Imperial College
London

M3S2

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May-June 2017

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science

Statistical Modelling II

Date: Friday 12 May 2017

Time: 10:00 - 12:00

Time Allowed: 2 Hours

This paper has 4 Questions.

Candidates should use ONE main answer book.

Supplementary books may only be used after the relevant main book(s) are full.

All required additional material will be provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Credit will be given for all questions attempted, but extra credit will be given for complete or nearly complete answers to each question as per the table below.

Raw Mark	Up to 12	13	14	15	16	17	18	19	20
Extra Credit	0	½	1	1 ½	2	2 ½	3	3 ½	4

- Each question carries equal weight.
- Calculators may not be used.

© 2017 Imperial College London

1. (a) Give the specification of a Generalized Linear Model (GLM) for n observations.

(b) Suppose we have a n-dimensional observation vector $\boldsymbol{Y}$ such that

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where X is a $n \times p$ design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$, where $\sigma^2 > 0$ is unknown. You may assume that the design matrix X has full rank.

(i) State why this model is a GLM.

(ii) State $\widehat{\boldsymbol{\beta}}$, the maximum likelihood estimator of $\boldsymbol{\beta}$. Further, state the covariance of $\widehat{\boldsymbol{\beta}}$.

(iii) Let $S(\boldsymbol{\beta}) = (\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta})$. Show that $S(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y}^T(I_n - P)\boldsymbol{Y}$, where P is a matrix to be defined.

(iv) Consider the particular case where

$$Y_i = \beta_1 + \beta_2 a_i + \beta_3 b_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\sum_{i=1}^n a_i = 0$, $\sum_{i=1}^n b_i = 0$.
Show that

$$P = \frac{1}{n}\mathbf{1}\mathbf{1}^T + \gamma_1 \boldsymbol{a}\boldsymbol{a}^T + \gamma_2(\boldsymbol{a}\boldsymbol{b}^T + \boldsymbol{b}\boldsymbol{a}^T) + \gamma_3 \boldsymbol{b}\boldsymbol{b}^T,$$

where, $\mathbf{1}$ is a column vector of ones, and γ_1, γ_2, γ_3, are constants to be found.

(v) Hence show that, if $\widehat{\boldsymbol{Y}} = X\widehat{\boldsymbol{\beta}}$ is the vector of fitted values, then

$$\frac{1}{\sigma^2}\text{var}(\widehat{Y}_i) = \frac{1}{n} + \gamma_1 a_i^2 + 2\gamma_2 a_i b_i + \gamma_3 b_i^2, \quad i = 1, \ldots, n.$$

2. Consider the independent random variables $Y_1, \ldots, Y_n$ where $Y_i \sim$ Binomial(n_i, π_i), $n_i \in \mathbb{Z}$, $0 \leq \pi_i \leq 1$. The probability mass function for Y_i is

$$P(Y_i = y_i; n_i, \pi_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i},$$

where $y_i = 0, 1, \ldots, n_i$. Consider a Binomial Generalized Linear Model (GLM) with canonical link function.

(a) Show that the Binomial distribution is a member of the exponential family. State the canonical link function and the variance function.

(b) Suppose that

$$\log\left(\frac{\mu_i}{n_i - \mu_i}\right) = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$$

where x_i are non-random covariates.

(i) Derive the log-likelihood for the model in terms of the components of $\boldsymbol{\beta}$.

(ii) Based on the log-likelihood derived in (i), show that

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \widehat{\mu}_i, \quad \sum_{i=1}^n y_i x_i = \sum_{i=1}^n \widehat{\mu}_i x_i, \quad \text{and} \quad \sum_{i=1}^n y_i x_i^2 = \sum_{i=1}^n \widehat{\mu}_i x_i^2$$

where $\widehat{\mu}_i$ are the fitted means of the responses based on the maximum likelihood estimates $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$.

[Question 2 continues on the next page.]

(c) Recall the birth-weight data set used in this course. This dataset contains information regarding 189 births at a US hospital. Here is a printout of a random 10 lines from the (reduced) dataset:

low	age	race	ftv
0	24	115	2+
0	23	128	0
0	22	158	2+
0	25	140	1
0	22	169	0
0	26	160	0
1	19	91	0
1	16	130	1
1	25	92	0
1	21	100	2+

The definitions of the variables in this dataset are shown in Table 1.

Table 1: Variables in birth-weight dataset.

Variable	Description	Remarks
`low`	birth weight less than 2.5 kg	0 or 1 reported
`age`	age of mother in years	None
`lwt`	weight of mother in pounds	None
`ftv`	number of physician visits in the first trimester	0, 1 , 2+ reported

[Question 2 continues on the next page.]

