Imperial College
London

M4S2

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May-June 2016

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science

Statistical Modelling II

Date: **Friday 20th May 2016**

Time: **14.00 – 16.30**

Time Allowed: **2 Hours 30 Mins**

This paper has Five Questions.

Candidates should use ONE main answer book.

Supplementary books may only be used after the relevant main book(s) are full.

Statistical tables will not be provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Credit will be given for all questions attempted, but extra credit will be given for complete or nearly complete answers to each question as per the table below.

Raw Mark	Up to 12	13	14	15	16	17	18	19	20
Extra Credit	0	½	1	1 ½	2	2 ½	3	3 ½	4

- Each question carries equal weight.
- Calculators may not be used.

1. Consider the independent random variables $Y_1, \ldots, Y_n$ where each $Y_i \sim$ Binomial(n_i, π_i), $n_i \in \mathbb{Z}$, $0 \leq \pi_i \leq 1$. The probability mass function for Y_i is

$$P(Y_i = y_i; n_i, \pi_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i},$$

where $y_i = 0, 1, \ldots, n_i$. Consider a Binomial Generalized Linear Model (GLM) with canonical link function. Recall that the canonical link function for a Binomial GLM, is the logit function, defined in terms of π_i as:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right).$$

a) Write out an expression for the log-likelihood for the Binomial GLM in terms of $\pi_1, \ldots, \pi_n$.

b) For the two-parameter model with linear predictors

$$\eta_i = \beta_1 + \beta_2 x_i \quad (i = 1, \ldots, n),$$

where x_i is a real-valued covariate, show that the log-likelihood can be written in terms of the parameters, β_1 and β_2, as

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i(\beta_1 + \beta_2 x_i) - n_i \log(1 + \exp\{\beta_1 + \beta_2 x_i\}) + \log \binom{n_i}{y_i} \right].$$

c) Using the result in the part b), show that the score vector for β_1 and β_2 is

$$\boldsymbol{U} = \begin{pmatrix} \sum_{i=1}^{n} (y_i - n_i \pi_i) \\ \sum_{i=1}^{n} x_i (y_i - n_i \pi_i) \end{pmatrix}.$$

d) Using the score vector $\boldsymbol{U}$ in part c), show that Fisher's Information matrix for β_1 and β_2, $\mathcal{J} = \text{cov}(\boldsymbol{U})$, is

$$\mathcal{J} = \begin{pmatrix} \sum_{i=1}^{n} n_i \pi_i (1 - \pi_i) & \sum_{i=1}^{n} n_i x_i \pi_i (1 - \pi_i) \\ \sum_{i=1}^{n} n_i x_i \pi_i (1 - \pi_i) & \sum_{i=1}^{n} n_i x_i^2 \pi_i (1 - \pi_i) \end{pmatrix}.$$

2. a) Suppose $Y \sim \text{Poisson}(\lambda)$. The probability mass function of Y is

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad \text{for } y = 0, 1, \ldots \text{ and } \lambda > 0.$$

Show that the Poisson distribution with parameter λ is a member of the exponential family. Further, show that the canonical link function for the Poisson distribution is

$$g(u) = \log(u).$$

b) Suppose that a Generalized Linear Model is fitted in R using the following command

```
myglm <- glm(y ~ x1+x2+x3,family=poisson)
```

where y $\in \{0, 1, 2, \ldots\}^n$ and x1, x2 and x3 are n dimension vectors with real-valued entries. Carefully specify the model fitted by the command. Further, provide an expression for the log-likelihood of the fitted model in terms of the mean of the response.

c) Recall that the R command

```
predict(myglm,type="response")
```

reports the fitted mean response for the data provided. Explain why the command

```
sum(y) - sum(predict(myglm,type="response"))
```

returns the value 0, using and arguing based on the log-likelihood derived in part b).

Hint: Consider the maximum likelihood equations. However, it is unnecessary to perform the maximisation.

[CONTINUED]

d) Consider the dataset `elephant` presented in Table 1 below — only the first 5 data are presented.

	Age	Matings
1	27	0
2	28	1
3	28	1
4	28	1
5	28	3
⋮	⋮	⋮

Table 1: Elephant Dataset — first 5 data only.

