Main Examination period 2020 – May/June – Semester B
Online Alternative Assessments

MTH5120: Statistical Modelling I

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

In completing this assessment, you may use books, notes, and the internet. You may use calculators and computers, but you should show your working for any calculations you do. You must not seek or obtain help from anyone else.

At the start of your work, please **copy out and sign** the following declaration:

> I declare that my submission is entirely my own, and I have not sought or obtained help from anyone else.

All work should be **handwritten**, and should **include your student number**.

You have **24 hours** in which to complete and submit this assessment. When you have finished your work:

- scan your work, convert it to a **single PDF file** and upload this using the upload tool on the QMplus page for the module;
- e-mail a copy to maths@qmul.ac.uk with your student number and the module code in the subject line;
- with your e-mail, include a photograph of the first page of your work together with either yourself or your student ID card.

You are not expected to spend a long time working on this assessment. We expect you to spend about **2 hours** to complete the assessment, plus the time taken to scan and upload your work. Please try to upload your work well before the end of the assessment period, in case you experience computer problems. **Only one attempt is allowed – once you have submitted your work, it is final**.

IFoA exemptions
This module counts towards IFoA actuarial exemptions. For your submission to be eligible for IFoA exemptions, you must submit within the first **3 hours** of the assessment period. You may then submit a second version later in the assessment period if you wish, which will count only towards your degree. There are two separate upload tools on the QMplus page to enable you to submit a second version of your work.

Examiners: L I Pettit, D S Coad

Question 1 [15 marks].
You should answer this question using a calculator. You should show all working.

(a) For the data given below find the values of $\bar{x}$, $\bar{y}$, S_{xx}, S_{yy} and S_{xy}. **[5]**

x	7	6	5	1	5	4	7	3	4
y	97	86	78	10	75	62	101	39	y_9

Note: The value of y_9 is equal to 43 plus the sum of the seventh and ninth digits of your student number.

(b) Hence find the least squares estimates of β_0 and β_1 in the simple linear regression model of y on x. **[3]**

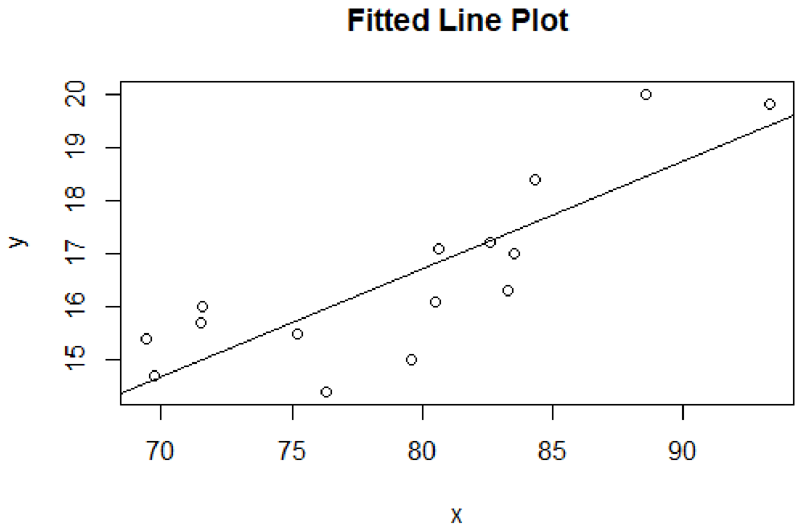
(c) Find the Analysis of Variance table to test the hypothesis that $\beta_1 = 0$ against a two sided alternative using a 5% significance level. **[7]**

You are given the following values from R:
`qf(0.95,1,6)= 5.99`, `qf(0.95,1,7)= 5.59`, `qf(0.95,1,8)= 5.32`, `qf(0.95,1,9)= 5.12`, `qf(0.95,2,6)= 5.14`, `qf(0.95,2,7)= 4.74`, `qf(0.95,2,8)= 4.46`.

Question 2 [34 marks]. Crickets are insects which make a characteristic chirping sound. It has been observed that the frequency of chirps seems to be related to the temperature. A biologist measured the number of chirps (y) per second a cricket made at various temperatures (x) measured in degrees Fahrenheit.