A Binomial GLM was fitted to the data using the following command:

```
myglm <- glm(low~lwt+ftv+age,family=binomial)
```

The R summary for this model (shortened and redacted) is given below:

```
Coefficients:
             Estimate Std. Error s value   ??????
(Intercept)  1.722745   0.997434   1.727   0.0841
lwt         -0.012898   0.006191  -2.083   0.0372
ftv1        -0.533598   0.415797  -1.283   0.1994
ftv2+       -0.180336   0.416657  -0.433   0.6651
age         -0.030864   0.033529  -0.921   0.3573
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

(i) Explain why there is no `ftv0` parameter in the summary. How is this parameter accounted for in the model?

(ii) In the R summary above there is a column marked by "??????". State how the values in this column are computed using the other information in the R summary.

(iii) Suppose we wish to assess whether the average number of low birth weights differs between the `ftv` groups `ftv0` years and `ftv2+` years. Explain how this question can be answered at a 5% significance level.

(iv) When comparing nested Binomial GLMs, which test should be used? Explain your reasoning.

3. Suppose Y is Exponential distributed with probability density function

$$P(Y = y; \lambda) = \frac{1}{\lambda} \exp\left(-\frac{y}{\lambda}\right) \quad \lambda > 0, \ y \geq 0.$$

(a) Show that Y is a member of the exponential family. Further, find $\mathrm{E}(Y)$ and var(Y). You may use results from lectures without proof, but must state any that you use.

(b) Consider a GLM with response variables $Y_1, \ldots, Y_n$ where $Y_i \sim$ Exponential(λ_i), canonical link function and given linear predictor.

(i) Find the maximal achievable value for the log-likelihood for this model.

(ii) Derive the deviance for this model in terms of the fitted responses, $\widehat{\mu}_1, \ldots, \widehat{\mu}_n$.

(c) Suppose an Exponential GLM is fitted with linear predictor

$$\eta_i = \beta_1 + \beta_2 x_i, \quad i = 1, \ldots, n.$$

and the log link function: $g(u) = \log(u)$. It is known that $\sum_{i=1}^n x_i = 2$ and $\sum_{i=1}^n x_i^2 = 6$.

(i) Define the adjusted dependent variable, z_i and weights w_{ii} used in the Iterative Weighted Least Squares algorithm for this particular GLM.

(ii) Compute the estimated covariance matrix for $\widehat{\boldsymbol{\beta}}$. Comment on the result.

(iii) Suppose the estimated linear parameter is

$$\widehat{\boldsymbol{\beta}} = (0.2, 0.17)^T.$$

Give a prediction for the mean response for an observation with $x = 10$.

4. Consider the balanced one-way random effects model

$$Y_{ij} = \mu + \nu_j + \epsilon_{ij}, \quad \text{for } j = 1, \dots, m; \ i = 1, \dots, K,$$

where μ is the fixed effect, $\nu_j \sim N(0, \sigma_\nu^2)$ are the random effects and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The ν_j and ϵ_{ij} are all independent.

(a) Let

$$\overline{Y}_{\bullet j} = \frac{1}{m} \sum_{i=1}^{K} Y_{ij} \quad \text{and} \quad \overline{Y} = \frac{1}{mK} \sum_{i=1}^{K} \sum_{j=1}^{m} Y_{ij}.$$

State the distribution of $\overline{Y}_{\bullet j}$ and $\overline{Y}$.

(b) (i) Prove that the correlation between two observations in the same group is

$$\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\epsilon^2}$$

(ii) Suppose we have a 95% confidence interval, (c_1, c_2), for $\sigma_\nu^2 / \sigma_\epsilon^2$. Derive a confidence interval for ρ.

(c) Recall that in the derivation of the ANOVA estimators for the variance-components, σ_ϵ^2 and σ_ν^2, we introduced the definitions

$$SSE := \sum_{i=1}^{K} \sum_{j=1}^{m} (Y_{ij} - \overline{Y}_{\bullet j})^2 \quad \text{and} \quad SSA := \sum_{i=1}^{K} \sum_{j=1}^{m} (\overline{Y}_{\bullet j} - \overline{Y})^2,$$

where

$$\mathrm{E}(SSE) = m(K-1)\sigma_\epsilon^2 \quad \text{and} \quad \mathrm{E}(SSA) = (m-1)(K\sigma_\nu^2 + \sigma_\epsilon^2)$$

