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 2010–2011**

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 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. Figure 1 shows a plot of tooth length versus dose by delivery method. Consider the following linear model:

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

where y_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 1 if the dose was administered by orange juice and zero 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     2     3     4     5     6     7     8     9    10    11
0.052 0.052 0.052 0.052 0.052 0.052 0.052 0.052 0.052 0.052 0.035
   12    13    14    15    16    17    18    19    20    21    22
0.035 0.035 0.035 0.035 0.035 0.035 0.035 0.035 0.035 0.063 0.063
   23    24    25    26    27    28    29    30    31    32    33
0.063 0.063 0.063 0.063 0.063 0.063 0.063 0.063 0.052 0.052 0.052
   34    35    36    37    38    39    40    41    42    43    44
0.052 0.052 0.052 0.052 0.052 0.052 0.052 0.035 0.035 0.035 0.035
   45    46    47    48    49    50    51    52    53    54    55
0.035 0.035 0.035 0.035 0.035 0.035 0.063 0.063 0.063 0.063 0.063
   56    57    58    59    60
0.063 0.063 0.063 0.063 0.063
```

 Question 1 continued on next page

1 (continued)

(i) For the `tooth.lm` linear model fitted above, the residual for observation 6 is -0.454; provide an `R` command that would give the residual for observation 6. ***(1 mark)***

(ii) Calculate the 6th scaled residual. ***(2 marks)***

(iii) Calculate the 6th standardized residual. ***(2 marks)***

(iv) Calculate the 6th standardized deletion residual. ***(2 marks)***

(v) Fig 2 shows some diagnostic plots for the model fitted. Comment on whether the plots indicate that the assumptions of the linear model are met. ***(2 marks)***

(vi) Calculate a 95% confidence interval for β_1. ***(4 marks)***

(vii) Let X be the design matrix in which the first column has a one in every position, the second column contains the vitamin C dose and the third column contains the indicator variable values for delivery method as described earlier in the question. Given that the variance-covariance matrix of $\hat{\beta}$ given in the `R` output above is $\hat{\sigma^2}(X^TX)^{-1}$, perform a single hypothesis test to test whether both $\beta_0 = 2\beta_1$ and $\beta_2 = 0$. You may find the following helpful:

