The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2017–18**

Linear and Generalised Linear Models **2 hours**

Attempt all the questions. The allocation of marks is shown in brackets.
RESTRICTED OPEN BOOK EXAMINATION
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator that conforms to University regulations.
There are 60 marks available on the paper.

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 (i) A car manufacturer conducts a study to investigate the proportion of cars manufactured requiring maintenance over time. It is suggested that the proportion of cars requiring a certain amount of maintenance increases with the age of the car. A sample of 100 cars was taken at each age point and the table below summarises the proportion requiring maintenance (NB two samples of 8 year old cars are included).

age (years)	1	4	8	8	10	15
maintenance (maint, in %)	5	8	20	23	30	45

The statistician at the company decided to fit a quadratic linear model to this data set, with the maintenance proportion as the response variable and age (in years) as the explanatory variable. The model is:

$$\text{maint} = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{age}_i^2 + \epsilon_i,$$

where ϵ_i follows the normal distribution $N(0, \sigma^2)$, for some variance σ^2 and ϵ_i is independent of ϵ_j, for $i \neq j$.

In answering this question you may find the following quantiles useful:
$t_{3,0.95} = 2.35$, $t_{3,0.995} = 5.84$, $t_{4,0.995} = 4.60$.

(a) Write down the design matrix X of this model. ***(1 mark)***

(b) The model was fitted in R and gave the following output:

```
Call:
lm(formula = maint ~ age + I(age^2))

Residuals:
     1       2       3       4       5       6
  1.38   -2.58   -1.61    1.39    2.13   -0.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.55292    2.85323   0.544   0.6241
age          2.00786    0.77705   2.584   0.0815 .
I(age^2)     0.06239    0.04686   1.331   0.2753
---

Multiple R-squared:  0.9833,Adjusted R-squared: 0.9721
F-statistic: 88.11 on 2 and 3 DF,  p-value: 0.002166
```

Write down the maximum likelihood estimates of β_0, β_1, β_2 and σ^2. ***(4 marks)***

 Question 1 continued on next page

1 (continued)

(c) Given

$$(X^TX)^{-1} = \begin{pmatrix} 1.34740071 & -0.314622963 & 0.0157196852 \\ -0.31462296 & 0.099935973 & -0.0057645042 \\ 0.01571969 & -0.005764504 & 0.0003635087 \end{pmatrix},$$

calculate the Pearson correlation coefficient between $\hat{\beta}_1$ and $\hat{\beta}_2$ (the maximum likelihood estimators of β_1 and β_2, respectively).

(2 marks)

(d) Find 99% confidence intervals for β_1 and β_2. ***(2 marks)***

(e) Based on the output above, comment on how well the quadratic model fits these data. ***(4 marks)***

(ii) Consider the linear model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is a vector of n observations, X is a $n \times p$ design matrix, β is a vector of p coefficients and $\boldsymbol{\epsilon}$ is the usual error vector satisfying the Gauss-Markov conditions. Let the residual vector be $\boldsymbol{e}$ and define $\mathbf{1}_n$ to be the column vector with n units, i.e.

$$\boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \quad \text{and} \quad \mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Assume that the first column of X is $\mathbf{1}_n$ and that X has rank p.

(a) State the distribution of $\boldsymbol{e}$; find the distribution of $\mathbf{1}_n^T\boldsymbol{e}$.

(2 marks)

(b) Given the general result

$$\operatorname{tr}\left[X(X^TX)^{-1}X^T\mathbf{1}_n\mathbf{1}_n^T\right] = n,$$

where $\operatorname{tr}(A)$ denotes the trace of square matrix A, show

$$\operatorname{Var}(\mathbf{1}_n^T\boldsymbol{e}) = 0.$$

(4 marks)

(c) Hence show $\sum_{i=1}^{n} e_i = 0$. ***(1 mark)***

2 (i) Suppose that data $y_1, y_2, \ldots, y_n$ are independently generated by the exponential family of distributions

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

where θ is the natural parameter, $b(\theta)$ a function of θ, which is assumed to be twice differentiable, ϕ the dispersion parameter, w_i the weights and $c(y_i, \phi)$ a function which depends on y_i and ϕ, but not on θ.

(a) Write down the log-likelihood of θ based on data $y_1, \ldots, y_n$.

(1 mark)

(b) If the canonical link is used and only the null model (comprising of the intercept only) is considered for the linear predictor, then show that the fitted values of y_i are all the same and equal to

$$\hat{y}_i = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}.$$

(4 marks)

 Question 2 continued on next page

2 (continued)

(ii) Consider that Y_i follows a Poisson distribution with rate λ_i and probability mass function

$$P(Y_i = y_i) = \exp(-\lambda_i)\frac{\lambda_i^{y_i}}{y_i!}, \tag{1}$$

for $\lambda_i > 0$ and $y_i = 0, 1, 2, \ldots$.