(i) It can be shown that

$$\frac{SSE}{\sigma_\epsilon^2} \sim \chi_{d_1}^2 \quad \text{and} \quad \frac{SSA}{K\sigma_\nu^2 + \sigma_\epsilon^2} \sim \chi_{d_2}^2$$

independently. What are the values of d_1 and d_2?

(ii) Consider testing the hypothesis

$$H_0 : \sigma_\nu^2 = 0 \quad \text{vs} \quad H_1 : \sigma_\nu^2 > 0$$

Using part (i), carefully suggest a test statistic for this hypothesis test and any related distribution result. Using the observed responses, $\boldsymbol{y}$, explain how to carry out the hypothesis test at a 5% significance level.

5. **Mastery Question** This mastery question is based on the material referred to in

Abbaszadeh, Rouzbeh, et al. "Evaluation of watermelons texture using their vibration responses." Biosystems engineering 115.1 (2013): 102-105. APA

and

Wasserman, L. (2013). All of Statistics: A Concise Course in Statistical Inference. Springer. pages 165 - 167.

(a) Summarise the objective and conclusion made in the Watermelon paper.

(b) Comment on the subjective nature of the data.

(c) Discuss what the authors mean when they say a "multiple linear regression" method was used. What is the response and predictors in this model?

(d) Suppose the number of different frequencies used was 1500. A stepwise method is used to reduce the number of variables. This stepwise method, known as forward selection, works as follows:

Start with no variables, just an intercept term. At each step, each variable (frequency level) that is not already in the model is tested for inclusion in the model. The most significant of these variables is added to the model, as long as its p-value is below a pre-specified value.

State and describe how this stepwise method is related to multiple testing. Provide example calculations illustrating the problems.

You may use the following estimate to support your answer:

$$0.95^{1500} \approx 4 \times 10^{-34}$$

(e) Consider the following simplified problem; testing 5 independent hypothesis with corresponding p-values:

$$P_1 = 0.1, \quad P_2 = 0.001, \quad P_3 = 0.05, \quad P_4 = 0.025, \quad P_5 = 0.15.$$

Conduct a procedure, specifying which hypothesis are rejected, such that the false discovery rate is controlled below 5%.

Imperial College
London

IMPERIAL COLLEGE LONDON

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2017

This paper is also taken for the relevant examination for the Associateship.

M3S2/M4S2

Statistical Modelling II (Solutions)

Setter's signature	Checker's signature	Editor's signature
....	....	....

1. a) The components of a GLM are $\boxed{\text{seen} \Downarrow}$

   $*$ The components of $\boldsymbol{Y}$ are independent and have the same distribution. This distribution is a member of the exponential family with $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$. $\boxed{1}$

   $*$ The linear predictor

$$\boldsymbol{\eta} = X\boldsymbol{\beta}$$

   where X is a design matrix (non-random) and $\boldsymbol{\beta}$ is the unknown linear parameter vector. $\boxed{1}$

   $*$ The link between the random and systematic components is

$$\eta_i = g(\mu_i) \quad \text{for} \quad i = 1, \dots, n.$$

   where g is a differentiable and monotonic function. $\boxed{1}$

b) i) A Normal Linear Model is a GLM since;

   · The observations $Y_1, \dots, Y_n$ are independent are normally distributed. The normal distribution is a member of the exponential family. $\boxed{1}$

   · It has a linear predictor of the form

$$\boldsymbol{\eta} = X\boldsymbol{\beta}$$

$\boxed{1}$

   · The link function is the identity function $g(u) = u$ which is clearly differentiable and monotonic. $\boxed{1}$

ii) The maximum likelihood estimator is:

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \boldsymbol{Y}.$$

The covariance of $\boldsymbol{\beta}$ is $\boxed{1}$

