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

MAS 6003

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2011–2012**

Linear Models **3 hours**

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.
All answers will be marked but credit will be given only for the best ***FIVE*** *answers. All questions are worth 20 marks. Total marks 100.*
Corner point constraints (treatment contrasts) are used in all `R` *output.*

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. ***(3 marks)***

(ii) Show that the Cook's distance for the i^{th} guinea pig can be written as $\frac{e_i^2 h_{ii}}{p\hat{\sigma}^2(1-h_{ii})^2}$ 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. ***(4 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. ***(3 marks)***

(iv) Hence calculate the Cook's distance for the first guinea pig in the dataset. ***(2 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? ***(3 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. ***(5 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. ***(3 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. ***(6 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. ***(5 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. ***(2 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
```

(2 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. ***(2 marks)***

3 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. ***(5 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. ***(2 marks)***

(iii) State the distribution used to model the n-vector of errors in a linear model. ***(2 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. ***(3 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 $\hat{\boldsymbol{\beta}}$? ***(2 marks)***

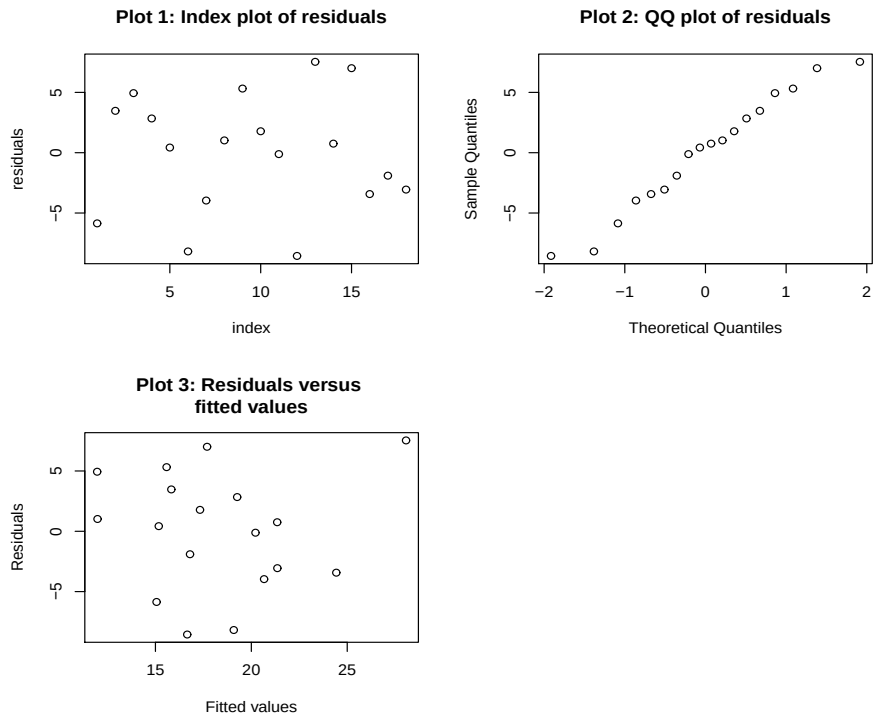


Figure 1: Residual plots for the theophylline data

3 (continued)

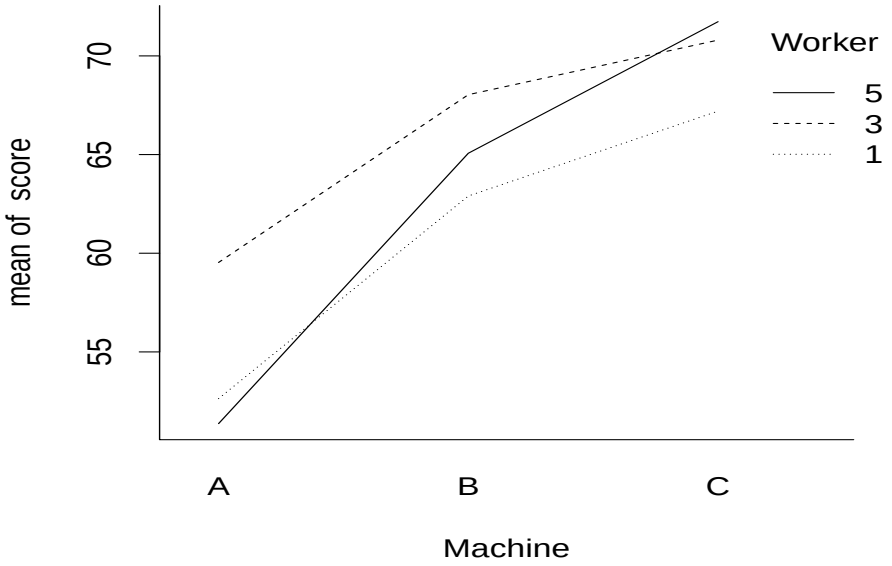
(vi) Let g_{ij} be the element in row i and column j of $(X^TX)^{-1}$ where X is the design matrix for the linear model with parameter vector $\begin{pmatrix} \beta_0 & \beta_1 & \dots & \beta_{p-1} \end{pmatrix}^T$ with $(p \geq 3)$. State the distribution of the 2-vector $\begin{pmatrix} \hat{\beta}_0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

(2 marks)

(vii) State the distribution of $2\hat{\beta}_0 + 3\hat{\beta}_1$ and hence give 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}. ***(4 marks)***

4 Figure 1 shows mean productivity scores for each of three randomly chosen workers tested on each of three types of machine. Each worker used each machine three times giving three sets of replicates at each set of conditions.

Figure 1



Below is annotated output from an `R` session in which linear mixed effects models were investigated.

