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

MAS 473

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2010–2011**

Extended Linear Models **2 hours**

Restricted Open Book Examination.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) and a calculator which conforms to University regulations.
All answers will be marked, but credit will be given for only the best ***THREE*** *answers.*
All questions carry equal weight. Total marks 60.
Corner point constraints (treatment contrasts) are used in all `R` *output.*

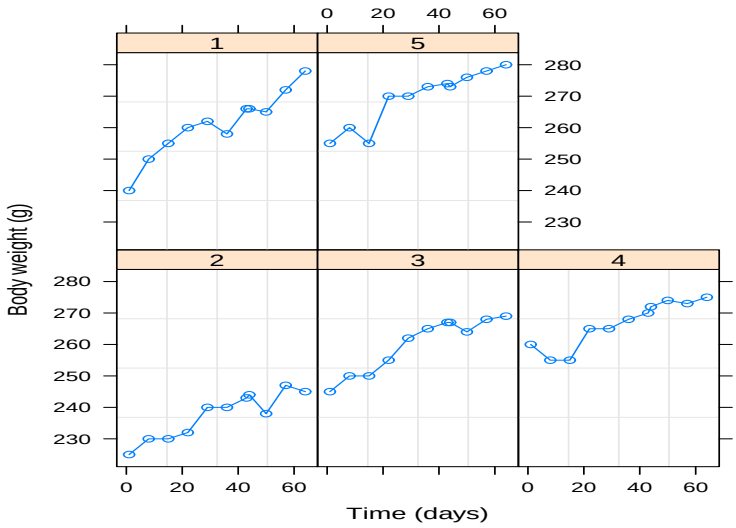
Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 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)***

(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)

1 (continued)

(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)

2 Suppose $Y_1, ..., Y_N$ are independent response variables in which $Y_i \sim binomial(n_i, \pi_i)$ for $1 \leq i \leq N$.

(i) If we observe the response values $y_1, ..., y_N$, show that the log likelihood $l(\beta; y_1, ..., y_N)$, where $\beta = (\pi_1, ..., \pi_N)^T$, is given by

$$l(\beta; y_1, ..., y_N) = \sum_{i=1}^{N} \left[y_i log\pi_i + (n_i - y_i) log(1 - \pi_i) + log \begin{pmatrix} n_i \\ y_i \end{pmatrix} \right]$$

(3 marks)

(ii) For the saturated model, what is the maximum likelihood estimate of π_i? ***(2 marks)***

(iii) For non-saturated models (with less than N parameters), if $\hat{y}_i$ is the fitted value, what is $\hat{\pi}_i$ in terms of n_i and $\hat{y}_i$? ***(1 mark)***

(iv) Hence show that the deviance for this model is

$$2 \sum_{i=1}^{N} \left[y_i log \left(\frac{y_i}{\widehat{y}_i} \right) + (n_i - y_i) log \left(\frac{n_i - y_i}{n_i - \widehat{y}_i} \right) \right]$$

. ***(3 marks)***

(v) Suppose also that we want to fit a generalized linear model such that $log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_i$, where x_i is a covariate of interest. Show that the log likelihood $l(\beta; y_1, ..., y_N)$ where $\beta = (\beta_0, \beta_1)^T$ is given by

$$l(\beta; y_1, ..., y_N) = \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta_1 x_i) - n_i log(1 + e^{\beta_0 + \beta_1 x_i}) + log \begin{pmatrix} n_i \\ y_i \end{pmatrix} \right]$$

(3 marks)

(vi) Derive the information matrix for $(\widehat{\beta}_0, \widehat{\beta}_1)^T$. ***(8 marks)***

3 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)***

4 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)

(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)

 Question 4 continued on next page

4 (continued)

(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