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 2011–2012**

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.*

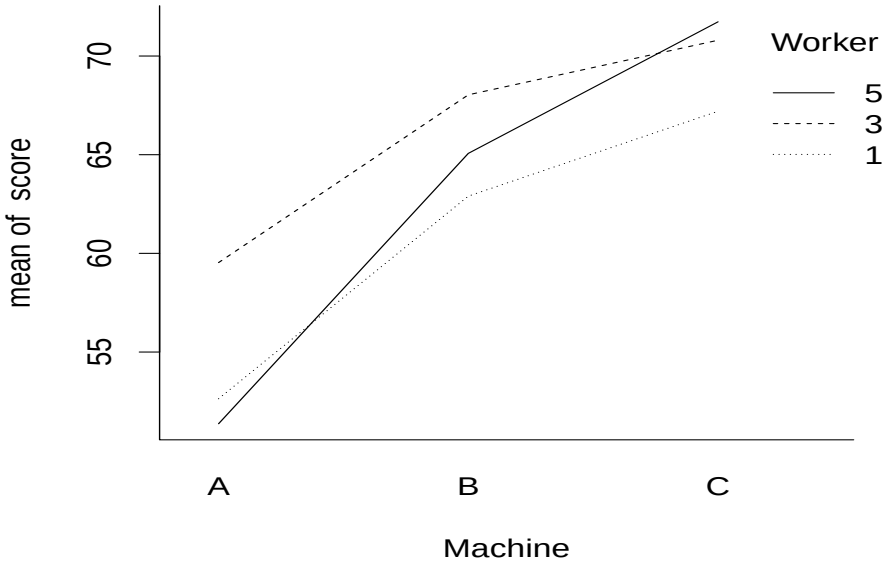
Please leave this exam paper on your desk
Do not remove it from the hall

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

Blank

1 Figure 1 shows mean productivity scores for each of three randomly chosen workers tested on each of three types of machine. Each worker used each machine three times giving three sets of replicates at each set of conditions.

Figure 1



Below is annotated output from an R session in which linear mixed effects models were investigated.

```
mach1.lme=lme(score~Machine, random = ~1 |Worker)
mach2.lme=lme(score~Machine, random = ~1 |Worker/Machine)

anova(mach1.lme,mach2.lme)
    Model    df      logLik     Test     L.Ratio   p-value
mach1.lme     5  -55.44817
mach2.lme     6  -44.83272   1 vs 2    21.23089    <.0001

mach3.lme=lme(score~Machine,random=pdCompSymm(~Worker))
```

(i) Write down the algebraic specification of the model `mach2.lme` stating clearly all assumptions and defining any terms that you use. ***(7 marks)***

(ii) Justify the choice of random and fixed effects in model `mach2.lme` ***(3 marks)***

(iii) Describe how you would check the normality assumptions within the `mach2.lme` model. State the value that the `levels` option would take in R for each set of residuals you'd check. ***(3 marks)***

(iv) Describe what is being tested by the `anova(mach1.lme,mach2.lme)` command in the R output above specifying the null hypothesis and state what the conclusion of the test is. ***(3 marks)***

1 (continued)

(v) For the `mach3.lme` model state the algebraic form for the covariance matrix of the worker random effects that is being specified. How does it differ to that fitted in the `mach2.lme` model? ***(4 marks)***

2 Moderately deaf people often find it difficult to use a telephone. Listening devices are designed to help moderately deaf people overcome this problem. A study is undertaken to assess the effect of a particular hearing device (D) and hearing score (S) on the probability of a moderately deaf person using a telephone (T). A total of 173 partially deaf people are surveyed for the study.

- Device (D) is a binary variable: 1=device A, 0=other device
- Hearing score (S) is a continuous variable (measured in some unit)
- The response variable is use of telephone (T): 1=use telephone, 0=do not use telephone.

Table 1 provides the residual deviances and degrees of freedom for 5 models fitted to the data using the logit link function.

Table 1

Model	Res. Deviance	df
1) T ~ 1	233.50	172
2) T ~ D	219.01	171
3) T ~ S	218.5	171
4) T ~ D + S	199.20	170
5) T ~ D*S	197.31	169

(i) What does the analysis of models 1-5 in Table 1 tell us about the dependence of the probability of using a telephone on hearing score and device type? ***(7 marks)***

(ii) More detailed analysis of model 4 is provided in Table 2 below. What does the numerical value of 0.026 represent in Table 2 in terms of the odds of using a telephone? ***(4 marks)***

(iii) Using the information in Table 2, calculate the odds ratio of a moderately deaf person using a telephone with a hearing score of 20 using device type A compared to a moderately deaf person with a hearing score of 20 not using device type A. ***(2 marks)***

Table 2

Coefficients:	Estimate	Std. Error
(Intercept)	-4.165	0.803
S	0.026	0.007
D	2.148	0.577

2 (continued)

