Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2023	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. Students must use the mathematics specific answer book.

Revision:	

ED01/2023

SECTION A

Q1 (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.

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

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

where $\pi \in [0,1]$, y is a non-negative integer, and k 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).

(c) Using only properties of the exponential dispersion family, derive the mean and variance of the negative binomial distribution (with fixed parameter k).

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

Q2 We are given diagnostic measurements, y_{ij}, taken from two individuals $i = 1, 2$ at time points t_j, $j = 1, \ldots, 4$, where the individual indexed $i = 2$ was unavailable at time point t_4. That is, a total of seven measurements are available. After the measurement at time t_2 was taken from each individual, a clinical intervention was applied to both individuals, the effect of which one is interested in. Naïvely, one could describe this problem by a linear model

$$y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 \mathbf{1}\{j > 2\} + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and $\mathbf{1}\{\cdot\}$ denotes the indicator function.

(a) Write this model in the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, giving explicit expressions for $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$.

(b) Explain why this model is likely to lead to incorrect inferences.

(c) One approach to improve the inference is to apply a marginal model. Inferences for such a model can be produced through generalized estimating equations $\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = 0$, where $\boldsymbol{\mu}$ is the vector of marginal means according to the model. Provide explicit expressions for the 7×7 matrix $\boldsymbol{\Sigma}$, assuming independence of measurements between individuals, and

(i) an exchangeable correlation structure;

(ii) an AR(1) correlation structure

within individuals.

(d) A second approach to deal with this problem is to set up a mixed model of type $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$, with random effects $\boldsymbol{u} = (u_1, u_2)^T$. Give an explicit expression for $\boldsymbol{Z}$ in the context of our modelling problem, and suggest an appropriate distributional assumption for $\boldsymbol{u}$. Then find an expression for $\boldsymbol{V} := \mathrm{Var}(\boldsymbol{Y})$ which only depends on $\boldsymbol{Z}$ and the variance parameters of $\boldsymbol{\epsilon}$ and $\boldsymbol{u}$.

(e) Can you identify any choice of correlation structures in part (c), and any choice of distributional assumption for $\boldsymbol{u}$ in part (d), for which the matrix $\boldsymbol{\Sigma}$ corresponds to $\boldsymbol{V}$?

(f) In the introduction to this question it was mentioned that observation y_{24} is not available. Does this cause any complications (in whatever respect) for the analysis?

SECTION B

Q3 Suppose a horde of 216 aliens (beings from a far-off planet) arrive in Durham and are surveyed about the suitability of building structures were they to make Durham their home (they need to report back to the rest of their species). Suppose the rest of their kind is interested in knowing whether the opinion about Suitability (Suitable, Unsuitable) is associated with Height (Short, Average, Tall) and Type (A,B). They therefore cross-classify these variables in the following contingency table:

		Suitability (Y)	
Type (Z)	Height (X)	Suitable	Unsuitable
A	Short	12	5
	Average	34	13
	Tall	25	22
B	Short	12	12
	Average	60	3
	Tall	14	4

3.1 Calculate the marginal YZ contingency table of the observed counts. Calculate an estimate of, along with a 95% confidence interval for, the marginal odds ratio of Suitability and Type. Based on this, infer whether Suitability and Type are dependent or not. What does the estimated odds ratio tell us about the relation between Suitability and Type?

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

3.3 Write down an appropriate log-linear model expression assuming homogeneous associations between Height, Type and Suitability. Assuming corner-point constraints and Poisson sampling, calculate and explain the number of free parameters in this model. Rearrange the log-linear model expression to obtain expressions for λ, λ_i^X and λ_{ik}^{XZ} assuming corner-point constraints. Explain the meaning of these parameters in relation to the expected counts. What does the assumption of homogeneous associations imply about the conditional (on Height) odds ratio between Type and Suitability?

[Question 3 continues on the next page]

3.4 Perform a suitable statistical test to assess whether Suitability and Type are conditionally independent given Height. Justify the choice of test. In order to assist you, the log-likelihood values for each possible hierarchical model of the three variables are provided below.

