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, S. Coad

 Continue to next page

Question 1 [20 marks]. This question is similar to those on exercise sheets.

(a) Taking log of the original expression gives:

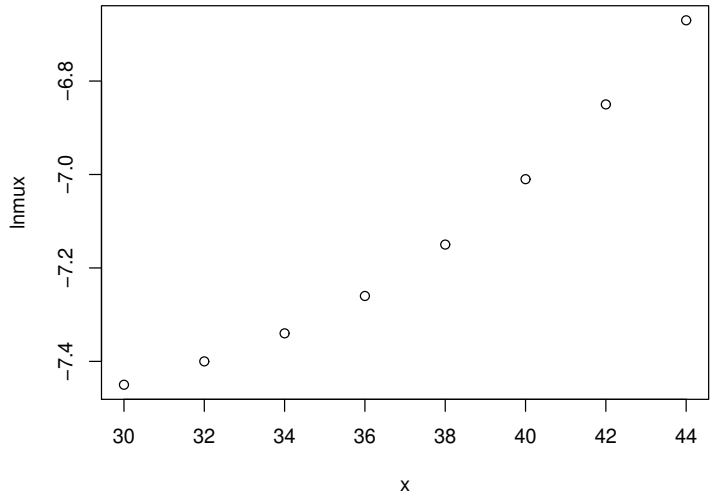
$$\log \mu_x = \log B + x \log c$$

This expression is now linear in x- Comparing the expression with $Y = \alpha + \beta x$ gives:

$$Y = \log \mu_x, \qquad \alpha = \log B, \qquad \beta = \log c$$

[**4**]

(b) The graphs appears to show an approximately linear relationship and this support the logarithmic transformation. However, it does appear to have a slight curve and this would warrant closer inspection of the model to see if it is appropriate for the data. [**2**]



(c) Obtaining the estimates of α and β using

$$S_{xx} = \sum x_i^2 - n\bar{x}^2 = 11,120 - 8\left(\frac{296}{8}\right)^2 = 168$$

$$S_{xy} = \sum xy - n\bar{x}\bar{y} = -2,104.5 - 8\left(\frac{296}{8}\right)\left(\frac{-57.129}{8}\right) = 9.273$$

Thus, the estimates are

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{9.273}{168} = 0.055196$$

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x} = \frac{-57.129}{8} - 0.0055196 \times \frac{296}{8} = -9.1834$$

 Continue to next page

Therefore, we obtain

$$B = e^{\alpha} = e^{-9.1834} = 0.000103$$
$$C = e^{\beta} = e^{0.055196} = 1.06$$

[4]

(d) The coefficient of determination is given by:

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{9.273^2}{168 \times 0.53467} = 95.7\%$$

where $S_{yy} = \sum y^2 - n\bar{y}^2 = 408.5 - 8 \times \left(\frac{-57.129}{8}\right)^2 = 0.53467$.
This tell us that 95.7% of the variation in the data can be explained by the model and so indicates an extremely good overall fit of the model. Obviously the value of the R^2 is different from the adjusted R^2, since the adjusted R^2 is equal to

$$R^2(adj) = \left(1 - (n-1)\frac{MS_E}{SS_T}\right) \neq \left(1 - \frac{SS_E}{SS_T}\right) = R^2$$

In particular, the adjusted R^2 takes into account the number of predictors in the model and can be useful for comparing models with different numbers of parameters. [4]

(e) The completed table of residuals using $\widehat{e}_i = y_i - \widehat{y}_i$ is Since the residuals for age

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

32, 36 and 40 are respectively computed as

$$\widehat{e}_{32} = -7.40 - (-9.1834 + 0.055196 \times 32) = 0.02$$
$$\widehat{e}_{36} = -7.26 - (-9.1834 + 0.055196 \times 36) = -0.06$$
$$\widehat{e}_{40} = -7.01 - (-9.1834 + 0.055196 \times 40) = -0.03$$

The residuals should be patternless when plotted against x, however it is clear to see that some pattern exists - this indicates that the linear model is not a good fit and that there is some other variable at work here. [2]

(f) Using the formula on lecture notes, the variance of the mean predicted response is:

$$\left\{\frac{1}{n} + \frac{(x_0 - \bar{x}^2}{S_{xx}}\right\}\widehat{\sigma}^2 = \left\{\frac{1}{8} + \frac{(35-37)^2}{168}\right\} \times 0.0038056 = 0.0005663$$

where $\widehat{\sigma}^2 = \frac{1}{6}\left(0.53467 - \frac{9.273^2}{168}\right) = 0.0038056$