$$\left(\begin{pmatrix} 1 & -2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1.64 & -0.90 & -0.60 \\ -0.90 & 0.77 & 0 \\ -0.60 & 0 & 1.20 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -2 & 0 \\ 0 & 1 \end{pmatrix}\right)^{-1} = \begin{pmatrix} 0.12 & 0.06 \\ 0.06 & 0.87 \end{pmatrix}$$

. ***(7 marks)***

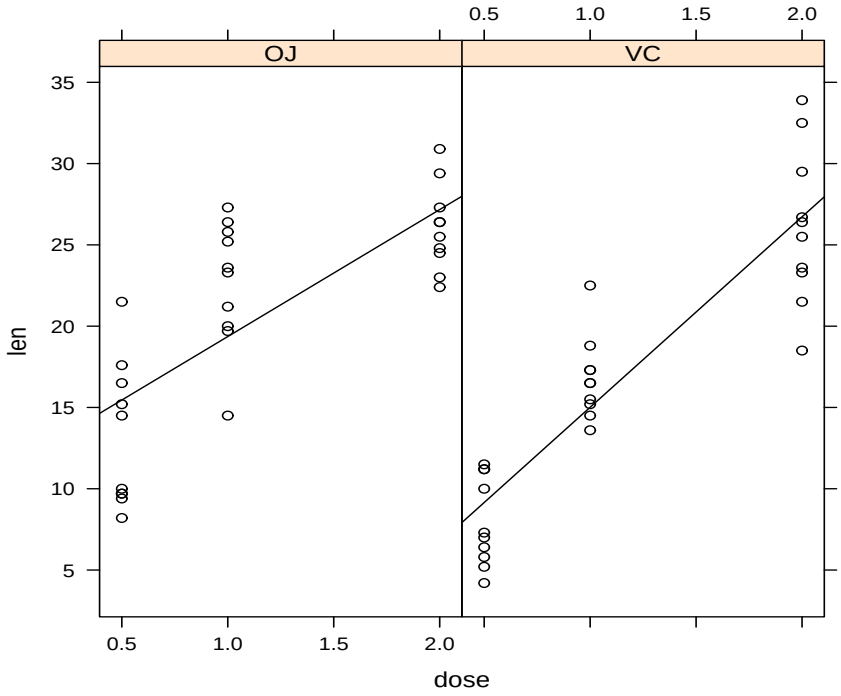


Figure 1: Tooth length against vitamin C dose by delivery method

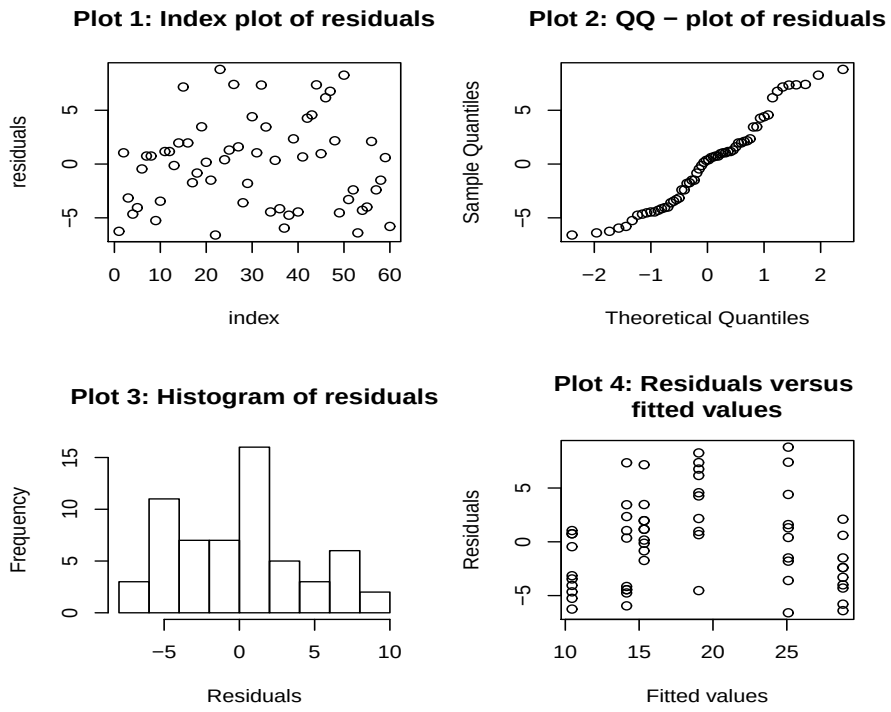


Figure 2: Residual plots for the guinea pig data

2 The linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is to be fitted by the least squares method using n observations; $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters and it is assumed that $\boldsymbol{X}$ is of full rank p. The errors $\boldsymbol{\varepsilon}$ are assumed to be independent and normally distributed and to have zero mean vector and covariance matrix $\sigma^2 I_n$, where $\sigma^2 > 0$. Let $\widehat{\boldsymbol{\beta}}$ be the LS estimate of $\boldsymbol{\beta}$.

(i) Given that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ show that $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$. ***(2 marks)***

(ii) Let $Q = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. By noting that

$$Q = \{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^T\{\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}$$

show that

$$Q = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

(3 marks)

(iii) Use Q to derive the distributions of Q/σ^2, $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/\sigma^2$ and $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{X}^T\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2$, stating any general results to which you appeal. ***(5 marks)***

The data for the rest of this question relate to how protein concentration (conc) can be predicted from the optical density of an experimental assay. Output from an `R` session is given below. Some of the numerical values have been replaced by letters.

