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

MAS6003

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2012–2013**

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.*

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.

The following R output is available in which '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:

```
> tooth1.lm<-lm(len~dose+I(dose^2)+supp)
> summary(tooth1.lm)

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

Coefficients:
            Estimate  Std. Error  t-value   Pr(>|t|)
(Intercept)  -0.6400      2.9094   -0.220   0.826690
dose         30.1550      5.5467    5.437   1.23e-06 ***
I(dose^2)    -7.9300      2.1349   -3.714   0.000471 ***
supp         -3.7000      0.9883   -3.744   0.000429 ***
---
Residual standard error: 3.828 on 56 degrees of freedom
Multiple R-squared: 0.7623,     Adjusted R-squared: 0.7496
F-statistic: 59.88 on 3 and 56 DF,  p-value: < 2.2e-16

> tooth2.lm<-lm(len~dose+supp)
> anova(tooth2.lm)
Analysis of Variance Table

Response: len
          Df  Sum Sq Mean Sq F value    Pr(>F)
dose       1 2224.30 2224.30 123.989 6.314e-16 ***
supp       1  205.35  205.35  11.447  0.001301 **
Residuals 57 1022.56   17.94
```

(i) With reference to the R output, discuss the fit of the model `tooth1.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 residual plots for the `tooth1.lm` linear model. State the underlying assumptions for this linear model and comment on whether the plots support these assumptions. ***(3 marks)***

(iii) The plots in Figure 1 are based on the raw residuals $(y_i - \hat{y}_i)$. State what other residuals might be more appropriate and why. ***(3 marks)***

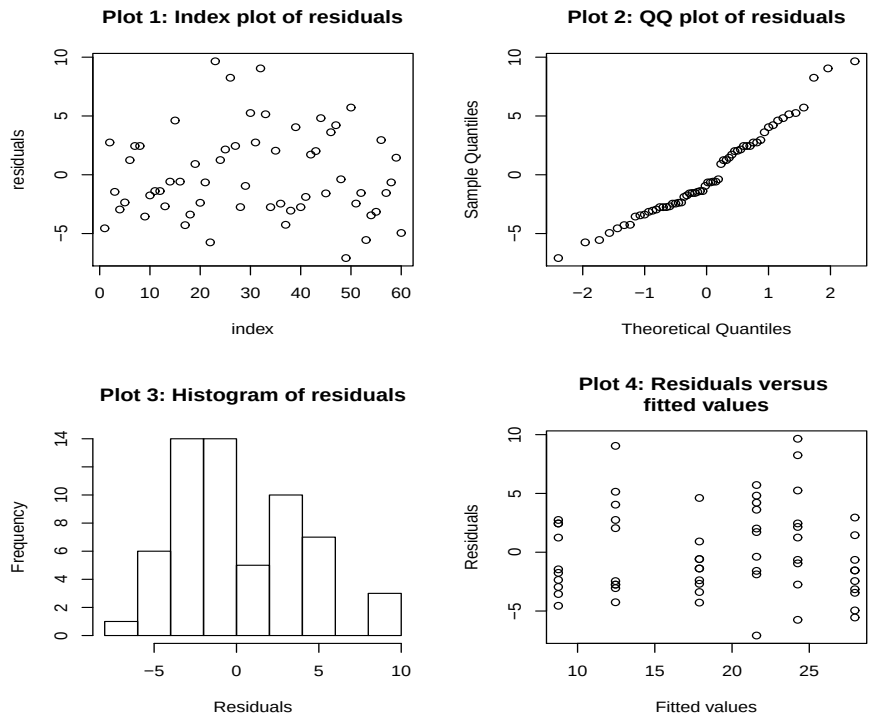

Figure 1: Residual plots for the `tooth1.lm` model

1 (continued)

(iv) Figure 2 shows the log-likelihood for the Box-Cox family of transformations for model `tooth1.lm` for values of λ between -1 and 3. Explain what the parameter λ represents and comment on what Figure 2 tells you about the need for a transformation of the response for the `tooth1.lm` model.

(2 marks)

(v) For the `anova(tooth2.lm)` command, state the null hypothesis for the two tests performed and describe the conclusion of each hypothesis test.

