The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS Spring Semester 2018–2019

MAS474 Extended linear models **2 hours**

Restricted Open Book Examination.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator which conforms to University regulations.
Answer all questions. Total marks 60.

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 survey is carried out to determine the vitamin C concentration in a specific variety of apple. Ten apple trees are selected at random from an orchard, and then three apples are selected from each tree. Two sample pieces of each apple are then selected at random, and the vitamin C concentration in each piece is then measured in a laboratory. The structure of the dataset is as follows:

```
>
> str(appledata)
'data.frame': 60 obs. of  3 variables:
 $ vitC : num  1.203 1 0.35 0.371 -0.296 ...
 $ tree : Factor w/ 10 levels "1","2","3","4",..: 1 1 1 1 1 1 2 2 2 2 ...
 $ apple: Factor w/ 3 levels "1","2","3": 1 1 2 2 3 3 1 1 2 2 ...
>
> head(appledata, 10)
       vitC tree apple
1  27.36133    1     1
2  27.23714    1     1
3  26.47704    1     2
4  26.49594    1     2
5  26.52272    1     3
6  26.57639    1     3
7  26.08871    2     1
8  26.00395    2     1
9  27.63607    2     2
10 27.78275    2     2
>
```

The following R command was used to fit a mixed effect model (model 1) to the data.

```
>
> library(lme4)
> model1 <- lmer(vitC ~ 1 + (1| tree/apple), data=appledata)
>
```

(i) (a) Let V_{ijk} be the vitamin C concentration measured in piece k for apple j for tree i. Write down the algebraic specification of the model that has been fitted to the data making sure you give the distribution of any random effects.

(4 marks)

1 (continued)

(b) The summary of the fitted model (model 1) is given below. Use the R output to give the estimated values for each parameter in your answer to part (a).

```
>
> summary(model1)
Linear mixed model fit by REML ['lmerMod']
Formula: vitC ~ 1 + (1 | tree/apple)
   Data: appledata

REML criterion at convergence: 103

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.80846 -0.53225 -0.03586  0.48503  1.87329

Random effects:
 Groups     Name        Variance  Std.Dev.
 apple:tree (Intercept) 1.098e+00  1.04774
 tree       (Intercept) 1.417e+02 11.90539
 Residual               7.933e-03  0.08907
Number of obs: 60, groups:  apple:tree, 30; tree, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)    18.80   REMOVED   4.987
>
```

(2 marks)

(ii) A simpler model (model 2) is fitted using the command

```
>
> model2 <- lmer(vitC~ 1 + (1| tree), data=appledata, REML=F)
>
```

Using the R output below, conduct a hypothesis test to choose between model 1 and model 2. Justify your approach.

```
> model1 <- lmer(vitC~ 1 + (1| tree/apple), data=appledata, REML=F)
> logLik(model1)
'log Lik.' -53.70089 (df=4)
> logLik(model2)
'log Lik.' -115.3253 (df=3)
```

(4 marks)

1 (continued)

(iii) Interest lies in the average vitamin C content for the population of apples from this particular type of tree. The standard error of this estimate has been removed from the R output above. Compute the standard error (for model 1) using the variance estimates provided.

Hint: The estimate of the fixed effect is given by the sample mean of the data.

(7 marks)

(iv) Briefly describe what diagnostic checks you could use to validate this mixed effect model.

(3 marks)

2 (i) The University are interested in whether the number of emails received by different lecturers varies between the different faculties. They collect data from n lecturers on the number of emails received each day. Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$ be the recorded number of emails received by each lecturer.

A simple statistical model would be that the number of emails received has a Poisson distribution. However, there is reason to believe that lecturers in Arts-based subjects receive emails at a different rate to lecturers in Science-based subjects, and so we will assume that

$$x_i \sim \begin{cases} Poisson(\lambda) & \text{for Arts subject lecturers} \\ Poisson(\mu) & \text{for Science lecturers.} \end{cases}$$

Unfortunately, the information about which faculty each lecturer worked in was lost. Let $0 \leq w \leq 1$ be the unknown proportion of Science-based lecturers in the data set.

We may thus assume that the x_i are samples from the following mixture distribution:

$$x_i \sim \begin{cases} Poisson(\lambda) & \text{with probability } 1-w, \\ Poisson(\mu) & \text{with probability } w. \end{cases}$$

The corresponding 'missing' variables $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ are defined as follows:

$$Y_i = \begin{cases} 0 & \text{if lecturer i is in the Arts faculty,} \\ 1 & \text{if lecturer i is in the Science faculty.} \end{cases}$$

Define $\theta = (w, \mu, \lambda)$.

(a) Show that the log-likelihood of θ given the complete data $(\mathbf{x}, \mathbf{Y})$ is

$$l(\theta; \mathbf{x}, \mathbf{Y}) = -\mu \sum Y_i - \lambda \sum (1 - Y_i) + \log \lambda \sum (1 - Y_i) x_i + \log \mu \sum Y_i x_i \\ - \sum \log x_i! + \log w \sum Y_i + \log(1-w) \sum (1 - Y_i)$$

where all sums are over $i = 1, \ldots, n$.

Hint: the probability mass function of a Poisson(λ) random variable is

$$\mathbb{P}(X = x) = \frac{\mathrm{e}^{-\lambda} \lambda^x}{x!} \quad \text{for } x = 0, 1, 2, \ldots$$

(5 marks)

(b) Using Bayes' theorem, show that

$$\mathbb{E}(Y_i | x_i, \theta) = \frac{w \mu^{x_i} \mathrm{e}^{-\mu}}{w \mu^{x_i} \mathrm{e}^{-\mu} + (1-w) \lambda^{x_i} \mathrm{e}^{-\lambda}}.$$

Denote this quantity by p_i.

