The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS Spring Semester 2017–2018

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 Nine guinea pigs are each weighed every 5 days during the first 50 days after their birth. During this time they have unrestricted access to food. The R data frame `Guinea` contains the following columns:
- `ID`: an identification number unique to each guinea pig in the study
- `days`: the age of the guinea pig in days
- `bwt`: the body weight of the guinea pig in grams.

The aim of the study is to understand weight changes in the guinea pig population.

The following R commands were used to fit two different models

```
> model1 <- lmer(bwt~days+(1|ID), Guinea, REML=F)
> model2 <- lmer(bwt~days+(days-1|ID)+(1|ID), Guinea, REML=F)
```

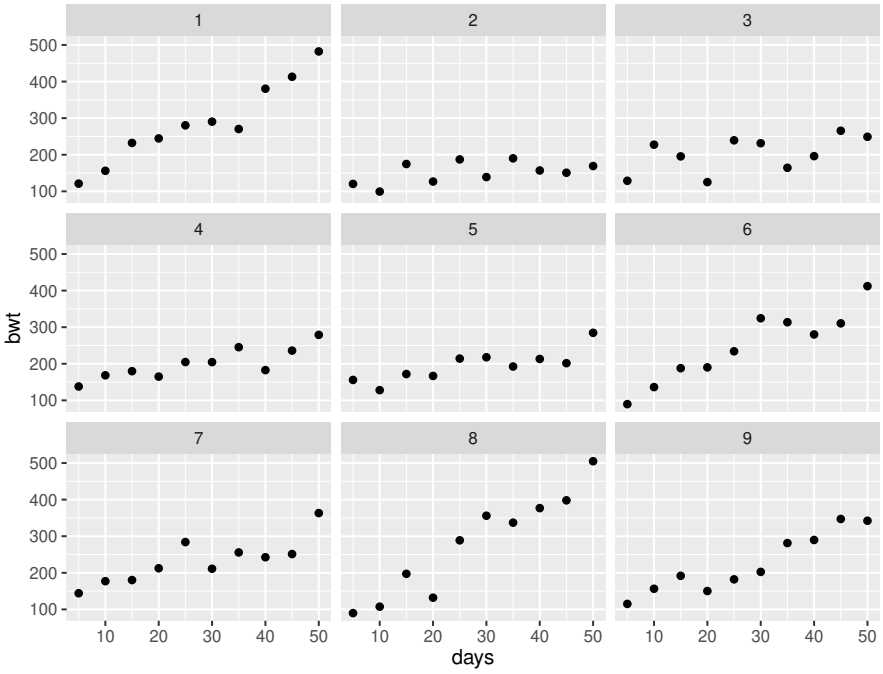

Figure 1: Change in body weight with time by guinea pig

(i) For each of the models `model1` and `model2`, write down the algebraic specification of the model, defining any terms that you use.

(4 marks)

(ii) Discuss which of these two models seems the most sensible given the information in Figure 1. ***(2 marks)***

1 (continued)

(iii) Using the R output below, conduct a hypothesis test to formally determine which of the two models is a better fit to the data. Justify your approach.

```
> logLik(model1)
'log Lik.' -487.5469
> logLik(model2)
'log Lik.' -473.1
```

(5 marks)

(iv) Describe a procedure for estimating a 95% confidence interval for the fixed effect parameters using bootstrap resampling. ***(4 marks)***

(v) Given the R output below, what is the best prediction for the weight of a 10-day old guinea pig chosen at random from the population? Derive a 95% prediction interval for this weight.

You may ignore the uncertainty in the parameter estimates when deriving this interval.

```
> model2
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: bwt ~ days + (days - 1 | ID) + (1 | ID)
   Data: Guinea
REML criterion at convergence: 918.7168
Random effects:
 Groups   Name        Std.Dev.
 ID       days         2.097
 ID.1     (Intercept) 20.839
 Residual             33.986
Number of obs: 90, groups:  ID, 9
Fixed Effects:
(Intercept)         days
    107.172        4.268
```

(5 marks)

2 (i) Suppose that $Y = (y_{ij})$ is a $n \times k$ matrix containing data, and that $M = (m_{ij})$ is the missing data indicator matrix, with

$$m_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is missing} \\ 0 & \text{otherwise.} \end{cases}$$

We write $Y = (Y_{obs}, Y_{mis})$ to denote the observed and missing components of Y respectively.

