The
University
Of
Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2016–17**

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 The level of light output of bulbs covered with two types of coating (A and B) is studied. Data are collected and given in the following table.

Length of operation (hours)	Coat	Drop in light output (percent of original output)
0	A	0
400	A	6
800	A	22
1200	A	27
1600	A	32
2000	A	36
2400	A	38
0	B	0
400	B	4
800	B	6
1200	B	9
1600	B	10
2000	B	11
2400	B	12

(i) Write down, assess and interpret the model for which `R` output is given in Table 1 overleaf. ***(6 marks)***

(ii) Assuming that the rate of drop in output per hour of duration for both types of bulbs is identical, set up a general hypothesis and a test procedure for checking whether the relationship of output to length of operation is the same for both types of bulbs. Do not carry out the test, but provide a detailed description of how this would be done. ***(7 marks)***

(iii) A new variable is created, each element of which is the result of averaging the two values of the drop in light output for coats A and B (see table below).

Length of operation (hours)	Averaged Drop in light output (percent of original output)
0	0
400	5
800	13
1200	18
1600	21
2000	23.5
2400	25

Then, a new linear model is set up by regressing the new Averaged Drop in light output on the length of operation. The `R` output of this model is given in Table 2. Comment on the performance of this model and state its limitations, if any. ***(5 marks)***

1 (continued)

(iv) Compare the models obtained in (i) and (iii) and suggest an overall best model. *(2 marks)*

Table 1

```
Call: lm(formula = Drop ~ Length + factor(Coat) - 1)

Residuals:
    Min     1Q Median    3Q   Max
 -10.25 -4.116  2.536 4.375 5.321

Coefficients:
                Value Std. Error  t value Pr(>|t|)
       Length   0.0106   0.0020     5.2080   0.0003
       Coat A  10.2500   3.3646     3.0464   0.0111
       Coat B  -5.3214   3.3646    -1.5816   0.1421

 Residual standard error: 6.107 on 11 degrees of freedom
 Multiple R-Squared: 0.9256
 F-statistic: 45.59 on 3 and 11 degrees of freedom,
 the p-value is 1.703e -006
```

Table 2

```
Call: lm(formula = Drop ~ Length)

Residuals:
      1    2     3     4     5       6      7
 -2.214 -1.5 2.214 2.929 1.643 -0.1429 -2.929

Coefficients:
              Value   Std. Error  t value Pr(>|t|)
  (Intercept) 2.2143   1.7219     1.2860  0.2548
  Length      0.0107   0.0012     8.9742  0.0003

 Residual standard error: 2.527 on 5 degrees of freedom
 Multiple R-Squared: 0.9415
 F-statistic: 80.54 on 1 and 5 degrees of freedom,
 the p-value is 0.00028 66
```

2 Consider the exponential family of distributions for the random variable Y, with probability function

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{\phi/w} + c(y,\phi)\right], \tag{1}$$

where θ is the natural parameter, ϕ is the dispersion parameter, w is the weight and the function $b(\theta)$ is assumed to be twice differentiable.

(i) Using the result that the variance of the score statistic is equal to minus the expectation of the second partial derivative of the log-likelihood of θ, with respect to θ, show that the variance of Y is

$$\mathrm{Var}(Y) = \frac{\phi}{w} b''(\theta),$$

where $b''(\theta)$ is the second derivative of $b(\cdot)$ with respect to θ. ***(5 marks)***

(ii) Show that the mode $\hat{Y} = \hat{y}$ of the distribution of Y, evaluated at observation y, satisfies the differential equation

$$\frac{\partial c(\hat{y},\phi)}{\partial y} = -\frac{w\theta}{\phi},$$

where Y is continuously distributed and $c(y,\phi)$ is assumed to be differentiable with respect to y. ***(4 marks)***

(iii) Consider the two-parameter inverse Gaussian distribution, with p.d.f.

$$f(y;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right],$$

where $\mu, \lambda > 0$ are the two parameters and $f(y;\mu,\lambda)$ is positive for $y > 0$.

(a) Write $f(y;\mu,\lambda)$ in the exponential form (1), hence determine θ, ϕ, w, $b(\theta)$ and $c(y,\phi)$ in terms of μ and λ. ***(6 marks)***

(b) Using the result of part (i) find the variance of Y. ***(2 marks)***

(c) Write down the canonical link of the distribution $f(y;\mu,\lambda)$. Explain why the canonical link may not be a good choice if we wish to fit a generalised linear model with that link. Hence, suggest another link function which is a better choice. ***(3 marks)***

3 The following table summarizes the response to chemotherapy of 30 patients suffering from lymphoma, a form of cancer.

	Gender of patient			
	Male ($j=1$)		Female ($j=2$)	
	Tumour type		Tumour type	
	Nodular	Diffuse	Nodular	Diffuse
	$k=1$	$k=2$	$k=1$	$k=2$
No response ($i=1$)	1	12	2	3
Response ($i=2$)	4	1	6	1

Gender and tumour type are controlled variables. Generalized linear models with Poisson errors and a log link function fitted to these data gave the following results.

Model fitted	Deviance	df
`G*T+R`	14.83	3
`G*T+R*G`	12.02	2
`G*T+R*T`	0.81	?
`G*T+R*T+R*G`	0.65	†

`R` = no response/response
`G` = gender
`T` = type of tumour

(i) Identify the degrees of freedom entries represented by ? and † in these results. ***(3 marks)***

(ii) Check that $\pi_{\bullet jk} = 1$ for all $j, k \in \{0, 1\}$. ***(2 marks)***

(iii) Show (without using the `R` output) that the fitted value $\hat{\mu}_{112}$ for the `G*T+R*G` model is 9.4. ***(2 marks)***

(iv) By first finding the fitted value using $\mu_{ijk} = n_{jk}\pi_{ijk}$, calculate the Pearson and deviance residuals for observation y_{221} for the `G*T+R*T` model. You must not use the `R` output in this question. ***(4 marks)***

(v) Using **only** the `R` output on the next page, verify the fitted value in (iii) above. ***(2 marks)***

(vi) Using changes in scaled deviance in the table above, what conclusions about the dependence of response on gender and type of tumour would you draw? Is the model you decide on a good fit? ***(5 marks)***

(vii) What model-checking would you perform? ***(2 marks)***

3 (continued)

```
 Call:
glm(formula = count ~ factor(gender) * factor(type) + factor(response) *
    factor(gender), family = poisson(log), data = data)

Deviance Residuals:
      1       2       3       4       5       6       7       8
-1.6292  0.8166 -0.7895  0.9274  1.8000 -1.6292  0.5909 -0.9859

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        1.28402    0.47050   2.729  0.00635 **
factor(gender)2                   -0.08004    0.68046  -0.118  0.90636
factor(type)2                      0.95551    0.52623   1.816  0.06941 .
factor(response)2                 -0.95551    0.52623  -1.816  0.06941 .
factor(gender)2:factor(type)2     -1.64866    0.80742  -2.042  0.04116 *
factor(gender)2:factor(response)2  1.29198    0.78726   1.641  0.10077
```

The following R output is additionally provided

```
> qchisq(0.95, 2)
[1] 5.991465
```

End of Question Paper