Main Examination period 2022 – May/June – Semester B

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**.

The exam is available for a period of **24 hours**. Upon accessing the exam, you will have **2 hours** in which 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.

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. To be eligible for IFoA exemption, **your must submit your exam within the first 3 hours of the assessment period**.

Examiners: L. Shaheen, A. Zincenko

Continue to next page

Question 1 [25 marks]. (Unseen, similar Questions)

(a) $\widehat{y} = \widehat{\beta_0} + \widehat{\beta_1}x = 429.048 + 18.244x.$ [**2**]

(b) For $\widehat{\beta_1}$, with $\widehat{se(\beta_1)}$ and $n = 12$, a 95% confidence interval is given by [**3**]

$$\hat{\beta_1} \pm t_{.025,12-2}\,\widehat{se(\beta_1)} = \left[\widehat{\beta_1} + t_{0.025,10}\,\widehat{se(\beta_1)}, \widehat{\beta_1} - t_{0.025,10}\,\widehat{se(\beta_1)}\right]$$

(c) For $\widehat{\beta_1} = 18.244, \widehat{se(\beta_1)} = 5.643, n = 12$, a 95% confidence interval is given by [**4**]

$$18.244 \pm t_{.025,10}(5.643) = 18.244 \pm (2.22814)(5.643) = \left[5.6706, 30.82\right]$$

(d) [**5**]

Source of Variation	DF	Sum Square	Mean Square	F Value
Regression	1	SSR =16059.8	MSR = $\frac{16058.9}{1}$ = 16058.9	
Residual	12-2 = 10	SSE= 15366	MSE= $\frac{15366}{(10)}$ = 1536.6	F = 10.451

(i) The null hypothesis is that there is no increase in mean sales from increasing the amount of shelf space. So it will be $H_0 : \beta_1 = 0, H_A : \beta_1 \neq 0$. [4]

(ii) Test Statistics: F= 10.451, for $\alpha = .01$ and $\mathcal{F}^1_{10}(0.01) = 10.044$. [3]

(iii) As F> $\mathcal{F}^1_{10}(0.01)$, we will reject the null hypothesis and conclude that $\beta_1 \neq 0$. There is a significant effect on mean weekly sales when we increase the shelf space. [4]

Question 2 [15 marks].

(a) Viewing at `R` outputs, the fitted linear regression model is
$\widehat{y} = -1298.282 + 61.127x.$ [**1**]

The R^2 of the model is 92.77%, which shows that linear fit of the model explains much of the variation. [**2**]

(b) Figures show the plots of the standardized residuals versus the data (left) and the standardized residuals versus the fitted values (right). [**1**]
A random scatter in standardized residualsl vs x and vs fits plots suggests that the assumption of equal variances holds. [**4**]
A funnel shape suggests the variance is increasing with the mean. In particular, we have that the plot has a funnel shape or more a trumpet shape in this case. [**1**]

(c) We see outliers in the Q-Q plot, which is an indication that normality assumption of errors is violated. This is further confirmed by Shapiro-Wilk test, where p-values are less than 0.05. [**2**]

In fact, a small p value means that the assumption of normality is not supported by the data. [**1**]

(d) We have seen that Shapiro-Wilk test have a small p-value, which implies that the assumption of normaility is not supported by data, also the increasing variance suggests that we should transform the dependent variable to $\log y$ or $\sqrt{y}$, for example. [**3**]

Question 3 [20 marks].

(a) Let $\widehat{\mu}$ and $\widehat{\sigma^2}$ be the maximum likelihood estimates for μ and σ^2. The probability density function of X_i is given by

$$f(x_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

for $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$.