The estimate is $Y = \log \mu_{35} = -9.1834 + 0.055196 \times 35 = -7.251$. Using the t_6 distribution, a 95% confidence interval for $Y = \ln \mu_{35}$ is

$$-7.251 \pm 2.447\sqrt{0.0005663} = (-7.309, -7.193)$$

The corresponding 95% confidence interval for μ_{35} is

$$(0.000669, 0.000752)$$

[**4**]

Question 2 [28 marks]. This question is similar to examples in the lecture notes

(a) Looking at the R output and relative summary command, we obtain that the intercept of the fitted line is 7.05518, with a standard error of 0.48407 and the estimated slope is -0.41088 with a standard error of 0.08477. In the table, we have also the t statistics and the p values corresponding to individual tests of the hypothesis "true coefficient equals to 0". Here, both the p-values are tiny, indicating that the regressors explains a substantial part of the variation in the data (thus statistically significant) and the intercept is significant different from zero at any reasonable level. Looking at the R^2, we have that the model explains only 11% of the variation in the data, thus the model seems not to be the best model to be considered.

Finally, the F statistic corresponds to an F test of the hypothesis that all regressors (excluding the intercept term) are jointly significant. Here with a single regressor, the p-value is of course identical to that of the t test for the coefficient `log(price)` [**6**]

(b) We need to run a t test for the null hypothesis $H_0 : \beta_2 = -0.6$ against $H_1 : \beta_2 \neq -0.6$. Thus we have the following t value:

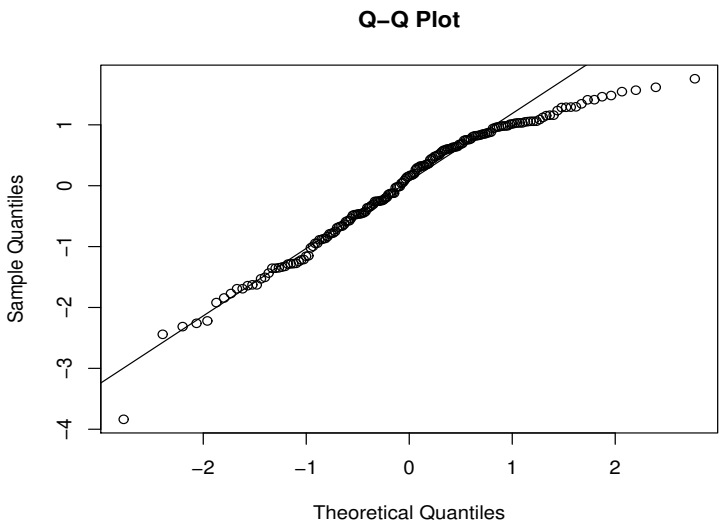
$$t = \frac{-0.41088 + 0.6}{0.08477} = \frac{0.18912}{0.08477} = 2.230978$$

This value in absolute value is bigger than the critical value. Thus we reject the null hypothesis $H_0 : \beta_2 = -0.6$. [**4**]

(c) We can complete the ANOVA Table by using the equations available in the lecture notes [**10**]

Source	df	SS	MS	F
`log(price)`	1	26.347	26.347	25.571
Residual	178	199.643	1.122	
Lack of fit	145	165.641	1.142	1.109
Pure Error	33	34.002	1.030	
Total	179	225.990		

 Continue to next page



(d) The Q-Q plot of the standardized residuals is used to check the normality assumption. In this case, we can see that the plot has some heavy tails in the upper part of the plot, thus we should reject the normality assumption. As a further test to check the normality assumption, one can run the Shapiro-Wilk test and if the p-value is smaller than the 5% significance level, we can reject the null hypothesis. [**5**]

(e) The 95% confidence interval for the intercept parameter β_1 is equal to

$$\left[\widehat{\beta}_1 - t_{n-2,\frac{\alpha}{2}}\widehat{se(\widehat{\beta}_1)}, \widehat{\beta}_1 + t_{n-2,\frac{\alpha}{2}}\widehat{se(\widehat{\beta}_1)}\right]$$

$$\left[7.05518 - t_{178,\frac{\alpha}{2}} \times 0.48407, 7.05518 + t_{178,\frac{\alpha}{2}} \times 0.48407\right]$$

$$[7.05518 - 1.973 \times 0.48407, 7.05518 + 1.973 \times 0.48407]$$

$$[6.10011, 8.01025]$$

[**3**]

Question 3 [20 marks]. This question is similar to examples in the lecture notes.

