University of Glasgow

13th December 2019
9.30 – 11.00 a.m.

EXAMINATION FOR THE DEGREES OF B.Sc., M.Sci., M.Sc. and M.Res.

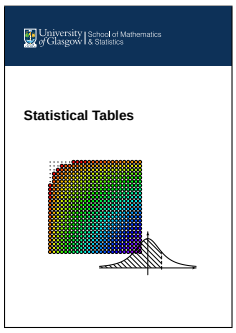
Regression Models (Level M)

This paper consists of 16 pages and contains 5 questions.
Candidates should attempt all questions.

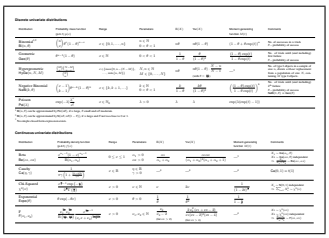
Question 1	10 marks
Question 2	17 marks
Question 3	9 marks
Question 4	18 marks
Question 5	6 marks
Total	60 marks

The following material is made available to you:

Statistical tables*



Probability formula sheet



"An electronic calculator may be used provided that it is allowed under the School of Mathematics and Statistics Calculator Policy. A copy of this policy has been distributed to the class prior to the exam and is also available via the invigilator."

CONTINUED OVERLEAF/

1. State whether the following statements are TRUE or FALSE. Provide an explanation for your answer

   (a) The parameters to be estimated in the simple linear regression model $y_i = \beta x_i + \gamma x_i^2 + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ for $i - 1, \ldots, n$ are γ, β and σ^2. **[2 MARKS]**

   TRUE. There are only two unknown regression parameters γ, β and we need to estimate the variance of the errors σ^2.

   (b) In a regression model $y_i = \alpha + \beta x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$, a 95% confidence interval for β was $(-2.33, -0.05)$. In a hypothesis test $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ you would accept the null hypothesis at a 5% signficance level. **[2 MARKS]**

   FALSE. The interval does not contain zero and so we would reject the null hypothesis.

   (c) If we compare two nested models using R^2 and $R^2(adj)$ then R^2 will prefer the model with more variables more often than $R^2(adj)$. **[2 MARKS]**

   TRUE. $R^2(adj)$ penalised for the number of parameters estimated in the model and therefore will be smaller than R^2.

   (d) The main purpose of plotting residuals against fitted values is to assess the assumptions that the errors are normally distributed. **[2 MARKS]**

   FALSE. The main purpose is checking to see if the residuals have constant variance and zero mean.

   (e) If $Cor(X, Y) = 0$ then we can conclude that there is no relationship between variables X and Y. **[2 MARKS]**

   FALSE. We can conclude that there is no linear relationship between X and Y only.

   Markers Comments: This question is bookwork and tests a basic understanding of concepts within this course.

2. Consider the following regression model

CONTINUED OVERLEAF/

Data: (y_{ij}, x_{ij}) for $j = 1, \ldots, n_i$, $i = 1, 2$

Model: $y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}$

(a) Write the model above in vector-matrix notation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ clearly identifying the elements of $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ **[4 MARKS]**

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

$$\mathbf{Y} = \begin{pmatrix} y_{11} \\ . \\ . \\ y_{1n_1} \\ y_{21} \\ . \\ . \\ y_{2n_2} \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & x_{11} \\ . & . & . \\ . & . & . \\ 1 & 0 & x_{1n_1} \\ 0 & 1 & x_{21} \\ . & . & . \\ . & . & . \\ 0 & 1 & x_{2n_2} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix}$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ . \\ . \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ . \\ . \\ \epsilon_{2n_2} \end{pmatrix}$$

(b) Using the re-parameterisation $y_{ij} = \alpha_i + \beta(x_{ij} - \bar{x}_{i.}) + \epsilon_i$ with $\bar{x}_{i.}$ the average value of covariate x within group i, write down the matrices $\boldsymbol{X}$ and $\boldsymbol{\beta}$. **[2 MARK]**

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

CONTINUED OVERLEAF/

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & (x_{11} - \bar{x}_{1.}) \\ . & . & . \\ . & . & . \\ 1 & 0 & (x_{1n_1} - \bar{x}_{1.}) \\ 0 & 1 & (x_{21} - \bar{x}_{2.}) \\ . & . & . \\ . & . & . \\ 0 & 1 & (x_{2n_2} - \bar{x}_{2.}) \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix}$$

