Main Examination period 2019

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 unauthorised 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 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. 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, W Yoo

Question 1. [30 marks] A chemist studied the concentration of a solution (Y) over time (x). Fifteen identical solutions were prepared. The solutions were randomly divided into five sets of three, and the five sets were measured, respectively after 1, 3, 5, 7, and 9 hours.

(a) Without making any plots the chemist entered the data into R, fitted a simple linear regression model and then carried out a goodness of fit test. The following is the Analysis of Variance table he produced but with some figures missing.

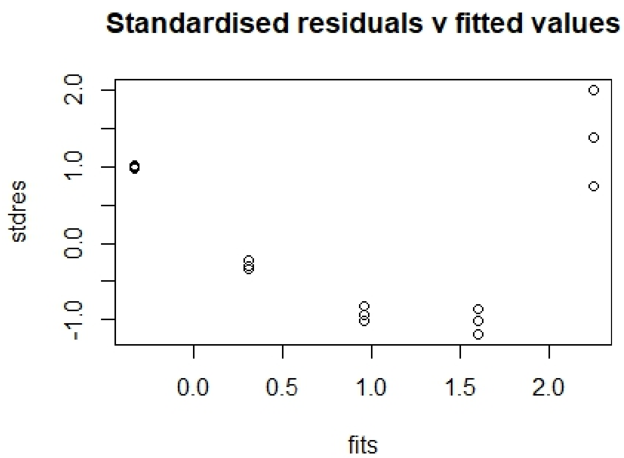
```
Analysis of Variance Table

Response: y
              Df  Sum Sq Mean Sq F value
x              1 12.5971
Residuals     13
  Lack of fit     2.770
  Pure error
Total         14 15.5218
```

(i) State the model for simple linear regression including the assumptions made about the errors. **[4]**

(ii) Copy and complete the Analysis of Variance Table. **[10]**

(iii) Carry out two possible F tests, write down the corresponding null hypotheses and state your conclusions. **[6]**

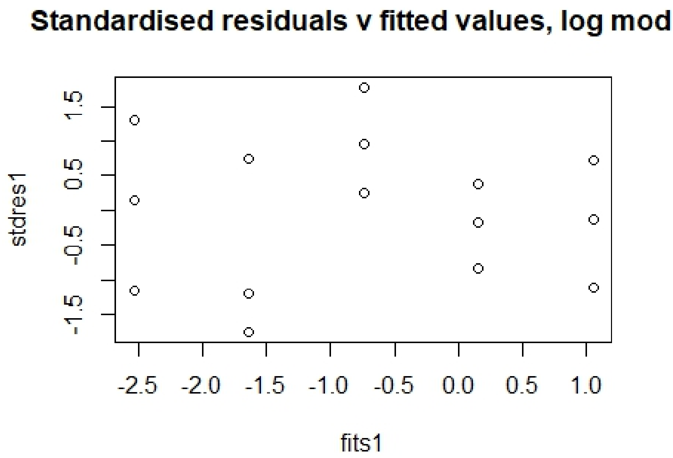
(b) The chemist decided to look at the plot of residuals versus fitted values.



(i) The chemist saw there were two possible model assumptions which were not satisfied from this plot. Identify these assumptions. **[2]**

(ii) The chemist decided to change the model by transforming the dependent variable from y to $\log_e y$. Explain why this choice was made. **[3]**

(iii) Having made this transformation the plot of residuals versus fitted values was plotted and is shown on the next page. Are the model assumptions now satisfied? **[2]**



(iv) What other plot or test would you advise the chemist to make and why? [**3**]

Question 2. [16 marks]
Consider the no intercept regression model

$$Y_i = \beta x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

where the usual assumptions are made about the errors.

(a) Show that the least squares estimator of β is given by

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}.$$

[**4**]

(b) Show that $\widehat{\beta}$ is an unbiased estimator of β. [**3**]

(c) Find the variance of $\widehat{\beta}$. [**4**]

(d) We define the residual $e_i = Y_i - \widehat{\beta} x_i$.
Show that $\mathrm{E}(e_i) = 0$ and

$$\mathrm{Var}(e_i) = \sigma^2 \left(1 - \frac{x_i^2}{\sum_j x_j^2} \right).$$

[**5**]

Question 3. **[32 marks]** A researcher wished to study the relationship between the annual salaries (Y in thousands of dollars) of 24 Mathematics Professors in a large American University and an index of publication quality (x_1), number of years of experience (x_2), an index of success in obtaining grants (x_3) and an index based on teaching evaluations (x_4). The data were read into R and the following commands and output were initially found.

```
salary<-lm(y~x1+x2+x3+x4)
summary(salary)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5891 -1.6925 -0.6017  2.5454  4.7078

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.54908    6.66329   6.236 5.47e-06 ***
x1           2.09307    0.65199   3.210 0.004607 **
x2           0.64761    0.07387   8.767 4.19e-08 ***
x3           2.78690    0.59594   4.676 0.000164 ***
x4          -2.18893    1.82959  -1.196 0.246255
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.455 on 19 degrees of freedom
Multiple R-squared:  0.917,Adjusted R-squared:  0.8995
F-statistic: 52.47 on 4 and 19 DF,  p-value: 5.234e-10
```

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

(ii) What null hypothesis and alternative does the output
`F-statistic:  52.47 on 4 and 19 DF, p-value:  5.234e-10`
test? What is the conclusion? **[4]**

(b) The following commands were then entered.

```
stdres <- rstandard(salary)
hat<-hatvalues(salary)
i<- 1:24
plot(i,hat, main="Hat values versus i, Salary")
shapiro.test(stdres)
```

Explain briefly the meaning of each command and what output it gives. **[9]**

(c) Look at the following output

```
> library(car)
> vif(salary)
      x1       x2       x3       x4
1.365795 1.324020 1.162684 1.052740
```

The researcher finds the vif values to investigate multicollinearity.

(i) What does vif stand for? [**1**]

(ii) What is multicollinearity and what are its effects? [**5**]

(iii) Is there any problem with multicollinearity here? Explain your answer. [**2**]

(d) Look at the following output

```
> library(leaps)
> best.subset <- regsubsets(y~x1+x2+x3+x4, salary, nvmax=4)
> best.subset.summary <- summary(best.subset)
> best.subset.summary$outmat
         x1  x2  x3  x4
1  ( 1 ) " " "*" " " " "
2  ( 1 ) " " "*" "*" " "
3  ( 1 ) "*" "*" "*" " "
4  ( 1 ) "*" "*" "*" "*"
> best.subset.summary$adjr2
[1] 0.7182713 0.8494270 0.8973512 0.8995185
```

(i) Define **adjusted** R^2. [**1**]

(ii) Explain briefly what this output shows. [**4**]

(e) Discuss, based on all the output above, whether the variable `x4` should be dropped from the model. [**4**]

Question 4. [22 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) Prove that the expectation of $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{\beta}$. [**4**]

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

(c) Show that the vector of fitted values is given by $\boldsymbol{HY}$ where $\boldsymbol{H}$ is the hat matrix which you should define. [**3**]

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

(e) Express the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i \qquad i = 1, 2, \ldots, 5$$

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 $\boldsymbol{\varepsilon}$. [**6**]

End of Paper.