MAS463

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2013–14**

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 The level of light output of light bulbs covered with two types of coating (A and B) is studied. Data on coating, length of operation and drop in light output are given in the following table.

Length of operation (hours)	Coating	Drop in light output (% of original output)
0	A	0
400	A	6
800	A	22
1200	A	27
1600	A	32
2000	A	36
2400	A	38
0	B	0
400	B	4
800	B	6
1200	B	9
1600	B	10
2000	B	11
2400	B	12

The following edited `R` output is available:

```
> model1.lm<-lm(drop~operation*factor(coating))
> summary(model1.lm)

Call:
lm(formula = drop ~ operation * factor(coating))

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.285714   2.243699   1.464  0.17380
operation                    0.016429   0.001556  10.560 9.62e-07
factor(coating)B            -1.642857   3.173069  -0.518  0.61589
operation:factor(coating)B -0.011607   0.002200  -5.276  0.00036
---
Residual standard error: 3.293 on 10 degrees of freedom
Multiple R-squared: 0.9522,     Adjusted R-squared: 0.9379
F-statistic: 66.46 on 3 and 10 DF,  p-value: 6.591e-07
```

(i) With reference to the `R` output, discuss the fit of the model `model1.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 plots of residuals for the `model1.lm` linear model. Specify any model assumptions for this linear model and discuss whether these assumptions are supported by the plots. ***(3 marks)***

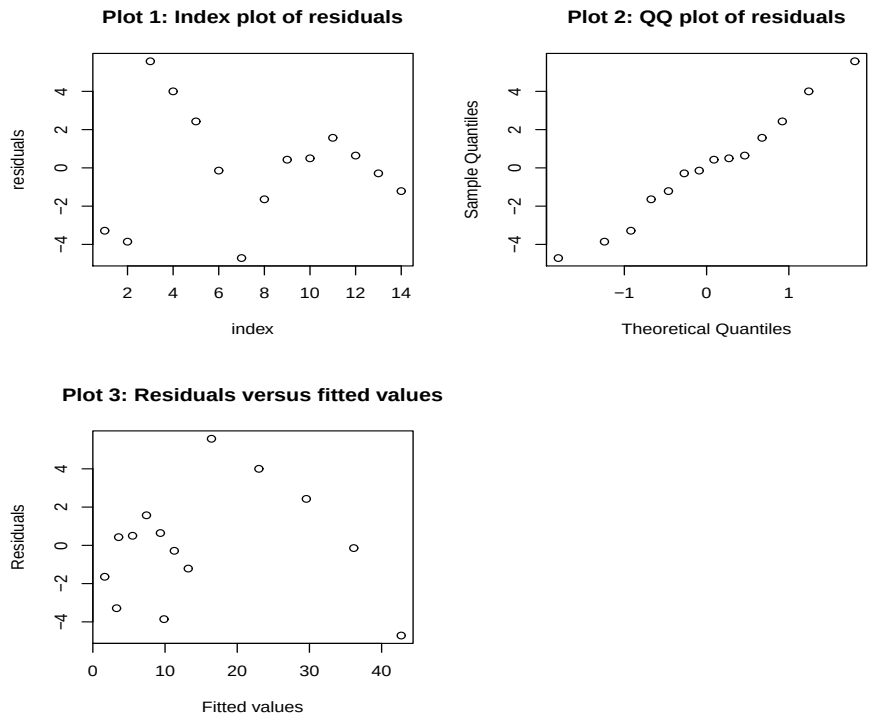

Figure 1: Residual plots for the `model1.lm` model.

1 (continued)

(iii) The statistical model for the ith observation corresponding to `model1.lm` can be written as

$$y_i = \beta_0 + \beta_1 I(\mathrm{B}) + \beta_2 x_i + \beta_3 x_i I(\mathrm{B}) + \epsilon_i$$

where $I(\mathrm{B})$ is an indicator variable taking the value 1 if the ith observation if from coating B and is zero otherwise; x_i is the operation time of the ith observation and $\epsilon_i \sim N(0, \sigma^2)$. Plot the expected value of y_i against x_i for this statistical model, specifying the value of the intercepts and gradients. ***(3 marks)***

(iv) For the statistical model in part (iii) the β_3 parameter is usually called the interaction parameter. Explain what this parameter represents in terms of the expected drop in output. ***(3 marks)***

1 (continued)

(v) Suppose a statistician wants to perform constrained least squares with the constraint $C\boldsymbol{\beta} - \boldsymbol{d} = 0$ where C is a full rank $m \times p$ matrix, $\boldsymbol{d}$ is a column vector of length m and $\boldsymbol{\beta}$ is a column vector of length p containing the parameters. They use the Lagrange multiplier $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_m)^T$ to perform the constrained least squares.