(a) State the assumptions made about the errors in a simple linear regression model. **[2]**

(b) The biologist fitted a simple linear regression model to the data and obtained the following plot. Comment on whether a simple linear regression model seems to fit the data. **[3]**

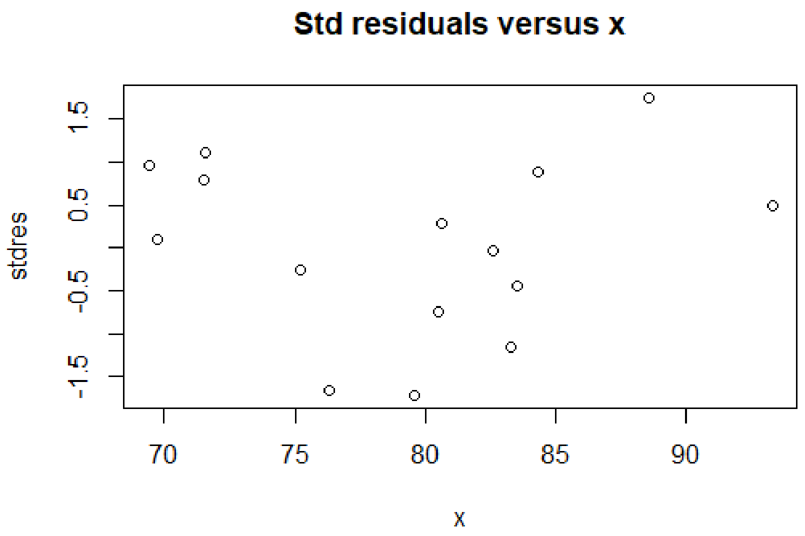


(c) The following output was obtained from R. Two quantities in the Table of Coefficients are recorded as A and B. Find their values. [**4**]

```
> crickets<-lm(y~x)
> summary(crickets)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.62746 -0.56464  0.08213  0.76762  1.54563

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46951    2.96747   A      0.876716
x            B          0.03727   5.447 0.000112 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9799 on 13 degrees of freedom
Multiple R-squared:  0.6953,Adjusted R-squared:  0.6719
F-statistic: 29.67 on 1 and 13 DF,  p-value: 0.0001119
```

(d) Two t tests can be carried out using this output. For each write down the null and alternative hypotheses, the distribution of the test statistic if H_0 is true and state your conclusions. [**8**]

(e) The biologist looked at the plot of standardised residuals versus x.



(i) Comment on this plot. [**3**]

(ii) The biologist obtained the following output. Explain what assumption he is checking and the conclusion. What graph could he have looked at to examine this assumption? [**4**]

```
> shapiro.test(stdres)

Shapiro-Wilk normality test

data:  stdres
W = 0.96628, p-value = 0.7996
```

(iii) The biologist is interested in predicting the mean number of chirps per second when the temperature is 85 degrees Fahrenheit. He obtains the following output. Explain what it shows. [**5**]

```
> pred1 <- predict(crickets, newdata=data.frame(x=85), interval='confidence')
> pred1
       fit      lwr     upr
1 17.72361 17.01161 18.4356
```

(iv) The biologist wonders if it would be possible to predict the temperature based on the number of chirps. Discuss how you could do that. Looking at the data comment on how good the prediction would be. [**5**]

Question 3 [19 marks]. 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, the formula for the least squares estimator $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

(a) Write the regression model

$$Y_i = \beta_1 x_i + \beta_2 z_i + \varepsilon_i, \qquad i = 1, 2, \ldots, 5,$$

where the ε_i have mean zero, variance σ^2 and are uncorrelated, as a general linear model in matrix form by specifying $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$. [**5**]

(b) Find expressions for the least squares estimators of β_1 and β_2,

(i) by minimising

$$S(\beta_1, \beta_2) = \sum_{i=1}^{5}\{Y_i - (\beta_1 x_i + \beta_2 z_i)\}^2,$$

[**6**]

(ii) by using the formula for $\hat{\boldsymbol{\beta}}$ above. [**5**]

(c) The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Find $\text{Var}(\hat{\beta}_1)$ and $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$. [**3**]

Question 4 [10 marks].
A random variable X has probability density function

$$p(x) = \theta^3 x^2 \exp(-\theta x), \quad x > 0,$$

and zero otherwise. A random sample $x_1, x_2, \ldots, x_n$ is collected from this distribution.

(a) Find the likelihood function. [**3**]

(b) Hence find the maximum likelihood estimator $\hat{\theta}$. Check that you have found a maximum. [**7**]

Question 5 [7 marks]. For each of the following say if it is a linear model or not. If it is not a linear model say if it is linearisable. If it is give the linearised model.

(a) $Y_i = \exp(\beta_0 + \beta_1 x_i) + \varepsilon_i$ [**2**]

(b) $Y_i = \beta_0 + \beta_1 \sqrt{x_{1i}} + \beta_2 \cos(x_{2i}) + \varepsilon_i$ [**2**]

(c) $Y_i = 3 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i)$ [**3**]

Question 6 [15 marks].