(iv) Using the information in Table 2, calculate the log odds of using a telephone for someone using device type A with a hearing score of 15 for model 4. ***(3 marks)***

(v) The R output below gives the estimated covariance matrix of the estimated coefficients of model 4.

```
             (Intercept)   S             D
(Intercept)  0.644009133  -3.969181e-03 -0.3451017
S           -0.003969181   4.480828e-05  0.0005948
D           -0.345101718   5.948000e-04  0.3328265
```

Use the above R output and your answer to part (iv) to calculate a 95% confidence interval for the log odds of using a telephone for someone using device type A with a hearing score of 15 for model 4. ***(4 marks)***

3 A spatial interaction model for the number F_{ij} of people flowing between origins i and destinations j takes F_{ij} to be Poisson distributed with mean $\mu_{ij} = \alpha_i \beta_j e^{\gamma d_{ij}}$ where d_{ij} is the known distance between sites i and j and α_i, β_j and γ are unknown parameters. In an R data set collected for this flow system, F denotes flow, A and B denote factor variables for origins and destinations respectively, and D the origin-destination distances.

(i) By writing down appropriate R commands involving F, A,B and D show how such a model could be fitted to the observed data using a Poisson model with a log link. Explain what the commands do and define any further notation used. ***(4 marks)***

(ii) Interpret the generalized linear models whose model formulae in R in this context are

(a) R ~ A+B

(b) R ~ D

(3 marks)

(iii) Explain how you would test the hypothesis that flows depend only on distance between origin and destination but not otherwise on characteristics of origin and destination. ***(4 marks)***

(iv) If the observed flow between origin i and destination j is r_{ij}, show that the log-likelihood for this Poisson flow model is $l = -\sum_{ij} \alpha_i \beta_j e^{\gamma d_{ij}} + \sum_{ij} r_{ij} log\{\alpha_i \beta_j e^{\gamma d_{ij}}\} + k$ where k is a constant you do not need to specify. ***(3 marks)***

3 (continued)

(v) Obtain the equations satisfied by the maximum likelihood estimators $\widehat{\alpha_i}, \widehat{\beta_j}$ and $\widehat{\gamma}$, and show that the fitted values $\widehat{r_{ij}} = \widehat{\alpha_i}\widehat{\beta_j}e^{\widehat{\gamma}d_{ij}}$ based on the model satisfy

$$\sum_i r_{ij} = \sum_i \widehat{r_{ij}}, \quad \sum_j r_{ij} = \sum_j \widehat{r_{ij}} \text{ and } \sum_{ij} r_{ij}d_{ij} = \sum_{ij} \widehat{r_{ij}}d_{ij}.$$

(6 marks)

4 A study was carried out to assess the effects of mothers' drinking history and diet on the birth weight of their babies. A birth weight of less than 2.5 kg was considered low for the purposes of this study. Table 3 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`. For notational convenience we use

Table 3

	Diet					
	Vegan		Vegetarian		Neither	
	drink		drink		drink	
Low	0	1	0	1	0	1
0	34	29	15	12	43	12
1	5	21	9	8	23	11

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.

4 log-linear models with Poisson errors were fitted with the results below:

Model	Res. Deviance	degrees of freedom
1) DT*DK	34.00	6
2) DT*DK+L	12.84	5
3) DT*DK+L*DK	10.85	4
4) DT*DK+L*DT	7.96	3

(i) Write down an algebraic form for the linear predictor for model 3.
(5 marks)

(ii) With reference to your answer to part (i), justify why the degrees of freedom for model 3 are 4.
(2 marks)

(iii) 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)

4 (continued)

(iv) Below is some further R output for model 3.

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

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.70835    0.14865  24.947  < 2e-16 ***
dietveg         -1.01865    0.27942  -3.646 0.000267 ***
dietvegan       -0.39029    0.22886  -1.705 0.088121 .
drink           -1.05416    0.24214  -4.354 1.34e-05 ***
low             -0.48097    0.21816  -2.205 0.027476 *
dietveg:drink    0.87184    0.38768   2.249 0.024522 *
dietvegan:drink  1.30262    0.32291   4.034 5.48e-05 ***
dietveg:low      0.01835    0.37875   0.048 0.961361
dietvegan:low   -0.40407    0.31926  -1.266 0.205648
```

Calculate the expected number of low birth weight babies whose mothers were vegan and did not drink during pregnancy based on the output above. *(3 marks)*

(v) Define `n` to be the vector `(39,39,50,50,24,24,20,20,66,66,23,23)` and `proportion` to be the vector of proportions given by

$(34/39, 5/39, 29/50, 21/50, 15/24, 9/24, 12/20, 8/20, 43/66, 23/66, 12/23, 11/23)$.

The following command is fitted in R
`glm(proportion~low*diet,weights=n,family=binomial)`.
Explain what the `n` and `proportion` vectors represent and what the R command is fitting. *(4 marks)*

End of Question Paper