UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE	:	**STATM001**
ASSESSMENT PATTERN	:	**STATM001C**
MODULE NAME	:	**Statistical Models and Data Analysis (Masters Level)**
DATE	:	**Tuesday 15 May 2018**
TIME	:	**14:30**
TIME ALLOWED	:	**2 hrs**

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

Year

Suitable for all candidates

EXAMINATION PAPER CANNOT BE REMOVED FROM THE EXAM HALL. PLACE EXAM PAPER AND ALL COMPLETED SCRIPTS INSIDE THE EXAMINATION ENVELOPE

2016/17-STATM001C-001-EXAM-Statistical Science *4*

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 (14), A2(8), A3 (18), B1 (13), B2 (16), B3 (17), B4 (14). 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 y_i is the response, n is the sample size, 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 independent and follow the Normal distribution $\mathcal{N}(0, \sigma^2)$, with σ^2 unknown. Recall that the model above can be written in a compact way as

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

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)^T$, $\mathbf{e} = (e_1, e_2, \dots, e_n)^T$, $\mathbf{X}$ is a full rank design matrix (containing a column vector of 1's and the covariates x_{ij}), and ^{T} denotes the transpose operator. Also, recall that the vector of predicted values, indicated by $\hat{\boldsymbol{\mu}}$ or $\hat{\mathbf{y}}$, is given by $\mathbf{X}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is a parameter estimator for $\boldsymbol{\beta}$.

(a) 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]

(b) Derive an expression for the hat (or projection) matrix $\mathbf{H}$ in terms of $\mathbf{X}$. [3]

(c) Show that $\mathbf{H}$ is symmetric. [3]

(d) Show that $\mathbf{H}$ is idempotent, i.e. $\mathbf{H}^2 = \mathbf{H}$. [3]

A2 Consider the linear regression model

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + e_i,$$

where $\bar{x}$ is the sample mean, and e_i is the usual error term.

Find h_{ii} (the i^{th} diagonal element of $\mathbf{H}$) for the above model. [8]

A3 In a study of infant feeding, 50 infants aged approximately 2 months were analysed to determine their intake of breast milk. This amount, together with five potential explanatory variables, were used for a statistical analysis. The variables are **dl.milk** (breast milk intake, dl/24 hr), **sexvar** (0 for boy and 1 for girl), **weight** (weight of infant, kg), **ml.suppl** (amount of milk substitute given to infant in a period before the breast milk intake measurement, ml/24 hr), **mat.weight** (weight of mother, kg), **mat.height** (height of mother, cm).

(a) Below you find the correlation matrix between the five explanatory variables.

	dl.milk	weight	ml.suppl	mat.weight	mat.height	sexvar
dl.milk	1.00	0.63	-0.06	0.43	0.50	-0.29
weight	0.63	1.00	0.12	0.40	0.38	-0.22
ml.suppl	-0.06	0.12	1.00	-0.07	0.18	-0.07
mat.weight	0.43	0.40	-0.07	1.00	0.56	-0.05
mat.height	0.50	0.38	0.18	0.56	1.00	-0.11
sexvar	-0.29	-0.22	-0.07	-0.05	-0.11	1.00

On the basis of this information, which explanatory variables would you expect to be included in a good multiple regression model which has `dl.milk` as the response variable? Briefly explain your choices. [5]

(b) A multiple regression model (Model A) with `dl.milk` as the response variable and all five explanatory variables was fitted in R, resulting in the following output:

Model A

```
> modA <- lm(dl.milk ~ sexvar + weight + ml.suppl + mat.weight + mat.height,
            data = milkdata)
> summary(modA)

Call:
lm(formula = dl.milk ~ sexvar + weight + ml.suppl + mat.weight +
    mat.height, data = milkdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74201 -0.81173 -0.00926  0.78326  2.52646

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.681839   4.361561  -2.678 0.010363 *
sexvar       -0.499532   0.312672  -1.598 0.117284
weight        1.349124   0.322450   4.184 0.000135 ***
ml.suppl     -0.002233   0.001241  -1.799 0.078829 .
mat.weight    0.006212   0.023708   0.262 0.794535
mat.height    0.072278   0.030169   2.396 0.020906 *
---
Signif. codes:  0 **0.001 *0.01 0.05 0.1 1

Residual standard error: 1.075 on 44 degrees of freedom
Multiple R-squared:  0.5459,Adjusted R-squared:  0.4943
F-statistic: 10.58 on 5 and 44 DF,  p-value: 1.03e-06
```

