The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2020–2021**

Linear Modelling **3.5 hours**

This is an open book exam.

Answer ALL questions.

You can work on the exam during the 24 hour period starting from 10am (BST), and you must submit your work within 3.5 hours of accessing the exam paper or by the end of the 24 hour period (whichever is earlier).

Late submission will not be considered without extenuating circumstances. *Calculations should be performed by hand. A university-approved calculator may be used. The use of any other calculational device, software or service is not permitted. To gain full marks, you will need to show your working. 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.*

1 A fruit supplier is investigating the yields of four different breeds of its plants. The yields (in kilograms) of 128 plants were recorded, together with the soil type (clay, coded as `"C"`, or loam, coded as `"L"`), the breed of the plant (1,2,3 or 4) and the sun exposure (in hours per day) of the site.

The data are stored in a data frame `yield.data` in R, with the variables being `yield` for the yield of the plant, `soil` for the soil type, `breed` for the breed and `sun` for the sun exposure.

(a) A model was fitted in R using

```
model1 <- lm(yield~soil+factor(breed)+sun,data=yield.data)
```

(i) Describe a suitable design matrix for this model. You should illustrate your description by giving the rows corresponding to two observations: one with clay soil, breed 2, and 5.18 hours per day sun exposure, and one with loam soil, breed 4, and 5.34 hours per day sun exposure. ***(7 marks)***

(ii) Explain why you might prefer the code above to the alternative

```
model0 <- lm(yield~soil+breed+sun,data=yield.data)
```

(3 marks)

(iii) Some of the output from entering `summary(model1)` is given below.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.52480    0.39094  11.574  < 2e-16 ***
soilL            1.48463    0.17582   8.444 7.52e-14 ***
factor(breed)2   2.90102    0.25507  11.373  < 2e-16 ***
factor(breed)3   1.12247    0.25273   4.441 1.98e-05 ***
factor(breed)4   1.15865    0.24831   4.666 7.94e-06 ***
sun              0.61997    0.06645   9.330 5.98e-16 ***
```

Give interpretations of the numbers 1.48463, 2.90102 and 0.61997 which appear in the `Estimate` column. ***(3 marks)***

(b) A second model was fitted in R using

```
model2 <- lm(yield~soil+factor(breed)*sun,data=yield.data)
```

(i) Explain the difference between this model and `model1`. How many columns would there be in the design matrix? ***(3 marks)***

(ii) Some output from the code `anova(model1,model2)` is below.