(a) First of all, we define the AIC criterion, which is

$$AIC = 2(p+1) - 2\log L$$

where $p-1$ is the number of regressor variables and $\log L$ is the log likelihood. We want to minimize the AIC. We start with the full model and we compare the AIC for the full model to AIC for each model dropping one regressor. We find the model with the minimum value of the AIC. If this is the full model, then we stop. Otherwise, we start form the new best model and we repeat it. [**6**]

(b) At step 1, we omit variable `BallWt`. At the second step, we omit variable `BallDia`, while at the third step, we omit variable `Cond`. Thus the final model includes only the variable `Velocity` and `Angle`. [**2**]

(c) The variance inflation factor (VIF) can be used to indicate when multicollinearity may be a problem. Consider a regression problem with $p-1$ regressors. Suppose we fitted a regression model with x_j as the dependent variable and the remaining $p-2$ variables as the regressor variables. Let R_j^2 be the coefficient of determination (not expressed as a percentage) for this model. Then we define the j-th variance inflation factor as

$$VIF_j = \frac{1}{1-R_j^2}$$

Moreover, a VIF bigger than 10 indicate that the multicollinearity may cause problems. Looking at our results, we have the VIF is smaller than 2 for all the variables, thus we do not have collinearity problems for all the variables. [**5**]

(d) We should check the assumption of

- constant variance by using the plot of the standardized residuals versus the fitted values;
- non-linear terms by using the plot of the standardized residuals versus `Velocity` and `Angle` separately;
- normality by using the Q-Q plot of the standardized residuals.

[**4**]

(e) The plots on the left shows some indication that the variance is increasing and moreover we have a few influential points and we should check for the outliers, thus through the leverage or Cook's distance plots. Moving to the right plot, we have some problems in both tails, but not a presence of heavy tails. Also in this case, we should better check the Shapiro-Wilk test in order to check correctly the normality assumption. [**3**]

 Continue to next page

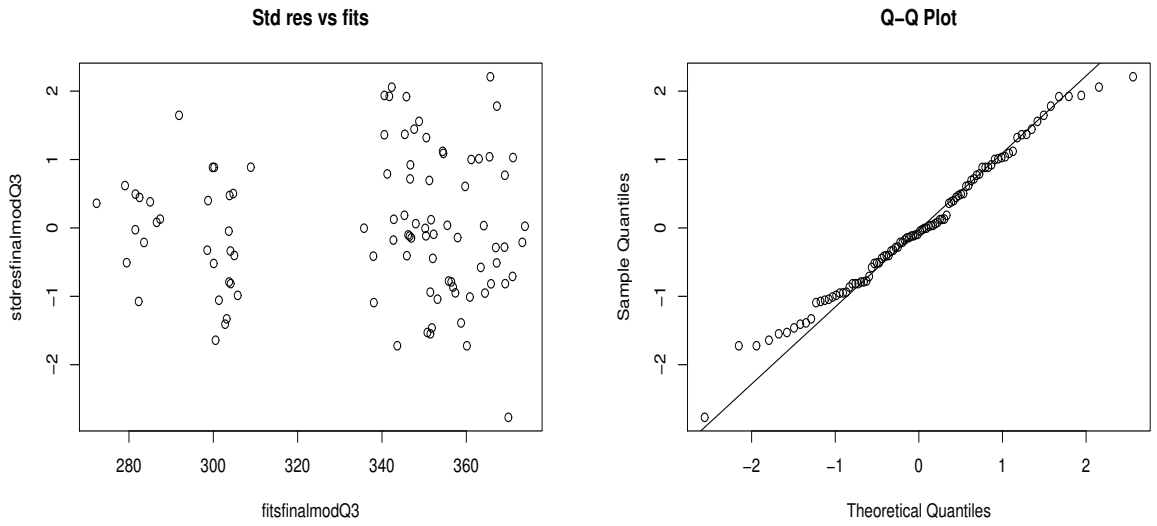


Figure 1: Plot of standardized residuals versus fitted values (left) and QQ plot (right) for the model with three explanatory variables.

Question 4 [20 marks]. This question is similar to examples in the lecture notes.

(a) The least square estimates $\widehat{\boldsymbol{\beta}}$ is written as the product of two different matrices:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y} \\ &= \begin{pmatrix} 30.930 & 4.811 & -6.679 \\ 4.811 & 3.945 & -1.177 \\ -6.679 & -1.177 & 1.449 \end{pmatrix}\begin{pmatrix} 223.001 \\ 45.428 \\ 1064.724 \end{pmatrix} = \begin{pmatrix} 4.300 \\ -1.338 \\ 0.172 \end{pmatrix}\end{aligned}$$

Thus the fitted model is equal to

$$y_i = 4.300 - 1.338x_{1i} + 0.172x_{2i}$$

[4]

(b) We can construct the ANOVA table by defining all the different elements of interest. We start from SS_T, which is

$$SS_T = \mathbf{Y}^{\mathsf{T}}\mathbf{Y} - n\bar{y}^2 = 1082.723 - 46 \times (4.848)^2 = 1.65$$

