Queen Mary
University of London

Main Examination period 2021 – 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 and notes.**
- **You may use calculators and computers, but you must show your working for any calculations you do.**
- **You may use the Internet as a resource, but not to ask for the solution to an exam question or to copy any solution you find.**
- **You must not seek or obtain help from anyone else.**

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

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

- scan your work, convert it to a **single PDF file**, and submit this file using the tool below the link to the exam;
- 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 expected 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 submission window, in case you experience computer problems. **Only one attempt is allowed – once you have submitted your work, it is final**.

IFoA exemptions. For actuarial students, this module counts towards IFoA actuarial exemptions. You are allowed two submissions for this exam—the first for your IFoA mark, and the second for your module mark. To be eligible for IFoA exemptions, **your IFoA submission must be within the first 3 hours of the assessment period**.

Examiners: L. Rossini, D. S. Coad

 Continue to next page

Question 1 [20 marks]. A life assurance company is examining the force of mortality, μ_x, of a particular group of policyholders. It is thought that it is related to the age, x, of the policyholders by the formula:

$$\mu_x = Bc^x$$

It is decided to analyze this assumption by using the linear regression model

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \text{where} \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

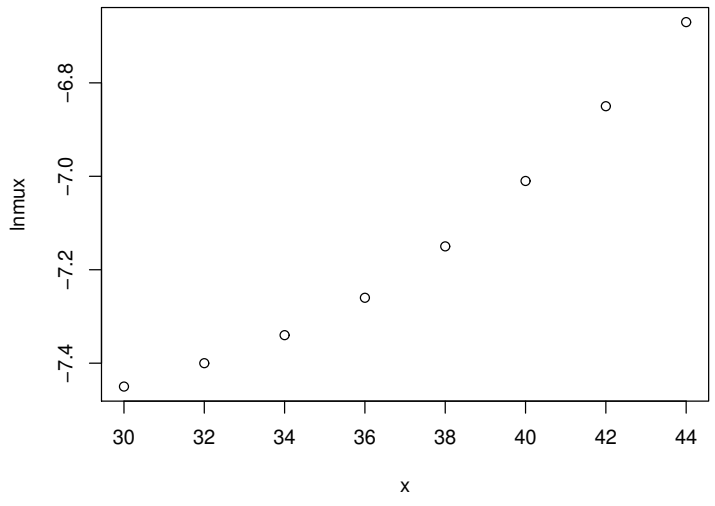
The results and summary statistics for eight ages were as follows where

Age, x	30	32	34	36	38	40	42	44
Force of mortality, μ_x ($\times 10^{-4}$)	5.84	6.1	6.48	7.05	7.87	9.03	10.56	12.66
$\ln \mu_x$	-7.45	-7.4	-7.34	-7.26	-7.15	-7.01	-6.85	-6.67

$$\sum_i x_i = 296 \qquad \sum_i x_i^2 = 11,120 \qquad \sum_i \ln \mu_{x_i} = -57.129$$
$$\sum_i (\ln \mu_{x_i})^2 = 408.5 \qquad \sum_i x_i \ln \mu_{x_i} = -2,104.5$$

(a) Apply a transformation to the original formula, $\mu_x = Bc^x$, to make it suitable for analysis by linear regression. Hence, write down expressions for Y, α and β in terms of μ_x, B and c. [**4**]

(b) Looking at the graph of $\log \mu_x$ against the age of the policyholder x, comment on the suitability of the regression model and state how this supports the transformation in part (a). [**2**]



 Continue to next page

(c) Use the data to calculate least squares estimates of B and c in the original formula. [**4**]

(d) Calculate the coefficient of determination between $\ln \mu_x$ and x. Hence comment on the fit of the model to the data. Is this value differ from the adjusted R^2? [**4**]

(e) Complete the table of residuals and use it to comment on the fit. [**2**]

Age, x	30	32	34	36	38	40	42	44
Residual, $\widehat{e}_i$	0.08		-0.03		-0.06		0.02	0.09

(f) Calculate a 95% confidence interval for the mean predicted response, $\ln \mu_{35}$, and hence obtain a 95% confidence interval for the mean predicted value of μ_{35}. [**4**]

Question 2 [28 marks]. The Journals data frame contains 180 observations (the journals) on 10 variables, among them the number of library subscriptions (`subs`), the library subscription price (`price`), and the total number of citations for the journal (`citations`). We make a logarithmic transformation of the variables. The goal is to estimate the effect of the subscription prices on the number of library subscriptions using the linear regression model

$$\log(\texttt{subs})_i = \beta_1 + \beta_2 \log(\texttt{price})_i + \varepsilon_i.$$

(a) We fit the model described above using R and the following output was obtained:

```
> jour_lm <- lm(log(subs) ~ log(price), data = Journals)
> summary(jour_lm)
Call:
lm(formula = log(subs) ~ log(price), data = Journals)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0527 -0.7058  0.1686  0.8705  1.8553

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.05518    0.48407  14.575  < 2e-16 ***
log(price)  -0.41088    0.08477  -4.847 2.72e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.059 on 178 degrees of freedom
Multiple R-squared:  0.1166,Adjusted R-squared:  0.1116
F-statistic: 23.49 on 1 and 178 DF,  p-value: 2.721e-06
```

