The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2013–2014**

MAS6003 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.*

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)***

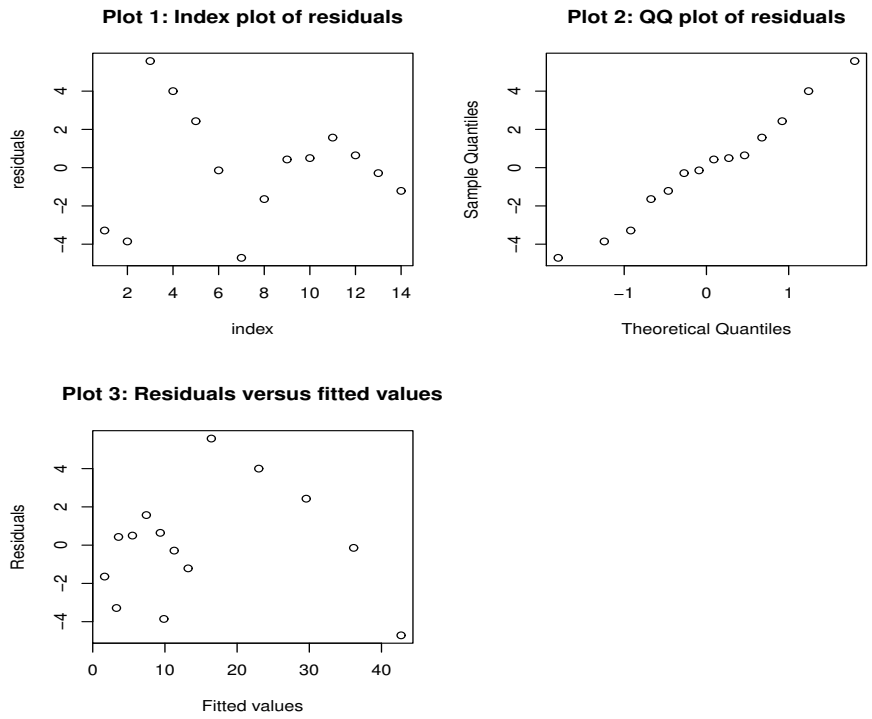

Figure 1: Residual plots for the `model1.lm` model.

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 An experiment has been conducted to investigate the time taken for a clot to form in a sample of blood, when treated with a drug. Three drugs are compared. There are ten volunteers in the study. Each volunteer donates six blood samples, and each drug is used on two of these samples, so that there are 60 observations in total. In an R dataset, the clot formation time (in seconds) is stored as a variable `clot.time` and the drug and volunteer labels are stored as factor variables `drug` and `volunteer` respectively. Below is some output from an R session.

```
> fm1<-lmer(clot.time~drug-1+(1|volunteer/drug))
> summary(fm1)
Linear mixed model fit by REML ['lmerMod']
Formula: clot.time ~ drug - 1 + (1 | volunteer/drug)

REML criterion at convergence: -44.3055

Random effects:
 Groups         Name        Variance Std.Dev.
 drug:volunteer (Intercept) 0.030827 0.17558
 volunteer      (Intercept) 0.001150 0.03391
 Residual                   0.008216 0.09064
Number of obs: 60, groups: drug:volunteer, 30; volunteer, 10

Fixed effects:
      Estimate Std. Error t value
drug1 10.07687    0.06007   167.8
drug2 10.99666    0.06007   183.1
drug3 12.04935    0.06007   200.6

Correlation of Fixed Effects:
      drug1 drug2
drug2 0.032
drug3 0.032 0.032
```

(i) Write down the equation of the model that has been fitted and assigned to the name `fm1`, defining your notation carefully. ***(3 marks)***

(ii) Give the estimated parameter values for each of the variance parameters in your model in (i). ***(1 mark)***

(iii) Calculate the estimated variance for any observation. ***(1 mark)***

(iv) Calculate the estimated correlation between any two different observations involving the same volunteer and the same drug. ***(2 marks)***

 Question 4 continued on next page

4 (continued)

(v) The estimator for the fixed effect for drug i is the sample mean of all the observations involving drug i.

(a) Derive an expression for the variance of this estimator, and hence verify that the estimated standard error is 0.06 (to 2 d.p.)
(3 marks)

(b) Derive an expression for the correlation between the fixed effect estimators of two different drugs, and hence verify that the estimated correlation is 0.03 (to 2 d.p.) ***(4 marks)***

(c) A fixed effects model is also fitted to the data using the command

```
lm(clot.time~drug*volunteer-1,contrasts=list(volunteer=contr.sum))
```

Edited output corresponding to `drug1` is given below.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
drug1          10.051984   0.020269 495.938  < 2e-16 ***

Residual standard error: 0.09064 on 30 degrees of freedom
```

The `drug1` estimate is again given by the mean of all the observations involving drug 1. Explain why the estimated standard error is smaller compared to model `fm1`. Explain the difference in interpretation of the `drug 1` term between the fixed and mixed effects models. ***(3 marks)***

(vi) The session is continued below.

