UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE	:	**STATM001**
ASSESSMENT PATTERN	:	**STATM001C**
MODULE NAME	:	**Statistical Models and Data Analysis (Masters Level)**
DATE	:	**06 May 2016**
TIME	:	**2:30 pm**
TIME ALLOWED	:	**2 hours**

This paper is suitable for candidates who attended classes for this module in the following academic year(s):

2015/16

TURN OVER

Answer ALL questions. Section A carries 40% of the total marks and Section B carries 60%. The relative weights attached to each question are as follows, given in total marks (the overall sum of marks is 100): A1 (23), A2(6), A3 (11), B1 (12), B2 (18), B3 (21), B4 (9). The numbers in square brackets indicate the relative weight attached to each part question.

Section A

A1 Consider the multiple linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + e_i, \quad i = 1, \dots, n, \quad j = 1, \dots, m,$$

where the x_{ij} represent fixed values of m covariates, $\beta_0, \beta_1, \dots, \beta_m$ are unknown regression coefficients, and the errors e_i, $i = 1, \dots, n$, are assumed to follow the distribution $\mathcal{N}(0, \sigma^2)$, with σ^2 unknown. Recall that the model above can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)^T$, $\mathbf{e} = (e_1, e_2, \dots, e_n)^T$ and $\mathbf{X}$ is a full rank design matrix. Also, recall that the predicted value vector $\hat{\boldsymbol{\mu}}$ is given by $\mathbf{X}\hat{\boldsymbol{\beta}}$.

(a) What are the distributional assumptions of $\mathbf{e}$ and $\mathbf{y}$? [4]

(b) Write down the residual sum of squares that $\hat{\boldsymbol{\beta}}$ (the least squares estimator of $\boldsymbol{\beta}$) must minimise and hence derive the normal equations which $\hat{\boldsymbol{\beta}}$ must satisfy. Finally, write down the expression for $\hat{\boldsymbol{\beta}}$. [5]

(c) Define the hat (or projection) matrix $\mathbf{H}$ such that $\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}$. [3]

(d) Show that $\mathrm{tr}(\mathbf{H}) = p$, where $p = m + 1$. [6]

(e) Give the distribution of $\hat{\boldsymbol{\beta}}$, and justify the use of a t-distribution for making inferences about the individual regression parameters from a fitted regression. [5]

A2 Consider a simplified version of the model in question **A1**. That is, assume that $m = 1$.

(a) Define the *residuals* r_i and show that they have the following properties:

$$\sum_{i=1}^{n} r_i = 0, \qquad \sum_{i=1}^{n} x_i r_i = 0.$$

[3]

(b) Which of the following common assumptions are required for the parameter estimator to be unbiased: the y_i are independent; the y_i all have the same variance; the y_i are normally distributed. [3]

A3 The dataset analysed in the computer output below is made up of 315 individuals and contains the following variables: age (in years), `quetelet` (which is a measure of obesity defined as weight divided by the square of height), plasma beta-carotene (`betaplasma`, ng/ml), fibre consumed (`fiber`, g per day), `cholesterol` (mg per day), and dietary beta-carotene (`betadiet`, mcg per day). The response variable is plasma beta-carotene `betaplasma`. The other variables are the predictors. The aim is to investigate how `betaplasma` depends on these explanatory variables.

(a) Write down algebraically the model that has been fitted and specify any statistical assumption made. Define your notation. [4]

(b) From the model above, what level of `betaplasma` can be expected if age=50, `quetelet`=30, `fiber`=15, `cholesterol`=300 and `betadiet`=2000? [3]

(c) What are the basic quantities required in a Fisher Scoring algorithm, and what does "Number of Fisher Scoring iterations" mean? (You are not required to define such quantities mathematically.) [4]

```
Call:
glm(formula = betaplasma ~ age + quetelet + fiber + cholesterol +
    betadiet, family = Gamma(link = "log"), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7469  -0.5347  -0.2282   0.1665   2.4370

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.596    2.919e-01  19.170  < 2e-16 ***
age           0.005135 3.158e-03   1.626  0.10491
quetelet     -0.03309  7.655e-03  -4.323 2.08e-05 ***
fiber         0.018611 9.869e-03   1.885  0.06031 .
cholesterol  -0.00098  3.747e-04  -2.627  0.00904 **
betadiet      0.000092 3.542e-05   2.612  0.00944 **
---
Signif. codes:  0 **0.001 *0.01 0.05 0.1 1

(Dispersion parameter for Gamma family taken to be 0.6472711)

    Null deviance: 181.80  on 312  degrees of freedom
Residual deviance: 145.21  on 307  degrees of freedom
AIC: 3791.4

Number of Fisher Scoring iterations: 6
```

Section B

B1 Let us consider the following simple linear model:

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i, \quad i = 1, \ldots, n.$$

(a) Write down $\mathbf{X}^\mathsf{T}\mathbf{X}$, where $\mathbf{X}$ is the design matrix for the above model. [2]

(b) If all x_i are equal, is $\mathbf{X}^\mathsf{T}\mathbf{X}$ invertible? Justify your answer. [5]

(c) In the case in which $\mathbf{X}^\mathsf{T}\mathbf{X}$ is not invertible, what possible values can $\hat{\beta}_1$ take? [5]

B2 Consider the case of count responses, y_i, which are assumed to follow a Poisson distribution with mean μ_i, and that you wish to employ the model

$$\log \mu_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$

Recall that the pmf of a Poisson is $\Pr(y_i) = \mu_i^{y_i} \exp(-\mu_i)/y_i!$.

(a) Show that log-likelihood function as a function of β_0 and β_1 (given the data) is

$$\ell = -\sum_{i=1}^{n} \mathrm{e}^{\beta_0 + \beta_1 x_i} + \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) + \text{constant}.$$

[6]

(b) Obtain the likelihood equations (by differentiating ℓ with respect to β_0 and β_1). Moreover, describe in words (i.e., without any mathematical derivation) how you could obtained the Fisher information matrix. [6]

(c) Generally speaking, what are the advantages and disadvantages of using the Fisher information matrix in optimisation? [6]

B3 Let $y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, be a sample of independent random variables and consider the linear predictor $\eta_i = \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}$.

(a) Show that the above model is a special case of the generalised linear model and give the corresponding link function. [5]

(b) Show that if σ^2 is known, the (scaled) deviance of the general linear model is given by

$$D = \frac{SS_R}{\sigma^2},$$

where $SS_R = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$. [5]

(c) State (without proof) the asymptotic distribution of D and hence derive an estimator for σ^2 when it is unknown. [4]

(d) The usual link function for the Gaussian distribution is the identity, yielding the normal linear model. Suppose instead that $y = \mu + \epsilon$, where

$$\mu = \beta_0 \frac{x}{\beta_1 + x},$$

where x is the explanatory variable and β_0 and β_1 are the parameters. State whether this model fits into the definition of generalised linear model and give the corresponding link function. [7]

B4 Consider the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, defined in question **A1**, and assume that all variables in $\mathbf{X}$ are standardised so that their regression parameters are of comparable size. One way to perform variable selection is to minimise the objective function $||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \sum_{j=1}^{m} |\beta_j|$, where λ is a tuning constant.

(a) Why β_0 is not covered by the constraint above? [3]

(b) Why are the absolute values of the β_j required to be of comparable size? [3]

(c) Is the resulting estimator biased? Justify your answer. [3]

END OF PAPER