(c) Using the re-parameterisation of $\boldsymbol{X}$ obtained in (b), show

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & S_{x_1x_1} + S_{x_2x_2} \end{pmatrix}$$

where $S_{x_ix_i}$ is the corrected sum of squares of x within group i. **[4 MARKS]**

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \\ (x_{11} - \bar{x}_{1.}) & \dots & (x_{1n_1} - \bar{x}_{1.}) & (x_{21} - \bar{x}_{2.}) & \dots & (x_{2n_2} - \bar{x}_{2.}) \end{pmatrix} \begin{pmatrix} 1 & 0 & (x_{11} - \bar{x}_{1.}) \\ . & . & . \\ . & . & . \\ 1 & 0 & (x_{1n_1} - \bar{x}_{1.}) \\ 0 & 1 & (x_{21} - \bar{x}_{2.}) \\ . & . & . \\ . & . & . \\ 0 & 1 & (x_{2n_2} - \bar{x}_{2.}) \end{pmatrix}$$

$$= \begin{pmatrix} n_1 & 0 & \sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1.}) \\ 0 & n_2 & \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_{2.}) \\ \sum_{j=1}^{n_1}(x_{1j} - \bar{x}_{1.}) & \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_{2.}) & \sum_i \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})^2 \end{pmatrix}$$

$$= \begin{pmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & S_{x_1x_1} + S_{x_2x_2} \end{pmatrix}$$

since $\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.}) = 0$ and $S_{x_ix_i} = \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})^2$ for $i = 1, 2$.

(d) Using $\boldsymbol{\beta}$ obtained in (b), show that

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{y}_{1.} \\ \bar{y}_{2.} \\ \frac{S_{x_1y_1} + S_{x_2y_2}}{S_{x_1x_1} + S_{x_2x_2}} \end{pmatrix}$$

CONTINUED OVERLEAF/

where $S_{x_i y_i}$ is the corrected sum of products of x and y and $\bar{y}_{i.}$ is the average value of y within group i. **[2 MARKS]**

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & 0 \\ 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & \frac{1}{S_{x_1x_1}+S_{x_2x_2}} \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{j=1}^{n_2} y_{2j} \\ \sum_i \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.}) y_{ij} \end{pmatrix}$$

$$= \begin{pmatrix} n_1\bar{y}_{1.} \\ n_2\bar{y}_{2.} \\ S_{x_1y_1} + S_{x_2y_2} \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \bar{y}_{1.} \\ \bar{y}_{2.} \\ \frac{S_{x_1y_1}+S_{x_2y_2}}{S_{x_1x_1}+S_{x_2x_2}} \end{pmatrix}$$

An architect is interested in English medieval cathedrals and measures the height and length of 25 cathedrals, 16 of which were Gothic and 9 were Romanesque. The architect wants to know if the relationship between height and length differs between the two styles of cathedral (Gothic or Romanesque).

Below are the results from a regression analysis fitting two parallel lines, one for each style.

Data: (y_{ij}, x_{ij}) for $i = 1, 2$, $j = 1, \ldots, n_i$, $n_1 = 16$, $n_2 = 9$
where the index $i = 1$ represents a cathedral of Romanesque style and $i = 2$ represents a cathedral of Gothic style.

y_{ij} =length of cathedral j of style i

x_{ij} =height of cathedral j of style i.

Model: $y_{ij} = \alpha_i + \beta(x_{ij} - \bar{x}_{i.}) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ are independent and identically distributed and $\bar{x}_{i.}$ is the average height of cathedrals of style i for $i = 1, 2$.