```
> fm2<-lmer(clot.time~drug-1+(1|volunteer))
> logLik(fm1)
'log Lik.' 22.15275 (df=6)
> logLik(fm2)
'log Lik.' 9.289661 (df=5)
> qchisq(0.99,1)
[1] 6.634897
```

Compare the models defined as `fm1` and `fm2` using a generalised likelihood ratio test. State clearly what the null hypothesis is, in terms of the parameters of model `fm1`, and interpret the result. ***(3 marks)***

5 A car tyre manufacturing company wants to assess the relationship between car tyre thickness (labelled as `thickness` in `R`) and the probability of a tyre splitting after 20,000 miles of use on two different road surfaces A and B (labeled as `surface` in `R`). For various tyre thicknesses and for each road surface they determine what proportion of the tyres split. The number of tyres tested at each tyre thickness is labelled as `tested` and the number splitting at each tyre thickness is labeled as `split`. The information is given in the table below.

(i) The following command is used to fit a model to the data:

```
lm1<-glm(split/tested~surface*thickness,weights=tested,
family=binomial(logit))
```

Write down, in terms of the linear predictor η_i, the statistical model for $E(y_i)$ that is fitted to the data. Specify η_i in terms of the variables and parameters in the model. ***(3 marks)***

(ii) The deviance in a glm with a binomial response is given by

$$D(y,\hat{\mu}) = -2\sum_i n_i \left\{ y_i log\left(\frac{\hat{\mu}_i}{y_i}\right) + (1-y_i) log\left(\frac{1-\hat{\mu}_i}{1-y_i}\right)\right\}.$$

For surface A, given the information in the table below, write down the numerical values of n_i and y_i for all values of i. ***(2 marks)***

Surface A			Surface B		
thickness (mm)	number tested	number split	thickness (mm)	number tested	number split
2.3	100	75	2.3	100	88
2.9	50	25	2.9	50	26
3.4	40	11	3.4	40	14
3.9	50	10	3.9	50	12
4.3	50	3	4.3	50	13

5 (continued)

(iii) Some further edited R commands and output are provided below

```
> summary(lm1)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         5.09125    0.64787   7.858 3.89e-15
surfaceB            0.04737    0.90763   0.052    0.958
thickness          -1.74354    0.21124  -8.254  < 2e-16
surfaceB:thickness  0.16636    0.28716   0.579    0.562

> anova(lm1)
Analysis of Deviance Table

Model: binomial, link: logit

Response: split/tested

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev
NULL                                  9    194.483
surface            1    5.822         8    188.662
thickness          1  176.614         7     12.048
surface:thickness  1    0.337         6     11.711

> print(vcov(lm1),digits=3)
                   (Intercept) surfaceB thickness surfaceB:thickness
(Intercept)              0.420   -0.420   -0.1336             0.1336
surfaceB                -0.420    0.824    0.1336            -0.2543
thickness               -0.134    0.134    0.0446            -0.0446
surfaceB:thickness       0.134   -0.254   -0.0446             0.0825

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

(a) By performing a series of hypothesis tests, assess the effects of road surface and tyre thickness on the response. ***(4 marks)***

(b) Using the output above, calculate the Pearson residual for 3.4 mm thick tyres tested on surface A. ***(4 marks)***

(c) Using the output above, calculate the odds ratio of a tyre with a thickness of 5mm splitting on surface A compared to a tyre with a thickness of 3.4mm splitting on surface A. Give a 95% confidence interval for this odds ratio and interpret the result. ***(5 marks)***

5 (continued)

(d) Give an example, in words, of an odds ratio of a tyre splitting involving the variables in the output above that requires the covariance between the parameter estimate for `surface` and the parameter estimate for the interaction between `surface` and `thickness` but does not require any of the covariances involving the parameter estimate for `thickness`. ***(2 marks)***

6 Data were collected relating to collisions involving cyclists in a particular city over a given period of time. The interest was in the factors affecting the outcome for cyclists (indicated by the variable `outcome`). The two main factors of interest were whether the cyclist was wearing a helmet (indicated by the variable `helmet`) and what the cyclist collided with (indicated by the variable `collision`). `Collision` is either `lorry`, `car` or `pedestrian`. `Outcome` is either `serious` or `minor`. The data is stored in the data frame `cycling` in `R`. The first few lines of the data frame are shown below.

```
> head(cycling)
  count collision helmet outcome
1    25     lorry      1       1
2    23     lorry      1       0
3    18     lorry      0       1
4    12     lorry      0       0
5    10       car      1       1
6    44       car      1       0
```

The full data are shown below.

helmet			no helmet		
collision	serious	minor	collision	serious	minor
lorry	25	23	lorry	18	12
car	10	44	car	17	20
pedestrian	1	8	pedestrian	1	6

(i) Explain why `outcome` is a response whilst `helmet` and `collision` are controlled factors. What is the minimal model when fitting linear models to these data? ***(2 marks)***

(ii) By looking at the proportions of cyclists suffering serious accidents in each category, make some brief initial observations about the relationship between the probability of having a serious cycling accident and helmet use and type of collision in this data set. ***(3 marks)***

6 (continued)

(iii) Various models are fitted to the data. A summary of the residual deviances for the models fitted and some quantiles are given in the table below.

Model	Residual Deviance	Df
helmet*collision+outcome	25.643	5
helmet*collision+outcome*helmet	20.751	4
helmet*collision+outcome*collision	8.371	3
helmet*collision+outcome*helmet+outcome*collision	2.318	2

```
> qchisq(0.95,1)
[1] 3.841459
> qchisq(0.95,2)
[1] 5.991465
```

Specify the terms in the model with linear predictor given by `helmet*collision+outcome` and hence explain why the degrees of freedom is 5. ***(2 marks)***

(iv) Specify an algebraic form for the linear predictor `helmet*collision+outcome*helmet` ***(2 marks)***

(v) With reference to the residual deviances in the table above, describe the most suitable model for the data. Discuss whether the model you select based on the residual deviances is consistent with your observations from part (ii). ***(7 marks)***

(vi) Calculate the estimated expected number of serious injuries for cyclists wearing helmets in collision with a car for the model with linear predictor `helmet*collision+outcome*collision`. ***(4 marks)***

End of Question Paper