We consider a generalised linear model with response model (1), with the canonical link and the linear predictor

$$\eta_i = \alpha + \beta x_i,$$

where α and β are subject to estimation and x_i is a known covariate, for $i = 1, 2, \ldots, n$.

(a) Write down the log-likelihood of $(\alpha, \beta)^T$, based on data $y_1, y_2, \ldots, y_n$. ***(3 marks)***

(b) Write down the partial derivatives of the log-likelihood with respect to α and β and show that the maximum likelihood estimates (MLEs) $\hat{\alpha}$ and $\hat{\beta}$ of α and β satisfy the following equations

$$\exp(\hat{\alpha}) = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n \exp(\hat{\beta} x_i)}$$

$$\exp(\hat{\alpha}) \sum_{i=1}^n x_i \exp(\hat{\beta} x_i) = \sum_{i=1}^n x_i y_i$$

(3 marks)

(c) If $n = 2$, $x_1 = 1$, $x_2 = -1$, $y_1 = 1$ and $y_2 = 3$, then show that the equations of part (b) above can be solved analytically and the MLEs are given by

$$\hat{\alpha} = \frac{1}{2}\log 3 \qquad \text{and} \qquad \hat{\beta} = -\frac{1}{2}\log 3$$

(5 marks)

(d) Using (c) calculate the fitted values of y_1 and y_2. Show that these fitted values indicate a perfect fit. Discuss how the above model is related to the saturated model. ***(4 marks)***

3 Data are collected on 80 individuals in a study assessing the effect of age and smoking status on lung cancer risk. The variables recorded are:

- X_1 - smoking status ($X_1 = 1$ for current smokers and $X_1 = 0$ for ex-smokers or non-smokers) abbreviated to `smoke` in the `R` analysis.
- X_2 - age.
- Y - lung cancer status ($Y = 1$ for people with lung cancer and $Y = 0$ for people without diagnosed lung cancer) abbreviated to `cancer` in the `R` analysis.

Various generalised linear models, with a logit link, are fitted where the binary variable Y is the response and X_1 and X_2 are the explanatory variables. Let η_i be the linear predictor for the i-th person. The four fitted models are:

- Model 1: $\eta_i = \beta_0 + \beta_1 X_{1i}$
- Model 2: $\eta_i = \beta_0 + \beta_2 X_{2i}$
- Model 3: $\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$
- Model 4: $\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i}$

In answering this question, you may find the following quantiles useful: $\chi^2_{1,0.95} = 3.84$, $\chi^2_{2,0.95} = 5.99$, $\chi^2_{3,0.95} = 7.81$, $\chi^2_{76,0.95} = 97.35$ $\chi^2_{77,0.95} = 98.48$

(i) What is the relationship between $E(Y_i)$ and η_i for the logit link? What would be the relationship if a probit link were used instead? ***(2 marks)***

(ii) The residual deviances for the four models are given in Table 1. By considering the changes in residual deviance determine the relationship between lung cancer risk, age and smoking status. For any hypothesis tests that you do, state clearly the null hypothesis and the degrees of freedom of the relevant χ^2 distribution used to perform the test. ***(6 marks)***

Model	Residual deviance
Model 1	59.36
Model 2	80.41
Model 3	38.58
Model 4	32.26

Table 1: Residual deviances for Models 1 to 4.

(iii) Comment on the fit of the model you selected in part (ii) based on an appropriate χ^2 distribution. ***(2 marks)***

 Question 3 continued on next page

3 (continued)

(iv) Using R, the following output is obtained for Model 4:

```
Call: glm(formula = cancer ~ smoke * age, family = binomial)
Coefficients:
(Intercept)        smoke          age    smoke:age
   -121.456      118.886        1.854       -1.777

Degrees of Freedom: 79 Total (i.e. Null);  76 Residual
Null Deviance:     104.8
Residual Deviance: 32.26  AIC: 40.26
AIC: 20.36
```

Using this output, calculate the odds of lung cancer for a 70 year old current smoker. ***(2 marks)***

(v) Based on the output from (iv), at what age would the estimated odds of lung cancer for smokers and non-smokers be the same? ***(2 marks)***

(vi) For Model 1 write down an expression for the log-likelihood (l) in terms of β_0, β_1 and X_{1i} and hence calculate $\dfrac{\partial^2 l}{\partial {\beta_0}^2}$ in terms of η_i. How might $\dfrac{\partial^2 l}{\partial {\beta_0}^2}$ be useful when making inferences in a generalised linear model? ***(6 marks)***

End of Question Paper