Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2023	MATH4071-WE01

Title: Topics in Statistics IV

Time:	3 hours	
Additional Material provided:		
Materials Permitted:		
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Answer all questions. Section A is worth 20%, Section B is worth 60%, and Section C is worth 20%. Within Sections A and B, all questions carry equal marks. Students must use the mathematics specific answer book.

Revision:	

SECTION A

Q1 An extremely bored academic wants to investigate the relationship between the number of years it has been since students graduated from Durham, and whether or not they can remember the name of their academic advisor. The response variable takes values 0 and 1, corresponding to "No" and "Yes". There is one measured predictor variable, which is the number of years since graduation, expressed as a continuous measure.

A logistic regression model is calculated for these data, generating the following formula for the linear component

$$\eta_i = 1.386 - 0.277 x_i$$

where x_i represents the number of years since graduation for student i.

1.1 Explain why we cannot use a linear model to describe the relationship between whether a student remembers their academic advisor's name, and the number of years since graduation.

1.2 After how many years is a student estimated to be equally likely to remember their academic advisor's name as to not remember it?

1.3 A GLM has the following $(1-\alpha)100\%$ confidence interval formula for $E(Y|\beta, x_0)$, where x_0 is the predictor variable value(s) for a new observation:

$$\left[h\left(\hat{\beta}^T x_0 - z_{\frac{\alpha}{2}} \sqrt{x_0^T F^{-1}(\hat{\beta}) x_0} \right), h\left(\hat{\beta}^T x_0 + z_{\frac{\alpha}{2}} \sqrt{x_0^T F^{-1}(\hat{\beta}) x_0} \right) \right]$$

Give the definition of $F(\beta)$ in terms of the log-likelihood function of the data (you do not need to give the specific form of $F(\beta)$ for logistic regression).

1.4 The academic finds that

$$F(\hat{\beta}) = \begin{pmatrix} 54.6 & 704.8 \\ 704.8 & 10233.3 \end{pmatrix}$$

Find an estimate for the probability that, ten years after graduation, a Durham student can remember the name of their academic advisor. Find a 95% confidence interval for that probability.

Q2 A retail website gives its customers the option to receive regular discount codes either by text, or by email. The company that runs the website is interested in whether customers who request updates by phone use these discount codes at the same rate as those who request updates by email.

To explore this, a market research company spends a month collecting relevant data, displayed in the contingency table below. In the table, X represents how a discount code was sent (1="By text", 2="By email"). Y represents whether the code was used or not (1="No", 2="Yes").

		Y		Total
		1	2	
X	1	241	33	274
	2	572	64	636
Total		813	97	910

2.1 Find the maximum likelihood estimates for each probability π_{ij}, $i, j \in \{1, 2\}$, without assuming that X and Y are independent.

2.2 State, giving a justification for your answer, which of the following is the most likely sampling scheme used by the marketing company: a) Poisson sampling, b) multinomial sampling, c) product multinomial sampling.

2.3 Find the maximum likelihood estimate $\hat{\theta}$ for the odds ratio θ for the events $\{Y = 1|X = 1\}$ and $\{Y = 1|X = 2\}$, without assuming that X and Y are independent.

2.4 Use the Pearson statistic to test the null hypothesis that X and Y are independent against the alternative hypothesis that X and Y are **not** independent at the 5% level. Ensure you show both your working and your conclusion. **Hint**: the formula for the Pearson statistic is

$$X^2 = \sum_{ij} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

SECTION B

Q3 The R dataset `Seatbelts` gives monthly data (from January 1963 to December 1984) for injuries and fatalities resulting from car accidents on British roads. The data in the following question is a small subset of that data.

The response variable for this question is `Drivers`, the number of drivers who lost their lives in car accidents in a month. The predictor variables are given below.

