Queen Mary
University of London

Late-Summer Examination period 2023

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.

The exam is intended to be completed within **2 hours**. However, you will have a period of **4 hours** to complete the exam and submit your solutions.

For actuarial students only: This module also counts towards IFoA exemptions. For your submission to be eligible, **you must submit within the first 3 hours.**

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

You are allowed to bring **three A4 sheets of paper** as notes for the exam.

Only approved non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

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: C.Sutton, L.Shaheen

Question 1 [16 marks]. Ten observations for an explanatory variable (x_i) and a response variable (y_i) where i = 1, 2, ..., 10 are given in the table below.

x_i	3.6	3.5	2.9	1.8	3.4	2.3	3.6	3.7	1.6	2.0
y_i	9.6	9.4	8.0	7.3	9.2	6.5	9.6	6.9	7.8	6.8

(a) State the three usual assumptions made in a simple linear regression model giving each in terms of the response variable. **[3]**

You are given that $\sum$x$_i$= 28.4, $\sum$y$_i$= 81.1, $\sum$x$_i y_i$= 236.2 and $\sum$x$_i^2$= 86.92.

(b) Find the least squares estimates for the intercept (β_0) and slope (β_1) parameters. **[6]**

You are given that the standard error of the slope parameter estimator is 0.4074 and that $t_{0.025;n-2} = 1.8595$.

(c) Calculate a 95% confidence interval for β_1. **[4]**

(d) What does your answer in (c) above say about the hypothesis H_0: $\beta_1 = 0$? **[2]**

(e) State another way in which the same hypothesis could be tested. **[1]**

Question 2 [20 marks]. A simple linear regression model is constructed from 32 observations where the mean of the explanatory variable observations is 3.7 and the mean of the response variable observations is 4.75. The following model is fitted

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

The least squares estimates of the parameters β_0 and β_1 are 9.2928 and -1.2269 respectively. The model R^2 is 63.07%.

(a) If the Residual Sum of Squares (SS_E) is 60.977 complete the Analysis of Variance Table for this model. **[9]**

(b) Use your table in (a) to estimate the variance of the residuals. **[1]**

A new x_i observation is taken with a value of 5. The value of the response that corresponds to this new x_i is not known.

(c) Calculate a 95% prediction interval for the new value of the response variable if S_{xx}= 69.18 and $t_{0.025;n-2} = 1.812$. **[7]**

(d) If another 16 (x_i, y_i) observations were obtained and the regression model re-run with these in addition to the initial 32 observations, what would be the impact on the width of prediction intervals such as those in (c) above? **[3]**

Question 3 [20 marks]. A simple linear regression model is fitted to some observation data which contains explanatory (x_i) and response (y_i) variables and then residuals are calculated for analysis.

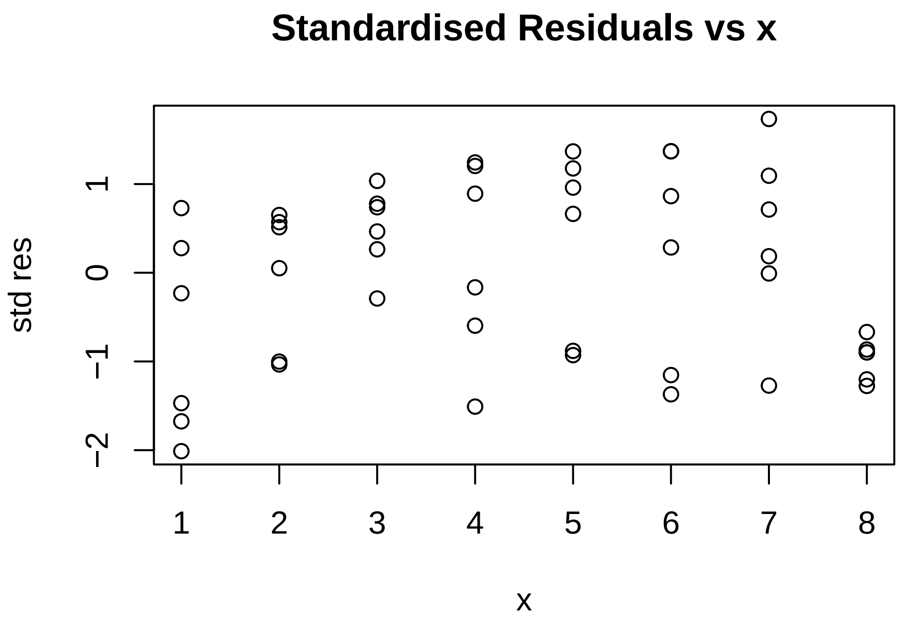
(a) Write down a formula for crude residuals in terms of observed and fitted values of the response. [**1**]

