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

 Continue to next page

Question 1 [23 marks]. Suppose that $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$ for $i = 1, 2, \ldots, n$, all independent, where $\mu_i = \beta_1 x_i + \beta_2 z_i$, x_i and z_i are known covariates, and σ is known.

(a) Write down the likelihood for the data $y_1, \ldots, y_n$. **[6]**

(b) Find the maximum likelihood estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ of β_1 and β_2. **[12]**

(c) Prove that $\hat{\beta}_1$ is an unbiased estimator of β_1. **[4]**

(d) Explain why $\hat{\beta}_1$ has a normal distribution. **[1]**

 Continue to next page

Question 2 [20 marks]. The numbers of babies surviving to discharge from a hospital (y) out of the number admitted to neonatal intensive care (r) for two epochs (w) and three gestational ages (x), in weeks, were recorded. Below are the data.

x	23	23	24	24	25	25
w	1	2	1	2	1	2
r	81	65	165	198	229	225
y	15	12	40	82	119	142

Let Y_{jk} denote the number of babies surviving to discharge out of the r_{jk} of gestational age x_k admitted to neonatal intensive care for epoch j. Then it is assumed that $Y_{jk} \sim \text{Bin}(r_{jk}, \pi_{jk})$ for $j = 1,2$ and $k = 1,2,3$, all independent, where $\log\{\pi_{jk}/(1-\pi_{jk})\} = \alpha_j + \beta_j x_k$. This model was fitted to the data using R and the following output was obtained:

```
Call:
glm(formula = p ~ w + w:x, family = binomial, weights = r)

Deviance Residuals:
      1        2        3        4        5        6
 1.1557  -0.3945  -1.3118   0.3665   0.4957  -0.1753

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.9574     3.6704  -6.255 3.98e-10 ***
w2           -0.5081     5.1116  -0.099    0.921
w1:x          0.9188     0.1499   6.128 8.88e-10 ***
w2:x          0.9611     0.1459   6.587 4.47e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 109.1191  on 5  degrees of freedom
Residual deviance:   3.6228  on 2  degrees of freedom
AIC: 42.76

Number of Fisher Scoring iterations: 4
```

(a) Plot the proportions of babies surviving to discharge against gestational age by epoch. What are your conclusions? **[6]**

(b) Write down the fitted logistic regression model for each epoch. **[6]**

(c) Use the above output to assess the goodness of fit of the model. **[4]**

(d) Give an approximate 95% confidence interval for $\beta_1 - \beta_2$. **[4]**

 Continue to next page

Question 3 [22 marks]. Suppose that $Y_i \sim \text{Bin}(r_i, \pi_i)$ for $i = 1, 2, \ldots, n$, all independent, where the r_i are known, $\log\{-\log(1-\pi_i)\} = \beta_0 + \beta_1 x_i$ and x_i is a known covariate.

(a) Explain why this is a generalised linear model. **[4]**

(b) Find the Fisher information matrix. **[8]**

(c) Obtain the asymptotic distribution of the maximum likelihood estimator $\hat{\beta}_1$ of β_1. **[8]**

(d) State the form of an approximate test for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. **[2]**

Question 4 [23 marks]. A study of 49 attending physicians and 71 surgical residents in training at a university hospital was carried out to investigate whether the two groups of surgeons were applying unnecessary blood transfusions at different rates. For each surgeon, the number of blood transfusions prescribed unnecessarily in one year was recorded. The contingency table below summarises the data.

	Unnecessary Blood Transfusion				
Surgeon	Frequent	Occasionally	Rarely	Never	Total
Attending	2	3	31	13	49
Resident	15	28	23	5	71

Let Y_{jk} denote the number of surgeons classified in row j and column k. Then it is assumed that the Y_{jk} for row j have a multinomial distribution with parameters $y_{j.}$ and θ_{jk} for $j = 1, 2$ and $k = 1, 2, 3, 4$, and that the rows are independent, where $y_{j.} = \sum_{k=1}^{4} y_{jk}$ and θ_{jk} is the probability that a surgeon is classified in row j and column k. The null hypothesis is that the distributions of unnecessary blood transfusions are the same for the two groups of surgeons.

(a) Briefly explain how you would enter these data into R. What commands would you use to fit a log-linear model to the data? **[4]**

(b) Explain why, under the null hypothesis, the expected frequency for cell (j,k) is $e_{jk} = y_{j.}y_{.k}/n$, where $n = 120$. **[4]**

(c) Obtain the expected values under the null hypothesis. Compare these with the observed values. **[5]**

(d) Find the deviance and the value of Pearson's goodness-of-fit test statistic. What is your conclusion about the numbers of unnecessary blood transfusions for the two groups of surgeons? **[10]**

Question 5 [12 marks]. Suppose that $T_i \sim \text{Exp}(\lambda_i)$ for $i = 1, 2, \ldots, n$, all independent, where $\lambda_i = \beta x_i$ and x_i is a known covariate.

(a) Explain what link is being used here. **[1]**

(b) Write down the likelihood for the data (t_i, δ_i) for $i = 1, 2, \ldots, n$, where δ_i is a censoring variable. **[4]**

(c) Show that the maximum likelihood estimator of β is $\hat{\beta} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} x_i T_i$. **[5]**

(d) Now assume that there is no censoring. Given that the vectors `t` and `x` in R contain the times and the covariate values, what commands would you use to obtain the details of the fitted model? **[2]**

End of Paper.