Regression output from `R`

```
parallel.model <- lm(length  ~  height+style, data=cathedral)
```

CONTINUED OVERLEAF/

```
summary(parallel.model)


Coefficients:
Estimate Std. Error t value
(Intercept)   44.298     81.648   0.543
height         4.712      1.058   4.452
style1        80.393     32.306   2.488


Analysis of Variance Table

Response: length
Df Sum Sq Mean Sq F value
height     1 116992  116992 19.4659
style      1  37217   37217  6.1924
Residuals 22 132223    6010
```

Summary statistics

$$\begin{array}{lllllll} S_{y_1y_1} & = & 41782.22 & S_{x_1x_1} = 308.22 & S_{x_1y_1} = 967.22 & \bar{x}_1 = 74.44 & \bar{y}_1 = 475.44 \\ S_{y_2y_2} & = & 209543.80 & S_{x_2x_2} = 5056.94 & S_{x_2y_2} = 24311.38 & \bar{x}_2 = 74.94 & \bar{y}_2 = 397.38 \end{array}$$

$$t(22, 0.975) = 2.073873$$

(e) Show that the mean difference in length between the two styles of cathedrals can be written as

$$\alpha_1 - \alpha_2 + \beta(\bar{x}_{2.} - \bar{x}_{1.})$$

at any height. **[1 MARKS]**

For a given height x, the difference between the two regression lines is

$$\begin{aligned} \alpha_1 + \beta(x - \bar{x}_{1.}) - [\alpha_2 + \beta(x - \bar{x}_{2.})] &= \alpha_1 + \beta x - \beta\bar{x}_{1.} - \alpha_2 - \beta x + \beta\bar{x}_{2.} \\ &= \alpha_1 - \beta\bar{x}_{1.} - \alpha_2 + \beta\bar{x}_{2.} \\ &= \alpha_1 - \alpha_2 + \beta(\bar{x}_{2.} - \bar{x}_{1.}) \end{aligned}$$

CONTINUED OVERLEAF/

(f) Using the summary statistics provided and the formula for the mean difference in length between the two styles of cathedrals obtained in (e), estimate a 95% confidence interval for the mean difference in length between the two styles of cathedrals. **[4 MARKS]**

A 95% C.I. for $\hat{\alpha}_1 - \hat{\alpha}_2 + \hat{\beta}(\bar{x}_{2.} - \bar{x}_{1.})$ is

$$\begin{aligned}
\hat{\alpha}_1 - \hat{\alpha}_2 + \hat{\beta}(\bar{x}_{2.} - \bar{x}_{1.}) &\pm t(n-p, 0.975)\sqrt{\left(\frac{RSS}{n-p}\right)\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{x}_{2.} - \bar{x}_{1.})^2}{S_{x_1x_1} + S_{x_2x_2}}\right)} \\
475.44 - 397.38 + 4.71(74.94 - 74.44) &\pm 2.074\sqrt{\left(\frac{132223}{25-3}\right)\left(\frac{1}{16} + \frac{1}{9} + \frac{(74.94-74.44)^2}{308.22 + 5056.94}\right)} \\
62.415 &\pm 2.074 * \sqrt{6010.136 * 0.1736577} \\
62.415 &\pm 2.074 * 32.30645 \\
62.415 &\pm 67.00357
\end{aligned}$$

$$(-4.59, 129.42)$$

Markers Comments: Parts (a) - (e) are bookwork. Part (f) is similar to examples seen in tutorials and lectures using a different data set.

3. Data are available giving daily air quality measurements in Glasgow. A researcher is particularly interested in the relationship between the mean Ozone concentration in parts per billion (Ozone) and solar radiation measured in Langleys (Solar). These data are plotted below.

(a) Describe the relationship between Ozone and Solar. **[1 MARK]**

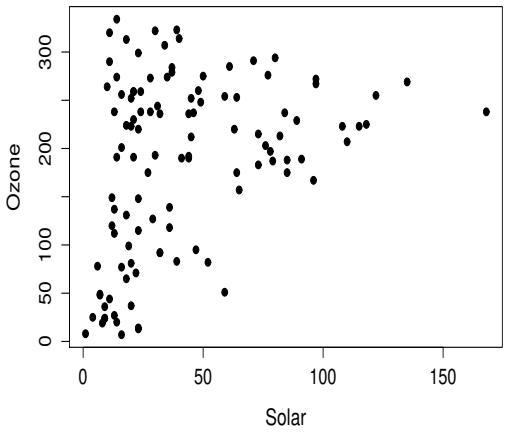
There could be a linear relationship between ozone and solar but there is clearly more variability in the data as solar increases.