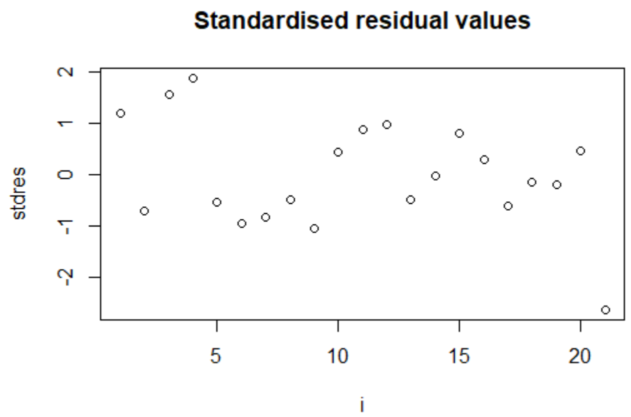
In a study of the efficiency of a plant which oxidises ammonia to nitric acid the dependent variable is stack loss and the independent variables are Airflow (flow of cooling air), Water.Temp (cooling water inlet temperature) and Acid.Conc (concentration of acid). The data were read into R and the commands and output are shown below.

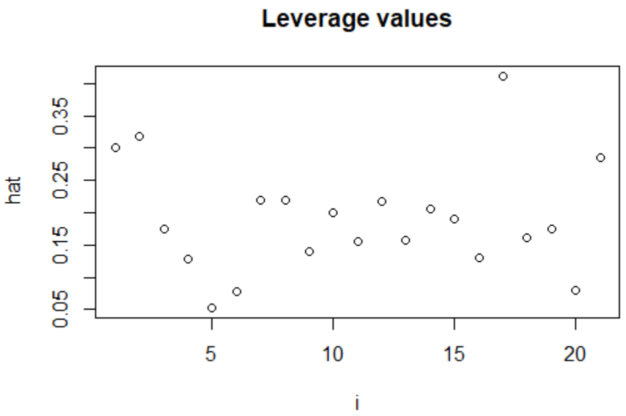
```
> stack <- lm(stack.loss ~ Airflow + Water.Temp + Acid.Conc)

> summary(stack)

Call:
lm(formula = stack.loss ~ Airflow + Water.Temp + Acid.Conc)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2377 -1.7117 -0.4551  2.3614  5.6978

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -39.9197    11.8960  -3.356  0.00375 **
Airflow             0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp          1.2953     0.3680   3.520  0.00263 **
Acid.Conc          -0.1521     0.1563  -0.973  0.34405
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared:  0.9136,Adjusted R-squared:  0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09

> stdres<-rstandard(stack)
> hat<- hatvalues(stack)
> cook<-cooks.distance(stack)
> i<- 1:21
> plot(i,stdres, main="Standardised residual values")
> plot(i,hat, main="Leverage values")
> plot(i,cook, main="Cooks distance values")
> qf(0.5, 4, 17)
[1] 0.8735735
```

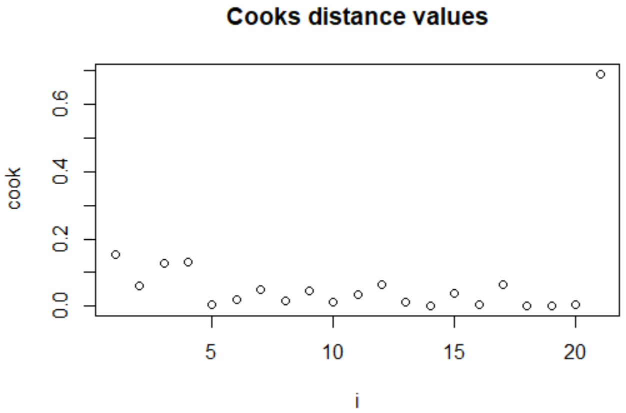
(a) Write down the fitted model. **[2]**

(b) Explain what is meant by an **outlier**, **leverage** and an **influential observation**. Include the relationship between these concepts and how they can be detected. **[8]**

(c) Comment on what the three plots on page 7 tell us about possible outliers, high leverage values and influential observations in these data. **[5]**


Standardised residual values
stdres
2
1
0
-1
-2
5
10
15
20
i


Leverage values
hat
0.05
0.15
0.25
0.35
5
10
15
20
i


Cooks distance values
cook
0.0
0.2
0.4
0.6
5
10
15
20
i

End of Paper.