The University Of Sheffield.

**SCHOOL OF MATHEMATICS AND STATISTICS** **Spring Semester 2019–2020**

**MAS474 Extended linear models** **1 hour (nominal)**

*This is an open book exam.*
*Answer* ***all*** *questions.*
*The submission deadline is 10 am (BST), twenty-four hours after it is released. Late submission will not be considered without extenuating circumstances. It is expected that you will be able to complete this exam in approximately one hour and it is recommended that you submit the work within four hours. You will not be penalised for taking longer, however.*
*Unless it is explicitly stated otherwise, it is intended that calculations are performed by hand (possibly with the aid of a calculator). To gain full marks, you will need to show your working. You will not get full marks if you simply write down output from a computer package.*
*By uploading your solutions you declare that your submission consists entirely of your own work, that any use of sources or tools other than material provided for this module is cited and acknowledged and that no unfair means have been used.*

*Total marks 30.*

**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 longitudinal study conducted in the USA by social scientists followed the salary of 85 individuals from 1968 to 1990. Interest lies in the effect of education, sex, and age on salary levels. The data are stored in the data frame `salaries`. The columns are

- `age`: age of the subject in 1968
- `educ`: years of education of the subject
- `sex`: sex of subject - a factor with levels 'F' or 'M'
- `income`: inflation adjusted annual income in dollars
- `year`: calendar year
- `person`: ID number for the subject

The following R code was used to fit a mixed effect model.

```
>
> library(lme4)
> head(salary)
  age educ sex income year person
1  31   12   M   8400   68      1
2  31   12   M   7420   69      1
3  31   12   M   7280   70      1
4  31   12   M   9660   71      1
5  31   12   M  10500   72      1
6  31   12   M  11200   73      1
>
> salary$cyear <- salary$year-78
> fit1 <- lmer(log(income) ~ sex +age+educ+(cyear|person),salary)
>
```

(i) For $i = 1, \ldots, 85$ and $j = 1, \ldots, 23$, let $y_{ij}$ be the income of individual $i$ in observation $j$, $\text{age}_i$ be the age of individual $i$ in 1968, $\text{educ}_i$ the years of education for individual $i$, and $\text{year}_{ij}$ the year observation $j$ was made on individual $i$.

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)

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

```
>
> summary(fit1)
Linear mixed model fit by REML ['lmerMod']
Formula: log(income) ~ sex + age + educ + (cyear | person)
   Data: salary

REML criterion at convergence: 3888.6

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1955  -0.2204   0.0764   0.4028   2.9207

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 person   (Intercept) 0.30072  0.54838
          cyear       0.00764  0.08741  0.28
 Residual             0.46669  0.68315
Number of obs: 1661, groups:  person, 85

Fixed effects:
            Estimate Std. Error t value
(Intercept)  6.82311    0.54850  12.440
sexM         1.18538    0.12093   9.802
age          0.01163    0.01365   0.852
educ         0.10635    0.02165   4.911

Correlation of Fixed Effects:
     (Intr) sexM   age
sexM -0.104
age  -0.874 -0.026
educ -0.598  0.009  0.167
>
```

*(4 marks)*

(iii) Explain, given the aim of the study, why the designation of explanatory variables as fixed and random is reasonable.

*(2 marks)*

(iv) What R command would you use to fit a model in which the two random effect terms are uncorrelated? *(2 marks)*

**1** (continued)

(v) Below is some R code and output. Describe the procedure performed and give the conclusions of the analysis.

```
>
> fit.reduced <-lmer(log(income) ~  sex+age+(cyear|person), salary, REML=F)
> fit.full <-lmer(log(income) ~  sex +age+educ+(cyear|person), salary, REML=F)
> obs.test.stat<- - 2*(logLik(fit.reduced)-logLik(fit.full))
>
> N<-1000
> boot.test.stats<-rep(0,N)
>
> for(i in 1:N){
+   new.y<-unlist(simulate(fit.full))
+   fm.reduced.new<-lmer(new.y~  sex+age+(cyear|person), salary, REML=F)
+   fm.full.new <-lmer(new.y~  sex +age+educ+(cyear|person), salary, REML=F)
+
+   boot.test.stats[i]<- -2*(logLik(fm.reduced.new)-logLik(fm.full.new))
+ }
>
> sum(boot.test.stats> obs.test.stat)
[1] 25
>
```

