The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2015–2016**

Linear Models **3 hours**

Marks will be awarded for your best ***five*** *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 100 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 A group of senior citizens who have never used the internet before are given training over a period of 6 months. A sample of 3 of them is chosen at random and their numbers of hours of internet use are recorded for the 6 months, as shown in Figure 1 below.

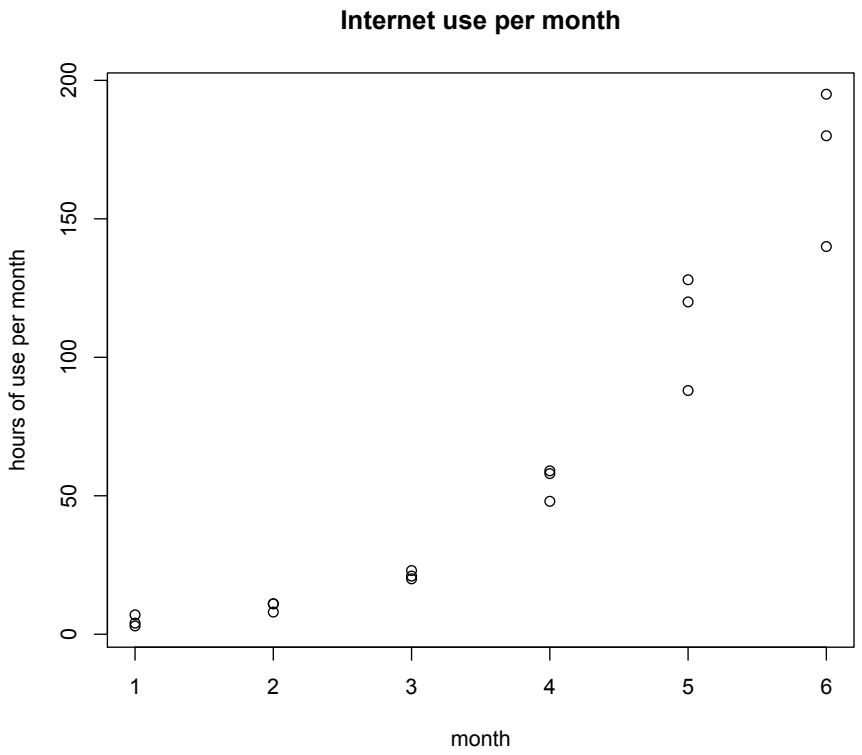

Figure 1: Plot of number of hours of internet use and months

(i) Describe briefly the data, discussing any interesting features. Based on Figure 1 only suggest the form of a possible linear model of the hours of use per month (as response variable) and month (as explanatory variable).
(2 marks)

1 (continued)

(ii) Let y be the hours of use per month and x be the month. An analysis in R gave the following output:

```
> summary(fit)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max
-33.393  -2.917   0.858   4.307  21.607

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.700     14.000   1.479   0.1599
x            -23.230      9.159  -2.536   0.0228 *
I(x^2)         8.113      1.281   6.334 1.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Residual standard error: 13.56
Multiple R-squared:  0.9602, Adjusted R-squared:  0.9549
F-statistic:   181 on 2 and 15 DF,  p-value: 3.152e-11
```

(a) Write down the fitted model. ***(2 marks)***

(b) Comment on the model and the quality of its goodness of fit, making appropriate reference to any goodness of fit diagnostics. State clearly any hypothesis you may use. ***(6 marks)***

(c) Using one of the following R extracts

```
> qnorm(0.95)
[1] 1.644854
> qt(0.95, df=15)
[1] 1.75305
```

```
> qt(0.95, df=14)
[1] 1.76131
> qt(0.995, df=15)
[1] 2.946713
```

calculate 90% confidence intervals for the coefficient of x and for the coefficient of x^2. ***(3 marks)***

