Queen Mary University of London

Main Examination period 2021 – January – Semester A

MTH6134 / MTH6134P: Statistical Modelling II

You should attempt ALL questions. Marks available are shown next to the questions.

In completing this assessment:

- **You may use books and notes.**
- **You may use calculators and computers, but you must show your working for any calculations you do.**
- **You may use the Internet as a resource, but not to ask for the solution to an exam question or to copy any solution you find.**
- **You must not seek or obtain help from anyone else.**

All work should be **handwritten** and should **include your student number**.

The exam is available for a period of **24 hours**. Upon accessing the exam, you will have **3 hours** in which to complete and submit this assessment.

When you have finished:

- scan your work, convert it to a **single PDF file**, and submit this file using the tool below the link to the exam;
- e-mail a copy to **maths@qmul.ac.uk** with your student number and the module code in the subject line;
- with your e-mail, include a photograph of the first page of your work together with either yourself or your student ID card.

You are expected to spend about **2 hours** to complete the assessment, plus the time taken to scan and upload your work. Please try to upload your work well before the end of the submission window, in case you experience computer problems. **Only one attempt is allowed – once you have submitted your work, it is final**.

Examiners: D. S. Coad, L. Rossini

Question 1 [23 marks]. This question is similar to those on exercise sheets.

(a) The likelihood is

$$\begin{aligned} L(\beta_1,\beta_2;\mathbf{y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i-\beta_1 x_i-\beta_2 z_i)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_1 x_i-\beta_2 z_i)^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_1 x_i-\beta_2 z_i)^2\right\}. \end{aligned}$$

[**6**]

(b) The log-likelihood is

$$\ell(\beta_1,\beta_2;\mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_1 x_i-\beta_2 z_i)^2.$$

Thus, we have

$$\frac{\partial \ell}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i-\beta_1 x_i-\beta_2 z_i)$$

and

$$\frac{\partial \ell}{\partial \beta_2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} z_i(y_i-\beta_1 x_i-\beta_2 z_i).$$

Setting these derivatives to zero, we obtain

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 - \hat{\beta}_2\sum_{i=1}^{n} x_i z_i = 0$$

and

$$\sum_{i=1}^{n} z_i y_i - \hat{\beta}_1\sum_{i=1}^{n} x_i z_i - \hat{\beta}_2\sum_{i=1}^{n} z_i^2 = 0.$$

Now, the first of these yields

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i z_i}.$$

Substituting this equation into the previous one, we have

$$\sum_{i=1}^{n} z_i y_i - \hat{\beta}_1\sum_{i=1}^{n} x_i z_i - \left(\sum_{i=1}^{n} x_i y_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2\right)\frac{\sum_{i=1}^{n} z_i^2}{\sum_{i=1}^{n} x_i z_i} = 0,$$

which may be rearranged to give

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} z_i y_i \sum_{i=1}^{n} x_i z_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} z_i^2}{\left(\sum_{i=1}^{n} x_i z_i\right)^2 - \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} z_i^2}.$$

[**12**]

Continue to next page

(c) We can write

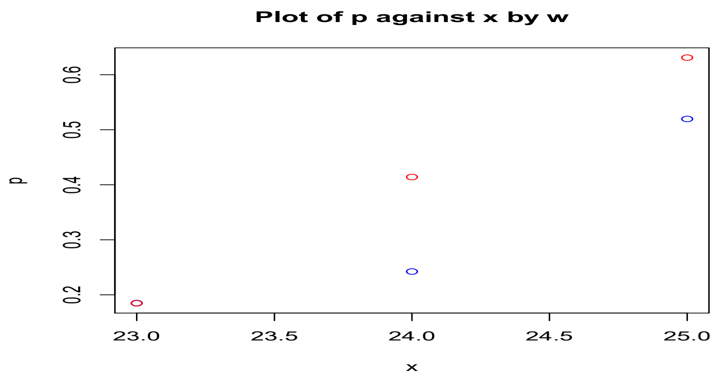
$$
\begin{aligned}
E(\hat{\beta}_1) &= \frac{E\left(\sum_{i=1}^n z_i Y_i\right)\sum_{i=1}^n x_i z_i - E\left(\sum_{i=1}^n x_i Y_i\right)\sum_{i=1}^n z_i^2}{\left(\sum_{i=1}^n x_i z_i\right)^2 - \sum_{i=1}^n x_i^2 \sum_{i=1}^n z_i^2} \\
&= \frac{\sum_{i=1}^n z_i E(Y_i)\sum_{i=1}^n x_i z_i - \sum_{i=1}^n x_i E(Y_i)\sum_{i=1}^n z_i^2}{\left(\sum_{i=1}^n x_i z_i\right)^2 - \sum_{i=1}^n x_i^2 \sum_{i=1}^n z_i^2} \\
&= \frac{\sum_{i=1}^n z_i(\beta_1 x_i + \beta_2 z_i)\sum_{i=1}^n x_i z_i - \sum_{i=1}^n x_i(\beta_1 x_i + \beta_2 z_i)\sum_{i=1}^n z_i^2}{\left(\sum_{i=1}^n x_i z_i\right)^2 - \sum_{i=1}^n x_i^2 \sum_{i=1}^n z_i^2} = \beta_1,
\end{aligned}
$$

and so $\hat{\beta}_1$ is an unbiased estimator of β_1. [**4**]

