Queen Mary University of London

B. Sc. Examination by course unit 2014

MTH5120 Statistical Modelling I

Duration: 2 hours

Date and time: 16 May 2014, 1000h–1200h

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 awarded are shown next to the questions.

Calculators ARE permitted in this examination. The unauthorized use of material stored in pre-programmable memory constitutes an examination offence. 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 workings in the answer book and cross through any work which is not to be assessed.

Important note: the Academic Regulations state that possession of unauthorized material at any time by a student who is under examination conditions is an assessment offence and can lead to expulsion from QMUL.

Please check now to ensure you do not have any notes, mobile phones or unauthorised electronic devices on your person. If you have any, then please raise your hand and give them to an invigilator immediately. Please be aware that if you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will 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. Disruption caused by mobile phones is also an examination offence.

Exam papers must not be removed from the examination room.

Examiner(s): B Bogacka, L I Pettit

Question 1 [26 marks] The Environmental Protection Agency has been interested in evaluating the fuel efficiency of cars. In an observational study $n = 38$ recordings of miles per gallon (Y) and two explanatory variables: weight (X_1) and displacement (X_2) of engine were taken. After the power transformation Y^λ with $\lambda = -0.5$, a second order model

$$Y_i^\lambda = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \varepsilon_i$$

was fitted, assuming $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. A part of the MINITAB output is given below.

```
The regression equation is
y^-0.5 = - 0.205 - 0.0377 x1 + 0.0142 x2 - 0.0147 x1^2 - 0.0220 x2^2 + 0.0371 x1x2

Coefficients
Predictor        Coef   SE Coef        T      P
Constant    -0.204908  0.002717   -75.41  0.000
x1          -0.037662  0.007113    -5.29  0.000
x2           0.014222  0.008231     1.73  0.094
x1^2         -0.01468   0.01242    -1.18  0.246
x2^2         -0.02199   0.01183    -1.86  0.072
x1x2          0.03707   0.02379     1.56  0.129

S = 0.00915622   R-Sq = 90.6%   R-Sq(adj) = 89.1%
PRESS = 0.00442312   R-Sq(pred) = 84.44%

Analysis of Variance
Source          DF         SS         MS      F      P
Regression       5  0.0257390  0.0051478  61.40  0.000
Residual Error  32  0.0026828  0.0000838
Total           37  0.0284218

Source   DF     Seq SS
x1        1  0.0242711
x2        1  0.0011307
x1^2      1  0.0000107
x2^2      1  0.0001230
x1x2      1  0.0002035
```

(a) Write down the null and alternative hypotheses for the parameter β_{22} tested in the table of `Coefficients`. What do you conclude? [4]

(b) Write down the null and alternative hypotheses tested in the `Analysis of Variance` table. What do you conclude? [4]

(c) Test, at the significance level $\alpha = 0.1$, the null hypothesis $H_0 : \beta_{11} = \beta_{22} = \beta_{12} = 0$ versus $H_1 : \neg H_0$. Write down the test statistic and explain your notation. [11]

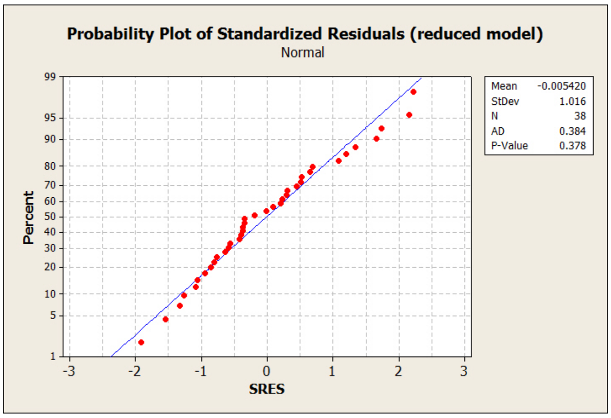
Question 1 continues on the next page.

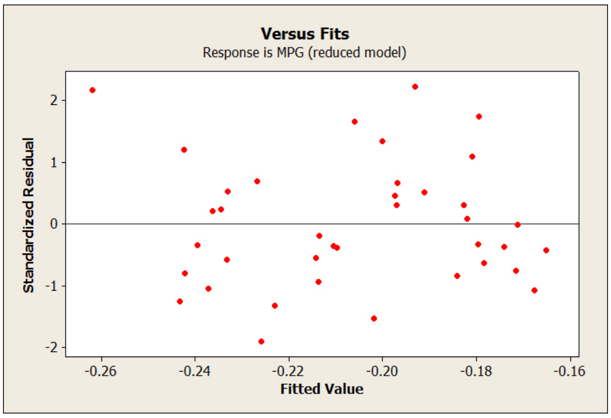
The reduced model for the transformed variable was fitted and a part of the MINITAB output is given below.

```
The regression equation is
y^-0.5 = - 0.206 - 0.0426 x1 + 0.0178 x2

Predictor        Coef   SE Coef        T      P
Constant    -0.206296  0.001507  -136.90  0.000
x1          -0.042571  0.004928    -8.64  0.000
x2           0.017837  0.004928     3.62  0.001

S = 0.00928900   R-Sq = 89.4%   R-Sq(adj) = 88.8%
PRESS = 0.00360448   R-Sq(pred) = 87.32%
```





(d) List three indicators shown in the numerical output which suggest an improvement in the model fit compared to the full model and briefly justify your choice. [3]

(e) Briefly comment on the residual plots shown above. [4]

Question 2 [24 marks] The Least Squares Estimator of β_1, $\widehat{\beta}_1$, in the Simple Linear Model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

can be written as

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i, \quad \text{where} \quad c_i = \frac{x_i - \overline{x}}{S_{xx}} \quad \text{and} \quad S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Show that $\widehat{\beta}_1$