CONTINUED

(i) Explain why there are not separate coefficients given for `boy` and `girl` for `sexvar`. What does the given coefficient for `sexvar` represent? [2]

(ii) From this model, what would be the expected breast milk intake in 24 hours for an infant boy weighing 5.5kg, who had no milk substitute in the period before measurement, and whose mother weighed 60kg and was 168cm tall? Use the further R output below to calculate both 95% confidence and prediction intervals for your estimate. [5]

```
> predict.modA <- predict(modA, data.frame(sexvar = 0, weight = 5.5, ml.suppl = 0,
+                         mat.weight = 60, mat.height = 168), se.fit = T)
> predict.modA$se.fit
[1] 0.2543874
> qt(0.975, 44)
[1] 2.015368
```

(c) The stepwise regression method provided by the `stepAIC` function in the R package MASS was applied to the data. Starting from the null model, Model B was obtained as follows:

```
library(MASS)
> mod0 <- lm(dl.milk ~ 1, data = milkdata)
> stepAIC(mod0,  ~ sex + weight + ml.suppl + mat.weight + mat.height, data = milkdata)

:
output omitted
```

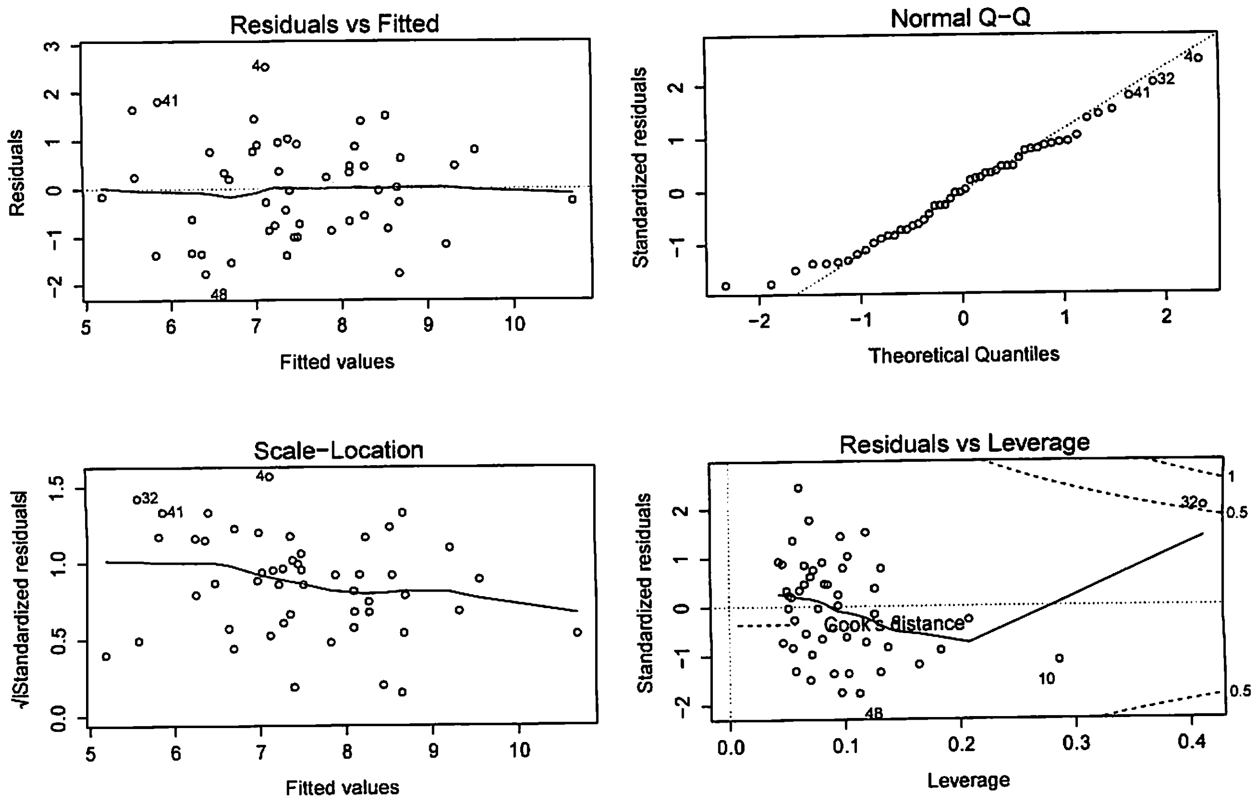
Model B

```
modB <- lm(dl.milk ~ sexvar + weight + ml.suppl + mat.height,
+          data = milkdata)
> summary(modB)

Call:
lm(formula = dl.milk ~ sexvar + weight + ml.suppl + mat.height,
    data = milkdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77312 -0.81196 -0.00683  0.76988  2.52240

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.112571   3.997860  -3.030  0.00405 **
sexvar       -0.494675   0.308875  -1.602  0.11626
weight        1.372524   0.306612   4.476 5.14e-05 ***
ml.suppl     -0.002313   0.001190  -1.943  0.05824 .
mat.height    0.076363   0.025560   2.988  0.00454 **
---
Signif. codes:  0 **0.001 *0.01 0.05 0.1 1

Residual standard error: 1.064 on 45 degrees of freedom
Multiple R-squared:  0.5452,Adjusted R-squared:  0.5047
F-statistic: 13.48 on 4 and 45 DF,  p-value: 2.658e-07
```