```
> DNase
           conc density
1    0.04882812   0.017
2    0.04882812   0.018
3    0.19531250   0.121
4    0.19531250   0.124
5    0.39062500   0.206
6    0.39062500   0.215
7    0.78125000   0.377
8    0.78125000   0.374
9    1.56250000   0.614
10   1.56250000   0.609
11   3.12500000   1.019
12   3.12500000   1.001
13   6.25000000   1.334
14   6.25000000   1.364
15 12.50000000   1.730
16 12.50000000   1.710
```

 Question 2 continued on next page

2(continued)

```
> DNase.lm<-lm(conc~density)

> summary(DNase.lm)
Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  -1.3010     0.5878  -2.213      0.044
density       6.5100     0.6586   9.885   1.08e-07
---
Residual standard error: A on 14 degrees of freedom
Multiple R-squared: 0.875,     Adjusted R-squared: 0.866
F-statistic: 97.71 on 1 and 14 DF,  p-value: 1.078e-07

> anova(DNase.lm)
Analysis of Variance Table
Response: conc
          Df  Sum Sq  Mean Sq  F-value     Pr(>F)
density    1 229.360  229.360   97.712  1.078e-07
Residuals 14  32.862  2.347

> DNase.lm$resid[1:8]
     1      2    3      4      5      6      7      8
 1.239  1.233    B  0.689  0.351  0.292 -0.372 -0.353

> vcov(DNase.lm)
            (Intercept)    density
(Intercept)   0.3455301 -0.2936573
density      -0.2936573  0.4337225

> influence(DNase.lm)$hat[2]
        2
        C
```

(iv) Calculate the value of A. *(3 marks)*

(v) Calculate the value of B. *(3 marks)*

(vi) Calculate the value of C. *(4 marks)*

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

- sex - the gender of the crab
- sp - the species (blue or orange)
- FL - frontal lobe size (mm)
- RW - rear width (mm)
- CL - shell length (mm)
- CW - carapace width (mm)
- BD - body depth (mm)

The interest is in how the frontal lobe size depends on the other variables collected. The investigator uses `R` to help choose a statistical model.

(i) In the following output, explain what the '`regsubsets`' command does and which of the various sized subsets you would select based on the outputs of the three '`summary`' commands.

```
> a<-regsubsets(FL~factor(sex)+factor(sp)+RW+CL+CW+BD)
> summary(a)
Subset selection object
Call:regsubsets.formula(FL~factor(sex)+factor(sp)+RW+CL+CW+BD)
6 Variables  (and intercept)

1 subsets of each size up to 6
Selection Algorithm: exhaustive
         factor(sex)M factor(sp)O RW  CL  CW  BD
1  ( 1 ) " "          " "         " " " " " " "*"
2  ( 1 ) " "          "*"         " " " " "*" " "
3  ( 1 ) "*"          "*"         " " "*" " " " "
4  ( 1 ) "*"          "*"         " " "*" "*" " "
5  ( 1 ) "*"          "*"         " " "*" "*" "*"
6  ( 1 ) "*"          "*"         "*" "*" "*" "*"
> summary(a)$rsq
[1] 0.975407 0.986325 0.987568 0.988003 0.988074 0.988087
> summary(a)$cp
[1] 202.433  27.548   9.404   4.354   5.213   7.000
> summary(a)$bic
[1] -730.466 -842.548 -856.314 -858.142 -854.021 -848.944
```

(5 marks)

3 (continued)

(ii) Describe what is being done at each stage of the R output below.

```
> crabs.lm<-lm(FL~1,data=crabs)
> step(crabs.lm,scope=list(upper=FL~factor(sex)+CW+BD+RW,
+ data=crabs),direction="both")
Start:  AIC=501.57
FL ~ 1

              Df Sum of Sq     RSS     AIC
+ BD           1   2371.45   59.79 -237.49
+ CW           1   2263.83  167.42  -31.57
+ RW           1   2000.00  431.24  157.67
<none>                     2431.24  501.57
+ factor(sex)  1      4.56 2426.68  503.19