The complete dataset consists 41 data points. The data reports the number of times each of the elephants successfully mates (`Matings`) and their age in years (`Age`). Mating is a rare event. Suppose we are interested in measuring the number of matings given the age of elephants. To do this, the following Poisson GLM was fitted in `R`.

```
myglm <- glm(Matings ~ Age, family=poisson, data=elephant)
```

The estimates from the fitted Poisson GLM are

```
> myglm$coeff
(Intercept)         Age
     -1.60        0.07
```

corresponding to $\widehat{\beta}$ and

```
> vcov(myglm)
             (Intercept)    Age
(Intercept)  3             -1
Age         -1              2
```

corresponding to the covariance matrix $\text{cov}(\widehat{\beta})$.

Using these `R` outputs, give a prediction for the expected number of Matings for a 27-year-old elephant. Provide a 95% asymptotic confidence interval for this prediction. Express your answer in a form which is as simple as possible, but you are not expected to perform any numerical calculations.

You may use the following approximate result: If $Z \sim N(0,1)$, then

$$P(Z \leq 1.96) = 0.975.$$

3. a) The deviance for a GLM with estimated means $\widehat{\mu}$ and dispersion parameter ϕ is defined as

$$D = 2\phi \left\{ \ell(y; y) - \ell(\widehat{\mu}; y) \right\}$$

where y is the observed response vector and $\ell(\cdot; y)$ is the log-likelihood of the model as a function of the response means μ. The log-likelihood evaluated at the observed data, $\ell(y; y)$, is related to the maximised log-likelihood for a saturated GLM. Explain, in no more than two sentences, the concept of a saturated GLM.

b) Consider the dataset called `bliss` presented in Table 2 below. These data record the number of insects that die (`dead`) and live (`alive`) using different concentration levels of an insecticide (`conc`). Each level of the experiment was performed using 30 insects.

	dead	alive	conc
1	2	28	0
2	8	22	1
3	15	15	2
4	23	7	3
5	27	3	4

Table 2: Bliss Dataset

Suppose we are interested in modelling the proportion of deaths given the concentration of insecticide. A Binomial GLM with canonical link function was fitted in `R` using the following command:

```
myGLM1 <- glm(cbind(dead,alive)~conc,family=binomial,data=bliss)
```

[CONTINUED]

The summary (shortened and redacted) of the fitted GLM is given below:

```
> summary(myGLM1)
Call:
glm(formula=cbind(dead,alive) ~ conc, family=binomial, data=bliss)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.32      0.4179  -5.561 2.69e-08 ***
conc          1.16      0.1814   6.405 1.51e-10 ||||||
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.75  on 4  degrees of freedom
Residual deviance:  0.57  on 3  degrees of freedom
AIC: ?????

Number of Fisher Scoring iterations: 4
```

Given that the log-likelihood under the null model is -40, calculate the log-likelihood of the fitted model. Hence, compute the AIC for this model which has been replaced with "?????".

c) In the R summary above there is "||||||" in the conc coefficient row. State the hypothesis test that can be conducted using the R output in this row and carefully explain the result of the test.

d) Another Binomial GLM is fitted to the same data using the following R command

```
myGLM2 <- glm(formula = cbind(dead, alive) ~ conc + exp(conc)
  + I(conc^2) + I(conc^5), family = binomial, data = bliss)
```

Which of the two fitted models, myGLM1 or myGLM2, should be preferred? Explain your reasoning.

4. Consider the balanced one-way random effects model

$$Y_{ij} = \mu + \nu_j + \epsilon_{ij} \quad \text{for } j = 1, \ldots, m; i = 1, \ldots, K,$$

where μ is the fixed effect, $\nu_j \sim N(0, \sigma_\nu^2)$ is the random effect and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The ν_j and ϵ_{ij} are all independent.

a) Rewrite the one-way random effects model in the general form given in lectures:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon},$$

specifying each component (including its dimension). Give an expression for the log-likelihood for this one-way random effects model.

b) Consider the pulp dataset presented in lectures. Recall that the data comes from an experiment to examine how paper brightness depends on a shift operator. The R output resulting from fitting the one-way random effects model to the pulp dataset is given below:

```
Linear mixed model fit by REML ['lmerMod']
Formula: bright ~ 1 + (1 | operator)
   Data: pulp

REML criterion at convergence: 18.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4666 -0.7595 -0.1244  0.6281  1.6012

Random effects:
 Groups   Name        Variance Std.Dev.
 operator (Intercept) 0.06     0.2609
 Residual             0.10     0.3260
Number of obs: 20, groups:  operator, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)  60.4000     0.1494   404.2
```

What is the estimate of the correlation between two observations in the same group under the fitted model?