```
mach1.lme=lme(score~Machine, random = ~1 |Worker)
mach2.lme=lme(score~Machine, random = ~1 |Worker/Machine)

anova(mach1.lme,mach2.lme)
    Model    df      logLik     Test     L.Ratio   p-value
mach1.lme     5  -55.44817
mach2.lme     6  -44.83272   1 vs 2    21.23089    <.0001

mach3.lme=lme(score~Machine,random=pdCompSymm(~Worker))
```

(i) Write down the algebraic specification of the model `mach2.lme` stating clearly all assumptions and defining any terms that you use. ***(7 marks)***

(ii) Justify the choice of random and fixed effects in model `mach2.lme` ***(3 marks)***

(iii) Describe how you would check the normality assumptions within the `mach2.lme` model. State the value that the `levels` option would take in `R` for each set of residuals you'd check. ***(3 marks)***

4 (continued)

(iv) Describe what is being tested by the `anova(mach1.lme,mach2.lme)` command in the `R` output above specifying the null hypothesis and state what the conclusion of the test is. ***(3 marks)***

(v) For the `mach3.lme` model state the algebraic form for the covariance matrix of the worker random effects that is being specified. How does it differ to that fitted in the `mach2.lme` model? ***(4 marks)***

5 Moderately deaf people often find it difficult to use a telephone. Listening devices are designed to help moderately deaf people overcome this problem. A study is undertaken to assess the effect of a particular hearing device (D) and hearing score (S) on the probability of a moderately deaf person using a telephone (T). A total of 173 partially deaf people are surveyed for the study.

- Device (D) is a binary variable: 1=device A, 0=other device
- Hearing score (S) is a continuous variable (measured in some unit)
- The response variable is use of telephone (T): 1=use telephone, 0=do not use telephone.

Table 1 provides the residual deviances and degrees of freedom for 5 models fitted to the data using the logit link function.

Table 1

Model	Res. Deviance	df
1) T ~ 1	233.50	172
2) T ~ D	219.01	171
3) T ~ S	218.5	171
4) T ~ D + S	199.20	170
5) T ~ D*S	197.31	169

(i) What does the analysis of models 1-5 in Table 1 tell us about the dependence of the probability of using a telephone on hearing score and device type?

(7 marks)

(ii) More detailed analysis of model 4 is provided in Table 2 below. What does the numerical value of 0.026 represent in Table 2 in terms of the odds of using a telephone?

(4 marks)

(iii) Using the information in Table 2, calculate the odds ratio of a moderately deaf person using a telephone with a hearing score of 20 using device type A compared to a moderately deaf person with a hearing score of 20 not using device type A.

(2 marks)

Table 2

Coefficients:	Estimate	Std. Error
(Intercept)	-4.165	0.803
S	0.026	0.007
D	2.148	0.577

(iv) Using the information in Table 2, calculate the log odds of using a telephone for someone using device type A with a hearing score of 15 for model 4.

(3 marks)

 Question 5 continued on next page

5 (continued)

(v) The R output below gives the estimated covariance matrix of the estimated coefficients of model 4.