(3 marks)

(vi) Briefly describe how the partition sum of squares property is used in the hypothesis tests performed in the `anova(tooth2.lm)` command.

(4 marks)

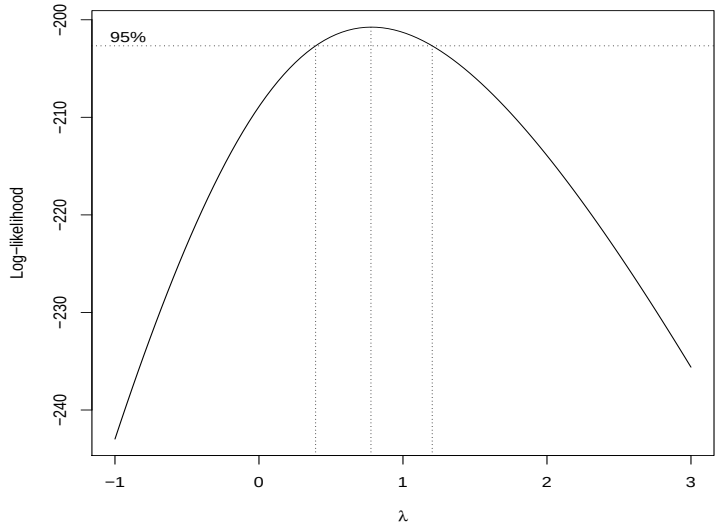

Figure 2: Log-likelihood function for the Box-Cox family of transformations for the `tooth1.lm` model.

2 The effect of the amount of a fertilizer ($F, grammes/m^2$) and the level of watering (W) on the yield ($Y, grammes/m^2$) of a tomato plant is studied. The exact level of watering is not known and is classified as low, medium and high. The data are listed in the first 3 columns of Table 1.

Table 1

Y	W	F	X	$Z1$	$Z2$	$Z3$
210	High	2.5	2	1	0	0
353	High	4.8	2	1	0	0
448	High	7.5	2	1	0	0
590	High	5.8	2	1	0	0
555	High	9.6	2	1	0	0
210	Med	5.2	1	0	1	0
352	Med	6.6	1	0	1	0
440	Med	10.1	1	0	1	0
621	Med	12.2	1	0	1	0
695	Med	14.8	1	0	1	0
247	Low	6.8	0	0	0	1
350	Low	9.8	0	0	0	1
346	Low	13.8	0	0	0	1
450	Low	15.3	0	0	0	1
560	Low	19.8	0	0	0	1

(i) A researcher analyses the data after first creating a new variable (X), as shown in Table 1, corresponding to the level of watering. She then uses the `R` command `lm1<-lm(Y~F+X)` to fit a linear model. Write down the statistical model for the ith observation being fitted by this `R` command.

(3 marks)

2 (continued)

(ii) A statistician recommends that the data should be reanalyzed. However, in her analysis she represents the different levels of watering by three new variables: $Z1$, $Z2$ and $Z3$ shown in Table 1. She uses the `R` command `lm2<-lm(Y~F+Z2+Z3)` to fit a linear model. Write down the statistical model for the ith observation being fitted by this `R` command and interpret the model parameters in terms of the expected yield of tomatoes. ***(7 marks)***

(iii) The researcher then uses the `R` command `lm3<-lm(Y~F+Z1+Z2+Z3)` to fit a linear model. Discuss what is wrong with this model. ***(3 marks)***

(iv) State, with justification, how the researcher could modify the linear predictor of model `lm3` to allow an additive combination of F, $Z1$, $Z2$ and $Z3$ to be included explicitly. Give an `R` command to do this. ***(2 marks)***

(v) Suppose another researcher records n observations at each of the three levels of watering (high, medium and low). The statistician fits a model with the `R` command `lm3<-lm(Y~Z1)`. Let $\bar{y_1}, \bar{y_2}$ and $\bar{y_3}$ represent the sample mean of the observations at high, medium and low levels of watering respectively. By first specifying the $3n$ by 2 design matrix X, derive an expression for the least squares estimate of the parameter for Z_1 in this model in terms of $\bar{y_1}, \bar{y_2}, \bar{y_3}$ and n. ***(5 marks)***

3 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
- conc - acid concentration
- yield - amount of ammonia produced

(i) Some R output generated by the statistician is given below. Describe what is being done and what the conclusions are in each part of the R output. What does the output say about the relationship between the amount of ammonia produced and the air flow, water temperature and acid concentration?