$$\operatorname{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}$$

$\boxed{1}$

iii)

$$\begin{aligned}
S(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{Y} - X\widehat{\boldsymbol{\beta}})^T (\boldsymbol{Y} - X\widehat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{Y} - X\left(X^T X\right)^{-1} X^T \boldsymbol{Y})^T (\boldsymbol{Y} - X\left(X^T X\right)^{-1} X^T \boldsymbol{Y}) \\
&= \boldsymbol{Y}^T (I_n - X\left(X^T X\right)^{-1} X^T)^T (I_n - X\left(X^T X\right)^{-1} X^T) \boldsymbol{Y} \\
&= \boldsymbol{Y}^T (I_n - P)^T (I_n - P) \boldsymbol{Y}
\end{aligned}$$

where $P = X\left(X^T X\right)^{-1} X^T$. Then $\boxed{1}$

$$(I_n - P)^T (I_n - P) = I_n - P^T - P + P^T P = I_n - P$$

since clearly $P^T = P$ and $P^T P = P$. Thus concluding

$$S(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y}^T (I_n - P) \boldsymbol{Y}$$

$\boxed{2}$

iv) For this part we need to compute $P = X\left(X^TX\right)^{-1}X^T$. The design matrix for this particular model is

$$X = \begin{pmatrix} 1 & a_1 & b_1 \\ 1 & a_2 & b_2 \\ \vdots & \vdots & \vdots \\ 1 & a_n & b_n \end{pmatrix}.$$

Then

$$X^TX = \begin{pmatrix} n & \sum_{i=1}^n a_i & \sum_{i=1}^n b_i \\ \sum_{i=1}^n a_i & \sum_{i=1}^n a_i^2 & \sum_{i=1}^n a_ib_i \\ \sum_{i=1}^n b_i & \sum_{i=1}^n a_ib_i & \sum_{i=1}^n b_i^2 \end{pmatrix} = \begin{pmatrix} n & 0 & 0 \\ 0 & \boldsymbol{a}^T\boldsymbol{a} & \boldsymbol{a}^T\boldsymbol{b} \\ 0 & \boldsymbol{a}^T\boldsymbol{b} & \boldsymbol{b}^T\boldsymbol{b} \end{pmatrix}.$$

and

$$(X^TX)^{-1} = \begin{pmatrix} 1/n & 0 & 0 \\ 0 & \alpha\boldsymbol{b}^T\boldsymbol{b} & -\alpha\boldsymbol{a}^T\boldsymbol{b} \\ 0 & -\alpha\boldsymbol{a}^T\boldsymbol{b} & \alpha\boldsymbol{a}^T\boldsymbol{a} \end{pmatrix},$$

where $\alpha = (\boldsymbol{a}^T\boldsymbol{a}\boldsymbol{b}^T\boldsymbol{b} - (\boldsymbol{a}^T\boldsymbol{b})^2)^{-1}$. Then $\boxed{4}$

$$P_{ij} = 1/n + \alpha a_j(a_i\boldsymbol{b}^T\boldsymbol{b} - b_i\boldsymbol{a}^T\boldsymbol{b}) + \alpha b_j(b_i\boldsymbol{a}^T\boldsymbol{a} - a_i\boldsymbol{a}^T\boldsymbol{b})$$

Thus we get

$$\gamma_1 = \alpha\boldsymbol{b}^T\boldsymbol{b}, \quad \gamma_2 = -\alpha\boldsymbol{a}^T\boldsymbol{b}, \quad \gamma_3 = \alpha\boldsymbol{a}^T\boldsymbol{a}.$$

$\boxed{3}$

v) First

$$\mathsf{cov}(\widehat{\boldsymbol{Y}}) = \mathsf{cov}(X\widehat{\boldsymbol{\beta}}) = X\mathsf{cov}(\widehat{\boldsymbol{\beta}})X^T = \sigma^2X(X^TX)^{-1}X^T = \sigma^2P$$

where we used the result stated in part b ii). Thus

$$\begin{aligned} \frac{1}{\sigma^2}\operatorname{var}(\widehat{Y}_i) &= i^{\text{th}} \text{ diagonal entry of } P \\ &= 1/n + \alpha a_i(a_i\boldsymbol{b}^T\boldsymbol{b} - b_i\boldsymbol{a}^T\boldsymbol{b}) + \alpha b_i(b_i\boldsymbol{a}^T\boldsymbol{a} - a_i\boldsymbol{a}^T\boldsymbol{b}) \\ &= 1/n + (\alpha\boldsymbol{b}^T\boldsymbol{b})a_i^2 + (\alpha\boldsymbol{a}^T\boldsymbol{a})b_i^2 + 2(-\alpha\boldsymbol{a}^T\boldsymbol{b})a_ib_i. \end{aligned}$$

$\boxed{2}$

$\boxed{\text{seen} \Downarrow}$

2. a) The Binomial probability mass function can be written as

$$\exp\left\{y\log\left(\frac{\pi}{1-\pi}\right)+n\log(1-\pi)+\log\binom{n}{y}\right\}.$$

We identify:

$$\theta=\log\left(\frac{\pi}{1-\pi}\right)$$

$\boxed{1}$

$$\begin{aligned}b(\theta)&=-n\log(1-\pi)\\&=n\log(1+e^{\theta})\end{aligned}$$

Must be written in term of θ to gain mark. $\boxed{1}$

$$a(\phi)=\phi=1$$

$\boxed{1}$

$$c(y,\phi)=\log\binom{n}{y}$$

$\boxed{1}$

The canonical link function, determined through the canonical parameter θ is

$$\log\left(\frac{\mu}{n-\mu}\right)$$

where $\mu\equiv E(Y)=n\pi$.

Must be written in terms of the mean μ. $\boxed{1}$

The variance function is obtained by rewriting $b''(\theta)$ in term of μ:

$$b'(\theta)=\frac{n}{1+e^{-\theta}},\quad b''(\theta)=n\frac{e^{-\theta}}{(1+e^{-\theta})^2}.$$

Thus the variance function is $V(\mu)=\dfrac{\mu(n-\mu)}{n}$. $\boxed{2}$