(4 marks)

2 (continued)

(c) The EM algorithm is to be used to obtain the maximum likelihood estimator $\hat{\theta} = (\hat{w}, \hat{\mu}, \hat{\lambda})$ of θ, given the data $\mathbf{x}$. Let the estimate of $\hat{\theta}$ after m iterations of the EM algorithm be denoted $\theta^{(m)}$. By maximising

$$Q(\theta|\theta^{(m)}) = \mathbb{E}[l(\theta; \mathbf{x}, \mathbf{Y})|\mathbf{x}, \theta^{(m)}],$$

show that the updated estimates of $\hat{\phi}, \hat{\mu}, \hat{\lambda}$ are

$$\begin{aligned} w^{(m+1)} &= \frac{\sum_{i=1}^{n} p_i}{n}, \\ \lambda^{(m+1)} &= \frac{\sum_{i=1}^{n} x_i(1-p_i)}{\sum_{i=1}^{n}(1-p_i)}, \\ \mu^{(m+1)} &= \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i}, \end{aligned}$$

where $p_i = \mathbb{E}(Y_i|x_i, \theta)$ is your expression derived in part (b). ***(5 marks)***

(d) Give an intuitive explanation of the three updates in part (c).
(3 marks)

(ii) A survey of voting intentions in conducted. Each person is asked for the political party they intend to vote for in the next general election, the city they live in, and their yearly income. Some values are missing for each covariate, but for different reasons in each case.

In each of the following cases, say whether the data are missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR):

(a) Some respondents refuse to say who they will vote for. The person conducting the interview suspects that people who intend to vote for far right parties are less likely to provide this information than the general population.

(b) Due to a corrupted file, the home city of every fourth person in the survey is lost.

(c) Some respondents refuse to reveal their income. It is known that older people are less likely than the general population to reveal their income, regardless of what their income is.

(3 marks)

3 (i) Consider a dataset on the record winning times of 35 hill races in Scotland. The covariates are

- `dist` - distance in miles
- `climb` - total height gained during the route, in feet.
- `time` - record time in minutes

The output below shows the structure of the dataset.

```
> str(hill)
'data.frame': 35 obs. of  3 variables:
 $ dist : num  NA 6 6 7.5 8 8 16 NA 5 NA ...
 $ climb: int  650 2500 NA 800 3070 NA 7500 800 800 650 ...
 $ time : num  16.1 NA 33.6 45.6 62.3 ...
>
> head(hill)
             dist climb   time
Greenmantle    NA   650 16.083
Carnethy      6.0  2500     NA
Craig Dunain  6.0    NA 33.650
Ben Rha       7.5   800 45.600
Ben Lomond    8.0  3070 62.267
Goatfell      8.0    NA 73.217
```

(a) The following R command is used.

```
> hill.mice <- mice(hill, m=5, method = c('norm', 'norm', 'mean'))
```

Describe in detail the statistical procedure that is used to fill in the missing values.

(6 marks)

3 (continued)

(b) Interest lies in the coefficient of `dist` (β_1) in the linear regression model

$$\texttt{time} = \beta_0 + \beta_1 \texttt{dist} + \beta_2 \texttt{climb} + \epsilon.$$

Use the R output below to calculate an expected value of β_1 and its standard error.

```
> fit.mice <- with(hill.mice, lm(time ~ dist+climb))
>
> (coefs = sapply(fit.mice$analyses, coef))
              [,1] [,2]  [,3]  [,4]  [,5]
(Intercept) -3.12 1.27 -1.14 -3.47 -2.29
dist         6.46 6.15  6.50  6.16  6.41
climb        0.01 0.01  0.01  0.01  0.01
>
> apply(coefs, 1, var)
(Intercept)        dist       climb
    3.7e+00     2.9e-02     6.9e-08
>
> sapply(fit.mice$analyses, function(x) vcov(x)[2,2])
[1] 0.49 0.56 0.47 0.52 0.37
>
```

(5 marks)

(ii) A longitudinal study is conducted on the weight gain of rats in the first two weeks of life. Let y_{ij} be the weight of the i^{th} rat at the j^{th} measurement where $i = 1, \ldots, 5$, and $j = 1, \ldots, 3$. Let x_j be the age of the rat in days at the j^{th} measurement with $x_1 = 1, x_2 = 7, x_3 = 14$. We will use the model

$$y_{ij} = \alpha + a_i + (\beta + b_i)x_j + \epsilon_{ij}$$

where $a_i \sim N(0, \sigma_a^2)$, $b_i \sim N(0, \sigma_b^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$, with all random effects independent of each other.

3 (continued)

(a) Write the model in matrix notation

$$\mathbf{Y} = X\boldsymbol{\beta} + Z_a\mathbf{a} + Z_b\mathbf{b} + \boldsymbol{\epsilon}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ \vdots \\ Y_{53} \end{pmatrix}.$$

Give the matrices X, Z_a and Z_b, and the vectors $\mathbf{a}$, $\mathbf{b}$ and $\boldsymbol{\epsilon}$.

(7 marks)

(b) The data are in the following format:

```
> head(data)
    Weight  Days RatID
1     9.56    1    1
2    19.70    7    1
3    35.80    14   1
4    12.45    1    2
5    25.85    7    2
>
```

What R command would you use to fit the model to the data?

(2 marks)

End of Question Paper