Model	Log Likelihood
$[X, Y, Z]$	-51.876
$[Z, XY]$	-42.324
$[Y, XZ]$	-44.405
$[X, YZ]$	-48.338
$[XY, XZ]$	-34.936
$[XY, YZ]$	-38.869
$[XZ, YZ]$	-39.950
$[XY, XZ, YZ]$	-31.898
$[XYZ]$	-26.478

Q4 For a generalized linear model with **Poisson**–distributed response, the log-likelihood function can be written as

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(-\mu_i + y_i \log \mu_i - \log(y_i!)\right)$$

where $\mu_i = h(\boldsymbol{x}_i^T \boldsymbol{\beta})$ are the mean functions for case $i = 1, \ldots, n$. Denote $D = \sum_{i=1}^{n} d_i = 2(L_{\text{sat}} - L(\hat{\boldsymbol{\beta}}))$ the deviance of the fitted model, where L_{sat} is the saturated log-likelihood and d_i are the deviance components. We assume that the data have not been grouped.

4.1 Explain the concept of a saturated log-likelihood, and provide its explicit expression for the situation outlined above.

4.2 Show that the deviance components can be written as

$$d_i = 2\left\{ y_i \log \frac{y_i}{\hat{\mu}_i} - y_i + \hat{\mu}_i \right\}.$$

Suggest a way of dealing with the case $y_i = 0$.

4.3 In the special case that $h(\cdot) = \exp(\cdot)$ and that the model contains an intercept, show that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\mu}_i$, and use this result to give a simplified expression for the deviance.

Hint: Equate the first row of the score-function to zero.

4.4 We are given data featuring $n = 32$ observations on faults in rolls of fabric. The only covariate z gives the logarithm of the length of the roll, and the response y gives the number of faults in the roll of fabric. We estimate a Poisson regression model

$$y_i | z_i \sim \text{Poi}\left(\exp\{\beta_1 + \beta_2 z_i\}\right)$$

yielding $\hat{\beta}_1 = -4.173$ and $\hat{\beta}_2 = 0.997$, with a deviance of 64.53. **On the next page** we provide a plot of the deviance residuals r_i^D versus fitted values $\hat{y}_i$, from which one observation, corresponding to the data point $(z_1, y_1) = (6.311, 6)$, was omitted.

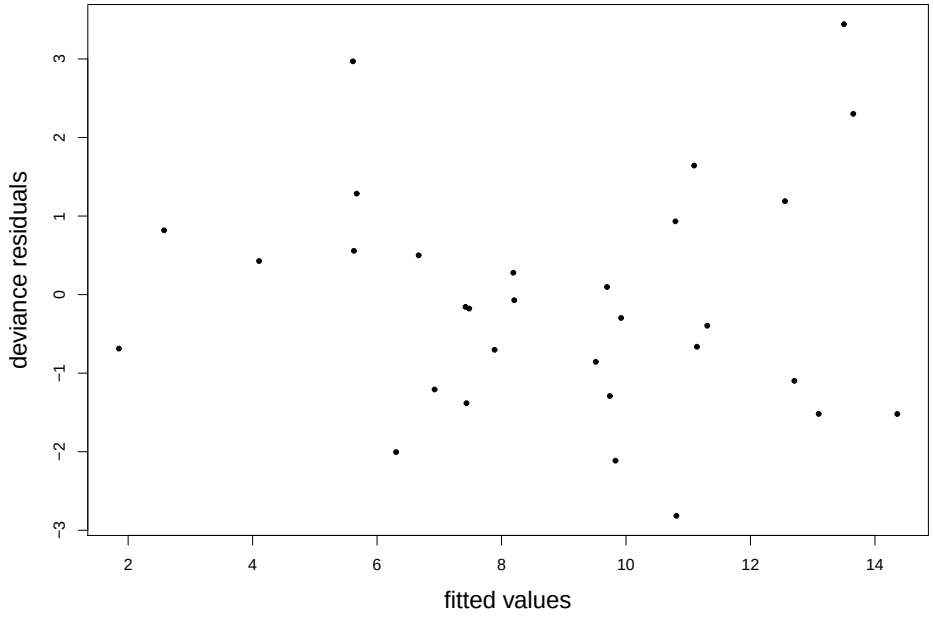
(i) Compute the missing value $(\hat{y}_1, r_1^D)$.

(ii) Give an interpretation of the residual plot.

(iii) Provide a rough estimate of the dispersion present in the data.

(iv) Carry out a test of goodness-of-fit of the Poisson model at the 5% level of significance, discussing any distributional assumption made.

[Question 4 continues on the next page]



END