[CONTINUED]

c) Explain how the REML procedure works. Further, explain how the estimates for the fixed effects are obtained using the REML procedure. Write down the equation for the REML estimator of the fixed effects. (It is not necessary to derive the estimator.) Your estimator for the fixed effects may depend on estimates for the variance components. You may assume that the design matrix, X, has full rank.

d) Explain why it does not make sense to estimate the random effects. Moreover, explain, in no more than two sentences, how the random effects, $\nu_1, \ldots, \nu_m$, are predicted when the fixed effects and variance components are both unknown.

e) Carefully explain how you would test whether or not to include the random effects, $\nu_1, \ldots, \nu_m$, in the one-way random effects model. Explain any problems with the test and briefly describe an alternative method.

5. **Mastery Question** This mastery question is based on the material referred to in Sections 2.9, 2.10, 3.1, 3.2, 3.3 and 3.4 in:

> Generalized Additive Models. Hastie, T. & Tibshirani, R. (1990). Chapman & Hall CRC Monographs on Statistics & Applied Probability.

Consider a model with one covariate and one smooth function:

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where Y_i is the response variable, x_i is a covariate, f is a smooth function and the ϵ_i are iid $N(0, \sigma^2)$ random variables. Suppose that $x_i \in [0, 1]$ for all i and $0 \leq x_1 \leq \cdots \leq x_n \leq 1$.

Upon choosing basis functions, b_j $(j = 1, \ldots, p)$, we assume that f takes the form

$$f(x) = \sum_{j=1}^{p} b_j(x)\beta_j.$$

This choice of representation of f means that (1) reduces to a linear model.

Given the observed responses, $y_1, \ldots, y_n$, the model can be fitted as follows: among all functions $f(x)$ with two continuous derivatives, find one that minimises the penalised residual sum of squares

$$\|y - X\beta\|^2 + \lambda \int_0^1 (f''(x))^2 \, dx, \tag{$\star$}$$

where $\lambda > 0$ is a fixed constant known as the smoothing parameter and $\|y - X\beta\|^2 \equiv \sum_{i=1}^{n}(y_i - f(x_i))^2$.

a) How is the natural cubic spline related to the penalised residual sum of squares $(\star)$?

b) Explain why it is preferable to fit the model by minimising $(\star)$ rather than by minimising $\|y - X\beta\|^2$.

c) The integral term in $(\star)$ is called a penalisation. Show that the penalisation can be written as

$$\int_0^1 (f''(x))^2 \, dx = \beta^T \Sigma \beta^T.$$

Be sure to specify the matrix Σ in terms of the basis functions b_j. You may assume that each b_j has at least two integrable derivatives.

d) Give an expression for the penalised least squares estimator of β i.e. the β that minimises $(\star)$.

[CONTINUED]

e) Explain why cross-validation is a reasonable approach to select the smoothing parameter λ.

f) Consider the cross-validation procedure in which each observation is omitted one at a time. A reasonable criterion is to select the value of λ that minimises the mean squared error:

$$M(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_\lambda(x_i) - f(x_i))^2,$$

where $\widehat{f}_\lambda(x_i)$ is the estimate of $f(x)$ obtained by minimising $(\star)$ using the smoothing parameter λ and evaluated at x_i. Since $f(x_i)$ is unknown, $M(\lambda)$ cannot be evaluated directly. However we can approximate $\mathrm{E}(M(\lambda)) + \sigma^2$ as follows. Let $\widehat{f}_\lambda^{(-i)}(x)$ denote the estimate of $f(x)$, fitted without the ith observation, and define the cross validation score:

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_\lambda^{(-i)}(x_i) - y_i)^2.$$

Show that

$$\mathrm{E}(\mathrm{CV}(\lambda)) \approx \mathrm{E}(M(\lambda)) + \sigma^2$$

with equality in the large-sample limit.

g) Explain one drawback of the cross validation procedure in which each observation is omitted one at a time.

Imperial College
London

IMPERIAL COLLEGE LONDON

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2016

This paper is also taken for the relevant examination for the Associateship.

M3S2/M4S2

Statistical Modelling II (Solutions)

Setter's signature	Checker's signature	Editor's signature
.....	.....	.....

