The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2012–2013**

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.
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 An experiment has been conducted to measure the fuel economy of four different models of car. Eight volunteer drivers have been recruited, and each volunteer drives each car three times along a specified route. The fuel economy (in miles per gallon) is recorded for each journey. The data are stored in R, with the variable names `econ` for fuel economy, `car` for the model of car, and `driver` for volunteer.

(i) A model is fitted in R using the command

```
> (fm1<-lmer(econ~car+(1|driver),REML=F))
```

Write down the equation of the model that has been fitted and assigned to the name `fm1`, defining your notation carefully.

(3 marks)

(ii) Define $\bar{Y}_{42\bullet}$ to be the mean of all the observations in which car 2 has been used by driver 4. Using your notation for the model specified in (i), derive expressions for the expectation and variance of $\bar{Y}_{42\bullet}$.

(3 marks)

(iii) Explain the difference between the commands

```
lmer(econ~car+(1|driver))
```

and

```
lm(econ~car+driver)
```

in terms of the models fitted to the data. State, with justification which model you believe to be more appropriate for analysing the data.

(3 marks)

 Question 1 continued on next page

1 (continued)

(iv) Some edited output from R is given below

```
> (fm1<-lmer(econ~car+(1|driver),REML=F))
Linear mixed model fit by maximum likelihood
Formula: econ ~ car + (1 | driver)
   AIC   BIC logLik deviance REMLdev
 169.1 184.4 -78.53    157.1   163.5
Random effects:
 Groups   Name        Variance Std.Dev.
 driver   (Intercept) 3.74157  1.93431
 Residual             0.19064  0.43662
Number of obs: 96, groups: driver, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)  38.0511     0.6897   55.17
car2          3.9785     0.1260   31.56
car3         -1.0796     0.1260   -8.57
car4         10.9470     0.1260   86.85
```

(a) Give the estimated parameter values for each parameter in your model in (i), including variance parameters.

(2 marks)

(b) Calculate the estimated variance for any observation, and the estimated covariance between any two different observations involving the same driver.

(2 marks)

(c) Give one criticism of the choice of R command with regard to estimating the residual variance, and suggest an alternative command.

(1 mark)

(v) State the model assumptions used in your model in part (i), and state how you would check them. State the gradients of any reference lines to be used in Q-Q plots.

(2 marks)

1 (continued)

(vi) The session is continued below.

```
> fm2<-lm(econ~car)
> fm3<-lmer(econ~car+(1|driver/car),REML=F)
> logLik(fm2)
'log Lik.' -201.9398
> logLik(fm3)
'log Lik.' -78.51752
> qchisq(0.999,1)
[1] 10.82757
> qchisq(0.95,1)
[1] 3.841459
```

Conduct any suitable hypothesis tests, stating clearly what the null hypothesis is in each case, to decide which model is most suitable, and interpret the result.

(4 marks)

2 A study has been conducted to find a suitable dose for a new drug. Each patient is given a particular dose, and the outcome is recorded at the end of the study as "patient responds" or "patient does not respond". It is suspected that a patient's genotype may affect the chances of responding. Five different doses are used (including a dose of 0 to give a control group), and the genotype of each patient is noted. The observed data are tabulated below.

genotype A			genotype B		
dose (mg)	number of patients	number responding	dose (mg)	number of patients	number responding
0	13	6	0	7	1
2	15	9	2	5	1
5	13	9	5	7	2
10	16	11	10	4	2
15	13	11	15	7	7

The data are stored in R for each patient with `response` and `genotype` binary indicator variables (1 indicating "patient responds" in `response` and 1 indicating genotype B in `genotype`) in each case, and `dose` representing the dose of the drug in mg. (The length of each vector is 100, corresponding to the 100 patients in the study).

(i) Defining your notation carefully, write down the model that is fitted to the data using the following command, including an equation for the linear predictor.

```
> lm1<-glm(response~dose*genotype,binomial(logit))
```

(2 marks)

(ii) Briefly describe what plots you would draw to choose between different possible link functions, and explain what you would look for in each plot.

