Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2026	MATH3411-WE01

Title: Advanced Statistical Modelling III

Time:	2 hours	
Additional Material provided:	Tables: Normal, t-distribution, chi-squared distribution.	
Materials Permitted:	None	
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Answer all questions. The indicative marks shown in brackets for the main parts of each question are given as a guide to the weighting the markers expect to apply. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

Revision:	

ED01/2026

SECTION A

1. (a) Define the form of the probability density/mass function corresponding to members of the exponential dispersion family of distributions. Explain, where appropriate, any parameters used. **[2]**

(b) The negative binomial distribution can be used to represent the probability of observing y "failures" before a fixed number n of "successes" occur. The probability mass function is given by

$$p(y|\pi) = \binom{y+n-1}{y} \pi^n (1-\pi)^y$$

where $\pi \in [0,1]$, y is a non-negative integer, and n is constant (assumed fixed and known).

Show that this probability distribution is a member of the exponential dispersion family of distributions, being sure to identify all the constituent parameters defined in part (a). **[3]**

(c) Using only properties of the exponential dispersion family, derive the mean and variance of the negative binomial distribution (with fixed parameter n). Assuming that $n = 1$, finding that $\mathrm{E}[Y] = 3$ corresponds to what value of π? **[3]**

(d) Hence, what is the natural link function when using a negative binomially distributed response in a generalised linear model? **[2]**

---

2. In a study of n subjects, for each subject $i = 1, \ldots, n$, there are m_i independent binary observations $\{y_{i1}, y_{i2}, \ldots, y_{im_i}\}$, with $y_{ir} \in \{0, 1\}$. Assume we group the data into the form $\{(\boldsymbol{x}_i, y_i)\}_{i\in\{1,\ldots,n\}}$, where $y_i = \frac{1}{m_i}\sum_{r=1}^{m_i} y_{ir}$. To model this grouped dataset, we use a generalised linear model with a rescaled binomial response and the logit link function g where:

$$y_i \sim \frac{1}{m_i}\text{Bin}(m_i, \mu_i) \quad \text{and} \quad g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right).$$

(a) Derive an expression for the log-likelihood of the data in terms of y_i, m_i, μ_i, $i = 1, \ldots, n$. **[2]**

(b) Derive the log-likelihood of the saturated model. **[2]**

(c) Instead of grouping the data, we can consider ungrouped data of the form $\{(\boldsymbol{x}_i, y_{ir})\}_{i\in\{1,\ldots,n\}, r\in\{1,\ldots,m_i\}}$ and fit a generalised linear model using the same link function but with a Bernoulli response where:

$$y_{ir} \sim \text{Bernoulli}(\mu_{ir}).$$

Derive an expression for the log-likelihood of the data in this case and compare with the one derived in part (a). **[3]**

(d) Show that the log-likelihood of the saturated model for part (c) could be different from the one obtained in part (b). **[3]**

SECTION B

3. The Council of Galactic Gardeners is testing cosmic seeds for growing space crops on distant planets. These seeds are crucial for terraforming projects and ensuring food security across the galaxy. The gardeners classify a sample of seeds according to Growth Success (Y: *Yes* or *No*), Seed Type (Z: *Alpha* or *Beta*), and Soil Type (X: *Rocky* or *Sandy*). The dataset is presented in the following table:

Table 1: Dataset.

Soil Type (X)	Growth Success (Y)	Seed Type (Z)	counts
Rocky	Yes	Alpha	21
Rocky	No	Alpha	3
Sandy	Yes	Alpha	55
Sandy	No	Alpha	13
Rocky	Yes	Beta	31
Rocky	No	Beta	43
Sandy	Yes	Beta	7
Sandy	No	Beta	18

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

(ii) Calculate an estimate of, along with a 90% confidence interval for, the marginal odds ratio comparing Growth Success between Soil Types.

(iii) Based on this, infer whether there is evidence that Soil Type and Growth Success are dependent or not.