Note that $\mu\equiv E(Y)=b'(\theta)$.

sim. seen ⇓

b) i) The log-likelihood, in terms of π_i, is

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i} \right]$$

Using the given link function, the log-likelihood in terms of $\boldsymbol{\beta}$ is 2

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i \sum_{j=1}^{3} \beta_j x_i^{j-1} - n_i \log\left(1 + \exp\left\{\sum_{j=1}^{3} \beta_j x_i^{j-1}\right\}\right) + \log\binom{n_i}{y_i} \right]$$

unseen ⇓

ii) The fitted estimated parameters are obtained by solving $\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{0}$. Taking each component separately we find:

$$\frac{\partial \ell}{\partial \beta_1} = 0 \implies \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{\mu}_i$$

$$\frac{\partial \ell}{\partial \beta_2} = 0 \implies \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \widehat{\mu}_i x_i$$

$$\frac{\partial \ell}{\partial \beta_3} = 0 \implies \sum_{i=1}^{n} y_i x_i^2 = \sum_{i=1}^{n} \widehat{\mu}_i x_i^2$$

4

c) i) `ftv` is a categorical variable with 3 levels. Thus it is included in the model using 2 dummy regressors. The intercept term accounts for the baseline level, namely `ftv0`. sim. seen ⇓ 2

ii) The "??????" column is obtained by

$$2(1 - \Phi(|\texttt{s value}|))$$

where Φ is the CDF of a standard Normal distribution. 1

iii) The question simply corresponds to testing $\widehat{\beta}_{\mathsf{ftv2+}} = 0$ which has been performed in the `R` summary (`ftv2+` row). Since the resulting p-value is 0.6651, there is insufficient evidence to reject the null that the `ftv` groups differ at the 5% level. 3

iv) The dispersion parameter of the Binomial distribution is equal to 1, therefore we can compare the models using the deviance test with the χ^2 as the reference distribution. 1

meth seen ⇓

3. a) The probability density function for the Exponential Distribution can be written in the form

$$\exp\left\{y(-1/\lambda) - \log(\lambda)\right\},$$

where we identify

$$\theta = -\frac{1}{\lambda}$$

$\boxed{1}$

$$b(\theta) = \log(\lambda) = -\log(-\theta)$$

$\boxed{1}$

$$a(\phi) = \phi = 1$$

$\boxed{1}$

$$c(y, \phi) = 0$$

$\boxed{1}$

From lectures we know that $E(Y) = b(\theta)$ and $\mathrm{var}(Y) = \phi b''(\theta)$. For the Exponential distribution we get:

$$E(Y) = b'(\theta) = -\frac{1}{\theta} = \lambda$$

$\boxed{1}$

$$\mathrm{var}(Y) = b''(\theta) = \frac{1}{\theta^2} = \lambda^2$$

$\boxed{1}$

b) i) The log-likelihood in terms of the mean of the response is:

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{y_i(-1/\mu_i) - \log(\mu_i)\right\}$$

From lectures we know the maximal achievable log-likelihood is obtained by setting $\boldsymbol{\mu} = \boldsymbol{y}$. Therefore, the answer is

$$\ell(\boldsymbol{y}; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{y_i(-1/y_i) - \log(y_i)\right\} = -n - \sum_{i=1}^{n} \log(y_i)$$

$\boxed{2}$

ii) The general definition of the deviance is

$$D = 2\phi\left(\ell(\boldsymbol{y}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\mu}}_i; \boldsymbol{y})\right)$$

For this particular case:

$$D = 2\sum_{i=1}^{n} \left\{\frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i} - \log\left(\frac{y_i}{\widehat{\mu}_i}\right)\right\}$$

where $\phi = 1$.

$\boxed{3}$

c) i) The adjusted dependent variable is:

$$z_i = \widehat{\eta}_i + \frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i}$$

The weights are w_{ii} such that

$$w_{ii}^{-1} = \left(\frac{\partial \eta}{\partial \mu}\right)^2 V(\mu)\Big|_{\mu=\widehat{\mu}_i} = 1$$

$\boxed{2}$

ii) From lectures, we know that for $\phi = 1$

$\boxed{\text{unseen} \Downarrow}$

$$\text{cov}(\widehat{\boldsymbol{\beta}}) \approx (X^T W X)^{-1}$$

where W is a matrix with diagonal entries w_{ii} given in the IWLS algorithm (and zeros everywhere else). Since $W = I_n$ we have that