*(3 marks)*

**2** (i) Consider the following dataset, where `NA` is used to denote missing observations:

| X | Y |
|---|---|
| NA | 5.2 |
| 2.0 | 1.6 |
| 2.3 | NA |
| 15.0 | 7.1 |
| 2.0 | 4.0 |

Interest lies in the expected value of $Y$, which we will denote by $\mu$. Using the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ as an estimator, compute an estimate of $\mu$ using

(a) Complete-case analysis

(b) Available-case analysis.

***(2 marks)***

(ii) Stochastic imputation with random sampling has been used to create 3 imputed datasets. These are shown below with the imputed values in **bold**. The bottom row gives the sample mean and the sample variance of each column.

| | Imputed 1 | | Imputed 2 | | Imputed 3 | |
|---|---|---|---|---|---|---|
| | X | Y | X | Y | X | Y |
| | **2.0** | 5.2 | **2.0** | 5.2 | **15.0** | 5.2 |
| | 2.0 | 1.6 | 2.0 | 1.6 | 2.0 | 1.6 |
| | 2.3 | **5.2** | 2.3 | **4.0** | 2.3 | **7.1** |
| | 15.0 | 7.1 | 15.0 | 7.1 | 15.0 | 7.1 |
| | 2.0 | 4.0 | 2.0 | 4.0 | 2.0 | 4.0 |
| sample mean | 4.62 | 4.62 | 4.66 | 4.38 | 7.26 | 5.00 |
| sample variance | 33.43 | 4.08 | 33.43 | 4.02 | 49.94 | 5.36 |

(a) Use these values to estimate the expected value $\mathbb{E}(\mu|Y_{obs})$.

(b) Estimate the standard error in your estimate of $\mu$.

**Hint:** The variance of the sample mean $\hat{\mu} = \frac{1}{n}\sum Y_i$ is

$$\mathbb{V}\text{ar}(\hat{\mu}) = \frac{\mathbb{V}\text{ar}(Y)}{n} \approx \frac{s_Y}{n}$$

where $s_Y$ is the sample variance of $Y$.

***(6 marks)***

**2** (continued)

(iii) Consider the mixed effect model

$$y_{ij} = \alpha + \beta_i + b_j + e_{ij}$$

where $i = 1, 2$ and $j = 1, \ldots, J$. The $e_{ij}$ are IID $N(0, \sigma^2)$ and the random effects have the distributions

$$b_j \sim N(0, \sigma_1^2)$$

with $b_j$ independent of $e_{ij}$ for all $i, j$. Sum to zero constraints are used for the fixed effects:

$$\beta_1 + \beta_2 = 0.$$

(a) The model can be written in matrix form as

$$\mathbf{Y} = X\boldsymbol{\beta} + Z\mathbf{b} + \boldsymbol{e}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{21} \\ Y_{12} \\ Y_{22} \\ \vdots \\ Y_{1J} \\ Y_{2J} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_J \end{pmatrix}, \quad \text{and } \boldsymbol{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{J2} \end{pmatrix}.$$

Give the matrices $X$ and $Z$, and the vector $\boldsymbol{\beta}$.

(b) Consider the following estimators of $\alpha$ and $\beta_1$:

$$\hat{\alpha} = \bar{Y} = \frac{1}{2J} \sum_{i,j} Y_{ij}, \qquad \hat{\beta}_1 = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{2}$$

where $\bar{Y}_{i\cdot} = \dfrac{1}{J} \displaystyle\sum_{j=1}^{J} Y_{ij}$. Calculate the correlation between $\hat{\alpha}$ and $\hat{\beta}_1$, i.e., $\text{cor}(\hat{\alpha}, \hat{\beta}_1)$.

***(7 marks)***

## End of Question Paper