(6 marks)

```
> int.lm<-lm(yield~1)
> step(int.lm,scope=list(upper=yield~air*temp+air*acid+acid*temp),
+ direction="forward")
Start:  AIC=98.4

yield ~ 1
       Df Sum of Sq     RSS    AIC
+ air   1    1750.1  319.12 61.142
+ temp  1    1586.1  483.15 69.852
+ acid  1     330.8 1738.44 96.741
<none>              2069.24 98.399
Step:  AIC=61.14

yield ~ air
       Df Sum of Sq    RSS    AIC
+ temp  1   130.321 188.80 52.119
<none>              319.12 61.142
+ acid  1     9.979 309.14 62.475
Step:  AIC=52.12

yield ~ air + temp
           Df Sum of Sq    RSS    AIC
+ air:temp  1    38.563 150.23 49.321
<none>                  188.79 52.119
+ acid      1     9.965 178.83 52.980
Step:  AIC=49.32
```

3 (continued)

```
yield ~ air + temp + air:temp
       Df Sum of Sq    RSS    AIC
<none>              150.23 49.321
+ acid  1   0.93534 149.30 51.190

Call:
lm(formula = yield ~ air + temp + air:temp)

Coefficients:
(Intercept)          air         temp     air:temp
   22.29030     -0.51551     -1.93006      0.05176
```

(ii) Further partial R output generated by the statistician is given below. Describe what this output says about the relationship between the amount of ammonia produced and the air flow, water temperature and acid concentration. ***(4 marks)***

```
> yield.amm<-regsubsets(yield~air*temp+air*acid+acid*temp)
> summary(yield.amm)
Subset selection object

1 subsets of each size up to 6
Selection Algorithm: exhaustive
         air temp acid air:temp air:acid temp:acid
1  ( 1 ) " " " "  " "  "*"      " "      " "
2  ( 1 ) " " " "  " "  "*"      " "      "*"
3  ( 1 ) " " "*"  " "  "*"      "*"      " "
4  ( 1 ) " " " "  "*"  "*"      "*"      "*"
5  ( 1 ) "*" "*"  "*"  "*"      "*"      " "
6  ( 1 ) "*" "*"  "*"  "*"      "*"      "*"

> summary(yield.amm)$rsq
[1] 0.9193685 0.9257151 0.9276716 0.9300364 0.9336477 0.9337217
> summary(yield.amm)$cp
[1] 0.031833 0.691239 2.277976 3.778440 5.015631 7.000000
> summary(yield.amm)$bic
[1] -46.78614 -45.46323 -42.97921 -40.63279 -38.70118 -35.68009
```

(iii) Other than using the `step` and `regsubsets` commands, what other statistical method(s) could the statistician use to assess the relationship between the amount of ammonia produced and the air flow, water temperature and acid concentration? Briefly outline the advantages and disadvantages of the `step` and `regsubsets` methods as well as the other method(s) you suggest. ***(5 marks)***

3 (continued)

(iv) The statistician decides that the data support the linear model $yield_i = \beta_0 + \beta_1(air_i) + \beta_2(temp_i) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$. Based on this model, the statistician forms a 95% confidence interval for β_1 and a 95% confidence interval for β_2. Let $(b_1 - a, b_1 + a)$ and $(b_2 - c, b_2 + c)$ represent the 95% confidence intervals for β_1 and β_2 respectively. Derive the 95% confidence interval for $\beta_1 + 2\beta_2$ in terms of b_1, b_2, a and c and any other terms you need. Explain how you could obtain the value of any other terms you need.

(5 marks)

4 An experiment has been conducted to measure the fuel economy of four different models of car. Eight volunteer drivers have been recruited, and each volunteer drives each car three times along a specified route. The fuel economy (in miles per gallon) is recorded for each journey. The data are stored in R, with the variable names `econ` for fuel economy, `car` for the model of car, and `driver` for volunteer.

