Main Examination period 2020 – January – Semester A

MTH6134 / MTH6134P: Statistical Modelling II

Duration: 2 hours

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

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

The New Cambridge Statistical Tables are provided.
Only non-programmable calculators that have been approved from the college list of non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

Complete all rough work in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any unauthorised notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the examination room.

Examiners: D. S. Coad, L. I. Pettit

Question 1 [20 marks]. Suppose that $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$ for $i = 1, 2, \ldots, n$, all independent, where $\mu_i = \boldsymbol{\beta}^\top \mathbf{x}_i$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top$, $\mathbf{x}_i = (1, x_{1i}, \ldots, x_{p-1,i})^\top$ and σ is known.

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

(b) Show that the maximum likelihood estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, where $\mathbf{X}$ is the $n \times p$ design matrix with ith row $\mathbf{x}_i^\top$. State any required assumptions on the design matrix. **[6]**

(c) Find the Fisher information matrix. **[4]**

(d) State the asymptotic distribution of $\hat{\boldsymbol{\beta}}$. Explain why, here, the distribution is exact. **[4]**

Question 2 [18 marks]. The number of deaths due to AIDS in Australia (y) per three-month period from January 1983 to June 1986 was recorded. The time (x) is measured in multiples of three months after January 1983. Below are the data.

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y	0	1	2	3	1	4	9	18	23	31	20	25	37	35

Let Y_i denote the number of deaths due to AIDS in period x_i. Then it is assumed that $Y_i \sim \text{Poisson}(\mu_i)$ for $i = 1, 2, \ldots, 14$, all independent, where $\log(\mu_i) = \beta_0 + \beta_1 x_i$. This model was fitted to the data using R and the following output was obtained:

```
Call:
glm(formula = y ~ x, family = poisson(link=log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2874  -1.1306  -0.6441   0.1341   2.8629

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.45622    0.24779   1.841   0.0656 .
x            0.24155    0.02197  10.997   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 188.084  on 13  degrees of freedom
Residual deviance:  33.627  on 12  degrees of freedom
AIC: 90.304

Number of Fisher Scoring iterations: 5
```

(a) Write down the fitted Poisson regression model, and the standard errors of the maximum likelihood estimates of β_0 and β_1. How are the standard errors calculated from the Fisher information matrix V? **[6]**

(b) Give the form of the test statistic for testing $H_0 : \beta_1 = 0$ and draw conclusions. **[4]**

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

(d) Is there evidence that this model is an improvement over the null model with just an intercept? Justify your answer. **[4]**

Question 3 [24 marks]. Suppose that $Y_i \sim \text{Bin}(1, \pi_i)$ for $i = 1, 2, \ldots, n$, all independent, where $\log\{\pi_i/(1-\pi_i)\} = \beta x_i$ and x_i is a known covariate.

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

(b) Obtain the likelihood equation. **[5]**

(c) Find the Fisher information. **[6]**

(d) Explain how the likelihood equation can be solved iteratively to find the maximum likelihood estimate of β using Fisher's method of scoring. **[7]**

Question 4 [26 marks]. Urine drug screening was performed on 2,537 applicants for positions in the U.S. Postal Service. The contingency table below shows the distribution of the results by drug present and gender. Those applicants who tested positive for more than one drug were classified under the more serious of the drugs, so that each individual only contributed to a single cell in the table.

	Drug Present				
Gender	None	Marijuana	Cocaine	Other Drugs	Total
Male	1,465	146	33	28	1,672
Female	764	52	22	27	865
Total	2,229	198	55	55	2,537

Let Y_{jk} denote the number of individuals 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 $k = 1, 2, 3, 4$, where $n = 2,537$ and θ_{jk} is the probability that an individual is classified in row j and column k. The null hypothesis is that gender and drug present 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) Given that the maximum likelihood estimate of θ_{jk} in the maximal model is y_{jk}/n and that under the null hypothesis is e_{jk}/n, where $e_{jk} = y_{j.}y_{.k}/n$, find the generalised likelihood ratio, $\Lambda(\mathbf{y})$, and hence obtain the deviance given by $D = -2\log\{\Lambda(\mathbf{y})\}$. **[12]**

(d) It was found that $D = 11.737$. What is your conclusion about the independence of gender and drug present? **[4]**

Question 5 [12 marks]. Suppose that the survival time $T > 0$ of a patient has probability density function $f(t)$ and distribution function $F(t)$.

(a) Define the survivor function $S(t)$ and the hazard function $h(t)$ in terms of $f(t)$ and $F(t)$. **[4]**

(b) Compute $S(t)$ and $h(t)$ when $T \sim \text{Exp}(\lambda)$. **[4]**

(c) Explain what is meant by saying that a survival time is **censored**. **[2]**

(d) Give two reasons why censoring might occur in practice. **[2]**

End of Paper.