Let $S = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$.
Show that

$$\frac{\partial}{\partial \boldsymbol{\lambda}}[S + \boldsymbol{\lambda}^T(\boldsymbol{d} - C\boldsymbol{\beta})] = \boldsymbol{d} - C\boldsymbol{\beta}$$

and that

$$\frac{\partial}{\partial \boldsymbol{\beta}}[S + \boldsymbol{\lambda}^T(\boldsymbol{d} - C\boldsymbol{\beta})] = 2(X^T X)\boldsymbol{\beta} - 2X^T\boldsymbol{y} - C^T\boldsymbol{\lambda}$$

stating any results for vector differentiation you have used. ***(2 marks)***

(vi) By setting the two partial derivatives in part(v) to zero, show that the constrained least squares solution for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \boldsymbol{b} + (X^T X)^{-1}C^T[C(X^T X)^{-1}C^T]^{-1}(\boldsymbol{d} - C\boldsymbol{b})$$

where $\boldsymbol{b} = (X^T X)^{-1}X^T\boldsymbol{y}$ is the usual least squares estimate. ***(4 marks)***

2 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
- acid - acid concentration
- yield - amount of ammonia produced

(i) The following `R` code is used to fit a linear model to this chemical data:

```
chem1.lm<-lm(yield~air+temp+acid)
```

Write down the statistical model for the ith response (y_i) that corresponds to this `R` command. ***(5 marks)***

(ii) Interpret all the parameters in your model in part (i) in terms of the expected yield. ***(4 marks)***

(iii) Some more `R code` is given below:

```
chem2.lm<-lm(yield~temp+acid)
anova(chem1.lm,chem2.lm)
Analysis of Variance Table

Model 1: yield ~ air + temp + acid
Model 2: yield ~ temp + acid
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     17 178.83
2     18 475.06 -1   -296.23 28.16 5.799e-05 ***
```

State the null and alternative hypotheses for the test performed and state the conclusion of the test. Based on the results of the test, the statistician tells the managing director of the company that increasing the air flow in the production facility will increase the yield. Is this conclusion supported by this statistical test? ***(5 marks)***

(iv) Suppose the statistician believes that the errors in a particular model are heteroscedastic. To allow for this they take a generalised least squares approach in which

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{and} \quad \text{var}(\boldsymbol{\epsilon}) = \sigma^2 V$$

where $V = CC^T$ is a known nonsingular matrix. Show that if C is a non-singular square matrix then $(C^{-1})^T = (C^T)^{-1}$. ***(2 marks)***

2 (continued)

(v) With the error covariance matrix given in part (iv) it was shown in the course notes that $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} \boldsymbol{y}$. Suppose that the statistician decides to transform the response and explanatory variables to remove the heteroscedasticity using the transformations $y^* = C^{-1}\boldsymbol{y}$ and $X^* = C^{-1}X$ with C as defined in part (iv) . Show that the least squares estimate of the parameter vector in this transformed model is the same as in part (iv) if we assume the errors in the transformed model are uncorrelated and have the same variance. ***(4 marks)***

3 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.

In the R output in this question '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.