(i) A model is fitted in R using the command

```
> (fm1<-lmer(econ~car+(1|driver),REML=F))
```

Write down the equation of the model that has been fitted and assigned to the name `fm1`, defining your notation carefully.

(3 marks)

(ii) Define $\bar{Y}_{42\bullet}$ to be the mean of all the observations in which car 2 has been used by driver 4. Using your notation for the model specified in (i), derive expressions for the expectation and variance of $\bar{Y}_{42\bullet}$.

(3 marks)

(iii) Explain the difference between the commands

```
lmer(econ~car+(1|driver))
```

and

```
lm(econ~car+driver)
```

in terms of the models fitted to the data. State, with justification which model you believe to be more appropriate for analysing the data.

(3 marks)

4 (continued)

(iv) Some edited output from R is given below

```
> (fm1<-lmer(econ~car+(1|driver),REML=F))
Linear mixed model fit by maximum likelihood
Formula: econ ~ car + (1 | driver)
   AIC   BIC logLik deviance REMLdev
 169.1 184.4 -78.53    157.1   163.5
Random effects:
 Groups   Name        Variance Std.Dev.
 driver   (Intercept) 3.74157  1.93431
 Residual             0.19064  0.43662
Number of obs: 96, groups: driver, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)  38.0511     0.6897   55.17
car2          3.9785     0.1260   31.56
car3         -1.0796     0.1260   -8.57
car4         10.9470     0.1260   86.85
```

(a) Give the estimated parameter values for each parameter in your model in (i), including variance parameters.

(2 marks)

(b) Calculate the estimated variance for any observation, and the estimated covariance between any two different observations involving the same driver.

(2 marks)

(c) Give one criticism of the choice of R command with regard to estimating the residual variance, and suggest an alternative command.

(1 mark)

(v) State the model assumptions used in your model in part (i), and state how you would check them. State the gradients of any reference lines to be used in Q-Q plots.

(2 marks)

4 (continued)

(vi) The session is continued below.

```
> fm2<-lm(econ~car)
> fm3<-lmer(econ~car+(1|driver/car),REML=F)
> logLik(fm2)
'log Lik.' -201.9398
> logLik(fm3)
'log Lik.' -78.51752
> qchisq(0.999,1)
[1] 10.82757
> qchisq(0.95,1)
[1] 3.841459
```

Conduct any suitable hypothesis tests, stating clearly what the null hypothesis is in each case, to decide which model is most suitable, and interpret the result.

(4 marks)

5 A study has been conducted to find a suitable dose for a new drug. Each patient is given a particular dose, and the outcome is recorded at the end of the study as "patient responds" or "patient does not respond". It is suspected that a patient's genotype may affect the chances of responding. Five different doses are used (including a dose of 0 to give a control group), and the genotype of each patient is noted. The observed data are tabulated below.

genotype A			genotype B		
dose (mg)	number of patients	number responding	dose (mg)	number of patients	number responding
0	13	6	0	7	1
2	15	9	2	5	1
5	13	9	5	7	2
10	16	11	10	4	2
15	13	11	15	7	7

The data are stored in R for each patient with `response` and `genotype` binary indicator variables (1 indicating "patient responds" in `response` and 1 indicating genotype B in `genotype`) in each case, and `dose` representing the dose of the drug in mg. (The length of each vector is 100, corresponding to the 100 patients in the study).

(i) Defining your notation carefully, write down the model that is fitted to the data using the following command, including an equation for the linear predictor.

```
> lm1<-glm(response~dose*genotype,binomial(logit))
```

(2 marks)

(ii) Briefly describe what plots you would draw to choose between different possible link functions, and explain what you would look for in each plot.

(4 marks)

5 (continued)

(iii) Some further commands and edited output from R are given below.

```
> summary(lm1)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.05467    0.38268   0.143  0.88641
dose           0.10019    0.05165   1.940  0.05242
genotype      -2.28760    0.88363  -2.589  0.00963
dose:genotype  0.20586    0.11589   1.776  0.07568
---

> anova(lm1)
Analysis of Deviance Table

Model: binomial, link: logit

Response: response

Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev
NULL                            99     135.37
dose           *   ******        98     121.35
genotype       *   ******        97     116.58
dose:genotype  *   ******        96     112.76
> qchisq(0.95,1)
[1] 3.841459
> qnorm(0.975)
[1] 1.959964
```

(a) Conduct suitable hypothesis tests to assess the effect of dose and genotype on response. Starting with the equation for the linear predictor in the fitted model, state the null hypothesis clearly in each case. Choose the most appropriate model for the data, state the linear predictor, and interpret the result.

