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 2011–12**

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 (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. Consider the following linear model:

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

where len_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 0 if the dose was administered by orange juice and 1 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]
   1
0.052
```

(i) Explain why Cook's distances are useful in linear models and how they are used. ***(5 marks)***

(ii) Show that the Cook's distance for the i^{th} guinea pig can be written as $\dfrac{e_i^2 h_{ii}}{p\hat{\sigma}^2(1-h_{ii})^2}$ where p is the number of parameters in the model, $\hat{\sigma}$ is the residual standard error and e_i and h_{ii} are the residual and leverage of the i^{th} guinea pig respectively. ***(6 marks)***

1 (continued)

(iii) Given that the first guinea pig in the dataset had a tooth length of 4.20 mm and was given a Vitaman C dose of 0.5 mg by supplement, calculate the residual for this guinea pig. ***(5 marks)***

(iv) Hence calculate the Cook's distance for the first guinea pig in the dataset. ***(3 marks)***

(v) Given that the maximum Cook's distance for all 60 guinea pigs is 0.103, what does this tell you about the leverage of the observations and about the presence of outliers in the data? ***(5 marks)***

(vi) Calculate a 95% prediction interval for the tooth length of a guinea pig given a vitamin C does of 1mg by orange juice. ***(9 marks)***

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

- sex - the gender of the crab (coded 0 for female, 1 for male)
- sp - the species (coded 0 for blue, 1 for orange)
- hl - head length (mm)
- bsl - body shell length (mm)

The interest is in how the head length depends on the other three variables.

(i) Explain why corner-point constraints are sometimes needed in linear models and how they are interpreted. ***(5 marks)***

(ii) Below is a command used to fit a linear model to the crab data in `R`.

```
model0.lm<-lm(hl~sex+sp*bsl,data=crabs)
```

Write down the statistical model that this `R` command fits for the i^{th} crab. Specify any distributional assumptions and clearly define all notation used. ***(10 marks)***

(iii) Explain what each parameter in `model0.lm` represents in terms of the change in head length as the value of the corresponding variable changes. ***(8 marks)***

(iv) Consider two nested linear models `modelA.lm` and `modelB.lm` fitted in `R`. Explain what it means for the two models to be nested and explain what the command `anova(modelA.lm,modelB.lm)` does, clearly stating the null hypothesis. ***(4 marks)***

(v) Below is some `R` output. State the null hypothesis of the test and what conclusions you would make based on the output.

```
model1.lm<-lm(hl~sex+sp,data=crabs)
model2.lm<-lm(hl~sex*sp,data=crabs)
anova(model1.lm,model2.lm)
Analysis of Variance Table

Model 1: FL ~ sex + sp
Model 2: FL ~ sex * sp
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    197 1960.3
2    196 1879.7  1    80.645 8.4091 0.00416
```

(3 marks)

(vi) State three different criteria for comparing the fit of two linear models that are **not** nested and how these criteria are used to compare the model fit. ***(3 marks)***

3 (i) A standard result is that if $\boldsymbol{y}$ is a vector of random variables with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then $E(\boldsymbol{y}\boldsymbol{y}^T) = \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^T$. Apply this result to the linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ to show that $E(\boldsymbol{y}\boldsymbol{y}^T) = \sigma^2 I_n + X\boldsymbol{\beta}\boldsymbol{\beta}^T X^T$.
(5 marks)

(ii) Assume that the true relationship between response $\boldsymbol{y}$ and $p+q$ explanatory variables is $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$ and $\boldsymbol{\beta}$ is a parameter vector of length $p+q$. If $M = I_n - X(X^TX)^{-1}X^T$, show that M is symmetric and idempotent and that $\hat{\sigma}^2 = \dfrac{\boldsymbol{y}^T M \boldsymbol{y}}{n-p-q}$ is an unbiased estimator of σ^2 (you can assume without proof that $\text{trace}(M) = n-p-q$).
(10 marks)

(iii) Now consider dropping the last q explanatory variables so that only the first p remain. The model can be written $\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$ where X_1 is the matrix consisting of the first p columns of X and $\boldsymbol{\beta}_1$ is a vector of p parameters. Our estimator of σ^2 based only on these p explanatory variables (which we denote as $\hat{\sigma_p}^2$) is $\hat{\sigma_p}^2 = \dfrac{\boldsymbol{y}^T M_1 \boldsymbol{y}}{n-p}$ where $M_1 = I_n - X_1(X_1^TX_1)^{-1}X_1^T$.
Using the fact that $\boldsymbol{y}^T M_1 \boldsymbol{y} = \text{trace}(\boldsymbol{y}^T M_1 \boldsymbol{y})$ and that $E(\text{trace}(A)) = \text{trace}(E(A))$ for any random matrix A, show that $(n-p)E(\hat{\sigma_p}^2) = \text{trace}\left[M_1 E(\boldsymbol{y}\boldsymbol{y}^T)\right]$.
(5 marks)

(iv) Using the results from parts (i) and (iii) or otherwise, show that $(n-p)E(\hat{\sigma_p}^2) = (n-p)\sigma^2 + \boldsymbol{\beta}^T X^T M_1 X \boldsymbol{\beta}$ (you can assume without proof that $\text{trace}(M_1) = n-p$).
(5 marks)

(v) Hence show that $E(\hat{\sigma_p}^2 - \hat{\sigma}^2) = (n-p)^{-1}\boldsymbol{\beta}^T X^T M_1 X \boldsymbol{\beta}$ and by writing $\boldsymbol{\beta}^T X^T M_1 X \boldsymbol{\beta}$ as a sum of squares discuss whether $\hat{\sigma_p}^2$ is an unbiased estimator of σ^2.
(8 marks)

4 An investigation was carried out to assess the effect of a drug, theophylline, on the blood flow in the brain of 18 hospital patients. Four measurements were recorded for each patient. These were:

- the blood flow (denoted by the variable B);
- the cardiac output (denoted by the variable C);
- the blood oxygen level (denoted by the variable O);
- whether the patient was on theophylline or not (denoted by the indicator variable T taking the value 1 if the patient was on theophylline and 0 if they were not).

A linear model is fitted resulting in the following edited `R` output.