Variable	Meaning
`front`	Number of fatalities among front passengers
`rear`	Number of fatalities among rear passengers
`PetrolPrice`	Petrol price in pence per millilitre
`Law`	Whether a national seatbelt law was in force (=1) or not (=0)

The R output for a Poisson GLM using the natural link is given on the next page, with one value removed.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0736  -1.0945  -0.1332   1.1997   3.6635

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.985e+00  9.502e-02  41.942  < 2e-16 ***
front        1.206e-03  7.654e-05  15.756  < 2e-16 ***
rear        -7.208e-04  1.244e-04  -5.792 6.96e-09 ***
PetrolPrice  7.423e-01  6.676e-01   1.112     [1]
law          1.394e-01  3.095e-02   4.505 6.63e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 984.5  on 191  degrees of freedom
Residual deviance: 448.5  on 187  degrees of freedom
AIC: 1731.3

Number of Fisher Scoring iterations: 4
```

3.1 For each of the predictor variables, state whether they are categorical or numerical.

3.2 Give a precise numerical interpretation of the estimated parameter for `PetrolPrice`.

3.3 Find the value labelled [1] in the output, and explain what precisely this value tells you about the variable `PetrolPrice`, in the context of the Poisson model presented.

3.4 Another GLM, this time using the Gamma distribution and its natural link, is fitted to the data.

(i) Comment critically on the decision to model this data using the gamma distribution rather than the Poisson distribution. What advantages/disadvantages could there be to such a change?

(ii) A friend suggests you scale the `Drivers` values by subtracting its mean value from each individual value, giving a variable which can now take negative or positive values. Your friend states that by doing so, you can now use a linear model rather than a Gamma model, as it is no longer the case that each response value must be non-negative.
State whether or not your friend is correct, justifying your answer by reference to the linear model.

Q4 **4.1** The inverse Gaussian distribution has probability density function:

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

for $y > 0$, for parameters $\mu > 0$ and $\lambda > 0$. Using the fact that an exponential dispersion family (EDF) of distributions must have a probability density function expressible in the form:

$$\exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

prove that the inverse Gaussian distribution is an EDF with $\phi = \lambda^{-1}$ by finding

(i) The natural parameter θ;

(ii) The log normaliser $b(\theta)$;

(iii) The function $c(y, \phi)$.

4.2 By using the properties of EDFs, use your answer to **4.1** to find the expectation and variance of an inverse Gaussian distribution, in terms of the natural and dispersion parameters.

4.3 A trial is conducted to test the efficacy of four detergent brands, labelled D1, D2, D3 and D4, to see which one most efficiently cleans clothes. Each of the four detergent brands is tested against three kinds of stain - coffee, mud, and tomato sauce - by deliberately staining a shirt and seeing how many washes of the shirt in the detergent are required before the stain can no longer be seen. Each combination of brand and stain is tested 5 times, for a total of 60 data points.

A Poisson GLM was fitted and an analysis of deviance carried out, yielding the (edited) `R` output below.

