Queen Mary
University of London

# B. Sc. Examination by course unit 2015

# MTH 5120: Statistical Modelling I

## Duration: 2 hours

## Date and time: 30.04.2015, 14.30-16.30

---

**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 unauthorised 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 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.**

**Examiner(s): B. Bogacka and L. I. Pettit**

---

**Question 1. (25 marks)**

(a) Show that the Least Squares Estimator $\widehat{\beta}$ of the parameter $\beta$ in the no-intercept model $Y_i = \beta x_i + \varepsilon_i$, where the random errors $\varepsilon_i$ are identically, independently normally distributed with zero mean and a constant variance $\sigma^2$, is

$$\widehat{\beta} = \frac{1}{a}\sum_{i=1}^{n} Y_i x_i, \qquad \text{where} \qquad a = \sum_{i=1}^{n} x_i^2.$$

[**8**]

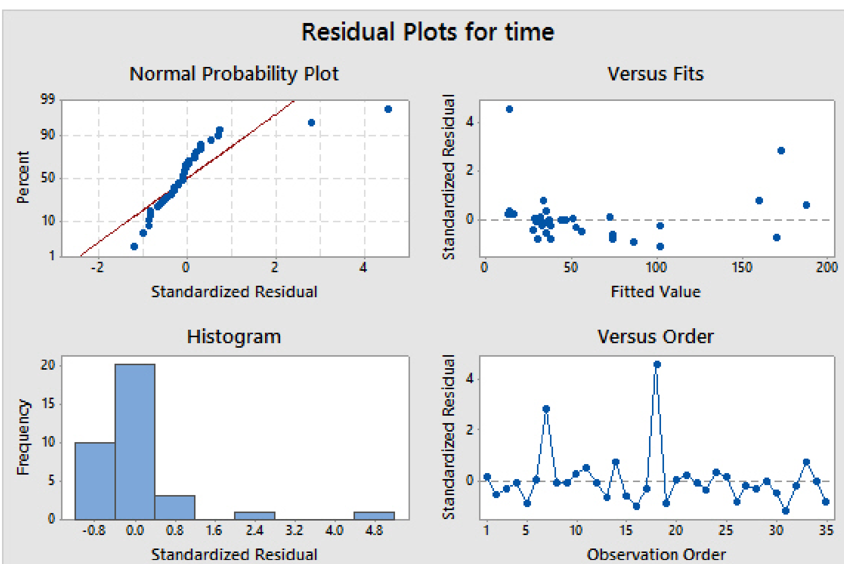
(b) Obtain the distribution of $\widehat{\beta}$ including the mean and the variance of the estimator. [**12**]

(c) Assuming that $\sigma^2$ is known, give a statistic and its distribution for testing the hypothesis $H_0 : \beta = 0$ versus the alternative $H_1 : \beta \neq 0$. [**5**]

**Question 2. (25 marks)**

The *winning times* (Y, [minutes]) in 1984 for 35 Scottish hill races were collected together with the *distance on the map* ($X_1$, [miles]) and the *total height gained during the route* ($X_2$, [feet]). A multiple linear regression analysis was performed and the results are given below.

(a) Briefly comment on the standardized residual plots. [**4**]



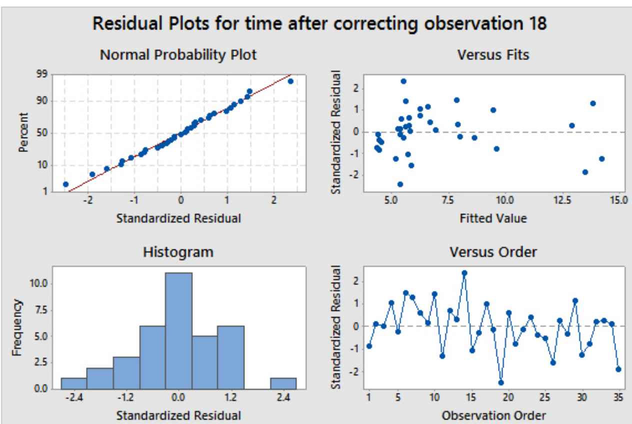

Question 2 continues on the next page.