(d) The distribution of $\hat{\beta}_1$ is normal because $\hat{\beta}_1$ is a linear combination of normal random variables. [**1**]

Question 2 [20 marks]. This question is similar to examples in the lecture notes.

(a) A plot of the proportions of babies surviving to discharge against gestational age by epoch is given below.



This suggests that the regression lines for the two epochs are not parallel. In particular, most of the survival rates have improved from the first epoch to the second. **[6]**

(b) Since the maximum likelihood estimates of α_1, α_2, β_1 and β_2 are $\hat{\alpha}_1 = -22.9574$, $\hat{\alpha}_2 = -23.4655$, $\hat{\beta}_1 = 0.9188$ and $\hat{\beta}_2 = 0.9611$, the fitted logistic regression model is

$$\hat{\pi}_{1k} = \frac{e^{-22.9574+0.9188x_k}}{1+e^{-22.9574+0.9188x_k}}$$

for those babies at the first epoch and

$$\hat{\pi}_{2k} = \frac{e^{-23.4655+0.9611x_k}}{1+e^{-23.4655+0.9611x_k}}$$

for those at the second. **[6]**

(c) In this case, we are fitting a logistic regression model with $p = 4$ parameters and the maximal model has $n = 6$ parameters. The data give $D = 3.6228$. Since $\chi^2_{2,0.1} = 4.605$, the p-value is $P > 0.1$, and so there is no evidence that the logistic regression model does not fit the data well. **[4]**

(d) An approximate 95% confidence interval for $\beta_1 - \beta_2$ is

$$\begin{aligned}
\hat{\beta}_1 - \hat{\beta}_2 \pm 1.96 \times \sqrt{\hat{v}^{33} + \hat{v}^{44}} &= -0.0423 \pm 1.96 \times \sqrt{0.1499^2 + 0.1459^2} \\
&= -0.0423 \pm 1.96 \times 0.2092 \\
&= -0.0423 \pm 0.4100
\end{aligned}$$

or $(-0.4523, 0.3677)$. [**4**]

Question 3 [22 marks]. Part (a) is bookwork, and parts (b), (c) and (d) are similar to questions on exercise sheets.

(a) We know that $\mu_i = E(Y_i) = r_i\pi_i$ and

$$\begin{aligned}\eta_i &= \log\{-\log(1-\pi_i)\}\\ &= \log\left\{-\log\left(\frac{r_i-\mu_i}{r_i}\right)\right\} = g(\mu_i).\end{aligned}$$

It follows that we can write the model in the form $g(\mu_i) = \boldsymbol{\beta}^\top \mathbf{x}_i$, where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ and $\mathbf{x}_i = (1, x_i)^\top$. Since the distribution of each Y_i is in canonical form and depends on a single parameter π_i, this is a generalised linear model. **[4]**

(b) We can write

$$\frac{\partial \eta_i}{\partial \mu_i} = -\frac{1}{(r_i-\mu_i)\log\left(\frac{r_i-\mu_i}{r_i}\right)} = -\frac{1}{r_i(1-\pi_i)\log(1-\pi_i)}.$$

Thus, since $\text{Var}(Y_i) = r_i\pi_i(1-\pi_i)$, the Fisher information matrix is

$$V = \begin{pmatrix} \sum_{i=1}^n \frac{r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2 & \sum_{i=1}^n \frac{x_i r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2 \\ \sum_{i=1}^n \frac{x_i r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2 & \sum_{i=1}^n \frac{x_i^2 r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2 \end{pmatrix}.$$

[8]

(c) We have

$$V^{-1} = \frac{1}{|V|}\begin{pmatrix} \sum_{i=1}^n \frac{x_i^2 r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2 & -\sum_{i=1}^n \frac{x_i r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2 \\ -\sum_{i=1}^n \frac{x_i r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2 & \sum_{i=1}^n \frac{r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2 \end{pmatrix},$$

where

$$\begin{aligned}|V| &= \sum_{i=1}^n \frac{r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2 \sum_{i=1}^n \frac{x_i^2 r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2\\ &\quad - \left[\sum_{i=1}^n \frac{x_i r_i(1-\pi_i)}{\pi_i}\{\log(1-\pi_i)\}^2\right]^2.\end{aligned}$$

