The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2015–16**

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

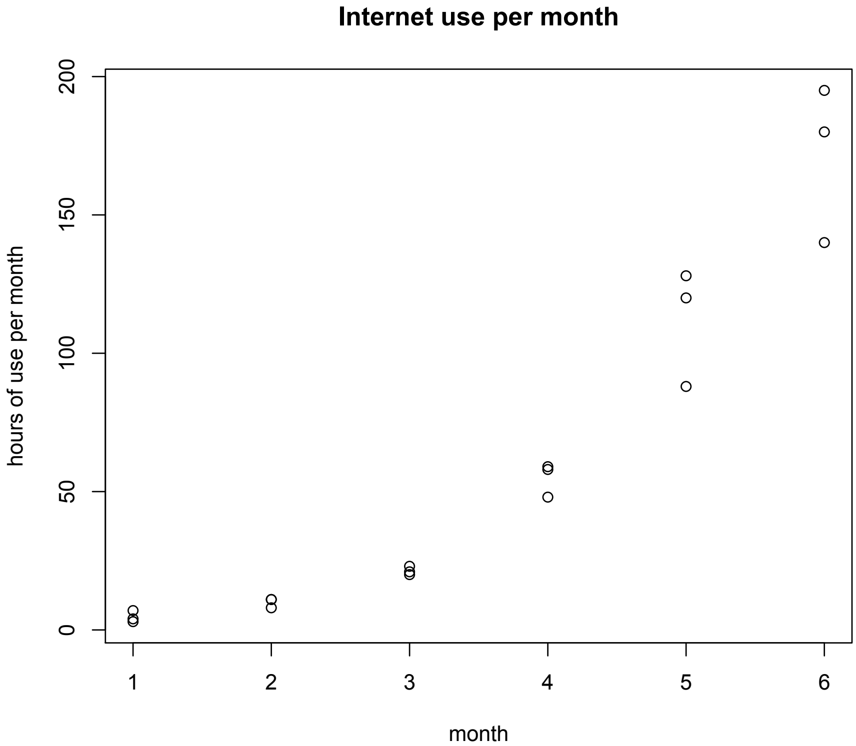

Figure 1: Plot of number of hours of internet use and months

1 A group of senior citizens who have never used the internet before are given training over a period of 6 months. A sample of 3 of them is chosen at random and their numbers of hours of internet use are recorded for the 6 months, as shown in Figure 1 above.

(i) Describe briefly the data, discussing any interesting features. Based on Figure 1 only suggest a possible linear model of the hours of use per month (as response variable) and month (as explanatory variable). ***(2 marks)***

 Question 1 continued on next page

1 (continued)

(ii) Let y be the hours of use per month and x be the month. An analysis in R gave the following output:

```
> summary(fit)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max
-33.393  -2.917   0.858   4.307  21.607

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.700     14.000   1.479   0.1599
x            -23.230      9.159  -2.536   0.0228 *
I(x^2)         8.113      1.281   6.334 1.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Residual standard error: 13.56
Multiple R-squared:  0.9602, Adjusted R-squared:  0.9549
F-statistic:   181 on 2 and 15 DF,  p-value: 3.152e-11
```

(a) Write down the fitted model. ***(2 marks)***

(b) Comment on the model and the quality of its goodness of fit, making appropriate reference to any goodness of fit diagnostics. State clearly any hypothesis you may use. ***(6 marks)***

(c) Using one of the following R extracts

```
> qnorm(0.95)                  > qt(0.95, df=14)
[1] 1.644854                  [1] 1.76131
> qt(0.95, df=15)              > qt(0.995, df=15)
[1] 1.75305                   [1] 2.946713
```

calculate 90% confidence intervals for the coefficient of x and for the coefficient of x^2. ***(3 marks)***

(d) For month $x = 1$ calculate a 90% predictive interval for the future observation y. You may use the following:

$$(X^T X)^{-1} = \begin{pmatrix} 1.066 & -0.650 & 0.083 \\ -0.650 & 0.456 & -0.063 \\ 0.083 & -0.063 & 0.009 \end{pmatrix},$$

where X is the design matrix of the linear model. ***(5 marks)***

 Question 1 continued on next page

1 (continued)

(e) A further R analysis gave

```
> vcov(fit)
            (Intercept)          x     I(x^2)
(Intercept)   196.00135 -119.43833  15.312606
x            -119.43833   83.89120 -11.484454
I(x^2)         15.31261  -11.48445   1.640636
```

Calculate the correlation coefficient of the estimator of the gradient (coefficient of x) and the estimator of the coefficient of x^2.