Step:  AIC=-237.49
FL ~ BD

              Df Sum of Sq     RSS     AIC
+ RW           1      9.62   50.17 -270.59
+ factor(sex)  1      5.03   54.76 -253.07
+ CW           1      3.19   56.60 -246.47
<none>                      59.79 -237.49
- BD           1   2371.45 2431.24  501.57

Step:  AIC=-270.59
FL ~ BD + RW

              Df Sum of Sq    RSS     AIC
+ CW           1      0.58  49.59 -270.91
<none>                      50.17 -270.59
+ factor(sex)  1      0.24  49.92 -269.56
- RW           1      9.62  59.79 -237.50
- BD           1    381.07 431.24  157.67

Step:  AIC=-270.91
```

3 (continued)

```
FL ~ BD + RW + CW

               Df Sum of Sq     RSS      AIC
<none>                      49.588 -270.913
- CW            1     0.579  50.167 -270.593
+ factor(sex)  1     0.034  49.554 -269.052
- RW            1     7.011  56.599 -246.466
- BD            1    99.143 148.732  -53.235

Call:
lm(formula = FL ~ BD + RW + CW, data = crabs)

Coefficients:
(Intercept)           BD           RW           CW
    0.71812      0.82993      0.16991      0.02901
```

(5 marks)

(iii) The AIC for the chosen model using the '`steps`' command above is -270.91; show how to calculate this value based on the 'RSS' values in the output.

(2 marks)

(iv) The investigator looks at a plot for a particular statistical model and concludes that it shows evidence of heteroscedasticity.

(a) Explain what heteroscedasticity means in a linear model.

(1 mark)

(b) Justify which plot would allow heteroscedasticity to be observed and what you would expect to see. ***(2 marks)***

(c) Explain why heteroscedasticity is a problem in linear models.

(1 mark)

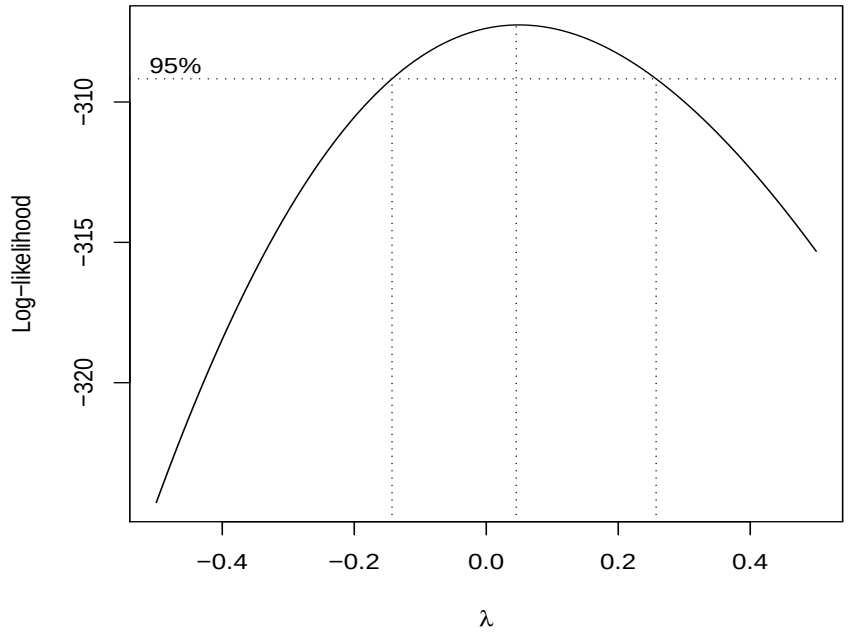


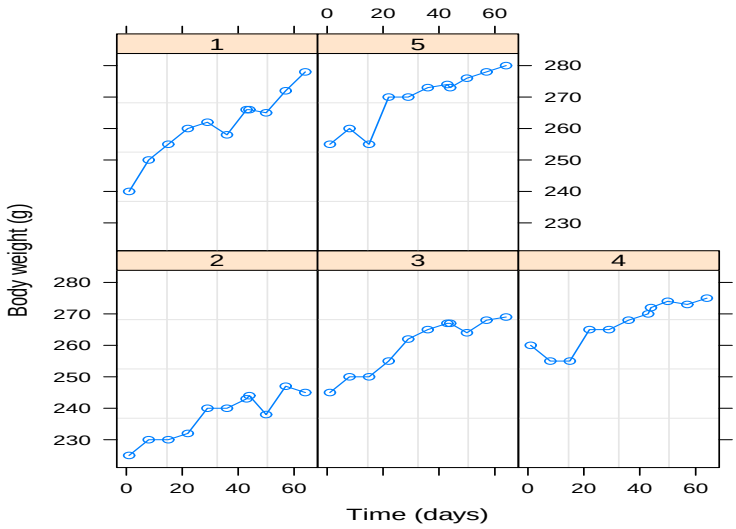
Figure 3: Log-likelihood function for the Box-Cox family of transformations

3 (continued)

(v) Figure 3 shows a plot of the log-likelihood as a function of the Box-Cox family parameter λ. Give a point estimate and a 95% confidence interval for λ based on the plot and hence explain whether the investigator's claim of heteroscedasticity is justified. ***(4 marks)***

4 This question relates to data on the body weights of 5 rats (selected from a large population of rats) measured at various time points over 64 days. The body weights of the rats (in grams) are measured on day 1 and every seven days thereafter until day 64, with an extra measurement on day 44. Several weeks before day 1, the rats were given a drug that was thought to slow the metabolism and hence increase body weight. The interest is in how the weight changes with time in the rat population.

Figure 1: Change in body weight with time by rat



The `R` commands used to fit two linear mixed effects models to the data are