(4 marks)

Question 2 continued on next page

2 (continued)

(iii) Some further commands and edited output from R are given below.

```
> summary(lm1)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.05467    0.38268   0.143  0.88641
dose           0.10019    0.05165   1.940  0.05242
genotype      -2.28760    0.88363  -2.589  0.00963
dose:genotype  0.20586    0.11589   1.776  0.07568
---

> anova(lm1)
Analysis of Deviance Table

Model: binomial, link: logit

Response: response

Terms added sequentially (first to last)

              Df Deviance Resid. Df Resid. Dev
NULL                             99     135.37
dose           *   ******        98     121.35
genotype       *   ******        97     116.58
dose:genotype  *   ******        96     112.76
> qchisq(0.95,1)
[1] 3.841459
> qnorm(0.975)
[1] 1.959964
```

(a) Conduct suitable hypothesis tests to assess the effect of dose and genotype on response. Starting with the equation for the linear predictor in the fitted model, state the null hypothesis clearly in each case. Choose the most appropriate model for the data, state the linear predictor, and interpret the result.

(7 marks)

(b) Out of 20 patients who are genotype B and given a dose of 7mg, calculate the expected number of responders, using your chosen model in part (a).

(2 marks)

Question 2 continued on next page

2 (continued)

(c) Calculate the odds ratio of a genotype A patient responding compared to a genotype B patient responding, if both patients are given the same dose, using your chosen model in part (a). Calculate also an approximate 95% confidence interval for this odds ratio.

(2 marks)

(d) For a genotype A patient, estimate the minimum required dose such that the probability of responding is at least 0.9, using your chosen model in part (a). Give one criticism of your estimate.

(3 marks)

3 Appleton, French and Vanderpump (1996) describe a 20-year follow-up study on the effects of smoking. In 1972-74, a sample of women was categorized by age and smoking status (smoker or non-smoker). Twenty years later, the investigators recorded whether each participant was still alive. Smokers who quit in the intervening period were excluded. The data are stored in an R dataframe `femsmoke`, and a subset is shown below.

```
> head(femsmoke)
   y smoker dead   age
1  2    yes  yes 18-24
2  1     no  yes 18-24
3  3    yes  yes 25-34
4  5     no  yes 25-34
5 14    yes  yes 35-44
6  7     no  yes 35-44
```

`y` is the observed count for each combination of smoking status, dead/alive outcome and age group.

Some edited output from an R session is given below.

```
> lm1<-glm(y~dead*smoker, poisson, femsmoke)
> anova(lm1)

Analysis of Deviance Table
Model: poisson, link: log
Response: y
            Df Deviance Resid. Df Resid. Dev
NULL                           27    1193.94
dead         1  261.274        26     932.66
smoker       1   17.161        25     915.50
dead:smoker  1    9.200        24     906.30

> qchisq(0.95,1)
[1] 3.841459
```

(i) Defining your notation carefully, write down the model that has been fitted to the data in R and assigned to `lm1`, including an equation for the linear predictor, stating any necessary parameter constraints. ***(3 marks)***

(ii) Assess whether there is evidence that the probability of death is dependent on smoking status. ***(1 mark)***

(iii) Below are the data tabulated by smoking status and dead/alive outcome only.

	dead	alive
non-smoker	230	502
smoker	139	433

Estimate the probability of death for smokers and non-smokers, and comment briefly on your estimated values. ***(2 marks)***

 Question 3 continued on next page

3 (continued)

(iv) The full data are given below.

smoker			non-smoker		
age	dead	alive	age	dead	alive
18-24	2	53	18-24	1	61
25-34	3	121	25-34	5	152
35-44	14	95	35-44	7	114
45-54	27	103	45-54	12	66
55-64	51	64	55-64	40	81
65-74	29	7	65-74	101	28
75+	13	0	75+	64	0

Based on this table, estimate the probability of death for smokers and non-smokers in each age group. Compare with your result in part (iii), and discuss briefly any apparent differences. ***(5 marks)***

(v) Some further edited R analysis of the data follow next.

```
> lm2<-glm(y~dead*age*smoker,poisson,femsmoke)
> anova(lm2)

Analysis of Deviance Table
Model: poisson, link: log
Response: y

                Df Deviance Resid. Df Resid. Dev
NULL                              27    1193.94
dead             *   ******       **     932.66
age              *   ******       **     752.16
smoker           *   ******       **     735.00
dead:age         *   ******       **     101.83
dead:smoker      *   ******       **      92.63
age:smoker       *   ******       **       2.38
dead:age:smoker  *   ******       **       ****

> qchisq(0.95,1:7)
[1]  3.841  5.991  7.814  9.487 11.070 12.591 14.067
```

Using this analysis of deviance table, decide on the most suitable model for the data. Defining your notation carefully, write down the equation of the linear predictor in your chosen model, stating any necessary parameter constraints. ***(9 marks)***

End of Question Paper