(2 marks)

2 A data-set on black cherry trees in the Allegheny National Forest, Pennsylvania, USA includes the height, radius (measured 4.5 feet above the ground) and volume, for each of 31 trees.

(i) A model

$$v_i = \beta_0 + \beta_1 r_i + \beta_2 h_i + \epsilon_i \tag{1}$$

has been proposed, where h_i, r_i, v_i are the logarithms of the height (in feet), radius (in feet) and volume (in cubic feet) of the ith tree, and $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. The following output summarizes the results of fitting this model in R.

```
> summary(cherry)

Call:
lm(formula = v ~ r + h)

Residuals:
      Min        1Q    Median        3Q       Max
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.33065    0.91031  -0.363    0.719
r            1.98265    0.07501  26.432  < 2e-16 ***
h            1.11712    0.20444   5.464 7.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared:  0.9777,    Adjusted R-squared:  0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16

> anova(cherry)
Analysis of Variance Table

Response: v
          Df Sum Sq Mean Sq F value    Pr(>F)
r          1 7.9254  7.9254 1196.53 < 2.2e-16 ***
h          1 0.1978  0.1978   29.86 7.805e-06 ***
Residuals 28 0.1855  0.0066
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Explain the hypothesis being tested by each of the three F statistics included in the output. What interpretation, if any, can be placed on their conclusions here? ***(6 marks)***

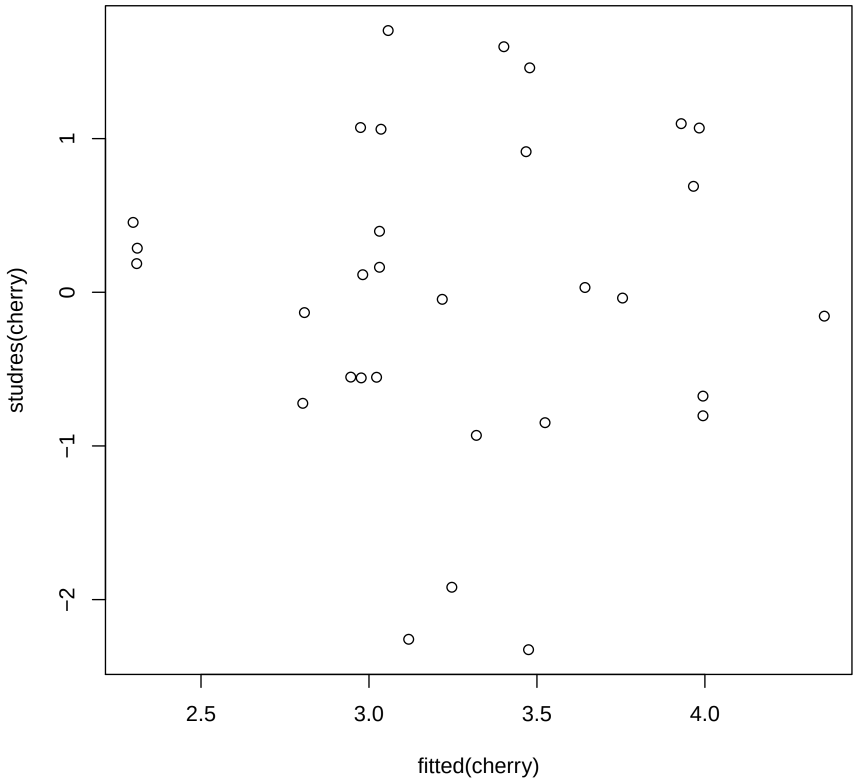

Figure 2: Standardized deletion residuals for the cherry tree model

2 (continued)

(ii) Figure 2 shows the standardized deletion residuals for the model above. The following calculations can be used as the basis of a test on the standardized deletion residuals, using the Šidák correction.