```
Model: poisson, link: log

Response: Washes

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev
NULL                         59    24.8476
Detergent [W]    [X]         [Y]       [Z]
Stain      2   2.4184         54     3.1235
```

(i) Complete the missing values for [W], [X], [Y], and [Z] in the `R` output above.

(ii) Test the model M0, which only contains an intercept term, against the full model M1 including an intercept term and both detergent brand and stain type, at the 5% level of significance. You may use without justification the fact that, if model $\tilde{M}$ is nested in model M, then the relevant test statistic is

$$sF = \frac{D(\tilde{M}, M)}{\hat{\phi}}$$

for the appropriate value of s.

Q5 The full (saturated) log-linear model for a three-way contingency table for categorical variables X, Y and Z is

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \ \forall(i,j,k).$$

5.1 Explain what constitutes a hierarchical model. State, giving a justification, whether each of the following two models are hierarchical or not:

(i) $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ijk}^{XYZ}$.

(ii) $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$.

5.2 Taking the term λ_{ij}^{XY} as your example, explain how corner point constraints are applied to a log-linear model, and how these constraints for λ_{ij}^{XY} specifically reduce the number of non-redundant parameters in the model.

5.3 Consider a $2 \times 2 \times 2$ contingency table containing 100 data points. The maximum log-likelihood is calculated for three models.

Model	**Maximum Log Likelihood**
$[X, Y, Z]$	-80.3
$[XY, Z]$	-29.4
$[XZ, YZ]$	-19.4

(i) Using a generalised likelihood ratio test at the 5% significance level, test the null hypothesis $H_0 : [X, Y, Z]$ against the alternative hypothesis $H_1 : [XZ, YZ]$.

(ii) The AIC for a model M has definition

$$AIC(M) = 2k - 2\log(\hat{L})$$

for k free parameters and likelihood function L. Use the AIC to compare the models $[XY, Z]$ and $[XZ, YZ]$.

Q6 Models generated with small data sets are extremely unreliable, but they can be useful to help understand the underlying mathematics. In this question, we use a data set where $n = 6$. There is a single predictor variable which is binary $A \in \{0, 1\}$. The response variable is also binary, $Y \in \{0, 1\}$. The full data set is given below.

i	A_i	y_i
1	0	0
2	0	0
3	0	1
4	1	1
5	1	0
6	1	1

6.1 Express this data as a 2×2 contingency table.

6.2 Assume $N_{ij} \sim \text{Pois}(\mu_{ij})$ independently for $i = 1, 2$ and $j = 1, 2$. Calculate the estimated parameter values for a log-linear model for Poisson sampling with sum-to-zero constraints for this data, under assumption $[AY]$. You may use without proof that the Lagrangian equations for such a model give us $\hat{\mu}_{ij} = n_{ij}$.

6.3 Predict the probability that, given we see one future observation, that observation has an A value of 1 and a Y value of 1. You may use without proof the result that if $N_{ij} \sim \text{Pois}(\mu_{ij})$ independently, then $N = \sum_{ij} N_{ij} \sim \text{Pois}(\sum_{ij} \mu_{ij})$.

6.4 Hence, or otherwise, find the multinomial distribution for a multinomial sampling scheme which has an identical likelihood function to the Poisson sampling scheme associated with the Poisson linear model found in **6.2**, assuming precisely N observations are seen in each case.

SECTION C

Q7 **7.1** Explain the meaning of the following notation in the context of bootstrap sampling:

(i) T.

(ii) t.

(iii) T_i^*.

7.2 A newspaper wants to start printing a new type of logic problem each day. To get a sense of how long this kind of problem might take their readers to solve, each member of the editing team tries a problem, and records how long in minutes it takes to complete. The 10 resulting scores are:

$$32.4, 29.5, 20.8, 42.8, 30.1, 50.2, 21.2, 29.4, 38.1, 40.2$$

The bootstrap will be used to make inferences about the population median time for completing a problem.

(i) Explain how, by using either dice or a uniform random number generator or `R`, you would take a bootstrap re-sample of size 10 from the data.

(ii) The following are the first two of 10000 bootstrap re-samples:

$$20.8, 20.8, 21.2, 21.2, 29.4, 29.4, 29.5, 29.5, 32.4, 40.2$$
$$29.5, 29.5, 30.1, 30.1, 32.4, 38.1, 42.8, 42.8, 42.8, 50.2$$

The first ten bootstrap statistics are:

$$b_1, b_2, 29.8, 29.75, 32.4, 29.8, 29.8, 35.25, 32.4, 35.15$$

Find the values b_1 and b_2.

(iii) The mean and standard deviation of the 10000 bootstrap statistics were respectively 32.7 and 3.81. The following table provides a number of percentiles:

Min	1%	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%	99%	Max
20.8	25.3	25.35	29.4	29.45	29.8	31.25	35.25	39.15	40.2	40.45	41.5	50.2

Showing your working, calculate 95% bootstrap confidence intervals for the population median, using: (i) the basic method; (ii) the normal approximation to the bootstrap sampling distribution.

7.3 Suppose that we want to make inferences about the population mean from a sample of size 3 and that the sampled values turn out to be: $a - b$, a, $a + b$, for some real numbers a and b.

Enumerate the possible re-samples from the non-parametric bootstrap, and find associated probabilities for each re-sample. Find the basic bootstrap confidence interval with approximately 90% nominal coverage probability for the population mean, assuming an effectively infinite resample size.

Compare this confidence interval to the normal approximation bootstrap confidence interval for the same nominal coverage probability. Comment briefly on the strengths and weaknesses of the two approaches, both in general and in this specific circumstance.