Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2025	MATH3411-WE01

Title: Advanced Statistical Modelling III

Time:	2 hours	
Additional Material provided:	Tables: Normal, t-distribution, chi-squared distribution.	
Materials Permitted:		
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Answer all questions. Section A is worth 40% and Section B is worth 60%. Within each section, all questions carry equal marks. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

Revision:	

ED01/2025

SECTION A

Q1 It is proposed to use a binary regression model, with logit link and linear predictor $\eta = \eta(z) = \beta_1 + \beta_2 z$, to model the connection between a single binary covariate $z \in \{0, 1\}$, and a binary response variable $y \in \{0, 1\}$. We observe n independent measurements corresponding to covariate values $z_1, ..., z_n$.

(a) Write down the response function and distributional assumption of the proposed logistic regression model.

(b) State the formula for the score function $S(\beta)$ of this logistic regression model, and hence solve the score equation $S(\hat{\beta}) = 0$ to show that

$$\hat{\beta}_1 = \log \frac{\bar{y}_0}{1 - \bar{y}_0} \qquad \hat{\beta}_2 = \log \frac{\bar{y}_1(1 - \bar{y}_0)}{\bar{y}_0(1 - \bar{y}_1)}$$

where $\bar{y}_b = \frac{1}{n_b} \sum_{i:z_i=b} y_i$, $n_b = \sum_{i:z_i=b} 1$, $b \in \{0, 1\}$.

A mobile phone company proposes to predict subscription cancellation. They hypothesise that certain *'interaction events'*, such as contacting customer support, downgrading their plan, or missing a payment, are indicators of potential cancellation. The company analyses a sample of 250 customers from the past year, checking whether each customer performed an interaction event over the course of that year, and whether they cancelled their subscription during that year. The data is shown in the table below.

	Cancellation?		
Interaction Event?	No(0)	Yes(1)	
No (0)	53	30	$n_1 = 83$
Yes (1)	28	139	$n_2 = 167$
Total	81	169	$n = 250$

Suppose the connection between interaction events and cancellation is modelled by a binary regression model, with logit link and linear predictor $\eta = \eta(z) = \beta_1 + \beta_2 z$, where covariate $z \in \{0, 1\}$ takes the value 1 if an interaction event occurred, and 0 otherwise.

(c) Calculate $\hat{\beta}_1$ and $\hat{\beta}_2$ in this case.

(d) Use this model to predict the probability that a customer that has not performed any interaction event in 2024 also cancelled their subscription during 2024.

Q2 We consider data on the number of deaths from coronary heart disease in 8 different age groups for men from 30 to 69 years old in a region of Australia. The dataset contains the following variables.

age	An integer variable representing the age groups: 1 (30-34 years), 2 (35-39 years), ..., 8 (65-69 years);
pop	The population size (i.e., the number of men in each age group);
death	The number of deaths in each age group.

The dataset is given in the table below.

age	1	2	3	4	5	6	7	8
pop	17742	16554	16059	13083	10784	9645	10706	9933
death	1	5	5	12	25	38	54	65

Below is the R code to fit a GLM to this dataset and the (reduced) model summary.

```
> model <- glm(death ~ age, offset = log(pop), family = poisson(),
               data = heart)
> summary(model)

Call:
glm(formula = death ~ age, family = poisson(), data = heart,
    offset = log(pop))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.01688    0.26188  -34.43   <2e-16 ***
age          0.52219    0.03905   13.37   <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: 257.51  on 7  degrees of freedom
Residual deviance:  14.69  on 6  degrees of freedom
```

(a) Write down the formulation for this GLM by specifying the linear predictor, the link function, the model parameters, and the distributional assumption.

(b) Provide an estimate of the dispersion present in the data.

(c) Using the estimated dispersion in part (b), perform a hypothesis test to determine whether we should include the age variable in the model.

(d) Show that if we fit the model above to any ungrouped dataset $\{(\boldsymbol{x}_i, y_i)\}_{i\in\{1,\ldots,n\}}$, then $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\mu}_i$, where $\hat{\mu}_i$ is the estimated mean of data point i. (Hint: Consider the row of the score equation that corresponds to the intercept component.)