Then we move to SS_R, which is equal to

$$SS_R = \widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\widehat{\boldsymbol{\beta}} - n\bar{y}^2 = 1081.574 - 1081.073 = 0.501$$

Thus, SS_E is

$$SS_E = SS_T - SS_R = 1.65 - 0.501 = 1.149$$

In conclusion, we can compute the ANOVA Table as

[9]

Source	df	SS	MS	F
Regression	2	0.501	0.25	9.378
Residuals	43	1.149	0.027	
Total	45	1.65		

(c) The average leverage is $\frac{p}{n}$. A case for which $h_{ii} > 2\frac{p}{n}$ is considered a high leverage case and one with $h_{ii} > 3\frac{p}{n}$ is considered a very high leverage case, where h_{ii} are the diagonal elements of the matrix $\mathbf{H}$.

In our case, n is equal to 46, while p is equal to 3, thus the high leverage happens for $h_{ii} > 2\frac{3}{46} = 0.13$ and the high leverage for $h_{ii} > 3\frac{3}{46} = 0.196$.

Looking at the diagonal elements of $\mathbf{H}$, we have that the fifth (0.135), fifteenth (0.198), twenty-seventh (0.131) and twenty-eighth (0.139) are bigger than 0.13. Moreover only the fifteenth element (0.198) is bigger than 0.196. [4]

Question 5 [12 marks]. This question is similar to examples in the lecture notes.

(a) We find that the general linear regression model could be defined as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

where

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} x_1 & z_1^2 \\ x_2 & z_2^2 \\ \vdots & \vdots \\ x_n & z_n^2 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

[5]

(b) We find the least square estimators by multiplying $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ and $\mathbf{X}^\mathsf{T}\mathbf{Y}$, thus

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}$$

In our case

$$(\mathbf{X}^\mathsf{T}\mathbf{X}) = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ z_1^2 & z_2^2 & \dots & z_n^2 \end{pmatrix} \begin{pmatrix} x_1 & z_1^2 \\ x_2 & z_2^2 \\ \vdots & \vdots \\ x_n & z_n^2 \end{pmatrix} = \begin{pmatrix} \sum x_i^2 & \sum x_i z_i^2 \\ \sum x_i z_i^2 & \sum z_i^4 \end{pmatrix}$$

We can compute the inverse as

$$(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \frac{1}{\sum x_i^2 \sum z_i^4 - (\sum x_i z_i^2)^2} \begin{pmatrix} \sum z_i^4 & -\sum x_i z_i^2 \\ -\sum x_i z_i^2 & \sum x_i^2 \end{pmatrix}$$

The product of $\mathbf{X}$ and $\mathbf{Y}$ is

$$\mathbf{X}^\mathsf{T}\mathbf{Y} == \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ z_1^2 & z_2^2 & \dots & z_n^2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum x_i y_i \\ \sum z_i^2 y_i \end{pmatrix}$$

Combining all the elements leads to the following least square estimator

$$\begin{aligned} \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} &= \frac{1}{\sum x_i^2 \sum z_i^4 - (\sum x_i z_i^2)^2} \begin{pmatrix} \sum z_i^4 & -\sum x_i z_i^2 \\ -\sum x_i z_i^2 & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \sum x_i y_i \\ \sum z_i^2 y_i \end{pmatrix} \\ &= \frac{1}{\sum x_i^2 \sum z_i^4 - (\sum x_i z_i^2)^2} \begin{pmatrix} \sum z_i^4 \sum x_i y_i - \sum x_i z_i^2 \sum z_i^2 y_i \\ \sum x_i^2 \sum z_i^2 y_i - \sum x_i z_i^2 \sum x_i y_i \end{pmatrix} \end{aligned}$$

[**6**]

(c) Based on the definition of the hat matrix, $\mathbf{H}$, initially, we prove that it is symmetric

$$\begin{aligned} \mathbf{H}^\mathsf{T} &= \left(\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\right)^\mathsf{T} = (\mathbf{X}^\mathsf{T})^\mathsf{T}((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1})^\mathsf{T}\mathbf{X}^\mathsf{T} \\ &= \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} = \mathbf{H} \end{aligned}$$

Thus $\mathbf{H}$ is symmetric, since $\mathbf{X}^\mathsf{T}\mathbf{X}$ is symmetric and $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ is symmetric.

Moving to the idempotent proof, we have

$$\begin{aligned} \mathbf{H}\mathbf{H} &= \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \underbrace{\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}}_{\mathbf{I}} \mathbf{X}^\mathsf{T} \\ &= \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} \\ &= \mathbf{H} \end{aligned}$$

Thus $\mathbf{H}$ is idempotent. [**4**]

End of Paper.