```
> alpha=0.05
> prob=1-(1-alpha)^(1/31)
> qt(prob/2,27)
[1] -3.495321
```

Explain the interpretation of the values `alpha` and `prob` used in the calculation, and carry out the test. ***(4 marks)***

2 (continued)

(iii) Thinking about the trunk of each tree as a cylinder, a simple geometric calculation suggests that

$$V_i \approx kR_i^2 H_i \qquad (2)$$

where $V_i = \exp(v_i)$ etc., and that $k \approx \pi$ (the usual circular constant). Explain why the model suggested by (2) can be represented as a special case of (1) under the null hypothesis that $\beta_1 = 2$ and $\beta_2 = 1$, and explain how that null hypothesis can be written in the general form

$$C\boldsymbol{\beta} = \boldsymbol{c}.$$

(3 marks)

Express the weaker hypothesis that $\beta_1 + \beta_2 = 3$ in a similar form, and calculate the corresponding F statistic, using the fact that

$$G = (X^T X)^{-1} = \begin{pmatrix} 125.1 & 5.839 & -28.07 \\ 5.839 & 0.8495 & -1.227 \\ -28.07 & -1.227 & 6.310 \end{pmatrix}.$$

What is the null distribution of this F statistic? ***(7 marks)***

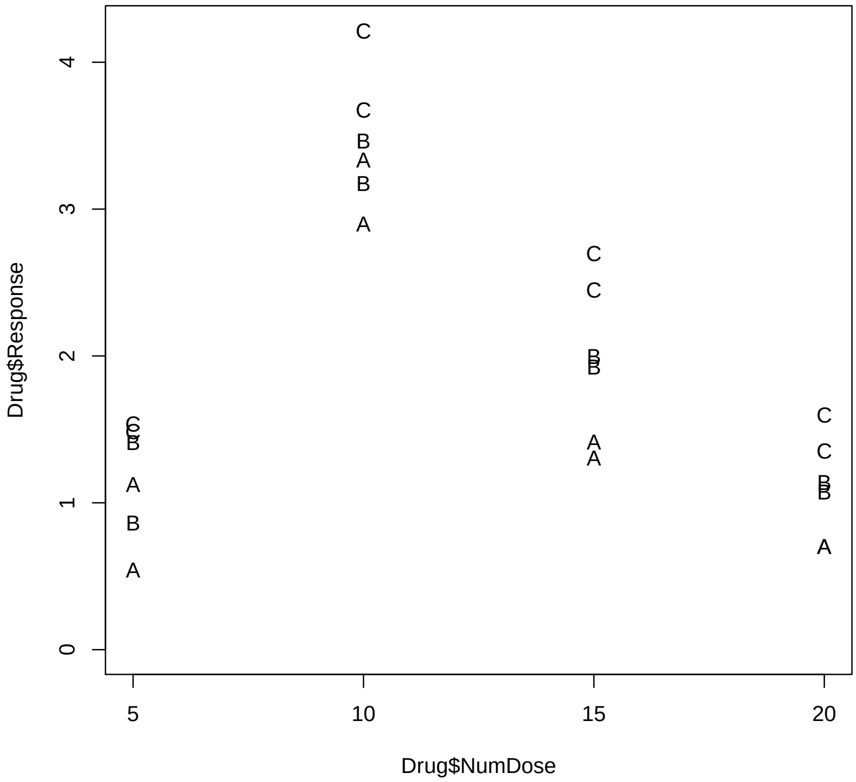

Figure 3: Raw data in the laboratory experiment in Question 3

3 A laboratory experiment is intended to investigate the effect of a drug on certain species of micro-organisms. Tissue cultures containing set amounts of one of three species of micro-organisms (A, B, C) are each exposed to doses of the drug being tested; there are four different doses used, and two replicates of each combination of species and dose. Figure 3 shows a plot produced in R of the dose and response for each run, the points being coded by species.

 Question 3 continued on next page

3 (continued)

(i) Various models are being considered for the response as a function of species and dose. The output below shows summaries of results for two models; `Response` and `Species` have the obvious meaning, `NumDose` refers to the dose as a quantitative variable, and `FacDose` refers to the dose as a factor variable.

```
> summary(FacModel)

Call:
lm(formula = Response ~ Species + FacDose, data = Drug)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25837 -0.15868  0.02226  0.07398  0.38053 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.74455    0.11078   6.721 2.66e-06 ***
SpeciesB     0.37777    0.11078   3.410  0.00312 ** 
SpeciesC     0.87320    0.11078   7.882 3.02e-07 ***
FacDose10    2.30045    0.12791  17.984 5.98e-13 ***
FacDose15    0.80540    0.12791   6.296 6.18e-06 ***
FacDose20   -0.06504    0.12791  -0.508  0.61730    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2216 on 18 degrees of freedom
Multiple R-squared:  0.9657,    Adjusted R-squared:  0.9562 
F-statistic: 101.3 on 5 and 18 DF,  p-value: 1.551e-12
```

 Question 3 continued on next page

3 (continued)

```
> summary(QuadModel)

Call:
lm(formula = Response ~ Species + NumDose + I(NumDose^2), data = Drug)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9074 -0.4361  0.1135  0.3315  0.9608