1. a) Take the logarithm of the given probability mass function and add the log-likelihoods for the individual observations : seen ⇓ 2

$$\ell(\boldsymbol{\pi}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i} \right].$$

b) Using the canonical link we have sim. seen ⇓

$$\eta = \log\left(\frac{\pi}{1-\pi}\right).$$

Rearranging to make π the subject, we obtain 1

$$\pi = \frac{1}{1+e^{-\eta}}.$$

For the given linear predictor we have 1

$$\pi_i = \frac{1}{1+e^{-\eta_i}} = \frac{1}{1+\exp\{-(\beta_1+\beta_2 x_i)\}}.$$

Start by rewriting the log-likelihood from part a): 2

$$\begin{aligned} &\sum_{i=1}^{n} \left[y_i\eta_i + n_i \log\left(1 - \frac{1}{1+e^{-\eta_i}}\right) + \log\binom{n_i}{y_i} \right] \\ =&\sum_{i=1}^{n} \left[y_i\eta_i - n_i \log\left(1+e^{\eta_i}\right) + \log\binom{n_i}{y_i} \right]. \end{aligned}$$

Next, plug in the linear predictors, η_i, given in the question: 2

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i(\beta_1+\beta_2 x_i) - n_i \log\left(1+\exp\{\beta_1+\beta_2 x_i\}\right) + \log\binom{n_i}{y_i} \right]$$

which is the stated result.

c) Recall that the score vector is given by: seen ⇓

$$\boldsymbol{U} = \begin{pmatrix} \frac{\partial \ell}{\partial \beta_1} & \frac{\partial \ell}{\partial \beta_2} \end{pmatrix}^T.$$ 1

To derive $\boldsymbol{U}$, take the partial derivatives in turn: meth seen ⇓

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_1} &= \sum_{i=1}^{n} \left[y_i - n_i \left(\frac{e^{\beta_1+\beta_2 x_i}}{1+e^{\beta_1+\beta_2 x_i}} \right) \right] \\ &= \sum_{i=1}^{n} \left[y_i - n_i \left(\frac{1}{1+e^{-(\beta_1+\beta_2 x_i)}} \right) \right] \\ &= \sum_{i=1}^{n} \left[y_i - n_i \pi_i \right] \end{aligned}$$

since 2

$$\frac{1}{1+e^{-(\beta_1+\beta_2 x_i)}} = \frac{1}{1+e^{-\eta_i}} = \pi_i.$$

1

Similarly,

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_2} &= \sum_{i=1}^{n} \left[y_i x_i - n_i x_i \left(\frac{e^{\beta_1+\beta_2 x_i}}{1+e^{\beta_1+\beta_2 x_i}} \right) \right] \\ &= \sum_{i=1}^{n} x_i (y_i - n_i \pi_i), \end{aligned}$$

using the same substitution used for the first partial derivative. 2

d) Recall that

meth seen ⇓

1

$$\mathcal{J} = -\,\mathrm{E}\begin{pmatrix} \dfrac{\partial^2 \ell}{\partial \beta_1^2} & \dfrac{\partial^2 \ell}{\partial \beta_1 \partial \beta_2} \\ \dfrac{\partial^2 \ell}{\partial \beta_2 \partial \beta_1} & \dfrac{\partial^2 \ell}{\partial \beta_2^2} \end{pmatrix} = -\,\mathrm{E}\begin{pmatrix} \dfrac{\partial}{\partial \beta_1} U_1 & \dfrac{\partial \ell}{\partial \beta_1} U_2 \\ \dfrac{\partial \ell}{\partial \beta_2} U_1 & \dfrac{\partial \ell}{\partial \beta_2} U_2 \end{pmatrix}.$$

Since U_1 and U_2 have already been computed we only need

2

$$\frac{\partial \pi_i}{\partial \beta_1} = \frac{e^{-(\beta_1+\beta_2 x_i)}}{(1+e^{-(\beta_1+\beta_2 x_i)})^2} = \pi_i(1-\pi_i)$$
$$\frac{\partial \pi_i}{\partial \beta_2} = \frac{x_i e^{-(\beta_1+\beta_2 x_i)}}{(1+e^{-(\beta_1+\beta_2 x_i)})^2} = x_i \pi_i(1-\pi_i)$$