(d) For month $x = 1$ calculate a 90% predictive interval for the future observation y. You may use the following:

$$(X^TX)^{-1} = \begin{pmatrix} 1.066 & -0.650 & 0.083 \\ -0.650 & 0.456 & -0.063 \\ 0.083 & -0.063 & 0.009 \end{pmatrix},$$

where X is the design matrix of the linear model. ***(5 marks)***

1 (continued)

(e) A further R analysis gave

```
> vcov(fit)
            (Intercept)           x     I(x^2)
(Intercept)   196.00135 -119.43833  15.312606
x            -119.43833   83.89120 -11.484454
I(x^2)         15.31261  -11.48445   1.640636
```

Calculate the correlation coefficient of the estimator of the gradient (coefficient of x) and the estimator of the coefficient of x^2.

(2 marks)

2 A data-set on black cherry trees in the Allegheny National Forest, Pennsylvania, USA includes the height, radius (measured 4.5 feet above the ground) and volume, for each of 31 trees.

(i) A model

$$v_i = \beta_0 + \beta_1 r_i + \beta_2 h_i + \epsilon_i \qquad (1)$$

has been proposed, where h_i, r_i, v_i are the natural logarithms of the height (in feet), radius (in feet) and volume (in cubic feet) of the ith tree, and $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ independently for different trees. The following output summarizes the results of fitting this model in R.

```
> summary(cherry)

Call:
lm(formula = v ~ r + h)

Residuals:
      Min        1Q    Median        3Q       Max
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.33065    0.91031  -0.363    0.719
r            1.98265    0.07501  26.432  < 2e-16 ***
h            1.11712    0.20444   5.464 7.81e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared:  0.9777,    Adjusted R-squared:  0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16

> anova(cherry)
Analysis of Variance Table

Response: v
          Df Sum Sq Mean Sq F value    Pr(>F)
r          1 7.9254  7.9254 1196.53 < 2.2e-16 ***
h          1 0.1978  0.1978   29.86 7.805e-06 ***
Residuals 28 0.1855  0.0066
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Explain the hypothesis being tested by each of the three F statistics included in the output. What interpretation, if any, can be placed on their conclusions here? ***(6 marks)***

2 (continued)

(ii) Figure 2 shows the standardized deletion residuals for the model above.

The following calculations can be used as the basis of a test on the standardized deletion residuals, using the Šidák correction.