```
             (Intercept)   S             D
(Intercept)  0.644009133  -3.969181e-03 -0.3451017
S           -0.003969181   4.480828e-05  0.0005948
D           -0.345101718   5.948000e-04  0.3328265
```

Use the above R output and your answer to part (iv) to calculate a 95% confidence interval for the log odds of using a telephone for someone using device type A with a hearing score of 15 for model 4. ***(4 marks)***

6 A study was carried out to assess the effects of mothers' drinking history and diet on the birth weight of their babies. A birth weight of less than 2.5 kg was considered low for the purposes of this study. Table 3 shows the results of the study. A value of 0 for `drink` indicates the mother did not drink during pregnancy. `Diet` is either `vegan`, `vegetarian` or `neither`. For notational convenience we use

Table 3

	Diet					
	Vegan		Vegetarian		Neither	
	drink		drink		drink	
Low	0	1	0	1	0	1
0	34	29	15	12	43	12
1	5	21	9	8	23	11

L, DT and DK to represent the variables low, diet and drink respectively. For this question, assume that L is a response factor and DT and DK are controlled factors.

4 log-linear models with Poisson errors were fitted with the results below:

```
Model              Res. Deviance    degrees of freedom
1) DT*DK           34.00            6
2) DT*DK+L         12.84            5
3) DT*DK+L*DK      10.85            4
4) DT*DK+L*DT      7.96             3
```

(i) Write down an algebraic form for the linear predictor for model 3.
(5 marks)

(ii) With reference to your answer to part (i), justify why the degrees of freedom for model 3 are 4. ***(2 marks)***

(iii) Referring to the residual deviances in the `R` output above, what would you conclude about the dependence of birth weight on diet and drinking habits of the birth mother. ***(6 marks)***

6 (continued)

(iv) Below is some further R output for model 3.

```
>  birth3.glm<-glm(count~diet*drink+low*diet,family=poisson)
> summary(birth3.glm)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.70835    0.14865  24.947  < 2e-16 ***
dietveg         -1.01865    0.27942  -3.646 0.000267 ***
dietvegan       -0.39029    0.22886  -1.705 0.088121 .
drink           -1.05416    0.24214  -4.354 1.34e-05 ***
low             -0.48097    0.21816  -2.205 0.027476 *
dietveg:drink    0.87184    0.38768   2.249 0.024522 *
dietvegan:drink  1.30262    0.32291   4.034 5.48e-05 ***
dietveg:low      0.01835    0.37875   0.048 0.961361
dietvegan:low   -0.40407    0.31926  -1.266 0.205648
```

Calculate the expected number of low birth weight babies whose mothers were vegan and did not drink during pregnancy based on the output above. *(3 marks)*

(v) Define `n` to be the vector `(39,39,50,50,24,24,20,20,66,66,23,23)` and `proportion` to be the vector of proportions given by

$(34/39, 5/39, 29/50, 21/50, 15/24, 9/24, 12/20, 8/20, 43/66, 23/66, 12/23, 11/23)$.

The following command is fitted in R
`glm(proportion~low*diet,weights=n,family=binomial)`.
Explain what the `n` and `proportion` vectors represent and what the R command is fitting. *(4 marks)*

End of Question Paper