SECTION B

Q3 Suppose a hospital is interested in knowing whether patient opinion about treatment satisfaction (Y: *Satisfied, Unsatisfied*) is associated with severity of condition (X: *Mild, Moderate, Severe*) and department visited (Z: *Emergency, Outpatient*). They therefore cross-classify these variables for a sample of patients in the following contingency table:

		Treatment Satisfaction (Y)	
Department (Z)	Severity (X)	Satisfied	Unsatisfied
Emergency	Mild	2	6
	Moderate	38	13
	Severe	20	28
Outpatient	Mild	11	13
	Moderate	33	13
	Severe	1	5

(a) (i) Calculate the marginal YZ contingency table of the observed counts.

(ii) Calculate an estimate of the marginal odds ratio comparing Treatment Satisfaction between Departments.

(iii) What does the estimated odds ratio tell us about the relation between Satisfaction and Department?

(b) Calculate a minimal set of marginal (over Department) global odds ratios between Severity and Satisfaction, along with a reasonable 95% confidence interval for each. What assumption about variable Severity is being made in order to do this? Interpret and explain the results in terms of associations between Severity and Satisfaction.

(c) (i) Write down an appropriate log-linear model expression assuming conditional independence of Department and Treatment Satisfaction given Severity.

(ii) Assuming sum-to-zero constraints and Poisson sampling, calculate and explain the number of free parameters in this model.

(iii) By rearranging the log-linear model expression, derive expressions for the following effect parameters in terms of expected counts, assuming sum-to-zero constraints:

- the constant effect, denoted λ in lectures,
- the main effect for Y, denoted λ_j^Y in lectures, and
- the interaction effect of X and Z, denoted λ_{ik}^{XZ} in lectures.

Explain the meaning of the parameters in relation to the expected counts.

Q4 We consider data on the reaction time of 18 subjects in a sleep deprivation study. On day 0, the subjects had their normal amount of sleep. Starting that night, they were restricted to 3 hours of sleep per night. The observations represent the average reaction time on a series of tests given each day to each subject. The variables in the dataset are given in the following table.

Subject	The subject ID;
Days	Number of days of sleep deprivation. This is an integer variable taking values from 0 to 9;
Reaction	The response variable, which is the average reaction time (ms) for each day.

The table below, which has already been sorted by subject IDs, shows part of the data. The whole dataset contains 180 observations, with 10 observations for each subject.

Subject	1	1	...	1	2	...	2	...	18
Days	0	1	...	9	0	...	9	...	9
Reaction	249.56	258.71	...	466.35	222.73	...	237.31	...	364.12

A linear mixed model is fitted to the data using the R code below.

```
> require(lme4)
> model <- lmer(Reaction ~ Days + (Days | Subject), data=sleepstudy)
```

(a) Let x_{ij} be the number of days of sleep deprivation for subject i on day j and let y_{ij} be the average reaction time of this subject on the same day. For the model fitted above, write the expression for y_{ij} in terms of x_{ij} and the model parameters. Clearly specify the parameters and all distributional assumptions.

(b) Recall that a linear mixed model can be written in the general form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$. Rewrite your formulation in part (a) into this form, clearly specifying $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{Z}$, $\boldsymbol{u}$, $\boldsymbol{\epsilon}$ and their dimensions.

(c) The summary of the fitted model is given on the next page. What are the estimated values of the model parameters and the estimated correlation between the random effects for the same subject?

(d) In this part, assume there is no correlation between the random effects for the same subject. For any i, j, k with $j \neq k$, derive an expression for $\mathrm{Corr}(y_{ij}, y_{ik})$, the correlation between y_{ij} and y_{ik}, in terms of x_{ij}, x_{ik}, and the variances of the random effects.

(e) Using your result in part (d), show that for any subject i and any days $j < k$, $\mathrm{Corr}(y_{i0}, y_{ik}) < \mathrm{Corr}(y_{i0}, y_{ij})$.

[Question 4 continues on the next page]

```
> model
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy
REML criterion at convergence: 1743.628
Random effects:
 Groups   Name        Std.Dev. Corr
 Subject  (Intercept) 24.741
          Days         5.922   0.07
 Residual             25.592
Number of obs: 180, groups:  Subject, 18
Fixed Effects:
(Intercept)         Days
     251.41        10.47
```