to obtain

1

$$\mathcal{J} = \mathrm{E}\begin{pmatrix} \displaystyle\sum_{i=1}^n n_i \frac{\partial}{\partial \beta_1}\pi_i & \displaystyle\sum_{i=1}^n x_i n_i \frac{\partial \ell}{\partial \beta_1}\pi_i \\ \displaystyle\sum_{i=1}^n n_i \frac{\partial \ell}{\partial \beta_2}\pi_i & \displaystyle\sum_{i=1}^n x_i n_i \frac{\partial \ell}{\partial \beta_2}\pi_i \end{pmatrix}$$
$$= \begin{pmatrix} \displaystyle\sum_{i=1}^n n_i \pi_i(1-\pi_i) & \displaystyle\sum_{i=1}^n n_i x_i \pi_i(1-\pi_i) \\ \displaystyle\sum_{i=1}^n n_i x_i \pi_i(1-\pi_i) & \displaystyle\sum_{i=1}^n n_i x_i^2 \pi_i(1-\pi_i) \end{pmatrix}$$

as required.

2

Note: the student may use an alternative method here and still gain full marks.

2. a) The probability mass function for the Poisson distribution can be written as: [seen ⇓]

$$\exp\{y\log(\lambda) - \lambda - \log(y!)\}.$$ [1]

Using the notation for an exponential family member given in the notes, we identify

$$\theta = \log(\lambda), a(\phi) = \phi = 1, b(\theta) = \lambda = \exp(\theta), c(y, \phi) = -\log(y!).$$

Therefore, the Poisson distribution is a member of the exponential family. The canonical link is found by setting $\theta = \eta$ (the linear predictor). The canonical link for the Poisson distribution is $g(u) = \log(u)$ as $\lambda = \mathrm{E}(Y) \equiv \mu$. [2]

b) The model being fitted by the R command is a Generalized Linear Model with [sim. seen ⇓]

* independent response variables $Y_1, \ldots, Y_n$ with each $Y_i \sim$Poisson(μ_i); and [1]
* linear predictors $\log(\mu_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$ for $i = 1, \ldots, n$ where $p = 4$ and [1]

$$X = (x_{ij}) = \begin{pmatrix} 1 & \uparrow & \uparrow & \uparrow \\ \vdots & \texttt{x1} & \texttt{x2} & \texttt{x3} \\ 1 & \downarrow & \downarrow & \downarrow \end{pmatrix}$$

and

* link function $g(u) = \log(u)$. [1]

The log-likelihood of the model is [2]

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} \{y_i \log(\mu_i) - \mu_i - \log(y_i!)\},$$

where $\mu_i = \exp\left\{\sum_{j=1}^{p} x_{ij}\beta_j\right\}$.

c) As we are using the canonical link function, we have

unseen ⇓

$$\log(\mu_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j.$$

[1]

Substituting this into the log-likelihood gives

[1]

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ y_i \left(\sum_{j=1}^{p} x_{ij}\beta_j \right) - \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j \right) - \log(y_i!) \right\}.$$

Because the first column of X corresponds to the intercept term in the model, $x_{i,1} = 1$ for all i, and $\widehat{\boldsymbol{\beta}}$ satisfies

[2]

$$\sum_{i=1}^{n} \left\{ y_i - \exp\left(\sum_{j=1}^{p} x_{ij}\widehat{\beta}_j \right) \right\} = 0$$

(consider $\frac{\partial \ell}{\partial \beta_1} \stackrel{!}{=} 0$), we have

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \exp\left(\sum_{j=1}^{p} x_{ij}\widehat{\beta}_j \right) = \sum_{i=1}^{n} \widehat{\mu}_i.$$

The `R` command computes $\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \widehat{\mu}_i$ which we have shown to be 0. [2]

d) Start by forming the linear predictor for a 27-year-old elephant:

sim. seen ⇓

$$\widehat{\eta}_\star = -1.60 + (0.07)(27) = 0.29.$$

[1]

The mean response is given by transforming $\widehat{\eta}_\star$ using the inverse link function [1]

$$\widehat{\mu}_\star = \exp(0.29).$$

In order to construct the required confidence interval, recall that asymptotically / approximately

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \Sigma),$$

where $\Sigma = \mathrm{cov}(\widehat{\boldsymbol{\beta}})$. Using a transformation and rearranging we have [2]

$$\frac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{c}^T\Sigma\boldsymbol{c}}} \sim N(0, 1),$$

for any $\boldsymbol{c} \in \mathbb{R}^2$. Selecting $\boldsymbol{c} = (1, 27)^T$ and plugging in the given values results in the interval on the linear predictor scale: [1]

$$\left(0.29 - 1.96\sqrt{1407}, 0.29 + 1.96\sqrt{1407}\right).$$

Finally, an asymptotic confidence interval for the prediction is [1]

$$\left(\exp(0.29 - 1.96\sqrt{1407}), \exp(0.29 + 1.96\sqrt{1407})\right).$$

The student may report a similar result, simplifying as far as possible.