(i) In the R output below, describe what the output shows in each part. What does the output say about the effect of vitamin C intake and delivery method on tooth length? ***(5 marks)***

```
> tooth.lm<-lm(len~1)
> step(tooth.lm,scope=list(upper=len~dose+I(dose^2)+supp))
Start:  AIC=245.15
len ~ 1
            Df Sum of Sq    RSS    AIC
+ dose       1   2224.30 1227.9 185.12
+ I(dose^2)  1   1993.42 1458.8 195.46
+ supp       1    205.35 3246.9 243.47
<none>                   3452.2 245.15

Step:  AIC=185.12
len ~ dose
            Df Sum of Sq    RSS    AIC
+ supp       1    205.35 1022.6 176.14
+ I(dose^2)  1    202.13 1025.8 176.33
<none>                   1227.9 185.12
- dose       1   2224.30 3452.2 245.15

Step:  AIC=176.14
len ~ dose + supp
            Df Sum of Sq    RSS    AIC
+ I(dose^2)  1    202.13  820.4 164.93
<none>                   1022.6 176.14
- supp       1    205.35 1227.9 185.12
- dose       1   2224.30 3246.9 243.47

Step:  AIC=164.93
len ~ dose + supp + I(dose^2)
            Df Sum of Sq     RSS    AIC
<none>                    820.43 164.93
- I(dose^2)  1    202.13 1022.56 176.14
- supp       1    205.35 1025.78 176.33
- dose       1    433.01 1253.44 188.36

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

3 (continued)

(ii) The best subsets method is then applied. The output is shown below. What does the output say about the effect of vitamin C intake and delivery method on tooth length? ***(6 marks)***

```
> tooth.growth<-regsubsets(len~dose+I(dose^2)+supp)
> summary(tooth.growth)
Subset selection object
Call: regsubsets.formula(len ~ dose + I(dose^2) + supp)
3 Variables  (and intercept)
          Forced in Forced out
dose          FALSE      FALSE
I(dose^2)     FALSE      FALSE
suppVC        FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: exhaustive
         dose I(dose^2) suppVC
1  ( 1 ) "*"  " "       " "
2  ( 1 ) "*"  " "       "*"
3  ( 1 ) "*"  "*"       "*"
> summary(tooth.growth)$rsq
[1] 0.6443133 0.7037969 0.7623478
> summary(tooth.growth)$cp
[1] 27.81349 15.79685  4.00000
> summary(tooth.growth)$bic
[1] -53.83361 -60.71957 -69.83945
```

(iii) Explain what a Box-Cox transformation is, when it is appropriate and how to interpret the likelihood plot. ***(4 marks)***

(iv) Suppose that a response variable y represents the proportion of counts with some property. In this case the variance of the response for individual i has the known form $\text{var}(y_i) = \dfrac{\mu_i(1-\mu_i)}{n_i}$ where n_i is the number of counts for individual i and $\mu_i = E(y_i)$. Find the variance stabilizing transform for this form of heteroscedasticity. ***(5 marks)***

4 Consider the 2 parameter linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. For this model the n by 2 design matrix is $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$.

(i) Using the course notes, write down $(\hat{\beta}_0, \hat{\beta}_1)^T$ in terms of $\bar{y}$, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and
$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})y_i$. ***(1 mark)***

(ii) Consider the centered model in which the mean of the covariate is subtracted from each covariate value. We can write this linear model as $\boldsymbol{y} = X_1\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)^T$ with $X_1 = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}$. Derive $(\hat{\gamma_0}, \hat{\gamma_1})^T$. How do these values compare to $(\hat{\beta}_0, \hat{\beta}_1)^T$? ***(7 marks)***

(iii) Calculate the variances of the parameters in the centered and non-centered models. You can state any results for simple linear regression from the notes without proof. ***(4 marks)***

(iv) Consider now the multiple regression linear model with a parameter for the intercept and p ($p \geq 2$) covariates in which the covariate-specific mean is subtracted from each covariate value

$$\boldsymbol{y} = \begin{pmatrix} \mathbf{1} & Z \end{pmatrix} \begin{pmatrix} \eta \\ \boldsymbol{\delta} \end{pmatrix} + \epsilon$$

where $\boldsymbol{\delta} = (\delta_1, ..., \delta_p)^T$ and $\mathbf{1_n}$ is an n by 1 vector of ones. Calculate $\text{cov}(\hat{\eta}, \hat{\boldsymbol{\delta}})$ ***(3 marks)***

(v) Using the result that the residual sum of squares is given by $y^T(I - M)y$ (where M is defined in the course notes) or otherwise, show that the residual sums of squares for the model in part (iv) is

$$\boldsymbol{y}^T[I - n^{-1}\mathbf{1_n}\mathbf{1_n}^T - Z(Z^TZ)^{-1}Z^T]\boldsymbol{y}$$

(5 marks)

End of Question Paper