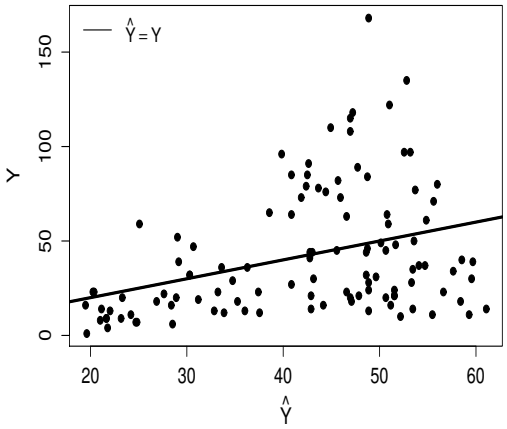
The researcher fits a linear regression model and examines the fitted values as a method of assessing model fit. Below is a plot of the fitted values ($\hat{Y}$) against the observed values (Y). The solid line shows $\hat{Y} = Y$.

(b) Briefly discuss the model fit. Comment specifically on whether or not the model assumptions appear valid. **[2 MARKS]**

There is a lot of variability around the fitted line. It looks like the assumption of constant variance in the error terms may be violated since the data are more spread out for higher values of $\hat{Y}$.

CONTINUED OVERLEAF/





(c) Assuming a linear relationship is appropriate, calculate the correlation between the observed values (Y) and fitted values ($\hat{Y}$) using the summary statistics

$$n = 110, \sum_{i=1}^{n} Y_i = 4673, \sum_{i=1}^{n} \hat{Y}_i = 4673$$

$$\sum_{i=1}^{n} Y_i^2 = 318531, \sum_{i=1}^{n} \hat{Y}_i^2 = 211508.8$$

$$\sum_{i=1}^{n} Y_i \hat{Y}_i = 211508.8$$

[2 MARKS]

CONTINUED OVERLEAF/

$$
\begin{aligned}
\hat{\rho}(Y,\hat{Y}) = r(Y,\hat{Y}) &= \frac{S_{y\hat{y}}}{\sqrt{S_{yy}S_{\hat{y}\hat{y}}}} \\
&= \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}} \\
&= \frac{\sum_{i=1}^{n} y_i\hat{y}_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n}\hat{y}_i}{n}}{\sqrt{\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\sum_{i=1}^{n}\hat{y}_i^2 - \frac{(\sum_{i=1}^{n}\hat{y}_i)^2}{n}}} \\
top &= 211508.8 - \frac{4673^2}{110} = 12991.26 \\
bottom &= \sqrt{(318531 - \frac{4673^2}{110})(211508.8 - \frac{4673^2}{110})} = 39485.78 \\
\frac{top}{bottom} &= 0.329
\end{aligned}
$$

(d) Use the statistical tables to perform a test (at a significance level of α=5%) of the null hypothesis that the population correlation coefficient ρ is 0

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

[2 MARKS]

From table 8, $c = 0.157$ at the 5% significance level. Since $0.329 > 0.157$ we can reject the null hypothesis and so the fitted and observd values are significantly correlated.

(e) Estimate the percent variability in the response variable Ozone explained by the fitted model (i.e. R^2). **[1 MARKS]**

We just need to find $r^2 = 0.108$

(f) Suggest a possible transformation that could be used in this context to improve the model fit. **[1 MARKS]**

The main issue may be with the constant variance assumptions of the residuals. We transforming the response variable Ozone using a log transformation or a square root transformation.

Markers Comments: Parts (a), (c), (d) and (f) of this question are similar to examples seen in tutorials and lectures uisng a different data set. Part (b) is different to questions seen in the course in that we had not assessed model fit by plotting observed and fitted values in this way. Part (e) was covered in the lecture material but students did not have to estimate R^2 from correlation in this way in any example questions.

CONTINUED OVERLEAF/

4. It is believed that children with too little or too much melatonin are more susceptible to depression. There is natural variation on levels of melatonin in children and melatonin levels are thought to be related to age. In order to investigate the relationship between melatonin and age, data are available giving the age (in months) and melatonin levels in 19 healthy children.

   The following model was fitted the the data

   Data: (y_i, x_i) for $i = 1, \ldots, 19$, y =melatonin, x =age

   Model: $y_i = \alpha + \beta x_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ independent.