Given any sample data set $x_1, x_2, , ..., x_n$, we can now write down the likelyhood function, which is given by:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i, \mu, \sigma^2) = (\sigma^2)^{-\frac{n}{2}}(2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

[**5**]

By applying log on both sides we get

$$\log L(\mu, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Differentiating with respect to μ and equating to 0, we will get [**2**]
$\sum_{i=1}^{n} x_i - n\hat{\mu} = 0$, this gives us

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Differentiating with respect to σ^2 and equating to 0, we will get [**3**]

$$-\frac{n}{2}\frac{1}{\widehat{\sigma^2}} + \frac{\sum_{i=1}^{n}(x_i - \widehat{\mu})^2}{2(\widehat{\sigma^2})^2} = 0$$

$$\sum_{i=1}^{n}(x_i - \widehat{\mu})^2 = n\widehat{\sigma^2}$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \widehat{\mu})^2}{n}.$$

(b) Let $Y_1, Y_2 \cdots Y_n$ are independent, normal random variables with

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2).$$

The conditional probability density function of Y_i for each x_i is given by

$$p(y_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\{y_i-(\beta_0+\beta_1 x_i)^2\}}{2\sigma^2}}.$$

Given any sample data set $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$, we can now write down the probability density, seeing that β_0, β_1 and σ^2 are unknown.

$$L(\beta_0, \beta_1, \sigma^2) = \log \prod_{i=1}^{n} p(y_i; \beta_0, \beta_1, \sigma^2)$$

[**4**]

$$= \sum_{i=1}^{n} \log p(y_i|x_i; \beta_0, \beta_1, \sigma^2)$$

$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2.$$

Differentiating with respect to β_0 and equating to zero we will have [**2**]

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}.$$

Differentiating with respect to β_1 and equating to zero we will have

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_x^2}.$$

[**2**]

Differentiating with respect to σ^2 and equating to zero we will have

$$-\frac{n}{2\widehat{\sigma^2}} + \frac{\sum_{i=1}^{n}(y_i - (\widehat{\beta_0} + \widehat{\beta_1}x_i))^2}{2(\widehat{\sigma^2})^2} = 0.$$

By simplifying the expressions we will have [**2**]

$$n\widehat{\sigma^2} = \sum_{i=1}^{n}(y_i - (\widehat{\beta_0} + \widehat{\beta_1}x_i)^2$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\widehat{\beta_0} + \widehat{\beta_1}x_i)^2.$$

Question 4 [25 marks].

(a) One starts with the null model, then one tries adding variables and adds the variable which has the smallest AIC or BIC. **[2]**
For AIC

$$AIC = 2k - 2l, BIC = 2(log(n)) - 2l$$

where l is the value of the logarithmic likelihood function of the constructed model, k is the number of parameters used (estimated), and n is the sample size that the model was built on. **[2]**
(b) BIC imposes a greater penalty on increasing the number of parameters compared to AIC,therefore the second model in the output corresponds to BIC. **[5]**
(c)

$$mpg = \beta_0 + \varepsilon$$

[1]
(d) In the *step* function, k=2 means that AIC is used. We have covered this in the lectures. **[3]**
(e) The statistician has employed backwards elimination procedure. Backward elimination starts with the full model, deleting the variable (if any) whose loss gives the minimal information criterion value, and repeating until nothing reduces the information criterion. **[6]**
(f) Multicollinearity is the presence of a linear relationship between explanatory variables. Multicollinearity is problematic because the mathematical regression model contains redundant variables. The main problem is that multicollinearity leads to unstable parameter estimates, which makes it very difficult to assess the influence of independent variables on dependent variables. **[4]**
(g) From the R output we see that there is no multicollinearity because vif is less than 5. **[2]**

Question 5 [15 marks].
(a) Due to the fact that the p - value is less than 0.05, we reject the null hypothesis, which means that we should include the extra parameters in the model. **[2]**
(b) One can build $2^3 = 8$ models. **[2]**
(c)(i) In the first model, male earns 329.56 less than female when controlling for age. **[2]**
In the second model, Males are estimated to make $\exp(-0.321)$ as much as female on average, while controlling for age. **[2]**
(c)(ii) It would not be correct to say that the second model is preferred over the first one because two models are on different scale. **[3]**
(c)(iii) Model (3) is better than Model (2) because R^2 is higher and σ_2 is lower. **[4]**

End of Paper.