```
> theo.lm<-lm(B~T+C+O, data=theo.data)
> summary(theo.lm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.131963   7.909480   2.292   0.0379 *
T            3.241456   2.554959   1.269   0.2252
C            0.018199   0.006958   2.616   0.0204 *
O           -0.246741   0.211732  -1.165   0.2634

Residual standard error: 5.336 on 14 degrees of freedom
Multiple R-squared: 0.4148,     Adjusted R-squared: 0.2894
F-statistic: 3.308 on 3 and 14 DF,  p-value: 0.05148
```

(i) With reference to the `R` output above, discuss the fit of the model and the need for the 4 parameters in the model. You should include discussion of the F-statistic and the associated p-value, the p-values for the 4 parameters and the multiple R-squared value. State the null hypothesis for any hypothesis tests you refer to. ***(8 marks)***

(ii) Briefly discuss the disadvantages of using the p-values for the 4 parameters in deciding whether the parameters are needed in the model. ***(3 marks)***

(iii) State the distribution used to model the n-vector of errors in a linear model. ***(3 marks)***

(iv) Figure 1 shows some plots used to verify the distribution of the errors in a linear model. State what model assumption each plot is used to check and what each plot tells you about whether the assumption is validated. ***(6 marks)***

(v) Let $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 & \hat{\beta}_1 & \dots & \hat{\beta}_{p-1} \end{pmatrix}^T$ be the p-vector of least squares estimates of the true but unknown p-vector of parameters $\boldsymbol{\beta}$ in a linear model. What is the distribution of $\hat{\boldsymbol{\beta}}$? ***(3 marks)***

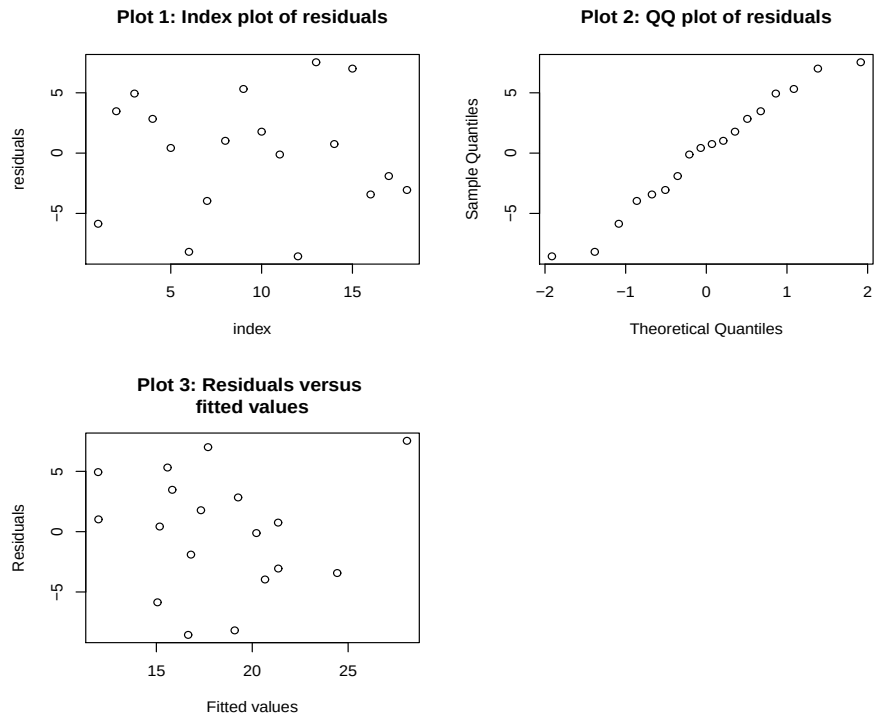

Figure 1: Residual plots for the theophylline data

4 (continued)

(vi) Let g_{ij} be the element in row i and column j of $(X^TX)^{-1}$ where X is the design matrix for the linear model with parameter vector $\begin{pmatrix} \beta_0 & \beta_1 & \dots & \beta_{p-1} \end{pmatrix}^T$ with $(p \geq 3)$. State the distribution of the 2-vector $\begin{pmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{pmatrix}^T$. State what standard statistical result you are using.

(3 marks)

(vii) What is the distribution of $2\hat{\beta}_0 + 3\hat{\beta}_1$? State what standard result you are using. ***(5 marks)***

(viii) Hence state an expression for a 95% confidence interval for $2\beta_0 + 3\beta_1$ in terms of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ and g_{ij}. ***(2 marks)***

End of Question Paper