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

```r
> 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)          20.70000    14.0000  1.4792  0.15992
x                   -23.23000     9.1590 -2.5357  0.02278 *
I(x^2)              8.1134700     1.2808  6.3339 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

```r
> qnorm(0.95)  > qt(0.95, df=14)
> qt(0.95, df=15)  > qt(0.995, df=15)
```

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:

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

where $X$ is the design matrix of the linear model. (5 marks)
(e) A further R analysis gave

\begin{verbatim}
> 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
\end{verbatim}

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

(2 marks)
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

\[ v_i = \beta_0 + \beta_1 r_i + \beta_2 h_i + \epsilon_i \]  

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 \( i \)th tree, and \( \epsilon_i \sim N(0, \sigma^2) \).

The following output summarizes the results of fitting this model in R.

> summary(cherry)

Call:
\( \text{lm(formula = v ~ r + h)} \)

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.168561</td>
<td>-0.048488</td>
<td>0.002431</td>
<td>0.063637</td>
<td>0.129223</td>
</tr>
</tbody>
</table>

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 ‘.’ 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

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>1</td>
<td>7.9254</td>
<td>7.9254</td>
<td>1196.53</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>h</td>
<td>1</td>
<td>0.1978</td>
<td>0.1978</td>
<td>29.86</td>
<td>7.805e-06 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>28</td>
<td>0.1855</td>
<td>0.0066</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

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

(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.

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

\[ V_i \approx k R_i^2 H_i \quad (2) \]

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

\[ C \beta = c. \]

(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)
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.

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

```r
> summary(FacModel)
```

**Call:**

```r
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
> 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)   -2.03632   0.69865  -2.915  0.00889 **  
SpeciesB       0.37777   0.29790   1.268  0.22008    
SpeciesC       0.87320   0.29790   2.931  0.00857 **  
NumDose        0.75892   0.12355   6.143  6.64e-06 ***  
I(NumDose^2)  -0.03171   0.00487  -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)
The output below shows the results of a possible approach to automated model selection for these data, using AIC.

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

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ FacDose</td>
<td>3</td>
<td>21.8001</td>
<td>3.9519</td>
</tr>
<tr>
<td>+ I(NumDose^2)</td>
<td>1</td>
<td>2.5444</td>
<td>23.2075</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td>25.7520</td>
<td>3.194</td>
<td></td>
</tr>
<tr>
<td>+ Species</td>
<td>2</td>
<td>3.0684</td>
<td>22.6836</td>
</tr>
<tr>
<td>+ NumDose</td>
<td>1</td>
<td>0.8570</td>
<td>24.8950</td>
</tr>
</tbody>
</table>

Step: AIC=-35.29
Response ~ FacDose

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Species</td>
<td>2</td>
<td>3.0684</td>
<td>0.8836</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td>3.9519</td>
<td>-67.245</td>
<td></td>
</tr>
<tr>
<td>- FacDose</td>
<td>3</td>
<td>21.8001</td>
<td>25.7520</td>
</tr>
</tbody>
</table>

Step: AIC=-67.24
Response ~ FacDose + Species

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>0.8836</td>
<td>-67.245</td>
<td></td>
</tr>
<tr>
<td>- Species</td>
<td>2</td>
<td>3.0684</td>
<td>3.9519</td>
</tr>
<tr>
<td>- FacDose</td>
<td>3</td>
<td>21.8001</td>
<td>22.6836</td>
</tr>
</tbody>
</table>

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)
Consider the linear model
\[ y_i = x_i^T \beta + \epsilon_i, \quad i = 1, 2, \ldots, n, \]  
(3)
where \( \epsilon_i \) is an i.i.d. sequence of random variables with zero mean and variance \( \text{Var}(\epsilon_i) = \sigma^2 c_i \), for some variance \( \sigma^2 \) and \( c_i > 0 \).

Discounted least squares considers the estimator \( \hat{\beta} \) that minimises the discounted sum of squares
\[ S_\delta(\beta) = \sum_{i=1}^{n} \delta^{n-i}(y_i - x_i^T \beta)^2, \]
for some discount factor \( \delta \) 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 \( \delta \). \hspace{1cm} (4 marks)

(ii) Using the relationship of discounted least squares and WLS as in (i), derive the variance of \( \epsilon_i \) as a function of \( \sigma^2 \) and \( \delta \). \hspace{1cm} (3 marks)

(iii) (a) Define \( y_i^*, x_i^* \) and \( \epsilon_i^* \) as functions of \( y_i, x_i, \epsilon_i \) and \( \delta \) so that the sum of squares \( S(\beta) \) of the linear model
\[ y_i^* = x_i^T \beta + \epsilon_i^*, \quad \text{Var}(\epsilon_i^*) = \sigma^2, \]
is the same as the discounted sum of squares \( S_\delta(\beta) \) of model (3). \hspace{1cm} (3 marks)

(b) Use part (a) and the standard least squares estimator to derive the estimator \( \hat{\beta} \) of \( \beta \) in model (3) as
\[ \hat{\beta} = \left( \sum_{i=1}^{n} \delta^{n-i} x_i x_i^T \right)^{-1} \sum_{i=1}^{n} \delta^{n-i} x_i y_i. \]
(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