The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2015–2016**

MAS473 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. Corner-point constraints are used in all R output.
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 study was conducted to investigate alcohol consumption in adolescents. For a period of three years, 82 different adolescents had their alcohol consumption recorded for the years in which they were 14, 15 and 16 years of age. In R, each alcohol measurement is recorded in the variable `alcuse`. Each subject is given a unique id (stored in the R vector `id`), and their age during a particular year is stored as the variable `age`. Interest lies in the average alcohol use and its change through adolescence, and in the differences between subjects. Below is some R output.

```
> age_15 = age -15
> fit1 <- lmer(alcuse ~ age_15 + (age_15|id))
> summary(fit1)
Linear mixed model fit by REML ['lmerMod']
Formula: alcuse ~ age_15 + (age_15 | id)

REML criterion at convergence: 643.2

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.48287 -0.37933 -0.07858  0.38876  2.49284

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 0.6480   0.8050
          age_15      0.1552   0.3939   0.26
 Residual             0.3373   0.5808
Number of obs: 246, groups:  id, 82

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.92195    0.09630   9.574
age_15       0.27065    0.06284   4.307

Correlation of Fixed Effects:
       (Intr)
age_15 0.169
```

(i) Let Y_{ij} be the alcohol consumption of the i^{th} individual during the j^{th} year for which they are observed, and let x_{ij} be the age of that person minus 15, i.e.

$$x_{ij} = \mathsf{age}_{ij} - 15$$

so that $x_{i1} = -1, x_{i2} = 0$, and $x_{i3} = 1$. Write down the equation of the model that has been fitted, defining your notation carefully and specifying appropriate distributions for the terms in the model. ***(3 marks)***

(ii) Briefly justify the choice of fixed and random effects in the model. ***(2 marks)***

1 (continued)

(iii) Give the estimated values for all 6 parameters in the model *(2 marks)*

(iv) Calculate the covariance between a subject's alcohol consumption at the age of 14, and their alcohol consumption at the age of 16. *(4 marks)*

(v) If $I = 82$ is the number of individuals in the dataset, and if σ^2 is the random error variance and σ_1^2 is the variance of the random intercepts, show that

$$\mathbb{V}ar(\hat{\beta}) = \frac{1}{3I}(3\sigma_1^2 + \sigma^2)$$

where $\hat{\beta}$ is the estimate of the average alcohol consumption for all adolescents. *(6 marks)*

(vi) The session is continued below.

```
> fit2 <- lmer(alcuse ~ age_15 + (age_15-1|id)+(1|id))
> fit2
Linear mixed model fit by REML ['lmerMod']
Formula: alcuse ~ age_15 + (age_15 - 1 | id) + (1 | id)
REML criterion at convergence: 645.5346
Random effects:
 Groups   Name        Std.Dev.
 id       age_15      0.3939
 id.1     (Intercept) 0.8050
 Residual             0.5808
Number of obs: 246, groups:  id, 82
Fixed Effects:
(Intercept)       age_15
     0.9220       0.2707
>
> x <- - 2*(logLik(fit1, REML=FALSE)-logLik(fit2, REML=FALSE))
> N<-100
> y<- c()
> for(i in 1:N){
+   z<-unlist(simulate(fit1))
+   fit1sim<-lmer(z ~ age_15 + (age_15|id))
+   fit2sim<-lmer(z ~ age_15 + (age_15-1|id)+(1|id))
+   y[i] <- - 2*(logLik(fit1sim, REML=FALSE)-logLik(fit2sim, REML=FALSE))
+ }
> mean(y>=x)
[1] 0.53
```

What is being tested here? Give the name of the procedure that has been used, and interpret the output. *(3 marks)*

2 Three models are proposed for the number T_{ij} of people moving between origin i and destination j. In each of these models T_{ij} is assumed to be Poisson distributed with mean μ_{ij}. The form of μ_{ij} for each of the three models is as follows

$$\begin{aligned} \text{Model 1} \quad \mu_{ij} &= {d_{ij}}^{\gamma} e^{\alpha_i + \beta_j} \\ \text{Model 2} \quad \mu_{ij} &= d_{ij} e^{\alpha_i} \\ \text{Model 3} \quad \mu_{ij} &= {d_{ij}}^{\gamma} e^{\alpha_i} \end{aligned}$$

where d_{ij} is the known distance between origin i and destination j and α_i, β_j and γ are unknown parameters. In an `R` data-set for this model:

- `T` denotes the number of people moving;
- `A` denotes the factor variable for origin;
- `B` denotes the factor variable for destination;
- `D` denotes the covariate origin-destination distances.