```
Model 1: yield ~ soil + factor(breed) + sun
Model 2: yield ~ soil + factor(breed) * sun
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    122 120.356
2    119  77.705  3    42.651 21.772 2.615e-11 ***
```

What is the null hypothesis of the test being performed here? What conclusion would you make? ***(4 marks)***

2 (a) Consider that observations y_i are generated by the exponential family of distributions

$$f(y_i) = \exp\left[w_i \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right], \qquad (1)$$

where w_i are the weights, ϕ is the scale parameter, θ_i is the natural parameter, $b(\theta_i)$ is a known function of θ_i and $c(y_i, \phi)$ is a known function of y_i, ϕ. Assume that y_i is independent of y_j, for $i \neq j$, and that the mean $\mu_i = E(y_i)$ is mapped to the linear predictor $\eta_i = x_i^T\beta$ via the link function $g(\cdot)$, so that $g(\mu_i) = \eta_i$.

(i) Based on a set of data $y = (y_1, y_2, \ldots, y_n)$, write down the log-likelihood function $\ell(\mu; y)$ of $\mu = [\mu_1, \mu_2, \ldots, \mu_n]^T$. ***(1 mark)***

(ii) Using the canonical link, calculate the partial derivatives of $\ell(\mu; y)$, with respect to β_k, for $k = 1, 2, \ldots, p$. ***(6 marks)***

(iii) Using part (ii) show that the MLE $\hat{\mu}$ of μ satisfies the matrix equality

$$X^T W y = X^T W \hat{\mu},$$

where X is the design matrix and W is the weight matrix, defined below

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \text{and} \quad W = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}.$$

(3 marks)

(b) Consider the discrete random variable $Y_i = y_i$ (taking values $1, 2, 3 \ldots$) generated by the following distribution

$$f(y_i) = \Pr[Y_i = y_i] = (1 - \pi_i)^{y_i - 1}\pi_i, \quad y = 1, 2, \ldots, \quad 0 \leq \pi_i \leq 1,$$

where $i = 1, 2, \ldots, n$.

(i) Write $f(y_i)$ in exponential form (1) and hence determine θ_i, $b(\theta_i)$, $c(y_i, \phi)$, ϕ and w_i. ***(2 marks)***

(ii) Use part (i) to calculate the mean $E(Y_i)$ and the variance $\mathrm{Var}(Y_i)$ of Y_i. ***(3 marks)***

(iii) Show that the canonical link of $f(y_i)$ is

$$g(\mu_i) = \log \frac{\mu_i - 1}{\mu_i},$$

where $\mu_i = E(Y_i)$. ***(3 marks)***

(iv) Show that the X^2 statistic is

$$X^2 = \sum_{i=1}^{n} \frac{(y_i\hat{\pi}_i - 1)^2}{1 - \hat{\pi}_i},$$

where $\hat{\pi}_i$ is the estimate of π_i. ***(2 marks)***

3 A study on potato yield was conducted in order to determine factors that affect yield on potato growth. Observations are recorded of potato `Yield` (in British pounds (£) per pound of potato seed planted) for 16 plots of land. The data also contains some factors, which are believed to influence potato yield. These are: `Variety` (the variety of the potatoes, labelled A, B and C), `Manure` whether manure is used as fertiliser or not, `Potash` whether potash is used and in what levels (0: none, 1: some and 2: a lot) and `Block` the block of the plot (labelled as 1 and 2 according to the location of the plot the potatoes were planted). A statistician was asked to fit a suitable model to this data in order to determine which factors affect potato yield. A small part of the data is given in the table below

	Yield	Variety	Manure	Potash	Block
1	12.1	A	None	0	1
2	11.4	A	None	0	2
3	14.1	A	None	1	1
4	19.0	A	None	1	2
5	17.5	A	None	2	1
6	19.6	A	None	2	2
7	17.0	A	Some	0	1
8	22.5	A	Some	0	2

(a) The statistician suggested a mixed effects model with `Yield` as response variable, where `Variety`, `Manure` and `Potash` are the fixed effects and `Block` are the random effects. Without making any reference to any R output justify the use of a mixed effects model and explain why `Block` should be the random effects factor. ***(2 marks)***

3 (continued)

(b) The mixed effects model in part (a) was fitted in R and part of the output is given below.

```
Linear mixed model fit by REML ['lmerMod']
Formula: Yield ~ factor(Variety) + factor(Manure) + factor(Potash) +
    (1 | Block)
   Data: Potatoes

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.97521 -0.70017  0.01956  0.52831  1.84556

Random effects:
 Groups   Name        Variance Std.Dev.
 Block    (Intercept) 3.750    1.936
 Residual             6.176    2.485
Number of obs: 36, groups:  Block, 2

Fixed effects:
                   Estimate Std. Error t value
(Intercept)          12.116    1.704   7.110
factor(Variety)B      3.258     1.014   3.212
factor(Variety)C     -5.400     1.014  -5.322
factor(Manure)Some    5.961     0.828   7.196
factor(Potash)1       4.983     1.014   4.912
factor(Potash)2       6.125     1.014   6.037
```

(i) Determine whether the above R output suggests that a linear model including only fixed effects, with `Yield` as the response variable and `Variety`, `Manure` and `Potash` as fixed effects would be adequate; whether it is necessary to include `Block` as a random effect.

(4 marks)

(ii) Provide a 95% confidence interval of the effect of `Variety`(B).

(3 marks)

(iii) The scaled residual corresponding to the first row in the table of observations is 1.29. Use this information and any relevant information of the R output provided above, in order to calculate the estimate of the random effects of `Block`. ***(5 marks)***

(iv) The statistician was considering to fit the mixed effects model of part (b) to the data using maximum likelihood estimation, rather than REML. Without making any calculations state what you expect the differences of REML and ML to be, if the statistician is going to go ahead with this. ***(2 marks)***

3 (continued)

(v) There was the suggestion that a simpler mixed effects model without `Potash` as a factor could be a better fit. This model was fitted and the ANOVA table of the two models provided in the R output below

```
Data: Potatoes
Models:
mod1: Yield ~ factor(Variety) + factor(Manure) + (1 | Block)
mod2: Yield ~ factor(Variety) + factor(Manure)
        + factor(Potash) + (1 | Block)
     npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
mod1:    6 207.94 217.44 -97.971   195.94
mod2:    8 181.87 194.54 -82.936   165.87 30.07  2  2.9e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Compare the two models and assess the evidence the models provide on whether `Potash` should be included in the model or not. You should state the null hypothesis and the distribution under the null hypothesis. ***(4 marks)***

End of Question Paper