(i) Give an explanation to why the `stepAIC` procedure has led to a more parsimonious model. [2]

(ii) Explain briefly why it may be preferable to use Model B than Model A. [2]

(iii) Comment on the residual plots (from Model B) reported below. Do the assumptions underlying the multiple regression model appear to be satisfied in this case? [2]



Section B

B1 Recall the model described in question **A1**. Also, let $\mathbf{z}$ be an $n \times 1$ vector of random variables and $\mathbf{A}$ an $n \times n$ symmetric matrix of constants and recall that if $\mathrm{E}(\mathbf{z}) = \boldsymbol{\theta}$ and $\mathrm{Cov}(\mathbf{z}, \mathbf{z}) = \boldsymbol{\Sigma}$ then $\mathrm{E}(\mathbf{z}^T \mathbf{A} \mathbf{z}) = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\theta}^T \mathbf{A} \boldsymbol{\theta}$.

(a) Prove that for the multiple linear regression model

$$\mathrm{E}(\epsilon^T \epsilon) = \sigma^2 \mathrm{tr}(\mathbf{I} - \mathbf{H}) + \boldsymbol{\beta}^T \mathbf{X}^T (\mathbf{I} - \mathbf{H}) \mathbf{X} \boldsymbol{\beta}.$$

where $\epsilon = (\mathbf{I} - \mathbf{H})\mathbf{y}$ is the vector of observed residuals, and $\mathbf{H}$ is the hat matrix. [6]

(b) Exploiting the result in part (a), show that an unbiased estimator of σ^2 is given by $\hat{\sigma}^2 = \epsilon^T \epsilon / (n - p)$, where $p = m + 1$ and m is the total number of covariates. [5]

(c) Explain briefly how the estimators $\hat{\boldsymbol{\beta}}$ (found in **A1**) and $\hat{\sigma}^2$ are related. [2]

B2 Let Y be a random variable that follows a negative binomial distribution whose density function can be written as

$$f(y; k, \mu) = \frac{\Gamma(y + k)}{\Gamma(k)\Gamma(y + 1)} \left(\frac{k}{k + \mu} \right)^k \left(1 - \frac{k}{k + \mu} \right)^y,$$

where $y \in \{0, 1, 2, \ldots\}$ is the number of successes, k is the number of failures, and $\mu > 0$ is the mean.