$$\text{cov}(\widehat{\boldsymbol{\beta}}) \approx (X^T X)^{-1} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} = \begin{pmatrix} n & 2 \\ 2 & 6 \end{pmatrix}^{-1} = \frac{1}{6n-4}\begin{pmatrix} 6 & -2 \\ -2 & n \end{pmatrix}$$

which does not depend on $\widehat{\boldsymbol{\beta}}$ and thus will not change over the iterations. $\boxed{4}$

iii) For the given covariate the linear predictor is

$\boxed{\text{sim. seen} \Downarrow}$

$$\widehat{\beta}_1 + 10\widehat{\beta}_2 = 0.2 + 10 \cdot (0.17) = 1.9$$

$\boxed{1}$

To get the prediction, we use the inverse of the link function

$$\exp(1.9)$$

$\boxed{2}$

unseen ⇓

4. a) By definition

$$\overline{Y}_{\bullet j} = \frac{1}{K}\sum_{i=1}^{K} Y_{ij} = \mu + \nu_j + \frac{1}{K}\sum_{i=1}^{K} \epsilon_{ij}.$$

and

$$\overline{Y} = \frac{1}{mK}\sum_{i=1}^{K}\sum_{j=1}^{m} Y_{ij} = \mu + \frac{1}{m}\sum_{j=1}^{m}\nu_j + \frac{1}{mK}\sum_{i=1}^{K}\sum_{j=1}^{m}\epsilon_{ij}.$$

Then

$$\mathrm{E}(\overline{Y}_{\bullet j}) = \mu \quad \text{and} \quad \mathrm{E}(\overline{Y}) = \mu$$

as $\mathrm{E}(\nu_j) = 0 = \mathrm{E}(\epsilon_{ij})$. Further,

$$\mathrm{var}(\overline{Y}_{\bullet j}) = \sigma_\nu^2 + \frac{1}{K}\sigma_\epsilon^2$$

and

$$\mathrm{var}(\overline{Y}) = \frac{1}{m}\sigma_\nu^2 + \frac{1}{mK}\sigma_\epsilon^2$$

as all ν_j and ϵ_{ij} are independent. Both $\overline{Y}_{\bullet j}$ and $\overline{Y}$ are normal since they are sums of normals; thus

$$\overline{Y}_{\bullet j} \sim N\left(\mu, \sigma_\nu^2 + \frac{1}{K}\sigma_\epsilon^2\right) \quad \text{and} \quad \overline{Y} \sim N\left(\mu, \frac{1}{m}\sigma_\nu^2 + \frac{1}{mK}\sigma_\epsilon^2\right)$$

4

b) i) The correlation for observations in the same group is:

seen ⇓

$$\text{corr}(Y_{1,j}, Y_{2,j}) = \frac{\mathrm{E}\left[(Y_{1,j} - \mathrm{E}(Y_{1,j}))\,(Y_{2,j} - \mathrm{E}(Y_{2,j}))\right]}{\sqrt{\mathrm{var}(Y_{1,j})\,\mathrm{var}(Y_{2,j})}}$$

We have that $\mathrm{var}(Y_{1,j}) = \mathrm{var}(Y_{2,j}) = \sigma_\epsilon^2 + \sigma_\nu^2$, and thus the denominator is $\sigma_\epsilon^2 + \sigma_\nu^2$. Next, we have $Y_{i,j} - \mathrm{E}(Y_{i,j}) = \nu_j + \epsilon_{ij}$, by noting that $\mathrm{E}(Y_{i,j}) = \mu$ and rearranging the model equation. Therefore the numerator is

$$\begin{aligned}\mathrm{E}\left[(Y_{1,j} - \mathrm{E}(Y_{1,j}))\,(Y_{2,j} - \mathrm{E}(Y_{2,j}))\right] &= \mathrm{E}\left[(\nu_j + \epsilon_{1,j}))\,(\nu_j + \epsilon_{2,j}))\right] \\ &= \mathrm{E}(\nu_j^2) + \mathrm{E}(\nu_j(\epsilon_{1,j} + \epsilon_{2,j})) + \mathrm{E}(\epsilon_{1,j}\epsilon_{2,j}) \\ &= \mathrm{E}(\nu_j^2) = \sigma_\nu^2,\end{aligned}$$

where we used the independence between the ν_j and ϵ_{ij}. Plugging in these results yields the required result.

4

unseen ⇓

ii) We have that

$$P\left(c_1 < \frac{\sigma_\nu^2}{\sigma_\epsilon^2} < c_2\right) = 0.95$$

Applying the increasing function $f(u) = u/(u+1)$ yields 2

$$P\left(\frac{c_1}{c_1+1} < \rho < \frac{c_2}{c_2+1}\right) = 0.95$$

Thus a 95% confidence interval for ρ is

$$\left(\frac{c_1}{c_1+1}, \frac{c_2}{c_2+1}\right)$$

2

unseen ⇓

c) i) We know that if $Z \sim \chi_d^2$ then $\mathrm{E}(Z) = d$. Then since

$$\mathrm{E}\left(\frac{SSE}{\sigma_\epsilon^2}\right) = m(K-1)$$

and

$$\mathrm{E}\left(\frac{SSA}{K\sigma_\nu^2 + \sigma_\epsilon^2}\right) = m-1$$

It follows that $d_1 = m(K-1)$ and $d_2 = m-1$. 2

ii) If H_0 is true then

unseen ⇓

$$\mathrm{E}\left(\frac{SSA}{m-1}\right) = K\sigma_\nu^2 + \sigma_\epsilon^2 = \sigma_\epsilon^2 = \mathrm{E}\left(\frac{SSE}{m(K-1)}\right).$$

On the other hand, if H_1 is true then:

$$\mathrm{E}\left(\frac{SSA}{m-1}\right) = K\sigma_\nu^2 + \sigma_\epsilon^2 > \sigma_\epsilon^2 = \mathrm{E}\left(\frac{SSE}{m(K-1)}\right).$$

Thus a reasonable test statistic is 2