This means that, for large n, $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, v^{22})$, where $v^{22} = \sum_{i=1}^n r_i(1-\pi_i)\{\log(1-\pi_i)\}^2/(\pi_i|V|)$. **[8]**

(d) For large n, under H_0, $Z = \hat{\beta}_1/\sqrt{\hat{v}^{22}} \sim \mathrm{N}(0,1)$. Consequently, the critical region for a test of $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$ with approximate significance level α is $R = \{\mathbf{y} : |z| > z_{\frac{\alpha}{2}}\}$. **[2]**

Question 4 [23 marks]. This question is similar to those on exercise sheets.

(a) The data would be entered into R column by column:

```
c1 <- c(2,15)
c2 <- c(3,28)
c3 <- c(31,23)
c4 <- c(13,5)
y <- c(c1,c2,c3,c4)
```

Then the levels of the row and column factors would be generated by

```
row <- gl(2,1,length=8)
column <- gl(4,2,length=8)
```

A log-linear model is fitted to the data using

```
blood <- glm(y ~ row + column,poisson)
```

[**4**]

(b) The null hypothesis states that $E(Y_{jk}) = y_{j.}\theta_{.k}$, where $\theta_{.k} = \sum_{j=1}^{2} \theta_{jk}$. By Birch's conditions, we know that the maximum likelihood estimate of $\theta_{.k}$ under the null hypothesis is $\hat{\theta}_{.k} = y_{.k}/n$, where $n = 120$. It follows that the expected frequency for cell (j,k) is

$$e_{jk} = y_{j.}\hat{\theta}_{.k} = \frac{y_{j.}y_{.k}}{n}.$$

[**4**]

(c) The expected values under the null hypothesis are given in the following table:

	Unnecessary Blood Transfusion				
Surgeon	Frequent	Occasionally	Rarely	Never	Total
Attending	6.942	12.658	22.050	7.350	49
Resident	10.058	18.342	31.950	10.650	71

By comparing these with the observed values, we see that surgical residents in training applied unnecessary blood transfusions more frequently than attending physicians. [**5**]

(d) The deviance is

$$D = 2\sum_{j=1}^{2}\sum_{k=1}^{4} y_{jk}\log\left(\frac{y_{jk}}{e_{jk}}\right) = 35.331$$

and the value of Pearson's goodness-of-fit test statistic is

$$X^2 = \sum_{j=1}^{2}\sum_{k=1}^{4} \frac{(y_{jk} - e_{jk})^2}{e_{jk}} = 31.881.$$

Since $\chi^2_{3,0.001} = 16.268$, the p-value is $P < 0.001$, and so there is very strong evidence that the distributions of unnecessary blood transfusions are not the same for the two groups of surgeons. [**10**]

Question 5 [12 marks]. This question is similar to those on exercise sheets.

(a) We know that $\mu_i = E(T_i) = 1/\lambda_i$. Consequently, we have $\eta_i = 1/\mu_i$, which corresponds to the reciprocal link. **[1]**

(b) The likelihood is

$$\begin{aligned} L(\beta;\mathbf{t}) &= \prod_{i=1}^{n} \left(\beta x_i e^{-\beta x_i t_i}\right)^{\delta_i} \left(e^{-\beta x_i t_i}\right)^{1-\delta_i} \\ &= \beta^{\sum_{i=1}^{n} \delta_i} \left(\prod_{i=1}^{n} x_i^{\delta_i}\right) e^{-\beta \sum_{i=1}^{n} x_i t_i}, \end{aligned}$$

where $\delta_i = 1$ if $T_i = t_i$ and $\delta_i = 0$ if $T_i > t_i$. **[4]**

(c) The log-likelihood is

$$\ell(\beta;\mathbf{t}) = \sum_{i=1}^{n} \delta_i \log(\beta) + \sum_{i=1}^{n} \delta_i \log(x_i) - \beta \sum_{i=1}^{n} x_i t_i.$$

Thus, we have

$$\frac{d\ell}{d\beta} = \frac{\sum_{i=1}^{n} \delta_i}{\beta} - \sum_{i=1}^{n} x_i t_i.$$

Setting this derivative to zero yields the maximum likelihood estimator $\hat{\beta} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} x_i T_i$. **[5]**

(d) The details of the fitted model are obtained using

```
model <- glm(t ~ x - 1,family=gamma)
summary(model,dispersion=1)
```

[2]

End of Paper.