```
> alpha=0.05
> prob=1-(1-alpha)^(1/31)
> qt(prob/2,27)
[1] -3.495321
```

Explain the interpretation of the values `alpha` and `prob` used in the calculation, and carry out the test. *(4 marks)*

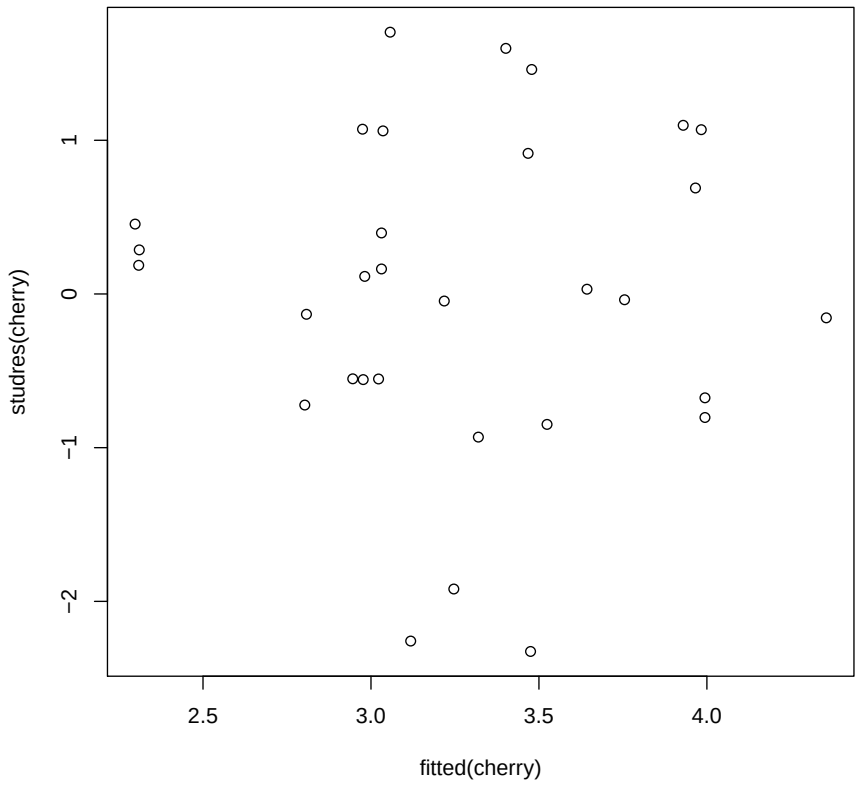

Figure 2: Standardized deletion residuals for the cherry tree model

2 (continued)

(iii) Thinking about the trunk of each tree as a cylinder, a simple geometric calculation suggests that

$$V_i \approx kR_i^2 H_i \tag{2}$$

where $V_i = \exp(v_i)$ etc., and that $k \approx \pi$ (the usual circular constant). Explain why the model suggested by (2) can be represented as a special case of (1) under the null hypothesis that $\beta_1 = 2$ and $\beta_2 = 1$, and explain how that null hypothesis can be written in the general form

$$C\boldsymbol{\beta} = \boldsymbol{c}.$$

(3 marks)

Express the weaker hypothesis that $\beta_1 + \beta_2 = 3$ in a similar form, and calculate the corresponding F statistic, using the fact that

$$G = (X^T X)^{-1} = \begin{pmatrix} 125.1 & 5.839 & -28.07 \\ 5.839 & 0.8495 & -1.227 \\ -28.07 & -1.227 & 6.310 \end{pmatrix}.$$

What is the null distribution of this F statistic? ***(7 marks)***

3 (i) A laboratory experiment is intended to investigate the effect of a drug on certain species of micro-organisms. Tissue cultures containing set amounts of one of three species of micro-organisms (A, B, C) are each exposed to doses of the drug being tested; there are four different doses used, and two replicates of each combination of species and dose. Figure 3 shows a plot produced in R of the dose and response for each run, the points being coded by species.

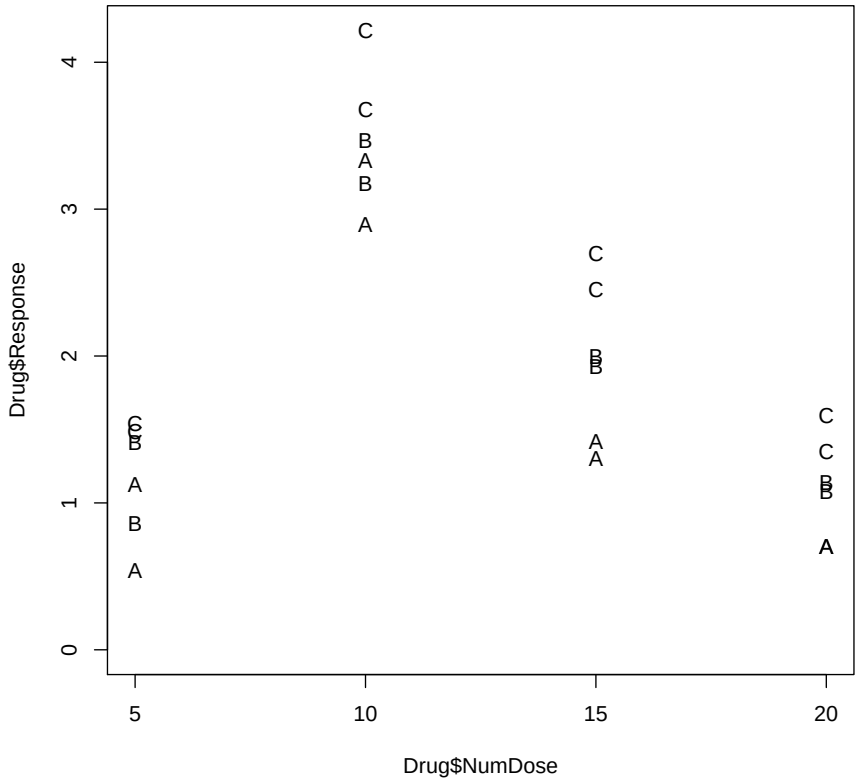

Figure 3: Raw data in the laboratory experiment in Question 3

3 (continued)

Various models are being considered for the response as a function of species and dose. The output below shows summaries of results for two models; `Response` and `Species` have the obvious meaning, `NumDose` refers to the dose as a quantitative variable, and `FacDose` refers to the dose as a factor variable.