The student may use the notation $\dot{\sim}$ to denote "asymptotically / approximating distributed as".

seen ⇓

3. a) A saturated GLM is a GLM (with the same response distribution and link function as the GLM of interest), with one parameter per (distinct) observation i.e. $p = n$ where p is the number of parameters and n is the number of (distinct) observations. [2]

unseen ⇓

b) In the R output, there are two deviances reported. The deviance of the null model, say D_{NULL}, is

$$D_{\text{NULL}} = 2\phi \left\{ \ell(\boldsymbol{y}; \boldsymbol{y}) - \ell_{\text{NULL}}(\widehat{\boldsymbol{\mu}}; \boldsymbol{y}) \right\}.$$

The deviance of the fitted model is

$$D = 2\phi \left\{ \ell(\boldsymbol{y}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\mu}}; \boldsymbol{y}) \right\}.$$

Taking the difference gives [3]

$$D_{\text{NULL}} - D = 2\phi \left\{ \ell(\widehat{\boldsymbol{\mu}}; \boldsymbol{y}) - \ell_{\text{NULL}}(\widehat{\boldsymbol{\mu}}; \boldsymbol{y}) \right\}.$$

Plugging in the reported values from the R output and the given log-likelihood under the null model yields: [3]

$$65.75 - 0.57 = 2 \left\{ \ell(\widehat{\boldsymbol{\mu}}; \boldsymbol{y}) - (-40) \right\}.$$

Rearranging gives

$$\ell(\widehat{\boldsymbol{\mu}}; \boldsymbol{y}) = \frac{65.18}{2} - 40 = -7.41.$$

The AIC is defined as AIC$= -2\ell(\widehat{\boldsymbol{\mu}}; \boldsymbol{y}) + 2p$ where p is the number of parameters. Thus, [2]

$$\text{AIC} = (-2) \cdot (-7.41) + (2) \cdot (2) = 18.82.$$

[1]

sim. seen ⇓

c) The null hypothesis in the conc row tested is $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ where β_2 is the corresponding coefficient for the conc term. R reports the p-value as 1.15×10^{-10}. Since this is much less than 0.1%, there is sufficient evidence to reject the null, at the 0.1% level — thus suggesting inclusion of the conc term in the model. [4]

sim. seen ⇓

d) The fitted model has as many parameters as observations, therefore it is a saturated GLM. (Thus the deviance of this model is 0.) [2]

A saturated GLM is not particularly useful since it merely predicts the death counts with their observed values, $y_1, \ldots, y_n$. [2]

Hence we prefer to use the model, myGLM1. [1]

sim. seen ⇓

4. a) Begin by casting the model in the general form

4

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\nu} + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{K,1} \\ Y_{1,2} \\ \vdots \\ Y_{K,2} \\ \vdots \end{pmatrix}, \quad X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

a matrix where column j is composed of $(j-1)K$ zeros,, followed by K ones, and then $(m-j)K$ zeros.

$$\boldsymbol{\beta} = \mu, \quad \boldsymbol{\nu} = \begin{pmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_m \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{K,1} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{K,2} \\ \vdots \end{pmatrix},$$

where $n = mK$, $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times 1}$, $Z \in \mathbb{R}^{n\times m}$, $\boldsymbol{\beta} \in \mathbb{R}$, $\boldsymbol{\epsilon}$ is an n-variate vector and $\boldsymbol{\nu}$ is an m-variate vector. The pdf of the marginal sampling distribution of $\boldsymbol{Y}$ is

$$\frac{1}{(2\pi)^{n/2}|\sigma_\epsilon^2 V_{\boldsymbol{\tau}}|^{1/2}} \exp\left\{\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T V_{\boldsymbol{\tau}}^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\right\},$$

where $V_{\boldsymbol{\tau}} := I_n + Z\Psi Z^T$, $\Psi = \frac{\sigma_\nu^2}{\sigma_\epsilon^2} I_m$ where $\boldsymbol{\tau} = (\sigma_\epsilon^2, \sigma_\nu^2)^T$ are the variance components. Thus, the log-likelihood function is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left|\sigma_\epsilon^2 V_{\boldsymbol{\tau}}\right| - \frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T V_{\boldsymbol{\tau}}^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}).$$

b) The correlation between two observations in the same group is

sim. seen ⇓

2

$$\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\epsilon^2}.$$

From the output we identify the estimates $\widehat{\sigma}_\nu^2 = 0.06$ and $\widehat{\sigma}_\epsilon^2 = 0.10$. Therefore the estimated correlation is