(a) Define in terms of the conditional distribution of M given Y what it means for data to be i) missing at random (MAR) and ii) not missing at random (NMAR).

(2 marks)

(b) Consider fitting a model of the form

$$f(Y, M|\theta, \psi) = f(Y|\theta)f(M|Y, \psi)$$

using maximum likelihood, where $f(Y|\theta)$ is the pdf of the data given an unknown parameter vector θ, and $f(M|Y, \psi)$ is the pdf of the missing-data mechanism given the data and unknown parameter ψ. Show that if interest lies solely in θ, and if the missing data mechanism is MAR, then the missingness can be ignored in the inference of θ and we can simply maximize

$$L_{ign}(\theta \mid Y_{obs}) = f(Y_{obs} \mid \theta).$$

(4 marks)

2 (continued)

(ii) A petrol company is interested in whether the two different petrol types it sells generate different fuel efficiencies in cars. To test this, two different cars of the same type have their fuel efficiency measured when using the two different types of petrol. The measurement is taken twice for each combination of car and petrol. Let y_{ijk} be the measured fuel efficiency of car j when given fuel type i in experiment k.

Consider the mixed effect model

$$y_{ijk} = \alpha + \beta_i + b_j + \epsilon_{ijk}$$

where $i = 1, 2$, $j = 1, 2$, $k = 1, 2$, with the sum to zero constraint $\beta_1 + \beta_2 = 0$, and

$$b_j \sim N(0, \sigma_1^2) \text{ and } \epsilon_{ij} \sim N(0, \sigma^2).$$

(a) Explain why this choice of random and fixed effects is natural for this problem. ***(2 marks)***

(b) If $y = (y_{111}, y_{112}, y_{121}, \ldots, y_{222})^\top$ and $\epsilon = (\epsilon_{111}, \epsilon_{112}, \epsilon_{121}, \ldots, \epsilon_{222})^\top$, write the model in matrix form. Be sure to specify both design matrices in full, as well as the vector of parameters and the vector of random effects.

(4 marks)

(c) Write down the least squares estimator for α and show that

$$\mathbb{V}\text{ar}(\hat{\alpha}) = \frac{\sigma_1^2}{2} + \frac{\sigma^2}{8}.$$

(4 marks)

(d) The least squares estimator of β_1 is

$$\hat{\beta}_1 = \frac{\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}}{2}.$$

Calculate the value of $\mathbb{C}\text{ov}(\hat{\alpha}, \hat{\beta}_1)$.

(4 marks)

3 Let $X_i \sim Exp(\lambda)$ be independent exponentially distributed random variables with mean $1/\lambda$. For $i = 1, \ldots, n$ we observe $X_i = x_i$ directly. For $i = n+1, \ldots, n+m$ we do not observe X_i, but only whether $X_i < h$ or not. We are told that from these m random variables exactly r have values less than h $\left(\text{i.e. } \sum_{i=1}^{m} \mathbb{I}_{X_{n+i}<h} = r\right)$.

We will now use the EM algorithm to estimate λ. We denote the complete dataset as $(x_1, \ldots, x_n, X_{n+1}, \ldots, X_{n+m})$.

(i) Show that the likelihood of the complete data is

$$(m+n)\log\lambda - \lambda\left(\sum_{i=1}^{n} x_i + \sum_{j=1}^{m} X_{j+n}\right).$$

(3 marks)

(ii) Show that

$$\mathbb{P}(X \leq x | X \leq h) = \frac{1 - e^{-\lambda x}}{1 - e^{-\lambda h}}.$$

(2 marks)

(iii) Prove that for a positive random variable Y with cumulative distribution function $F(y)$

$$\mathbb{E}(Y) = \int_0^\infty (1 - F(y))\mathrm{d}y.$$

Hint: you may find it useful to rewrite the right hand side as

$$\int_0^\infty \mathbb{P}(Y > y)\mathrm{d}y$$

and to swap the order of integration. ***(4 marks)***

(iv) Hence using parts (ii) and (iii) (or otherwise) derive an expression for

$$\mathbb{E}(X | X \leq h, \lambda).$$

(3 marks)

(v) Derive an expression for the quantity

$$Q(\lambda | \lambda^{(k)})$$

used in the E-step of the EM algorithm. ***(5 marks)***

(vi) Complete the M-step of the EM algorithm, i.e., derive a formula for the next estimate of λ, denoted by $\lambda^{(k+1)}$, in terms of $\lambda^{(k)}$. ***(3 marks)***

End of Question Paper