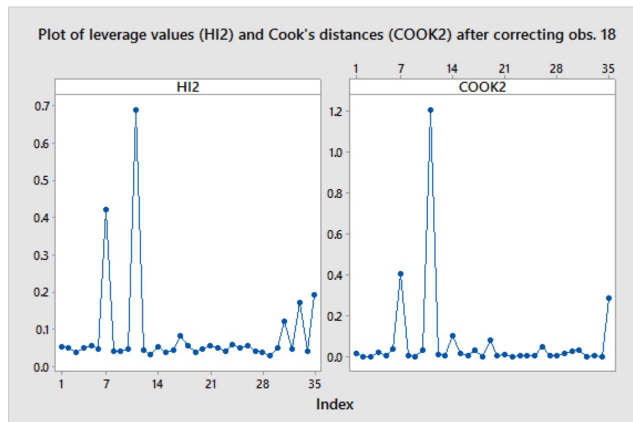
(b) It occurred that observation 18 was wrongly typed in. After correcting the mistake a new regression analysis was performed on the response transformed by power transformation with $\lambda = 0.5$. That is, the assumed model for the independent response variables $Y_i$ was

$$Y_i^{\lambda} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

(i) Briefly comment on the new plots of standardized residuals regarding the assumptions of normality, constant variance and linearity of the model. **[3]**

(ii) Based on the figures shown below, briefly comment on the unusual observations. **[6]**




Analysis of Variance for Transformed Response

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 2 | 261.250 | 130.625 | 553.01 | 0.000 |
| x1 | 1 | 71.003 | 71.003 | 300.60 | 0.000 |
| x2 | 1 | 24.573 | 24.573 | 104.03 | 0.000 |
| Error | 32 | 7.559 | 0.236 | | |
| Lack-of-Fit | 31 | 7.545 | 0.243 | 18.43 | 0.183 |
| Pure Error | 1 | 0.013 | 0.013 | | |
| Total | 34 | 268.809 | | | |

```
Regression Equation
time^0.5 = 3.107 + 0.3452 x1 + 0.000693 x2
```

(c) Based on the numerical output shown above do the following:

(i) Test the hypothesis regarding significance of the regression. Give the formula of the test statistic and explain your notation. **[8]**

(ii) Obtain an estimate of the expected record in a Scottish hill race where the distance on the map is 5 miles and the total height gained during the route is 1000 feet. **[2]**

(iii) Interpret in practical terms the meaning of the estimate of $\beta_1$ for a given total height gained during the route. **[2]**

**Question 3. (25 marks)**
Technicians measure the total heat flux as part of a solar thermal energy test. An energy engineer wants to determine how total heat flux ($Y$) is predicted by other variables: insolation ($X_1$), the position of the focal points in the east ($X_2$), south ($X_3$), and north ($X_4$) directions, and the time of day ($X_5$). The best subset regression was performed in Minitab and the results are shown below.

```
Best Subsets Regression: Heat Flux versus Insolation, East, ...

Response is Heat Flux
```

| Vars | R-Sq | R-Sq (adj) | PRESS | R-Sq (pred) | Mallows Cp | S | Insolation | East | South | North | Time of Day |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 72.1 | 71.0 | 4855.9 | 66.9 | 38.5 | 12.328 | | | | X | |
| 1 | 39.4 | 37.1 | 10822.6 | 26.3 | 112.7 | 18.154 | X | | | | |
| 2 | 85.9 | 84.8 | 2736.5 | 81.4 | 9.1 | 8.9321 | | | X | X | |
| 2 | 82.0 | 80.6 | 3786.4 | 74.2 | 17.8 | 10.076 | | | | X | X |
| 3 | 87.4 | 85.9 | 3089.7 | 79.0 | 7.6 | 8.5978 | | X | X | X | |
| 3 | 86.5 | 84.9 | 2725.9 | 81.4 | 9.7 | 8.9110 | X | | X | X | |
| 4 | 89.1 | 87.3 | 2847.2 | 80.6 | 5.8 | 8.1698 | X | X | X | X | |
| 4 | 88.0 | 86.0 | 3045.7 | 79.3 | 8.2 | 8.5550 | X | | X | X | X |
| 5 | 89.9 | 87.7 | 3109.9 | 78.8 | 6.0 | 8.0390 | X | X | X | X | X |

(a) Briefly explain the meaning of all the columns in the above numerical output. Give the formulae for the statistics used. [**8**]

(b) Based on the information in this output suggest the best, from the point of view of prediction, parsimonious subset of the explanatory variables. Briefly justify your choice. [**4**]

Question 3 continues on the next page.