Comment on the output produced by R. [**6**]

 Continue to next page

(b) Writing the slope parameter as β_2, what is the conclusion of a test of the null hypothesis $H_0 : \beta_2 = -0.6$ against a two sided alternative? [**4**]

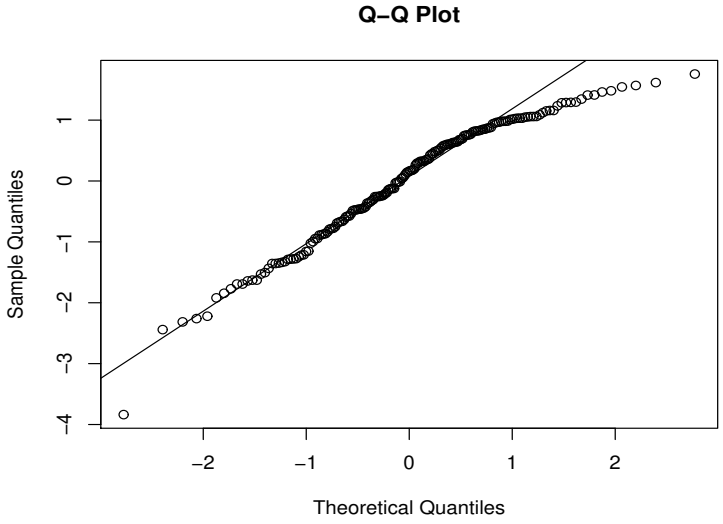
(c) Below is the Analysis of Variance Table that was produced with some figures missing.

```
Analysis of Variance Table
Response :   log(subs)
Predictor:   log(price)

                      Analysis of Variance Table
------------------------------------------------------------------------
                DF      Sum Sq     Mean Sq     F Value
------------------------------------------------------------------------
log(price)        1    26.34746
Residual        178
 Lack of fit              165.641
 Pure Error      33
 Total          179    225.9901
```

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

(d) A Q-Q plot of the standardized residuals is shown below. What assumption is this plot examining? What is your conclusion? What test could be carried out to check this assumption? [**5**]



(e) Find a 95% confidence interval for the intercept parameter β_1. [**3**]

Question 3 [20 marks]. Until 2010, the Minnesota Twins professional baseball team played its games in the Metrodome, an indoor stadium with a fabric roof. In addition to the large air fans required to keep the roof from collapsing, the baseball field is surrounded by ventilation fans that blow heated or cooled air into the stadium. Air is normally blown into the center of the field equally from all directions. To see if manipulating the fans could possibly make any difference, a group of students at the University of Minnesota and their professor built a "cannon" that used compressed air to shoot baseballs.
For the 96 observations, the possible regressor variables are: `Cond`, the condition, head or tail wind; `Velocity`, the actual velocity in feet per second; `Angle`, the actual angle; `BallWt`, the weight of the ball in grams used on that particular test; `BallDia`, the diameter in inches of the ball used on that test. The response variable considered is `Dist`, the distance in feet of the flight of the ball. To find the best fitting model the method of backward fitting was to be employed.

(a) Describe this method of fitting a multiple regression model as implemented in R, including a definition of the AIC. [**6**]

(b) The following R output was obtained. Which variables are dropped at each step and which retained in the final chosen model? [**2**]

```
> mymod <- lm(Dist ~ Velocity + Angle + BallWt + BallDia + Cond, data = domedata1)
> reduced.model <- step(mymod,direction="backward")
Start:  AIC=433.65
Dist ~ Velocity + Angle + BallWt + BallDia + Cond

           Df Sum of Sq   RSS    AIC
- BallWt    1         0  7758 431.65
- BallDia   1        17  7775 431.85
- Cond      1        75  7833 432.57
<none>                  7758 433.65
- Angle     1      4996 12754 479.37
- Velocity  1     69660 77418 652.49

Step:  AIC=431.65
Dist ~ Velocity + Angle + BallDia + Cond

           Df Sum of Sq   RSS    AIC
- BallDia   1        16  7775 429.85
- Cond      1        78  7837 430.62
<none>                  7758 431.65
- Angle     1      5211 12969 478.97
- Velocity  1     73431 81189 655.06

Step:  AIC=429.85
Dist ~ Velocity + Angle + Cond

           Df Sum of Sq   RSS    AIC
- Cond      1        80  7855 428.84
```

```
<none>                    7775 429.85
- Angle     1      5382 13157 478.36
- Velocity  1     73831 81606 653.55

Step:  AIC=428.84
Dist ~ Velocity + Angle

           Df Sum of Sq   RSS    AIC
<none>                    7855 428.84
- Angle     1      5478 13333 477.63
- Velocity  1     74045 81900 651.90
```

