Queen Mary
University of London

Main Examination period 2017

MTH5120: Statistical Modelling I

Duration: 2 hours

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

You should attempt ALL questions. Marks available are shown next to the questions.

Only non-programmable calculators that have been approved from the college list of non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.
Statistical functions provided by the calculator may be used provided that you state clearly where you have used them.
The New Cambridge Statistical Tables are provided.

Complete all rough work in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it shall be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the examination room.

Examiners: L I Pettit, A Gnedin

Question 1. [23 marks]
The following table gives the age in years (x) and total cholesterol level in mg/ml (y) for 19 patients suffering from hyperlipoproteinaemia.

x	y	x	y	x	y	x	y
46	3.5	20	1.9	52	4.0	30	2.7
57	4.5	25	3.0	28	2.9	36	3.8
22	2.1	43	3.8	57	4.1	33	3.0
22	2.5	63	4.6	40	3.2	48	4.2
28	2.3	49	4.0	52	4.3		

Summary statistics for these data are $\sum x_i = 751$, $\sum y_i = 64.4$, $S_{xx} = 3306.74$, $S_{xy} = 192.705$, $S_{yy} = 12.698$.

(a) The data are expected to be linearly related and the simple linear regression model

$$y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

is to be fitted. What assumptions are usually made about the errors (ε_i)? [**4**]

(b) Derive the least squares estimators of α and β by minimising a suitable function. Check that your solution does give a minimum. [**10**]

(c) Hence find the equation of the fitted model for the cholesterol data. [**4**]

(d) State the form of the 95% confidence interval for β. Find the numerical estimate of this interval. [**5**]

Question 2. [19 marks]
To investigate the effect of dose of a drug on a response, twenty patients were allocated at random to one of four doses (1,5,9,13mg) so that five patients received each dose. A simple linear regression model of response was fitted.

(a) Copy and complete the following Analysis of Variance table. [**11**]

(b) What two hypotheses can be tested? [**2**]

(c) Carry out these tests using a 1% significance level and make clear your conclusions. [**6**]

```
Analysis of Variance

Source           DF       SS       MS       VR
Regression        1   1387.6
Residual Error
  Lack of Fit           33.2
  Pure Error
Total            19   1454.8
```

Question 3. [35 marks]
Data were collected from 17 US Navy hospitals. The variables measured were `x1`, average daily patient load, `x2`, X-rays taken per month, `x3`, occupied bed days per month, `x4`, eligible population in thousands, `x5`, average length of stay in days, and Y the staff hours per month.

(a) The data were entered into Minitab and the best subsets regression procedure carried out. The output is shown below.

```
Best Subsets Regression: y versus x1, x2, x3, x4, x5

Response is y

                         Mallows           x x x x x
Vars  R-Sq  R-Sq(adj)        Cp        S   1 2 3 4 5
   1  97.2       97.0      20.4   957.86       X
   1  97.1       97.0      21.2   969.53   X
   2  98.7       98.5       4.9   685.17     X X
   2  98.6       98.4       5.7   700.42   X X
   3  99.0       98.8       2.9   614.78     X X   X
   3  98.9       98.7       3.7   634.99   X X     X
   4  99.1       98.8       4.0   615.49     X X X X
   4  99.1       98.7       4.3   622.09   X X   X X
   5  99.1       98.7       6.0   642.09   X X X X X
```

(i) Define the four statistics given in the table: R^2, $R^2(adj)$, Mallows C_p and S. [**4**]

(ii) Based on these statistics say which model you would choose for these data and justify your choice. [**8**]

(iii) In what way is $R^2(adj)$ an improvement on R^2? [**3**]

(b) The Minitab session output for the model with regressors `x2`, `x3` and `x5` is given below.

```
The regression equation is
y = 1523 + 0.0530 x2 + 0.978 x3 - 321 x5


Predictor      Coef  SE Coef      T      P     VIF
Constant     1523.4    786.9   1.94  0.075
x2          0.05299  0.02009   2.64  0.021   7.737
x3           0.9785   0.1052   9.31  0.000  11.269
x5           -321.0    153.2  -2.10  0.056   2.493


S = 614.779   R-Sq = 99.0%   R-Sq(adj) = 98.8%
```

(i) What is meant by **multicollinearity**? What problems can it cause. [**6**]

(ii) Define the **variance inflation factor** (VIF). [**3**]