```
rats1.lme<-lme(weight~Time,Bodyweight,random=~1|Rat)
rats2.lme<-lme(weight~Time,Bodyweight,random=~Time|Rat)
```

(i) For each of the models `rats1.lme` and `rats2.lme`, write down the algebraic specification of the model, defining any terms that you use. ***(4 marks)***

(ii) Explain why the designation of explanatory variables as fixed and random in the models is reasonable, given the aim of the study. ***(2 marks)***

(iii) Discuss which of these two models seems the most sensible based on Figure 1. ***(1 mark)***

4 (continued)

(iv) Explain what each of the `anova` commands below does and what conclusions you would draw based on them.

```
> anova(rats1.lme,rats2.lme)
          Model df      logLik   Test  L.Ratio p-value
rats1.lme     1  4  -154.0386
rats2.lme     2  6  -153.3843 1 vs 2 1.308685  0.5198

> anova(rats1.lme)
            numDF denDF   F-value p-value
(Intercept)     1    49 2134.6668  <.0001
Time            1    49  309.1238  <.0001
```

(4 marks)

(v) Describe the model checks that you would carry out to check the assumptions of the `rats1.lme` model. You do not need to provide the `R` code to do the checks but you should state what values you would assign to the `levels` option within `R` where appropriate. *(5 marks)*

(vi) For each of the three mixed effects models below, state the algebraic form of the variance-covariance matrix of the within-subject random effects:

```
rats2.lme<-lme(weight~Time,Bodyweight,random=~Time|Rat)
rats3.lme<-lme(weight~Time,Bodyweight,random=pdCompSymm(~Time))
rats4.lme<-lme(weight~Time,Bodyweight,random=pdDiag(~Time))
```

(4 marks)

5 The data in this question relate to the occurrence of pain in the chest of patients with heart disease after taking a single deep breath. The response, y, is the occurrence (y = 1) or non-occurrence (y = 0) of pain. The continuous explanatory variables are the volume and rate of air inhaled in a single breath. The only factor explanatory variable is the type of drug each subject was taking (each subject was on one of three different drugs labelled 1,2 and 3). The three explanatory variables are labelled volume, rate and drug.

Some edited R output is provided below:

```
> model1.glm<-glm(Y~Rate*factor(drug),binomial)
> summary(model1.glm)