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.036322   0.698646  -2.915  0.00889 **
SpeciesB      0.377774   0.297904   1.268  0.22008
SpeciesC      0.873198   0.297904   2.931  0.00857 **
NumDose       0.758921   0.123549   6.143 6.64e-06 ***
I(NumDose^2) -0.031709   0.004865  -6.518 3.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5958 on 19 degrees of freedom
Multiple R-squared:  0.7381,    Adjusted R-squared:  0.6829
F-statistic: 13.39 on 4 and 19 DF,  p-value: 2.378e-05
```

(a) Give the equations for these two models, explaining your notation and assumptions. ***(6 marks)***

(b) Calculate the BIC for each of these two models. Based on the BIC, explain which of the two models you would prefer and why. ***(5 marks)***

(c) What advantages and disadvantages do these two modelling approaches have for this experiment, beyond those taken into account in the BIC? ***(2 marks)***

(d) Explain what you would expect to see if the model

`Response ~ Species + NumDose + I(NumDose^2) + I(NumDose^3)`

were fitted. ***(2 marks)***

 Question 3 continued on next page

3 (continued)

(ii) The output below shows the results of a possible approach to automated model selection for these data, using AIC.

```
> NullModel <- lm(Response ~ 1, data = Drug)
> step(NullModel,Response~Species+NumDose+I(NumDose^2)+FacDose)
Start:  AIC=3.69
Response ~ 1

               Df Sum of Sq     RSS     AIC
+ FacDose       3   21.8001  3.9519 -35.293
+ I(NumDose^2)  1    2.5444 23.2075   3.194
<none>                      25.7520   3.691
+ Species       2    3.0684 22.6836   4.646
+ NumDose       1    0.8570 24.8950   4.879

Step:  AIC=-35.29
Response ~ FacDose

          Df Sum of Sq     RSS     AIC
+ Species  2    3.0684  0.8836 -67.245
<none>                  3.9519 -35.293
- FacDose  3   21.8001 25.7520   3.691

Step:  AIC=-67.24
Response ~ FacDose + Species

          Df Sum of Sq     RSS     AIC
<none>                  0.8836 -67.245
- Species  2    3.0684  3.9519 -35.293
- FacDose  3   21.8001 22.6836   4.646

Call:
lm(formula = Response ~ FacDose + Species, data = Drug)

Coefficients:
(Intercept)  FacDose10  FacDose15  FacDose20   SpeciesB  SpeciesC
    0.74455    2.30045    0.80540   -0.06504    0.37777   0.87320
```

Explain whether the results shown agree with your choice in (c), and in general what issues might lead these approaches to reach similar or different conclusions. ***(5 marks)***

4 Consider the linear model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{3}$$

where ϵ_i is an i.i.d. sequence of random variables with zero mean and variance $\mathrm{Var}(\epsilon_i) = \sigma^2 c_i$, for some variance σ^2 and $c_i > 0$.

Discounted least squares considers the estimator $\hat{\boldsymbol{\beta}}$ that minimises the discounted sum of squares

$$S_\delta(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta^{n-i} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2,$$

for some discount factor δ that satisfies $0 < \delta \leq 1$.

(i) Show that discounted least squares is a special case of weighted least squares (WLS) and calculate the weights of WLS as functions of δ. ***(4 marks)***

(ii) Using the relationship of discounted least squares and WLS as in (i), derive the variance of ϵ_i as a function of σ^2 and δ. ***(3 marks)***

(iii) (a) Define y_i^*, $\boldsymbol{x}_i^*$ and ϵ_i^* as functions of y_i, $\boldsymbol{x}_i$, ϵ_i and δ so that the sum of squares $S(\boldsymbol{\beta})$ of the linear model

$$y_i^* = \boldsymbol{x}_i^{*T} \boldsymbol{\beta} + \epsilon_i^*, \quad \mathrm{Var}(\epsilon_i^*) = \sigma^2,$$

is the same as the discounted sum of squares $S_\delta(\boldsymbol{\beta})$ of model (3). ***(3 marks)***

(b) Use part (a) and the standard least squares estimator to derive the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ in model (3) as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \delta^{n-i} \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \sum_{i=1}^{n} \delta^{n-i} \boldsymbol{x}_i y_i.$$

(4 marks)

(iv) For the simple linear regression model with no intercept and a near constant covariate $x_i \approx x$, i.e.

$$y_i \approx x\beta + \epsilon_i,$$

show that

$$\hat{\beta} = \frac{(1-\delta)}{x(1-\delta^n)} \sum_{i=1}^{n} \delta^{n-i} y_i.$$

(6 marks)

End of Question Paper