   The results of the fitted model are:

```
Coefficients:
Estimate
(Intercept)  5.9053
age         -0.042446


Analysis of Variance Table

           Df     Sum Sq   Mean Sq  F value     Pr(>F)
Regression    1   21.619   21.619   159.23      1.322e-09
Residuals     17  2.308    0.136
Total         18  23.927
```

   Writing this model in vector matrix notation, $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, you may assume

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} 0.3552770 & -0.0050221 \\ -0.0050221 & 0.0000833 \end{pmatrix}$$

   (a) Write down the assumptions of this normal linear model and describe how you could informally check each assumption if the data were available to you.

   [6 MARKS]

   - The regression model is linear in parameters. This is true since we have defined the regression model to be linear in the parameters
   - The deterministic part of the model captures all the non-random structure in the data. Look at a plot of residuals against fitted values to see if there is constant variance.
   - The scale of the variability of the errors is constant at all values of the explanatory variables. Look at a plot of residuals against fitted values to see if the residuals are randomly scattered around zero.

CONTINUED OVERLEAF/

- The errors are independent. Plot the residuals against age to see if there is any relationship between the residuals are age increases.
- The errors are Normally distributed. Look at a Q-Q plot or histogram of the residuals.
- The values of the explanatory variables are recorded without error. Unless we know otherwise, we will make this assumption. We could also check for any potential outliers that could be measurement errors and remove them.

(b) Calculate and comment on the R^2 value. **[2 MARKS]**

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{2.308}{23.927} = 0.904$$

This value is very close to one and so age explains 90% of the variation in melatonin. These variables are highly correlated.

(c) Calculate the estimated standard error for the constant term α and the estimated standard error for the coefficient of age β. **[4 MARKS]**

$$\begin{aligned}
ese(\boldsymbol{b}^T\hat{\boldsymbol{\beta}}) &= \sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}} \\
&= \sqrt{\frac{2.308}{17}\begin{pmatrix}0 & 1\end{pmatrix}\begin{pmatrix}0.3552770 & -0.0050221 \\ -0.0050221 & 0.0000833\end{pmatrix}\begin{pmatrix}0 \\ 1\end{pmatrix}} \\
&= \sqrt{\frac{2.308}{17}\begin{pmatrix}-0.0050221 & 0.0000833\end{pmatrix}\begin{pmatrix}0 \\ 1\end{pmatrix}} \\
&= \sqrt{\frac{2.308}{17}(0.0000833)} \\
&= 1.13092e^{-05}
\end{aligned}$$

$$\begin{aligned}
ese(\boldsymbol{b}^T\hat{\boldsymbol{\beta}}) &= \sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}} \\
&= \sqrt{\frac{2.308}{17}\begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix}0.3552770 & -0.0050221 \\ -0.0050221 & 0.0000833\end{pmatrix}\begin{pmatrix}1 \\ 0\end{pmatrix}} \\
&= \sqrt{\frac{2.308}{17}\begin{pmatrix}0.3552770 & -0.0050221\end{pmatrix}\begin{pmatrix}1 \\ 0\end{pmatrix}} \\
&= \sqrt{\frac{2.308}{17}(0.3552770)} \\
&= 0.04823408
\end{aligned}$$

CONTINUED OVERLEAF/

(d) Calculate a 95% prediction interval for a future child of age 55 months and interpret this interval.

[6 MARKS]

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t\,(n-p; 0.975)\sqrt{\frac{RSS}{n-p}(1+\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}$$

$$\begin{aligned}
\sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}} &= \sqrt{\frac{2.308}{17}\left[1+\begin{pmatrix}1 & 55\end{pmatrix}\begin{pmatrix}0.3552770 & -0.0050221\\ -0.0050221 & 0.0000833\end{pmatrix}\begin{pmatrix}1\\55\end{pmatrix}\right]}\\
&= \sqrt{\frac{0.2517}{17}[1+0.0548285]}\\
&= \sqrt{0.01561766}\\
&= 0.1249706
\end{aligned}$$

$$\begin{aligned}
\mathbf{b}^T\hat{\boldsymbol{\beta}} &\pm t\,(n-p; 0.975)\sqrt{\frac{RSS}{n-p}(1+\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}\\
3.57077 &\pm 2.109816(0.1249706)\\
3.57077 &\pm 0.263665
\end{aligned}$$

$$(3.307105, 3.834435).$$

A melatonin level for a future healthy child of age 55 months is highly likely to lie between 3.3 and 3.8

Markers Comments: Parts (a) is bookwork. Parts (b) - (d) are similar to questions seen in tutorials and lectures using a different data set.

