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

PAS 472

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2009–2010**

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.*

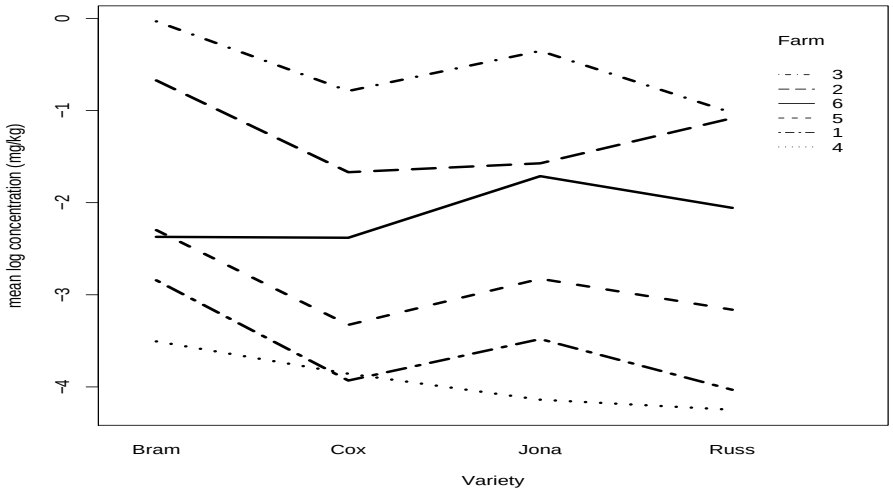
Please leave this exam paper on your desk
Do not remove it from the hall

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

Blank

1 It is common practice among commercial apple-growers to spray their crop with a pesticide during the growing season to protect against insect damage. To monitor pesticide residues after harvest, ten apples of each of four varieties were sampled from the produce of each of six randomly chosen growers who had used the pesticide *Chloropyrifos*. The apple varieties were *Bramley, Cox's Orange Pippin, Jonas Gold* and *Russet*. The concentration of *Chloropyrifos* was measured on each apple. Interest was in the extent to which different varieties retain different amounts of the pesticide, and in the variability of results to be expected in apples coming to market. Figure 1 shows mean log-concentrations for each grower and each apple variety.

Figure 1: Mean log-concentration of Chloropyrifos



Extracts from the results of fitting two linear mixed effects models to the log-concentrations are:

```
pest1.lme <- lme(logconc~Variety, random=~1|Farm)
pest2.lme <- lme(logconc~Variety, random=~1|Farm/Variety)

anova(pest1.lme, pest2.lme)
  Model          df      logLik     Test   L.Ratio    p-value
  pest1.lme 1     6     -324.64
  pest2.lme 2     7     -323.30   1 vs 2      2.68      0.102
```

1 (continued)

```
summary(pest1.lme)
 . . .
 Random effects:
  Formula: ~1 | Farm
      (Intercept)  Residual
  StdDev:    1.32      0.88

 Fixed effects: logconc ~ Variety
                Value   Std.Error    DF    t-value   p-value
 (Intercept)   -1.95        0.55   231      -3.55    0.0005
   VarietyC    -0.70        0.16   231      -4.38    0.0000
   VarietyJ    -0.39        0.16   231      -2.44    0.0153
   VarietyR    -0.65        0.16   231      -4.02    0.0001
```

(i) For each of the models `pest1.lme` and `pest2.lme`, write down the algebraic specification of the model, defining any terms that you use. Explain why the designation of explanatory variables as fixed and random in the models is reasonable, given the aim of the study. ***(6 marks)***

(ii) Assuming that checks on model fit prove satisfactory, discuss whether there is evidence of differences between farms in the patterns of pesticide retention experienced across different apple varieties. Justify your answer and comment on the evidence in relation to Figure 1. ***(5 marks)***

(iii) On the basis of the results for model `pest1.lme`, what can be said about differences between the effects of varieties on residue log-concentrations, and about the sources of variability in residues? ***(5 marks)***

(iv) Describe how you would find a confidence interval for the expected residue concentration on a Cox apple. If you need further R output, say how you would obtain it. ***(4 marks)***

2 A discrete distribution has probability function

$$f(y;\alpha) = \frac{(y+k-1)!}{y!(k-1)!}\frac{\alpha^y}{(1+\alpha)^{y+k}} \quad \alpha > 0, \quad y = 0, 1, ...$$

where k is a known positive constant.