```
> summary(FacModel)

Call:
lm(formula = Response ~ Species + FacDose, data = Drug)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25837 -0.15868  0.02226  0.07398  0.38053

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.74455    0.11078   6.721 2.66e-06 ***
SpeciesB     0.37777    0.11078   3.410  0.00312 **
SpeciesC     0.87320    0.11078   7.882 3.02e-07 ***
FacDose10    2.30045    0.12791  17.984 5.98e-13 ***
FacDose15    0.80540    0.12791   6.296 6.18e-06 ***
FacDose20   -0.06504    0.12791  -0.508  0.61730
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2216 on 18 degrees of freedom
Multiple R-squared:  0.9657,    Adjusted R-squared:  0.9562
F-statistic: 101.3 on 5 and 18 DF,  p-value: 1.551e-12
```

3 (continued)

```
> summary(QuadModel)

Call:
lm(formula = Response ~ Species + NumDose + I(NumDose^2), data = Drug)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9074 -0.4361  0.1135  0.3315  0.9608

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.036322   0.698646  -2.915  0.00889 **
SpeciesB      0.377774   0.297904   1.268  0.22008
SpeciesC      0.873198   0.297904   2.931  0.00857 **
NumDose       0.758921   0.123549   6.143 6.64e-06 ***
I(NumDose^2) -0.031709   0.004865  -6.518 3.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5958 on 19 degrees of freedom
Multiple R-squared:  0.7381,    Adjusted R-squared:  0.6829
F-statistic: 13.39 on 4 and 19 DF,  p-value: 2.378e-05
```

(a) Give the equations for these two models, explaining your notation and assumptions. ***(6 marks)***

(b) Calculate the BIC for each of these two models. Based on the BIC, explain which of the two models you would prefer and why. ***(5 marks)***

(c) What advantages and disadvantages do these two modelling approaches—dose as a factor, and dose as a numerical variable—have for this experiment, beyond those taken into account in the BIC? ***(2 marks)***

3 (continued)

(ii) Consider the linear model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{3}$$

where ϵ_i is an i.i.d. sequence of random variables with zero mean and variance $\mathrm{Var}(\epsilon_i) = \sigma^2 c_i$, for some variance σ^2 and $c_i > 0$.

Discounted least squares considers the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, which minimises the discounted sum of squares

$$S_\delta(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta^{n-i} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2,$$

for some discount factor δ that satisfies $0 < \delta \leq 1$.

(a) Show that discounted least squares is a special case of weighted least squares (WLS) and calculate the weights of WLS as functions of δ. ***(2 marks)***

(b) Using the relationship of discounted least squares and WLS as in (a), derive the variance of ϵ_i as a function of σ^2 and δ. ***(1 mark)***

(c) For the simple linear regression model with no intercept and a near constant covariate $x_i \approx x$, i.e.

$$y_i \approx x\beta + \epsilon_i,$$

show that

$$\hat{\beta} = \frac{(1-\delta)}{x(1-\delta^n)} \sum_{i=1}^{n} \delta^{n-i} y_i.$$

(4 marks)

4 A study was conducted to investigate alcohol consumption in adolescents. For a period of three years, 82 different adolescents had their alcohol consumption recorded for the years in which they were 14, 15 and 16 years of age. In R, each alcohol measurement is recorded in the variable `alcuse`. Each subject is given a unique id (stored in the R vector `id`), and their age during a particular year is stored as the variable `age`. Interest lies in the average alcohol use and its change through adolescence, and in the differences between subjects. Below is some R output.

