Queen Mary
University of London

Main Examination period 2022 – 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.

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, A. Guillaumin

 Continue to next page

Question 1 [24 marks]. Suppose that $Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$, all independent, where $\mu_i = \beta_1 x_i + \beta_2 x_i^2$, x_i is a known covariate and the σ_i are 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) Explain why the above is a generalised linear model. **[4]**

(d) State the iterative weights and working dependent variates for Fisher's method of scoring. **[2]**

 Continue to next page

Question 2 [19 marks]. The numbers of new melanoma cases (y) in 1969-1971 among white males in two areas (w) for six ages (x), in years, were recorded, where the ages are midpoints of intervals. Below are the data.

x	30	40	50	60	70	80	30	40	50	60	70	80
w	1	1	1	1	1	1	2	2	2	2	2	2
y	61	76	98	104	63	80	64	75	68	63	45	27

Let Y_{jk} denote the number of new melanoma cases for age x_k in area j. Then it is assumed that $Y_{jk} \sim \text{Poisson}(\mu_{jk})$ for $j = 1,2$ and $k = 1,2,\ldots,6$, all independent, where $\log(\mu_{jk}) = \alpha_j + \beta_j x_k$. This model was fitted to the data using R and the following output was obtained:

```
Call:
glm(formula = y ~ w + w:x, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.29127  -1.75130  -0.07461   1.19941   2.42769

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.264158   0.155300  27.458  < 2e-16 ***
w2           0.531125   0.232380   2.286   0.0223 *
w1:x         0.002206   0.002668   0.827   0.4084
w2:x        -0.014209   0.003225  -4.405 1.06e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 74.240  on 11  degrees of freedom
Residual deviance: 29.885  on  8  degrees of freedom
AIC: 110.11

Number of Fisher Scoring iterations: 4
```

(a) Plot the numbers of new melanoma cases against age by area. What are your conclusions? **[5]**

(b) Write down the fitted Poisson regression model for each area. **[5]**

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

(d) Test whether the regression lines are parallel. **[5]**

 Continue to next page

Question 3 [21 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, $\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_i$, x_i is a known covariate and Φ denotes the standard normal distribution function.

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

(b) Obtain the asymptotic distribution of the maximum likelihood estimator $\hat{\beta}_0$ of β_0. **[8]**

(c) Write down an approximate $100(1-\alpha)\%$ confidence interval for β_0. **[3]**

(d) Given that the vectors `y` and `x` in R contain the responses and the covariate values, what commands would you use to obtain the details of the fitted model? **[2]**

Question 4 [23 marks]. An experiment was conducted in which 141 fish were placed in a large tank for a period of time and some are eaten by large birds of prey. The fish are categorised by their level of parasitic infection. A summary of the data is provided in the contingency table below.

	Level of Infection			
	Uninfected	Lightly Infected	Highly Infected	Total
Eaten	1	10	37	48
Not Eaten	49	35	9	93
Total	50	45	46	141

Let Y_{jk} denote the number of fish classified in row j and column k. Then it is assumed that the Y_{jk} have a multinomial distribution with parameters n and θ_{jk} for $j = 1, 2$ and $k = 1, 2, 3$, where $n = 141$ and θ_{jk} is the probability that a fish is classified in row j and column k. The null hypothesis is that being eaten and infection status are independent.

(a) State the null hypothesis in terms of $E(Y_{jk})$. Express this as a log-linear model, explaining your notation and any additional constraints. **[6]**

(b) Write down the maximal model. **[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 independence of being eaten and infection status? **[8]**

Question 5 [13 marks]. Suppose that $T_1, \ldots, T_n$ are independent Weibull random variables with probability density function

$$f(t) = 3\lambda t^2 e^{-\lambda t^3},$$

where $\lambda > 0$.

(a) Show that this distribution is a member of the exponential family. **[4]**

(b) Explain why the distribution is not in canonical form. **[1]**

(c) 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]**

(d) Find the maximum likelihood estimator $\hat{\lambda}$ of λ. **[4]**

End of Paper.