1

$$\frac{0.06}{0.06 + 0.10} = 6/16 = 3/8.$$

sim. seen ⇓

c) The REML procedure uses a transformation that eliminates the fixed effects, then estimates the variance components, say $\widehat{\boldsymbol{\tau}}$. The variance components are then treated as known and fixed at their REML estimators in the full log-likelihood. The estimator of the fixed effects, $\widehat{\boldsymbol{\beta}}$, satisfies the equation [3]

$$X^T V_{\widehat{\boldsymbol{\tau}}}^{-1} \boldsymbol{Y} - X^T V_{\widehat{\boldsymbol{\tau}}}^{-1} X \widehat{\boldsymbol{\beta}} = 0.$$

Assuming X has full rank, the solution is

$$\widehat{\boldsymbol{\beta}} = \left(X^T V_{\widehat{\boldsymbol{\tau}}}^{-1} X\right)^{-1} X^T V_{\widehat{\boldsymbol{\tau}}}^{-1} \boldsymbol{Y}.$$

The student may write down the form of $\widehat{\boldsymbol{\beta}}$ without derivation.

seen ⇓

d) The random effects are **random variables** so it does not make sense to estimate them. The (vector) of random effects are predicted by $\widehat{\boldsymbol{\nu}} := \mathrm{E}(\boldsymbol{\nu} | \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{\tau} = \widehat{\boldsymbol{\tau}})$, where $\widehat{\boldsymbol{\tau}}$ is an estimator of the variance components $\boldsymbol{\tau}$. [3]

seen ⇓

e) We can use the difference of the deviances or the likelihood ratio test to compare mixed effects models. To compare two nested models M_0 (smaller) and M_1, the difference of the deviances is [2]

$$2\left\{\ell(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\tau}}_1; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{\tau}}_0; \boldsymbol{y})\right\},$$

where $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\tau}}_0$ are the MLEs of $\boldsymbol{\beta}$, $\boldsymbol{\tau}$ for M_0 and $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\boldsymbol{\tau}}_1$ are the MLEs of $\boldsymbol{\beta}$, $\boldsymbol{\tau}$ for M_1.

To test whether or not to include random effects in mixed effects model, we consider the null hypothesis: [1]

$$H_0 : \ \sigma_\nu^2 = 0.$$

The standard derivation of the asymptotic χ^2 distribution of the difference of the deviances depends on the null hypothesis lying on the interior of the parameter space. This assumption does not hold for this test. [1]

If the χ^2 approximation is used, then the test will tend to be conservative, in the sense that the p-values will tend to be larger than they should be. [1]

meth seen ⇓

A parameteric bootstrap method can be used to repeatedly sample from the null model to obtain a bootstrap p-value. We can then compare the distribution of the differences in deviances for the datasets simulated under the null model with that computed on the real data to obtain a bootstrap p-value. [2]

5. **Mastery Question**

seen ⇓

a) Quoting

2

> Generalized Additive Models. Hastie, T. & Tibshirani, R. (1990). Chapman & Hall CRC Monographs on Statistics & Applied Probability.

" The penalised residual sum of squares ($\star$) has an explicit, unique minimizer and the minimizer is a natural cubic spline with knots at the unique values of x_i."

sim.seen (material) ⇓

b) Minimising with the penalisation term introduces a trade off between model fit and model smoothness, which is controlled through the smoothness parameter λ. Fitting the model without the penalisation term would result in an overfitted model; a function that simply interpolates between the given data.