```
> age_15 = age -15
> fit1 <- lmer(alcuse ~ age_15 + (age_15|id))
> summary(fit1)
Linear mixed model fit by REML ['lmerMod']
Formula: alcuse ~ age_15 + (age_15 | id)

REML criterion at convergence: 643.2

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.48287 -0.37933 -0.07858  0.38876  2.49284

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 0.6480   0.8050
          age_15      0.1552   0.3939   0.26
 Residual             0.3373   0.5808
Number of obs: 246, groups:  id, 82

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.92195    0.09630   9.574
age_15       0.27065    0.06284   4.307

Correlation of Fixed Effects:
       (Intr)
age_15 0.169
```

(i) Let Y_{ij} be the alcohol consumption of the i^{th} individual during the j^{th} year for which they are observed, and let x_{ij} be the age of that person minus 15, i.e.

$$x_{ij} = \mathsf{age}_{ij} - 15$$

so that $x_{i1} = -1, x_{i2} = 0$, and $x_{i3} = 1$. Write down the equation of the model that has been fitted, defining your notation carefully and specifying appropriate distributions for the terms in the model. ***(3 marks)***

(ii) Briefly justify the choice of fixed and random effects in the model. ***(2 marks)***

4 (continued)

(iii) Give the estimated values for all 6 parameters in the model *(2 marks)*

(iv) Calculate the covariance between a subject's alcohol consumption at the age of 14, and their alcohol consumption at the age of 16. *(4 marks)*

(v) If $I = 82$ is the number of individuals in the dataset, and if σ^2 is the random error variance and σ_1^2 is the variance of the random intercepts, show that

$$\mathbb{V}ar(\hat{\beta}) = \frac{1}{3I}(3\sigma_1^2 + \sigma^2)$$

where $\hat{\beta}$ is the estimate of the average alcohol consumption for all adolescents. *(6 marks)*

(vi) The session is continued below.

