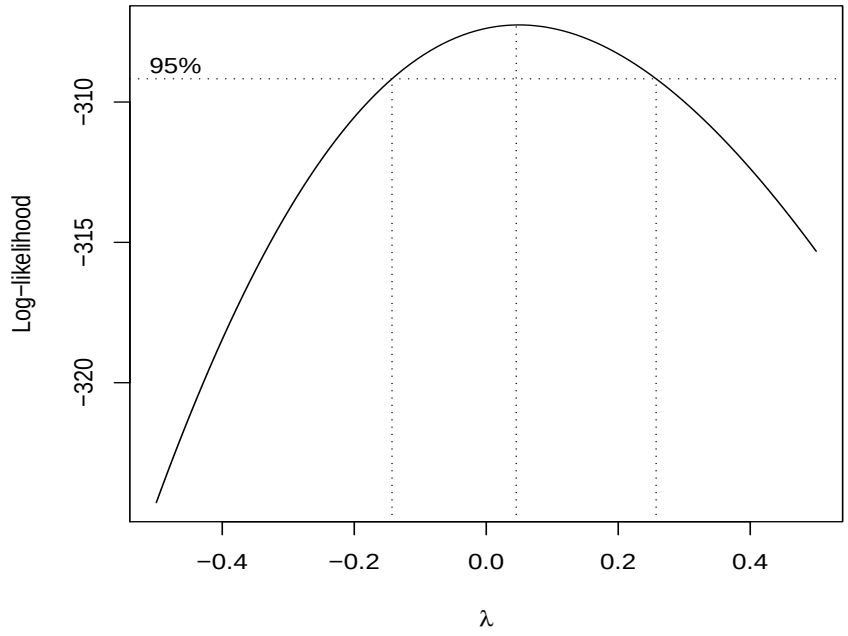

Figure 3: Log-likelihood function for the Box-Cox family of transformations

4 (continued)

(v) Figure **??** shows a plot of the log-likelihood as a function of the Box-Cox family parameter λ. Give a point estimate and a 95% confidence interval for λ based on the plot and hence explain whether the investigator's claim of heteroscedasticity is justified. ***(6 marks)***

End of Question Paper