$$\frac{\frac{SSA}{(m-1)(K\sigma_\nu^2+\sigma_\epsilon^2)}}{\frac{SSE}{m(K-1)(\sigma_\epsilon^2)}}$$

To conclude, under the null hypothesis $\sigma_\nu^2 = 0$:

$$T := \frac{\frac{SSA}{(m-1)}}{\frac{SSE}{m(K-1)}} \sim F_{m-1,m(K-1)}$$

which is a F-distribution with $m-1$ and $m(K-1)$ degrees of freedom since we have taken the ratio of independent χ^2 distributions (appropriately scaled by their degrees of freedom). 2

To conduct the test form the rejection region:

$$R = \{\boldsymbol{y} : T(\boldsymbol{y}) > u\}$$

where u is such that $P(A > u) = 0.05$ and $A \sim F_{m-1,m(K-1)}$. Thus if observed responses is such that $\boldsymbol{y} \in R$ the null is reject; otherwise the null is not rejected. 2

Alternatively, we may use the p-value approach.

sim.seen (material) ⇓

5. a) [**Suggested Solution**]

The objective in this paper is to find a method to predict the ripeness of a watermelon based on acoustic measurements. The authors claim that they have found a model that predicts watermelon ripeness with high accuracy. They state that acoustic measurements can predict 99.9% of the variation in ripeness! 4

b) [**Suggested Solution**]

The ripeness of the watermelon is rated by a panel of consumers. These ratings are subjective, which typically aren't consistent. For instance, if the consumer had to rerate the watermelon ripeness, they probably would not agree with their own rating with 99.9% accuracy.

3

c) Multiple linear regression is a linear regression model with more than 1 predictor. The response is the watermelons ripeness and the predictors are the different frequencies the watermelons were subjected to.

2

d) [**Suggested Solution**]

At a given step in the stepwise procedure, n independent hypotheses are being tested. 2

Suppose at the first step of the stepwise method, each hypothesis test was conducted at an α level. Further, suppose each of the null hypotheses are true. The probability of obtaining at least 1 false positive is

$$1 - (1 - \alpha)^n$$

At the 5% with $n = 1500$ we have that this probability is very close to 1 – which is undesirable. 4

e) Here is a list with the ordered p-values along with their corresponding l_i

$$\begin{aligned} P_2 = P_{(1)} &= 0.001 \quad l_1 = 0.01 \\ P_4 = P_{(2)} &= 0.025 \quad l_2 = 0.02 \\ P_3 = P_{(3)} &= 0.05 \quad l_3 = 0.03 \\ P_1 = P_{(4)} &= 0.1 \quad l_4 = 0.04 \\ P_5 = P_{(5)} &= 0.15 \quad l_5 = 0.05 \end{aligned}$$

Therefore $R = 1$ and so the BH rejection threshold is $T = 0.001$. Thus we reject the second hypothesis.

5

Imperial College London
Department of Mathematics

Examiner's Comments

Exam: M3S2 - Statistical Modelling 2 **Session: 2016-2107**

Question 1

Please use the space below to comment on the candidates' overall performance in the exam. A brief paragraph highlighting common mistakes and parts of questions done badly (or well) is sufficient. Do not refer to individual candidates. The purpose of this exercise is to provide guidance to the external examiners, and to the candidates themselves, on how you feel the cohort faired. Your comments will be available to students online.

A surprising amount of students were unable to give the specification of a GLM (which is the main model of the course!). Easy marks were available and many students were able to recall results from lectures. Part (iv) was tricky! Few students got this part totally correct. Many students were let down by their own handwriting leading to careless mistakes.

This question was well done, but I would have liked to see a bit higher marks!

Marker: Din-Houn Lau

Signature: Lau **Date:** 22/5/17

Please return with exam marks (one report per marker)

Imperial College London
Department of Mathematics

Examiner's Comments

Exam: M3S2 – Statistical Modelling 2. **Session: 2016-2107**

Question 2