(7 marks)

(b) Out of 20 patients who are genotype B and given a dose of 7mg, calculate the expected number of responders, using your chosen model in part (a).

(2 marks)

5 (continued)

(c) Calculate the odds ratio of a genotype A patient responding compared to a genotype B patient responding, if both patients are given the same dose, using your chosen model in part (a). Calculate also an approximate 95% confidence interval for this odds ratio.

(2 marks)

(d) For a genotype A patient, estimate the minimum required dose such that the probability of responding is at least 0.9, using your chosen model in part (a). Give one criticism of your estimate.

(3 marks)

6 Appleton, French and Vanderpump (1996) describe a 20-year follow-up study on the effects of smoking. In 1972-74, a sample of women was categorized by age and smoking status (smoker or non-smoker). Twenty years later, the investigators recorded whether each participant was still alive. Smokers who quit in the intervening period were excluded. The data are stored in an R dataframe `femsmoke`, and a subset is shown below.

```
> head(femsmoke)
   y smoker dead   age
1  2    yes  yes 18-24
2  1     no  yes 18-24
3  3    yes  yes 25-34
4  5     no  yes 25-34
5 14    yes  yes 35-44
6  7     no  yes 35-44
```

y is the observed count for each combination of smoking status, dead/alive outcome and age group.

Some edited output from an R session is given below.

```
> lm1<-glm(y~dead*smoker, poisson, femsmoke)
> anova(lm1)

Analysis of Deviance Table
Model: poisson, link: log
Response: y
           Df Deviance Resid. Df Resid. Dev
NULL                         27    1193.94
dead        1  261.274        26     932.66
smoker      1   17.161        25     915.50
dead:smoker 1    9.200        24     906.30

> qchisq(0.95,1)
[1] 3.841459
```

(i) Defining your notation carefully, write down the model that has been fitted to the data in R and assigned to `lm1`, including an equation for the linear predictor, stating any necessary parameter constraints. ***(3 marks)***

(ii) Assess whether there is evidence that the probability of death is dependent on smoking status. ***(1 mark)***

(iii) Below are the data tabulated by smoking status and dead/alive outcome only.

	dead	alive
non-smoker	230	502
smoker	139	433

Estimate the probability of death for smokers and non-smokers, and comment briefly on your estimated values. ***(2 marks)***

6 (continued)

(iv) The full data are given below.

smoker			non-smoker		
age	dead	alive	age	dead	alive
18-24	2	53	18-24	1	61
25-34	3	121	25-34	5	152
35-44	14	95	35-44	7	114
45-54	27	103	45-54	12	66
55-64	51	64	55-64	40	81
65-74	29	7	65-74	101	28
75+	13	0	75+	64	0

Based on this table, estimate the probability of death for smokers and non-smokers in each age group. Compare with your result in part (iii), and discuss briefly any apparent differences. ***(5 marks)***

(v) Some further edited R analysis of the data follow next.

```
> lm2<-glm(y~dead*age*smoker,poisson,femsmoke)
> anova(lm2)

Analysis of Deviance Table
Model: poisson, link: log
Response: y

                Df Deviance Resid. Df Resid. Dev
NULL                                27    1193.94
dead             *   ******         **     932.66
age              *   ******         **     752.16
smoker           *   ******         **     735.00
dead:age         *   ******         **     101.83
dead:smoker      *   ******         **      92.63
age:smoker       *   ******         **       2.38
dead:age:smoker  *   ******         **       ****

> qchisq(0.95,1:7)
[1]  3.841  5.991  7.814  9.487 11.070 12.591 14.067
```

Using this analysis of deviance table, decide on the most suitable model for the data. Defining your notation carefully, write down the equation of the linear predictor in your chosen model, stating any necessary parameter constraints. ***(9 marks)***

End of Question Paper