(a) is normally distributed, [4]

(b) is unbiased for β_1, [10]

(c) and has variance equal to σ^2/S_{xx}. [10]

Question 3 [24 marks] In the investigation of the dependence of heat rate (Y, in KJ/KW/h) of gas turbines on cycle speed (X_1, in revolutions per minute), cycle pressure ratio (X_2), inlet temperature (X_3, in C°) and exhaust gas temperature (X_4, in C°), $n = 67$ observations were recorded. X_1 - X_4 are potential explanatory variables for a multiple linear regression model of the response variable Y. A part of the MINITAB numerical output is displayed below.

```
Response is y
                       Mallows          x x x x
Vars  R-Sq  R-Sq(adj)       Cp       S  1 2 3 4
   1  71.2       70.8    156.9  862.01  X
   1  64.1       63.5    211.5  963.13      X
   2  87.3       86.9     36.5  578.32  X   X
   2  84.8       84.3     55.4  631.88  X X
   3  91.9       91.5      3.0  464.98  X   X X
   3  90.2       89.7     16.3  512.27  X X X
   4  91.9       91.4      5.0  468.63  X X X X
```

(a) Based on all the information above, suggest, with justification, the two best models for describing the relationship between the response and explanatory variables. Indicate which one of the models might be better and why. [8]

A part of the MINITAB numerical output for the model including all available explanatory variables is given below.

(b) Give the definition of the variance inflation factor (VIF). What impact on the statistical inference can a high variance inflation factor make? [4]

(c) Briefly comment on the values of the VIF shown in the output below. [3]

```
The regression equation is
y = 14398 + 0.106 x1 - 4.4 x2 - 9.03 x3 + 12.1 x4

Predictor      Coef  SE Coef      T      P     VIF
Constant    14397.9    784.6  18.35  0.000
x1          0.10551  0.01107   9.53  0.000   1.818
x2            -4.36    30.08  -0.15  0.885   4.897
x3           -9.033    1.529  -5.91  0.000  13.264
x4           12.054    3.308   3.64  0.001   6.407

S = 468.635   R-Sq = 91.9%   R-Sq(adj) = 91.4%
```

A part of the MINITAB numerical output for the model including X_1, X_3, X_4 is given on the next page.

(d) Briefly comment on the values of the VIF. [3]

(e) Based on the MINITAB output for both models compare the accuracy of estimation of the model parameters. [2]

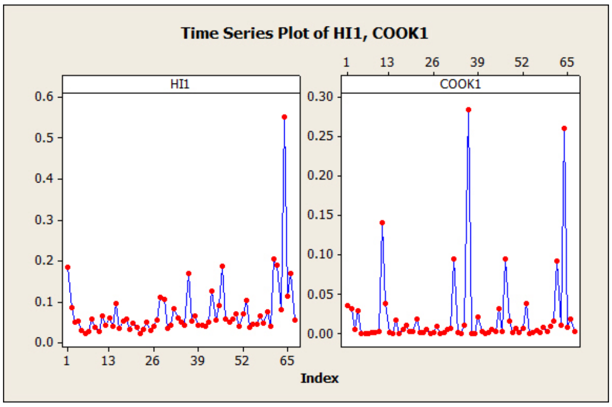
The regression equation is
y = 14360 + 0.105 x1 - 9.22 x3 + 12.4 x4

Predictor	Coef	SE Coef	T	P	VIF
Constant	14359.7	733.3	19.58	0.000	
x1	0.10515	0.01071	9.82	0.000	1.727
x3	-9.2226	0.7869	-11.72	0.000	3.570
x4	12.426	2.071	6.00	0.000	2.551

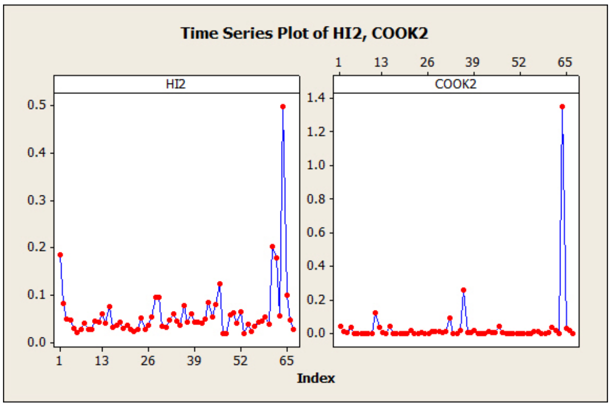
S = 464.980   R-Sq = 91.9%   R-Sq(adj) = 91.5%

Plots of leverage values and Cook's distance values for both models are given below.

(f) Briefly comment on the impact of removing X_2 from the full model on the potentially influential observations. [4]



Model including all four explanatory variables.
$F_{0.5;4,62} = 0.8484$



Model without X_2.
$F_{0.5;3,63} = 0.7973$

Question 4 [26 marks] Consider the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y}$ is an n-dimensional vector of response variables, $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown, constant parameters and $\boldsymbol{X}$ is an $(n \times p)$-dimensional design matrix.

We define the vector of residuals as $\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$, where $\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$. Show the following properties of matrix $\boldsymbol{H}$:

(a) $\boldsymbol{H}$ is symmetric and $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$, [3]

(b) $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{I} - \boldsymbol{H}$. [2]

Furthermore, show the following properties of the vector of residuals:

(c) $\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$, [3]

(d) $\mathrm{Var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$, [3]

(e) $SS_E = \boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, where $SS_E = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$, [3]

(f) $\mathrm{E}(SS_E) = (n - p)\sigma^2$, [10]

(g) $\mathrm{E}(MS_E) = \sigma^2$. [2]

End of Paper