(i) Give an appropriate `R` command (in terms of `T`, `A`, `B` and `D`) that would allow model 1 to be fitted using a generalized linear model. Explain what your instructions do and define any other notation you use. ***(4 marks)***

(ii) Given observations $\{t_{ij}\}$ show that the log-likelihood for all three models is $\sum_i \sum_j t_{ij} \log \mu_{ij} - \mu_{ij} + \text{constant}$. ***(3 marks)***

(iii) For Model 1 differentiate the log-likelihood with respect to α_i and β_j. If $\hat{\alpha}_i$, $\hat{\beta}_j$ and $\hat{\gamma}$ are the maximum likelihood estimators of α_i, β_j and γ respectively, show that the fitted values $\hat{t}_{ij} = {d_{ij}}^{\hat{\gamma}} e^{\hat{\alpha}_i + \hat{\beta}_j}$ must satisfy the equations $\sum_i t_{ij} = \sum_i \hat{t}_{ij}$ and $\sum_j t_{ij} = \sum_j \hat{t}_{ij}$. ***(5 marks)***

(iv) Explain how plots of $\log \mu_{ij}$ against $\log d_{ij}$ would differ for Model 1 and Model 2. ***(4 marks)***

(v) Explain how you would test the null hypothesis that the number of people moving depends on distance between origin and destination and on the origin itself but not the destination? ***(4 marks)***

3 A study was carried out to assess the effects of mothers' drinking history and diet on the birth weight of their babies. Table 1 shows the results of the study. A value of 0 for `drink` indicates the mother did not drink during pregnancy. `Diet` is either `vegan`, `vegetarian` or `neither`. A value of `No` for `Low` indicates the baby wasn't a low weight baby. For notational convenience we use `L`, `DT` and `DK` to represent the variables low, diet and drink respectively. For this question, assume that `L` is a response factor and `DT` and `DK` are controlled factors.

The residual deviances for 4 log-linear models with Poisson errors are given below along with some relevant quantiles:

Model	Residual Deviance	Degrees of Freedom
1) DT*DK	14.32	6
2) DT*DK+L	8.60	5
3) DT*DK+L*DK	2.88	4
4) DT*DK+L*DT	8.07	3

$\chi^2_{5,0.95} = 11.07$, $\chi^2_{4,0.95} = 9.49$, $\chi^2_{3,0.95} = 7.81$, $\chi^2_{2,0.95} = 5.99$, $\chi^2_{1,0.95} = 3.84$

Table 1

	Diet					
	Vegan		Vegetarian		Neither	
	drink		drink		drink	
Low	0	1	0	1	0	1
No	34	29	15	12	43	12
Yes	5	21	9	8	23	11

(i) Write down an algebraic form for the linear predictor for Model 4 and state the corner-point constraints used in this model. ***(5 marks)***

(ii) Specify the non-zero parameters in Model 4 and hence explain why the degrees of freedom for Model 4 is equal to 3. ***(2 marks)***

(iii) Specify the nested structure of the 4 models. Referring to the residual deviances in the `R` output above, what would you conclude about the dependence of birth weight on diet and drinking habits of the birth mother? ***(6 marks)***

(iv) For Model 3 calculate the expected number of low birth weight babies whose mothers were vegan and drank during pregnancy. ***(4 marks)***

(v) Explain how you would calculate a 95% confidence interval for the expected number of babies who didn't have a low birth weight and whose mothers were vegan and did not drink during pregnancy for Model 3. Assume you are given `R` output from the `summary` function in `R` for Model 3. ***(3 marks)***

End of Question Paper