Call:
glm(formula = Y ~ Rate * factor(drug),family = binomial)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.7027     1.8603  -0.915    0.360
Rate                    1.6388     1.3751   1.192    0.233
factor(drug)2          0.5562     2.4575   0.226    0.821
factor(drug)3          0.1082     2.2615   0.048    0.962
Rate:factor(drug)2    -1.0691     1.5464  -0.691    0.489
Rate:factor(drug)3    -0.7323     1.5752  -0.465    0.642

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.040  on 38  degrees of freedom
Residual deviance: 47.851  on 33  degrees of freedom
AIC: 59.851
```

(i) Write down an algebraic specification of the linear predictor (η_i) for patient i in the above model. ***(3 marks)***

(ii) How does $E(y_i)$, the expected value of the response for patient i, relate to η_i for the fitted model? ***(1 mark)***

(iii) Based on the above output, what proportion of patients with heart disease on drug 2 and with an inhalation rate of 1.1 would you expect to feel chest pain? ***(3 marks)***

(iv) Based on the above output, calculate the odds ratio of chest pain for a patient on drug 3 with an inhalation rate of 1.2 compared to a patient on drug 2 with an inhalation rate of 0.7 units. ***(4 marks)***

(v) Assess the model fit of the above model. ***(2 marks)***

(vi) Describe, in detail, how to use iteratively re-weighted least squares to estimate the parameters in the model fitted in the output above. You should derive the algebraic form of any vectors or matrices that you suggest using. ***(7 marks)***

6 Data is collected on 189 mothers for a study in which the interest is in how the race and smoking status of mothers affects the birth weight of their babies. The variables recorded in the study are 'race', 'smoke' and 'low'.

- 'race' is a categorical variable with 3 levels (A,B and C);
- 'smoke' is a binary variable (1 = smoked in pregnancy, 0 = did not smoke in pregnancy);
- 'low' is a binary variable that records whether the baby was low weight at birth (1=was low weight, 0=wasn't low weight).

The results are shown in Table 1.

	Race					
	A		B		C	
	smoke		smoke		smoke	
Low	0	1	0	1	0	1
0	40	33	11	4	35	7
1	4	19	5	6	20	5

Table 1: Birth weight of babies by race and smoking status of the mother.

For notational convenience we use L, R and S to represent the variables low, race and smoke respectively. For this question, assume that L is a response factor and R and S are controlled factors.

Several log-linear models with Poisson errors were fitted giving the following results:

Model fitted	Res. Deviance	df
R*S	45.19	6
R*S+L	17.85	5
R*S+L*R	12.844	
R*S+L*S	12.987	

(i) Explain why R*S appears in all the above fitted models. What is the effect of including R*S on row and column totals in the table of fitted values?
(3 marks)

(ii) State, with justification, the 2 missing degrees of freedom in the output above. ***(4 marks)***

(iii) Assuming appropriate model-checking diagnostics proved satisfactory, what would you conclude about the dependence of birth weight on race and smoking status during pregnancy? Refer to the residual deviances above.
(5 marks)

6 (continued)

(iv) For the R*S+L*R model, calculate the expected number of low birth weight babies with mothers having a race code of B that did not smoke during pregnancy. ***(3 marks)***

(v) Calculate the Pearson residual for low birth weight babies with mothers having a race code of B that did not smoke during pregnancy. ***(2 marks)***

(vi) Some edited `R` output is given below (where `count` is the variable recording the number of mothers in each cell):

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

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.51030    0.16127  21.766  < 2e-16
raceB       -1.28776    0.34163  -3.769 0.000164
raceC        0.03001    0.23038   0.130 0.896361
smoke        0.16705    0.20484   0.816 0.414759
low         -1.15497    0.23912  -4.830 1.36e-06
raceB:smoke -0.63706    0.45217  -1.409 0.158868
raceC:smoke -1.68948    0.37878  -4.460 8.18e-06
raceB:low    0.84481    0.46341   1.823 0.068301
raceC:low    0.63617    0.34783   1.829 0.067405
```

Calculate the expected number of low birth weight babies with mothers having a race code of C that did smoke during pregnancy. ***(3 marks)***

End of Question Paper