(iv) What does the estimated odds ratio tell us about the relation between Soil Type and Growth Success? **[5]**

(b) Based on the results of Part (a), the gardeners are convinced that the dataset provides at least some evidence that, regardless of Seed Type, seeds in Sandy soil have a greater chance of successfully growing than seeds in Rocky soil. Perform reasonable analyses to:

(i) illustrate to the gardeners why they have jumped to the wrong conclusion, and

(ii) discuss alternative inferences. **[5]**

Credit will be awarded for the clarity of your answer.

[Question 3 continues on the next page]

(c) (i) Write down an appropriate log-linear model expression assuming homogeneous associations between Soil Type, Seed Type and Growth Success.

(ii) Assuming corner-point 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 corner-point constraints:

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

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

(iv) What does the assumption of homogeneous associations imply about the conditional (on Soil Type) odds ratio between Seed Type and Growth Success? **[5]**

---

4. We consider data from a study of the efficacy of two programmes for discouraging young people from smoking. The subjects of the study were 1,600 students from 28 Los Angeles schools. The dataset contains the following variables:

school	The school ID of the student;
SC	A binary variable indicating whether the student was exposed to a school-based curriculum programme (1 = yes, 0 = no);
TV	A binary variable indicating whether the student was exposed to a television-based prevention programme (1 = yes, 0 = no);
PTHK	An integer variable taking values between 0 and 7, indicating the student's level of tobacco and health knowledge before the study (larger values indicate a higher level of knowledge);
y	The response variable, which also takes integer values between 0 and 7, indicating the student's level of tobacco and health knowledge after the study (larger values indicate a higher level of knowledge).

The study compared four groups designed according to the combinations of the SC and TV values, namely, (SC = 0, TV = 0), (SC = 0, TV = 1), (SC = 1, TV = 0) and (SC = 1, TV = 1). Each school was randomly assigned to one of the four groups for intervention. The table below, which has already been sorted by school IDs, shows part of the data.

school	SC	TV	PTHK	y
403	1	0	2	3
403	1	0	4	4
...				
515	0	0	3	2
515	0	0	3	3

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

```
> require(lme4)
> model <- lmer(y ~ SC + TV + PTHK + (1|school), data=Smoking)
```

(a) Let y_{ij} be the response value for student j in school i. Similarly, let SC_{ij}, TV_{ij}, and PTHK_{ij} be the SC, TV, and PTHK values for the same student. For the model fitted above, write the expression for y_{ij} in terms of SC_{ij}, TV_{ij}, PTHK_{ij}, and the model parameters by clearly identifying fixed effects, random effects and the error term. Clearly specify the set of parameters and state all distributional assumptions. **[3]**

(b) Derive the marginal mean, $\text{E}(y_{ij})$, the marginal variance, $\text{Var}(y_{ij})$, and the marginal covariance, $\text{Cov}(y_{ij}, y_{ik})$, where $j \neq k$. **[3]**

(c) 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. **[3]**

[Question 4 continues on the next page]

(d) The summary of the fitted model is given below. What are the estimated values of the model parameters? Using these values, compare the effects of the school-based curriculum and television-based prevention programmes on the tobacco and health knowledge of the students after the study. **[3]**

```
> model
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ SC + TV + PTHK + (1 | school)
   Data: Smoking
REML criterion at convergence: 5383.938
Random effects:
 Groups   Name        Std.Dev.
 school   (Intercept) 0.2315
 Residual             1.2861
Number of obs: 1600, groups:  school, 28
Fixed Effects:
(Intercept)           SC           TV         PTHK
    1.79268      0.46984      0.02036      0.30848
```

(e) Using the marginal variance and covariance derived in part (b), and the summary of the fitted model above, compute the intra-class correlation $\text{Corr}(y_{ij}, y_{ik})$ between responses of two students in the same school. Interpret this value. **[3]**