```
> fit2 <- lmer(alcuse ~ age_15 + (age_15-1|id)+(1|id))
> fit2
Linear mixed model fit by REML ['lmerMod']
Formula: alcuse ~ age_15 + (age_15 - 1 | id) + (1 | id)
REML criterion at convergence: 645.5346
Random effects:
 Groups   Name        Std.Dev.
 id       age_15      0.3939
 id.1     (Intercept) 0.8050
 Residual             0.5808
Number of obs: 246, groups:  id, 82
Fixed Effects:
(Intercept)       age_15
     0.9220       0.2707
>
> x <- - 2*(logLik(fit1, REML=FALSE)-logLik(fit2, REML=FALSE))
> N<-100
> y<- c()
> for(i in 1:N){
+   z<-unlist(simulate(fit1))
+   fit1sim<-lmer(z ~ age_15 + (age_15|id))
+   fit2sim<-lmer(z ~ age_15 + (age_15-1|id)+(1|id))
+   y[i] <- - 2*(logLik(fit1sim, REML=FALSE)-logLik(fit2sim, REML=FALSE))
+ }
> mean(y>=x)
[1] 0.53
```

What is being tested here? Give the name of the procedure that has been used, and interpret the output. *(3 marks)*

5 Three models are proposed for the number T_{ij} of people moving between origin i and destination j. In each of these models T_{ij} is assumed to be Poisson distributed with mean μ_{ij}. The form of μ_{ij} for each of the three models is as follows

$$\begin{aligned} \text{Model 1} \quad \mu_{ij} &= {d_{ij}}^{\gamma} e^{\alpha_i + \beta_j} \\ \text{Model 2} \quad \mu_{ij} &= d_{ij} e^{\alpha_i} \\ \text{Model 3} \quad \mu_{ij} &= {d_{ij}}^{\gamma} e^{\alpha_i} \end{aligned}$$

where d_{ij} is the known distance between origin i and destination j and α_i, β_j and γ are unknown parameters. In an `R` data-set for this model:

- `T` denotes the number of people moving;
- `A` denotes the factor variable for origin;
- `B` denotes the factor variable for destination;
- `D` denotes the covariate origin-destination distances.

(i) Give an appropriate `R` command (in terms of `T`, `A`, `B` and `D`) that would allow model 1 to be fitted using a generalized linear model. Explain what your instructions do and define any other notation you use. ***(4 marks)***

(ii) Given observations $\{t_{ij}\}$ show that the log-likelihood for all three models is $\sum_i \sum_j t_{ij} \log \mu_{ij} - \mu_{ij} + \text{constant}$. ***(3 marks)***

(iii) For Model 1 differentiate the log-likelihood with respect to α_i and β_j. If $\hat{\alpha}_i$, $\hat{\beta}_j$ and $\hat{\gamma}$ are the maximum likelihood estimators of α_i, β_j and γ respectively, show that the fitted values $\hat{t}_{ij} = {d_{ij}}^{\hat{\gamma}} e^{\hat{\alpha}_i + \hat{\beta}_j}$ must satisfy the equations $\sum_i t_{ij} = \sum_i \hat{t}_{ij}$ and $\sum_j t_{ij} = \sum_j \hat{t}_{ij}$. ***(5 marks)***

(iv) Explain how plots of $\log \mu_{ij}$ against $\log d_{ij}$ would differ for Model 1 and Model 2. ***(4 marks)***

(v) Explain how you would test the null hypothesis that the number of people moving depends on distance between origin and destination and on the origin itself but not the destination? ***(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. Table 1 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`. A value of `No` for `Low` indicates the baby wasn't a low weight baby. For notational convenience we use `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.

The residual deviances for 4 log-linear models with Poisson errors are given below along with some relevant quantiles:

Model	Residual Deviance	Degrees of Freedom
1) DT*DK	14.32	6
2) DT*DK+L	8.60	5
3) DT*DK+L*DK	2.88	4
4) DT*DK+L*DT	8.07	3

$\chi^2_{5,0.95} = 11.07$, $\chi^2_{4,0.95} = 9.49$, $\chi^2_{3,0.95} = 7.81$, $\chi^2_{2,0.95} = 5.99$, $\chi^2_{1,0.95} = 3.84$

Table 1

	Diet					
	Vegan		Vegetarian		Neither	
	drink		drink		drink	
Low	0	1	0	1	0	1
No	34	29	15	12	43	12
Yes	5	21	9	8	23	11

(i) Write down an algebraic form for the linear predictor for Model 4 and state the corner-point constraints used in this model. ***(5 marks)***

(ii) Specify the non-zero parameters in Model 4 and hence explain why the degrees of freedom for Model 4 is equal to 3. ***(2 marks)***

(iii) Specify the nested structure of the 4 models. 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)***

(iv) For Model 3 calculate the expected number of low birth weight babies whose mothers were vegan and drank during pregnancy. ***(4 marks)***

(v) Explain how you would calculate a 95% confidence interval for the expected number of babies who didn't have a low birth weight and whose mothers were vegan and did not drink during pregnancy for Model 3. Assume you are given `R` output from the `summary` function in `R` for Model 3. ***(3 marks)***

End of Question Paper