(a) For a member of the generalized exponential family, define the canonical parameter θ, dispersion parameter ϕ, $a(\phi)$, $b(\theta)$ and $c(y, \phi)$. [11]

(b) State the generic results for $E(Y)$ and $V(Y)$ (no derivation required). [2]

(c) Using the results stated above, and not otherwise, derive the mean and variance for the negative binomial distribution. [3]

B3 Data were collected to study a type of damage caused by waves to the forward section of certain cargo-carrying vessels. The values of the following variables were recorded: `type`: ship type (coded as A to E); `year`: year of construction (1960, 1965, 1970, 1975); `period`: period of operation (1960-74, 75-79); `service`: aggregate months of service; `incidents`: number of damage incidents. The output of the fitted model is:

```
Call:
glm(formula = incidents ~ offset(log(service)) + type + year + period,
    family = poisson, data = wdships)

Deviance Residuals:
Min      1Q       Median   3Q      Max
-1.6768  -0.8293  -0.4370  0.5058  2.7912

Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -6.40590  0.21744     -29.460  < 2e-16 ***
typeB       -0.54334  0.17759      -3.060  0.00222 **
typeC       -0.68740  0.32904      -2.089  0.03670 *
typeD       -0.07596  0.29058      -0.261  0.79377
typeE        0.32558  0.23588       1.380  0.16750
year65       0.69714  0.14964       4.659  3.18e-06 ***
year70       0.81843  0.16977       4.821  1.43e-06 ***
year75       0.45343  0.23317       1.945  0.05182 .
period75     0.38447  0.11827       3.251  0.00115 **

(Dispersion parameter for poisson family taken to be 1)

Null deviance:     146.328 on 33 degrees of freedom
Residual deviance: 38.695 on 25 degrees of freedom
```

(a) State carefully the model used in the above analysis explaining how the covariates have been treated. [4]

(b) Give the reason to use `offset(log(service))` in the above analysis. [5]

(c) Based on the information contained in the R output, does the model fit the data well? Justify your answer. [5]

(d) Should we consider allowing the dispersion parameter to vary? Justify your answer. [3]

B4 Consider the model

$$y_i = f(x_i) + e_i, \; i = 1, \ldots, n,$$

where $f(\cdot)$ is a smooth function of covariate x_i and e_i is defined as in **A1**. Given the regression spline representation of $f(\cdot)$, the model can be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{X}$ is a design matrix containing the basis functions associated with x_i ($i = 1, \ldots, n$) and $\boldsymbol{\beta}$ is the corresponding spline parameter vector. To avoid overfitting, a penalty matrix $\mathbf{S}$ is typically employed during model fitting, whilst a smoothing parameter $\lambda > 0$ is used to control the trade-off between goodness of fit and smoothness.

(a) Show that for any function f, which has a basis expansion

$$f(x) = \sum_j \beta_j b_j(x),$$

where $b_j(x)$ is the j^{th} basis function of x and β_j the corresponding parameter, it is possible to write

$$\int \left\{ f''(x) \right\}^2 dx = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{S} \boldsymbol{\beta},$$

where matrix $\mathbf{S}$ can be expressed in terms of the known basis functions b_j (assuming that these possess at least two (integrable) derivatives). $\boldsymbol{\beta}$ is a vector whose j^{th} element is given by the parameter β_j. [5]

(b) Using the least square method, write down the objective function that needs to be minimised in order to obtain a sensible estimator for $\boldsymbol{\beta}$. [3]

(c) For a fixed smoothing parameter λ, derive $\hat{\boldsymbol{\beta}}$. [2]

(d) Find $E(\hat{\boldsymbol{\beta}})$ and state whether $\hat{\boldsymbol{\beta}}$ is biased or unbiased. [2]

(e) Explain why it is not possible to jointly estimate $\boldsymbol{\beta}$ and λ. [2]

END OF PAPER