(i) By writing $lnf(y;\alpha)$ in generalized linear form

$$lnf(y:\theta,\phi) = \left\{\frac{w}{\phi}(y\theta - b(\theta)) + c(y,\phi)\right\}$$

specify the value of the scale parameter ϕ and weights w, state the value of θ in terms of α and show that $b(\theta) = -kln\left(1-e^{\theta}\right)$. ***(6 marks)***

2 (continued)

(ii) State the value of μ (the expected value of Y) in terms of $b(\theta)$ and hence derive an expression for μ in terms of α and k. ***(3 marks)***

(iii) Find an expression for the variance function $V(\mu)$ in terms of μ and k. ***(7 marks)***

(iv) Derive the canonical link function $g(\mu)$ in terms of μ and k. ***(4 marks)***

3 Data are collected on 80 individuals in a survey looking at the effect of age and education level on salary. Information is gathered on the following three variables:

- X_1 - age
- X_2 - education level ($X_2 = 1$ for graduates and $X_2 = 0$ for non-graduates)
- Y - tax level ($Y = 1$ for higher-rate tax payers and $Y = 0$ for non higher-rate tax payers)

A series of GLMs are fitted to the data in which the binary variable Y is the response and X_1 and X_2 are the explanatory variables. The canonical logit link is used. If η is the linear predictor, the five models are:

- Model 1: $\eta = \beta_0$
- Model 2: $\eta = \beta_0 + \beta_1 X_2$
- Model 3: $\eta = \beta_0 + \beta_2 X_1$
- Model 4: $\eta = \beta_0 + \beta_1 X_2 + \beta_2 X_1$
- Model 5: $\eta = \beta_0 + \beta_1 X_2 + \beta_2 X_1 + \beta_3 X_1 X_2$

(i) Sketch a diagram showing the nested structure of the 5 tested models. ***(1 mark)***

Model	Residual deviance
Model 1	87.7
Model 2	61.3
Model 3	78.6
Model 4	48.4
Model 5	48.3

Table 1: Residual deviances for Models 1 to 5.

3 (continued)

(ii) The residual deviances for each of the five models is given in Table 1. By considering the changes in residual deviance between nested models, which model would you choose to represent the relationship between tax level, age and education status? For any hypothesis tests that you do, state clearly the null hypothesis and the degrees of freedom used for the test distribution. ***(6 marks)***

(iii) Comment on the fit of your selected model based on an appropriate χ^2 distribution. ***(2 marks)***

(iv) Using `R`, the following output is obtained for Model 4:

```
> model4.glm<-glm(tax~edu+age,family=binomial,data=tax.data)
> model4.glm

Call:  glm(formula = tax ~ edu + age, family = binomial)

Coefficients:
(Intercept)          edu          age
   -25.0249      19.6910       0.1215

Degrees of Freedom: 79 Total (i.e. Null);  77 Residual
Null Deviance:      87.71
Residual Deviance: 48.36        AIC: 54.36
```

Using the output from Model 4 above, calculate the odds of being a higher-rate tax payer for a 55 year old graduate. ***(3 marks)***

(v) For the logit link, write down the relationship between the mean value for an individual observation (μ_i) and the linear predictor (η_i). Hence calculate the probability that a 55 year old graduate is a higher-rate taxpayer. ***(2 marks)***

(vi) Ten graduates aged 55 were surveyed, 8 of these were higher-rate tax payers. Calculate the Pearson residual for 55 year old graduates. ***(3 marks)***

(vii) An over-dispersed binomial is fitted to Model 4 with scale parameter ϕ. Using the residual deviance for Model 4, give an estimate of ϕ. Is fitting an over-dispersed binomial justified? ***(3 marks)***

	Gender			
	Male		Female	
	Tumour type		Tumour type	
Response	Nodular	Diffuse	Nodular	Diffuse
Yes	2	22	4	6
No	8	2	11	2

Table 2: Patient response to chemotherapy.

4 57 patients registered as having a particular form of cancer were tested for response to a particular treatment. The results are shown in Table 2.
Various log-linear models with Poisson errors were fitted with the following results:

Model fitted	Res. Deviance	df
G*T	28.85	
G*T + R	26.71	
G*T + R*G	22.52	2
G*T + R*T	1.496	2
G*T + R*G + R*T	1.295	

R is the response to treatment, G is gender and T the tumour type. For this question, assume that R is a response factor and G and T are controlled factors.

(i) What is meant by homogeneity, response factor, controlled factor and the minimal model in log-linear models? ***(4 marks)***

(ii) Fill in the 3 missing degrees of freedom, justifying your values. ***(3 marks)***

(iii) Let π_{1jk} represent the probability a patient responds to treatment given their gender ($j = 0$ for male, $j = 1$ for female) and tumour type ($k = 0$ for nodular and $k = 1$ for diffuse). For each of the three models G∗T, G∗T+R and G∗T+R∗G, state the value of π_{1jk} for the four combinations of values of j and k. ***(6 marks)***

(iv) Assuming appropriate model-checking diagnostics proved satisfactory, what would you conclude about the dependence of response to treatment on gender and type of tumour (refer to the residual deviances above)? ***(4 marks)***

(v) For the model G∗T + R∗T, calculate the expected number of Males with diffuse tumour types who responded to treatment. ***(3 marks)***

End of Question Paper