2

sim.seen (material) ⇓

c) Start with the function f and differentiate twice with respect to x:

$$f''(x) = \sum_{j=1}^{p} b_j''(x)\beta_j.$$

Then

$$\begin{aligned}
\int_0^1 \left(f''(x)\right)^2 dx &= \int_0^1 \left(\sum_{j=1}^{p} b_j''(x)\beta_j\right)^2 dx \\
&= \int_0^1 \sum_{j=1}^{p} b_j''(x)\beta_j \sum_{k=1}^{p} b_k''(x)\beta_k dx \\
&= \sum_{j=1}^{p}\sum_{k=1}^{p} \beta_j\beta_k \int_0^1 b_j''(x)b_k''(x)dx \\
&= \sum_{j=1}^{p}\sum_{k=1}^{p} \beta_j\beta_k\Sigma_{jk} \quad (\dagger)
\end{aligned}$$

where $\Sigma_{jk} := \int_0^1 b_j''(x)b_k''(x)dx$. (The form $(\dagger)$ has been seen in lectures before.) Therefore,

2

$$\int_0^1 \left(f''(x)\right)^2 dx = \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta},$$

where $\Sigma_{jk} := \int_0^1 b_j''(x)b_k''(x)dx$.

2

seen (material) ⇓

d) By part c) the penalised least squares is obtained by minimising

1

$$\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}^T.$$

meth.seen (material) ⇓

Expanding this yields

1

$$\begin{aligned}\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}^T &= (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}^T \\ &= \boldsymbol{y}\boldsymbol{y}^T - 2\boldsymbol{\beta}^T X^T \boldsymbol{y} + \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} + \lambda \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}^T \\ &= \boldsymbol{y}\boldsymbol{y}^T - 2\boldsymbol{\beta}^T X^T \boldsymbol{y} + \boldsymbol{\beta}^T (X^T X + \lambda\Sigma)\boldsymbol{\beta}.\end{aligned}$$

Differentiating this wrt $\boldsymbol{\beta}$, setting it to zero and solving yields

$$(X^T X + \lambda\Sigma)\widehat{\boldsymbol{\beta}} = X^T \boldsymbol{y}.$$

Finally, the regularised estimator of $\boldsymbol{\beta}$ is

1

$$\widehat{\boldsymbol{\beta}} = (X^T X + \lambda\Sigma)^{-1} X^T \boldsymbol{Y}.$$

unseen ⇓

e) Cross validation essentially works with two datasets — the training data set, which is used to fit the model and the test dataset which is used to assess the fitted models predictions. Since separate datasets are used to fit the model and assess its goodness-of-fit, in the sense of producing accurate out-of-sample predictions, the model fitted using cross validation does not suffer from overfitting / fidelity to the data.

3

If the same data are used to fit the model and measure its goodness-of-fit, then complicated, overfitted models are selected over simpler ones. These models overfit in the sense that they minimise the in-sample prediction errors by incorporating random fluctuations into the model predictions. Such random fluctuations are not present in new data and thus out-of-sample predictions will be poor.

seen (material) ⇓

f) Start by rewriting CV(λ) as

$$\begin{aligned}\text{CV}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_\lambda^{(-i)}(x_i) - f(x_i) - \epsilon_i)^2 \quad \text{since } y_i = f(x_i) + \epsilon_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_\lambda^{(-i)}(x_i) - f(x_i))^2 - 2(\widehat{f}_\lambda^{(-i)}(x_i) - f(x_i))\epsilon_i + \epsilon_i^2.\end{aligned}$$

Taking expectation yields [2]

$$\begin{aligned}\text{E}(\text{CV}(\lambda)) &= \text{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_\lambda^{(-i)}(x_i) - f(x_i))^2\right] - 2\,\text{E}\left[(\widehat{f}_\lambda^{(-i)}(x_i) - f(x_i))\epsilon_i\right] + \text{E}(\epsilon_i^2) \\ &= \text{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_\lambda^{(-i)}(x_i) - f(x_i))^2\right] + \sigma^2\end{aligned}$$

since $\epsilon_i \sim N(0, \sigma^2)$ and ϵ_i and $\widehat{f}_\lambda^{(-i)}(x_i)$ are independent. [1]

Now, $\widehat{f}_\lambda^{(-i)}(x_i) \approx \widehat{f}_\lambda(x_i)$ with equality as $n \to \infty$. So $\text{E}(\text{CV}(\lambda)) \approx \text{E}(M(\lambda)) + \sigma^2$, again with equality as $n \to \infty$. [1]

unseen ⇓

g) This cross validation procedure involves omitting one observation at a time. This is computationally inefficient since it involves fitting the model n times. [2]