Please use the space below to comment on the candidates' overall performance in the exam. A brief paragraph highlighting common mistakes and parts of questions done badly (or well) is sufficient. Do not refer to individual candidates. The purpose of this exercise is to provide guidance to the external examiners, and to the candidates themselves, on how you feel the cohort faired. Your comments will be available to students online.

The first parts were done well, as it was repeatedly seen in lectures. Some lost marks for not writing the link and variance functions in terms of the mean.

It was nice to see that the majority understood categorical variables which was stressed in lectures this year.

On the other hand, many forgot how to compute the p-value in part (c)(ii). Further, a fair few candidates lost easy marks in part (b).

Marker: Din-Houn Lau

Signature: Lau **Date:** 22/5/17

Please return with exam marks (one report per marker)

Imperial College London
Department of Mathematics

Examiner's Comments

Exam: ______________________ Session: 2016-2107

Question 3

Please use the space below to comment on the candidates' overall performance in the exam. A brief paragraph highlighting common mistakes and parts of questions done badly (or well) is sufficient. Do not refer to individual candidates. The purpose of this exercise is to provide guidance to the external examiners, and to the candidates themselves, on how you feel the cohort faired. Your comments will be available to students online.

Quite good performance on this question. Some sloppy handling of parameters of GLM. Occasional negative variances!.

Some confusion on IWLS – occasional attempts to simply regurgitate algorithm

Marker: NIALL ADAMS

Signature: Niall

Date: 15/6/17

Please return with exam marks (one report per marker)

Imperial College London
Department of Mathematics

Examiner's Comments

Exam: ______________________ **Session: 2016-2107**

Question 4

Please use the space below to comment on the candidates' overall performance in the exam. A brief paragraph highlighting common mistakes and parts of questions done badly (or well) is sufficient. Do not refer to individual candidates. The purpose of this exercise is to provide guidance to the external examiners, and to the candidates themselves, on how you feel the cohort faired. Your comments will be available to students online.

Mixed performance on this more challenging question

Sloppy manipulations, showing lack of understanding, often in first part.

Last part required more description from many students

Marker: NIALL ADAMS

Signature: Niall Ad **Date:** 15/7/17

Please return with exam marks (one report per marker)

Imperial College London
Department of Mathematics

Examiner's Comments

Exam: M4S2 – Statistical Modelling II Session: 2016-2107

Question ~~1~~ 5

Please use the space below to comment on the candidates' overall performance in the exam. A brief paragraph highlighting common mistakes and parts of questions done badly (or well) is sufficient. Do not refer to individual candidates. The purpose of this exercise is to provide guidance to the external examiners, and to the candidates themselves, on how you feel the cohort faired. Your comments will be available to students online.

Overall, this question was fairly well attempted by most students. I believe the 'summary' type questions may have confused some students. Almost all candidates noticed the subjective nature of the data – which was excellent!

Further, most students noticed the link with the method and the multiple comparison problem. However, some lost easy marks for not explaining the problem in detail.

Despite being provided the B-H method some students were not able to execute it correctly – easy marks lost here!

Marker: DIN-HOUN LAU

Signature: Lau Date: 22/5/17.

Please return with exam marks (one report per marker)