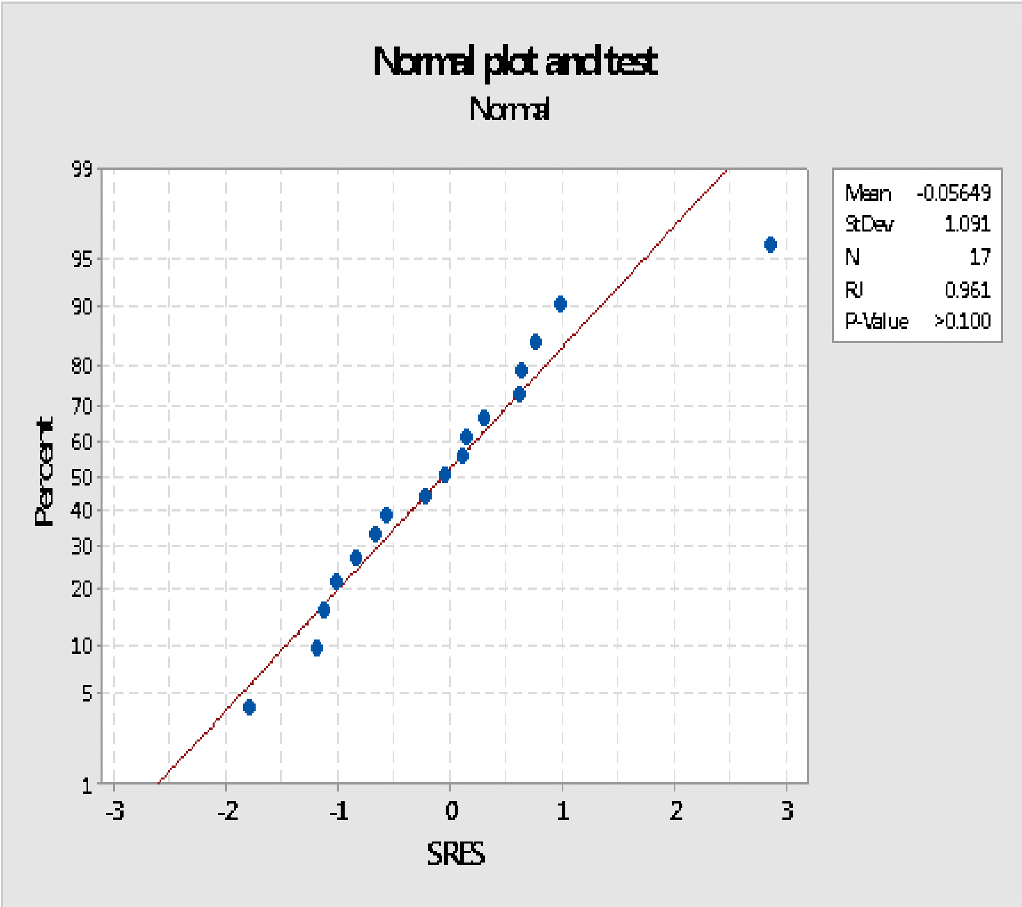
(iii) Why do we calculate variance inflation factors? [**2**]

(iv) Comment on the sizes of the variance inflation factors in this example. [**3**]

(c) The following normal plot and test of the standardised residuals was produced.

(i) Comment on what this tells us about the assumption of normally distributed errors. [**3**]

(ii) Name one other plot you would like to see to assess if the model is fitting well. Explain what it would tell you. [**3**]



Question 4. [23 marks]

(a) For the general linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is a vector of errors assumed to be uncorrelated with zero mean and constant variance σ^2, state the formula for the least squares estimator $\hat{\boldsymbol{\beta}}$. [**1**]

(b) Prove that the expectation of $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{\beta}$. [**4**]

(c) Derive a formula for the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, quoting any necessary results. [**6**]

(d) (i) Define the hat matrix $\boldsymbol{H}$. [**1**]

(ii) Show that the vector of fitted values is given by $\boldsymbol{H}\boldsymbol{Y}$. [**2**]

(e) Show that $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$. [**3**]

(f) Express the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

where the ε_i have mean zero, variance σ^2 and are uncorrelated, as a general linear model by writing down the vectors $\boldsymbol{Y}$ and $\boldsymbol{\beta}$ and the matrix $\boldsymbol{X}$. [**6**]

End of Paper.