(c) Describe what the VIF is and interpret the following R command: [**5**]

```
> vif(mymod)
Velocity    Angle   BallWt  BallDia     Cond
1.500821 1.510233 1.290144 1.076135 1.029261
```

(d) For the chosen model you wish to check the assumptions. Say what plots you would look at and why. [**4**]

(e) In this model, we add a quadratic term for the variable `Angle` and we run the linear regression. The plots in Figure 1 are obtained. Comment on the overall fit of this model. [**3**]

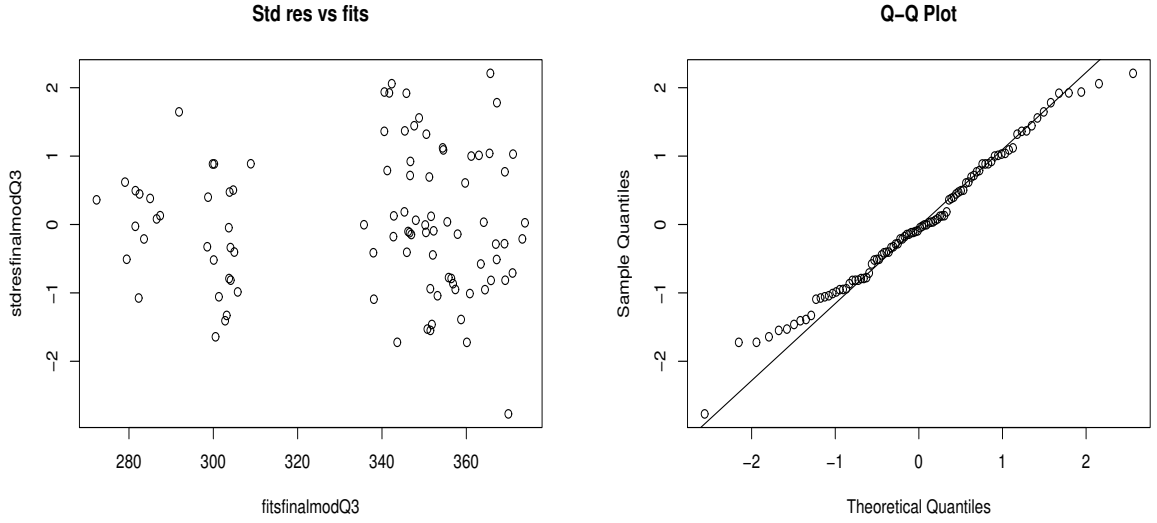


Figure 1: Plot of standardized residuals versus fitted values (left) and Q-Q plot (right) for the model with three explanatory variables.

Question 4 [17 marks]. We have the data for cigarette consumption for 46 US States for the year 1992 and we are interested in the relationship between the logarithm of cigarette consumption (in packs) per person of smoking age (> 16 years), the so-called $\mathbf{Y}$, the logarithm of real price of cigarettes in each state, $\mathbf{X}_1$, and the logarithm of real disposable income (per capita) in each state, $\mathbf{X}_2$. Data were collected for the 46 US States and the following computations for a multiple regression analysis of the model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

were obtained:

$$(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \begin{pmatrix} 30.930 & 4.811 & -6.679 \\ 4.811 & 3.945 & -1.177 \\ -6.679 & -1.177 & 1.449 \end{pmatrix}, \qquad \mathbf{X}^\mathsf{T}\mathbf{Y} = \begin{pmatrix} 223.001 \\ 45.428 \\ 1064.724. \end{pmatrix}$$

Also $\mathbf{Y}^\mathsf{T}\mathbf{Y} = 1082.723$ and $\bar{Y} = 4.848$ were computed.

(a) Find the least squares estimates $\widehat{\boldsymbol{\beta}}$ and hence write down the fitted model. [**4**]

(b) Use the results to construct the Analysis of Variance Table. [**9**]

(c) Describe the conditions for a high leverage and a very high leverage. Comment on the following output: [**4**]

$$\begin{aligned}\text{diag}(\mathbf{H}) = (&0.048, 0.031, 0.085, 0.098, 0.135, 0.033, 0.110, 0.043, 0.040, 0.041,\\ &0.040, 0.065, 0.031, 0.022, 0.198, 0.076, 0.055, 0.122, 0.086, 0.024, 0.086, 0.088,\\ &0.079, 0.031, 0.024, 0.065, 0.131, 0.139, 0.064, 0.089, 0.029, 0.042, 0.050, 0.026,\\ &0.050, 0.073, 0.040, 0.029, 0.055, 0.086, 0.024, 0.077, 0.087, 0.071, 0.025, 0.056).\end{aligned}$$

Question 5 [15 marks].

(a) Write the linear regression model

$$y_i = \beta_1 x_i + \beta_2 z_i^2 + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, as a general linear model. Thus identify the quantities $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$. [**5**]

(b) Find the least squares estimators of β_1 and β_2. [**6**]

(c) The hat matrix, $\mathbf{H}$, is defined as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}.$$

Show that it is symmetric and idempotent. [**4**]

End of Paper.