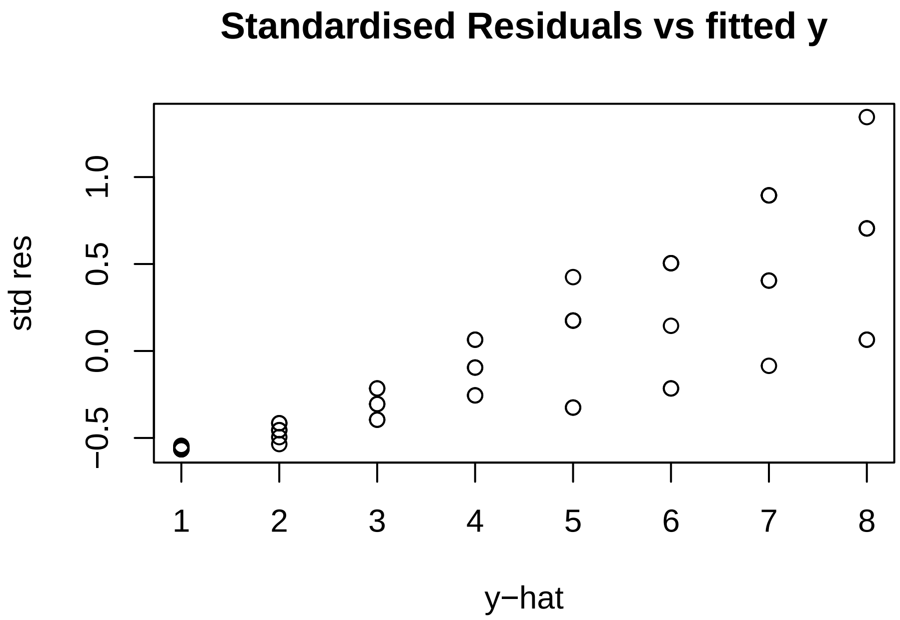
(b) Explain why we prefer to work with standardised residuals when analysing a simple linear regression model. [**3**]

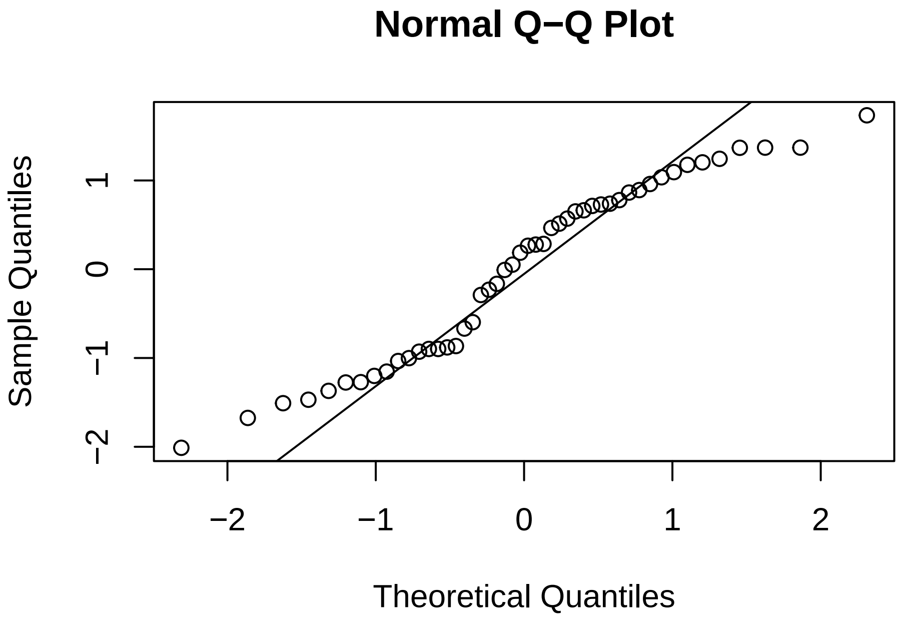
(c) Explain what each of the following lines of R code are doing. [**7**]

```
slr<-lm(y~ x)
di<-rstandard(slr)
yh<-fitted(slr)
plot(x, di)
plot(yh,di)
qqnorm(di)
qqline(di)
```