A regression analysis for the full model was performed and a part of the Minitab numerical output is given below.

```
Regression Analysis: Heat Flux versus Insolation, East, South, North, Time of Day

Analysis of Variance
Source         DF   Seq SS  Contribution   Adj SS   Seq MS  F-Value  P-Value
Regression      5  13195.5        89.88%  13195.5  2639.11    40.84    0.000
  Insolation    1   5783.8        39.39%    350.6  5783.78    89.50    0.000
  East          1    811.7         5.53%    270.1   811.72    12.56    0.002
  South         1   1181.5         8.05%    437.2  1181.51    18.28    0.000
  North         1   5303.0        36.12%   4656.6  5303.03    82.06    0.000
  Time of Day   1    115.5         0.79%    115.5   115.50     1.79    0.194
Error          23   1486.4        10.12%   1486.4    64.63
Total          28  14681.9       100.00%

Tests use the sequential sums of squares

Model Summary
      S    R-sq  R-sq(adj)    PRESS  R-sq(pred)
8.03902  89.88%     87.68%  3109.95      78.82%

Coefficients
Term            Coef  SE Coef        95% CI       T-Value  P-Value   VIF
Constant       325.4     96.1  ( 126.6,  524.3)      3.39    0.003
Insolation    0.0675   0.0290  (0.0075, 0.1275)      2.33    0.029  2.32
East            2.55     1.25  ( -0.03,   5.13)      2.04    0.053  1.36
South           3.80     1.46  (  0.78,   6.82)      2.60    0.016  3.18
North         -22.95     2.70  (-28.54, -17.36)     -8.49    0.000  2.61
Time of Day     2.42     1.81  ( -1.32,   6.16)      1.34    0.194  5.37

Regression Equation
Heat Flux = 325.4 +0.0675Insolation +2.55East +3.80South -22.95North
            +2.42TimeofDay
```

(c) State the null and the alternative hypotheses regarding significance of regression. Based on the information shown in the ANOVA table test the null hypothesis. [**4**]

(d) State the null hypotheses for the coefficients of the explanatory variables which are tested in the ANOVA table. [**3**]

(e) State the null hypotheses for the coefficients of the explanatory variables which are tested in the table of Coefficients. [**3**]

(f) Explain how you could improve the model fit. [**3**]

**Question 4. (25 marks)**

(a) For each of the following regression models, indicate whether it is a linear regression model (in the parameters $\beta$). If it is not, state whether it can be linearized by a suitable transformation of the response and write down the transformed model. [**8**]

(i) $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \log_{10} x_{2i} + \beta_3 x_{1i}^2 + \varepsilon_i$

(ii) $Y_i = \varepsilon_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$

(iii) $Y_i = \beta_0 \exp(\beta_1 x_{1i}) + \varepsilon_i$

(iv) $Y_i = \{1 + \exp(\beta_0 + \beta_1 x_{1i} + \varepsilon_i)\}^{-1}$

(b) Consider the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y}$ denotes the $n \times 1$ vector of responses, $\boldsymbol{X}$ denotes the $n \times p$ design matrix, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown parameters and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of uncorrelated random errors with zero mean and constant variance $\sigma^2$.

(i) Show that $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ denotes the least squares estimator of $\boldsymbol{\beta}$, is an unbiased estimator of the expectation of $\boldsymbol{Y}$. [**6**]

(ii) Obtain the variance-covariance matrix of $\widehat{\boldsymbol{\mu}}$. [**6**]

(iii) State the distribution of $\widehat{\boldsymbol{\mu}}$. [**5**]

---

**End of Paper.**