(d) Three residual plots from a simple linear regression model are shown below. What conclusions can we make about the model from these plots? You should explain what we are looking for in each plot and how that relates to the model. [**9**]




Standardised Residuals vs fitted y
std res
−0.5
0.0
0.5
1.0
1
2
3
4
5
6
7
8
y−hat


Normal Q−Q Plot
Sample Quantiles
−2
−1
0
1
−2
−1
0
1
2
Theoretical Quantiles

Question 4 [15 marks]. A simple linear regression model is fitted to 6 pairs of observations (x_i, y_i) where $i = 1, 2, 3, 4, 5, 6$. In matrix form the model is

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\epsilon}$$

(a) Write out each of the four vector or matrices with all of its components. **[4]**

(b) Explain why $\boldsymbol{X^T X}$ needs to be invertible for there to be a least squares estimator. **[3]**

(c) Show that $\boldsymbol{X^T X}$ is indeed invertible with the matrix $\boldsymbol{X}$ you wrote out in (a) above. **[5]**

Once the least squares estimator vector is found the fitted values can be written $\boldsymbol{HY}$.

(d) Write down a formula for $\boldsymbol{H}$ in terms of $\boldsymbol{X}$ and state two properties of $\boldsymbol{H}$. **[3]**

Question 5 [29 marks]. A company that owns a number of book shops across London wishes to model weekly total sales (y_i). The company director asks the manager of each book shop manager to a suggestion a quantifiable explanatory variable for sales. After dismissing duplicates data is collected for 11 explanatory variables and weekly sales across 16 shops for a period of 9 weeks. The explanatory variables are labelled x1, x2, . . . , x11. A multiple linear regression model is fitted.

(a) Write the lines of R code needed to compute and display the model plus its ANOVA table assuming observation data has already been assigned to explanatory (`x1, ..., x11`) and response (`y`) variables in R. **[3]**

The total sum of squares for the observed data is 15,890 and the residual sum of squares of the fitted model is 318.

(b) Calculate the Mean Squares for Regression and for Residuals and the Variance Ratio. **[5]**

(c) Under what conditions does the variance ratio in (b) follow a Fisher's F distribution? **[2]**

An analyst studies the results and suggests that a reduced model with 6 explanatory variables out of the original 11 would be more useful to the company.

(d) Explain why the company should consider reduced models. **[2]**

The 6 variable reduced model uses explanatory variables x1, x3, x5, x8, x9 and x11 from the original full model. The analyst wishes to compare the full and reduced model using the extra sum of squares principle and a F test.

(e) What is the total sum of squares of the reduced model? **[1]**

(f) List the steps needed to complete the comparison of the two models and the F test. You should include all the calculations you would make, what hypotheses are being tested and write out any R code needed to generate values needed to complete the test. [**12**]

The company director plans to give three manager-of-the-month prizes to the three shop managers whose explanatory variables improve the R^2 the most compared to a null model for sales. The analyst suggests using Adjusted R^2 instead.

(g) Write down the formula for Adjusted R^2 and explain the advantage of using this instead. [**4**]

End of Paper.