5. Nematodes are a major contribution to reduced productivity in livestock in the UK. Parasitic control is difficult in grazing livestock since infectious nematode larvae live on pasture (where livestock graze). Once larvae are ingested they reproduce in the gut of the host animal and the cycle continues.

   In order to quantify a nematode infection in an animal, we can count the number of nematode present in the gut or we can measure the size of the nematodes in the gut. Longer nematodes lay more eggs.

   We are interested in the relationship between nematode number and nematode length and have data detailing nematode infection in 485 sheep.

   The statistician employed thinks the relationship between nematode length and number is not linear and they fit three models

CONTINUED OVERLEAF/

Data: (y_i, x_i) for $i = 1, \ldots, 485$, y =nematode length, x =nematode number

Mod1: $y_i = \alpha + \beta x_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ independent.

Mod2: $y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ independent.

Mod3: $y_i = \alpha + \beta x_i + \gamma x_i^2 + \phi x_i^3 + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ independent.

(a) The results of the fitted models are given below. Which model would you choose? Give an explanation for your answer. **[2 MARKS]**

Based on the model output, mod3 has the lowest AIC value. However mod2 and mod3 shows the quadratic term to not be significant and so you may choose mod1 or mod3 at this stage.

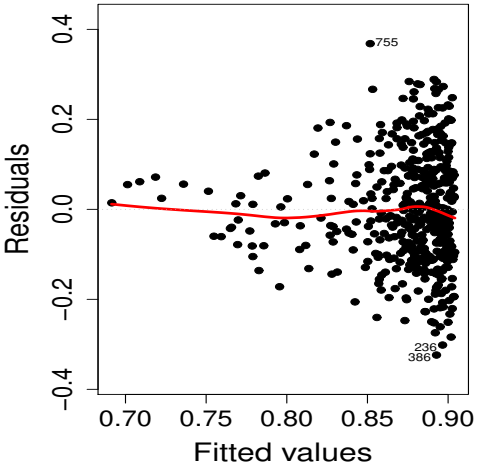
```
Mod1
AIC=664.4734

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.056e-01  7.470e-03 121.228  < 2e-16
number      -7.540e-06  1.219e-06  -6.185 1.32e-09

Mod2
AIC=665.2243

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.874427   0.005517 158.493  < 2e-16
poly(number, 2)1   -0.750747   0.121502  -6.179 1.38e-09
poly(number, 2)2   -0.005498   0.121502  -0.045    0.964

Mod3
AIC=663.2264

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.874427   0.005504 158.859  < 2e-16
poly(number, 3)1   -0.750747   0.121222  -6.193 1.27e-09
```

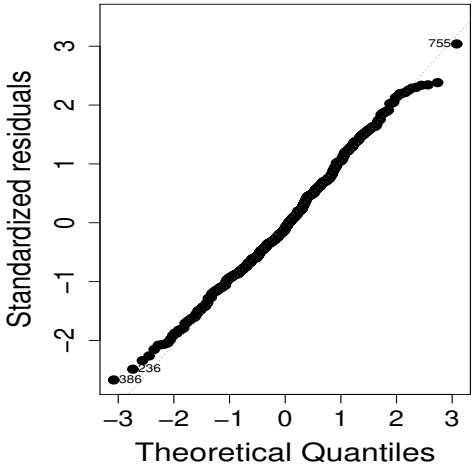
CONTINUED OVERLEAF/

```
poly(number, 3)2   -0.005498   0.121222  -0.045   0.9638
poly(number, 3)3    0.217898   0.101222   1.798   0.0229
```

(b) Below are plots of residuals vs fitted values and a Q-Q plot based on Mod1. Comment on both plots in relation to model assumptions. **[2 MARKS]**

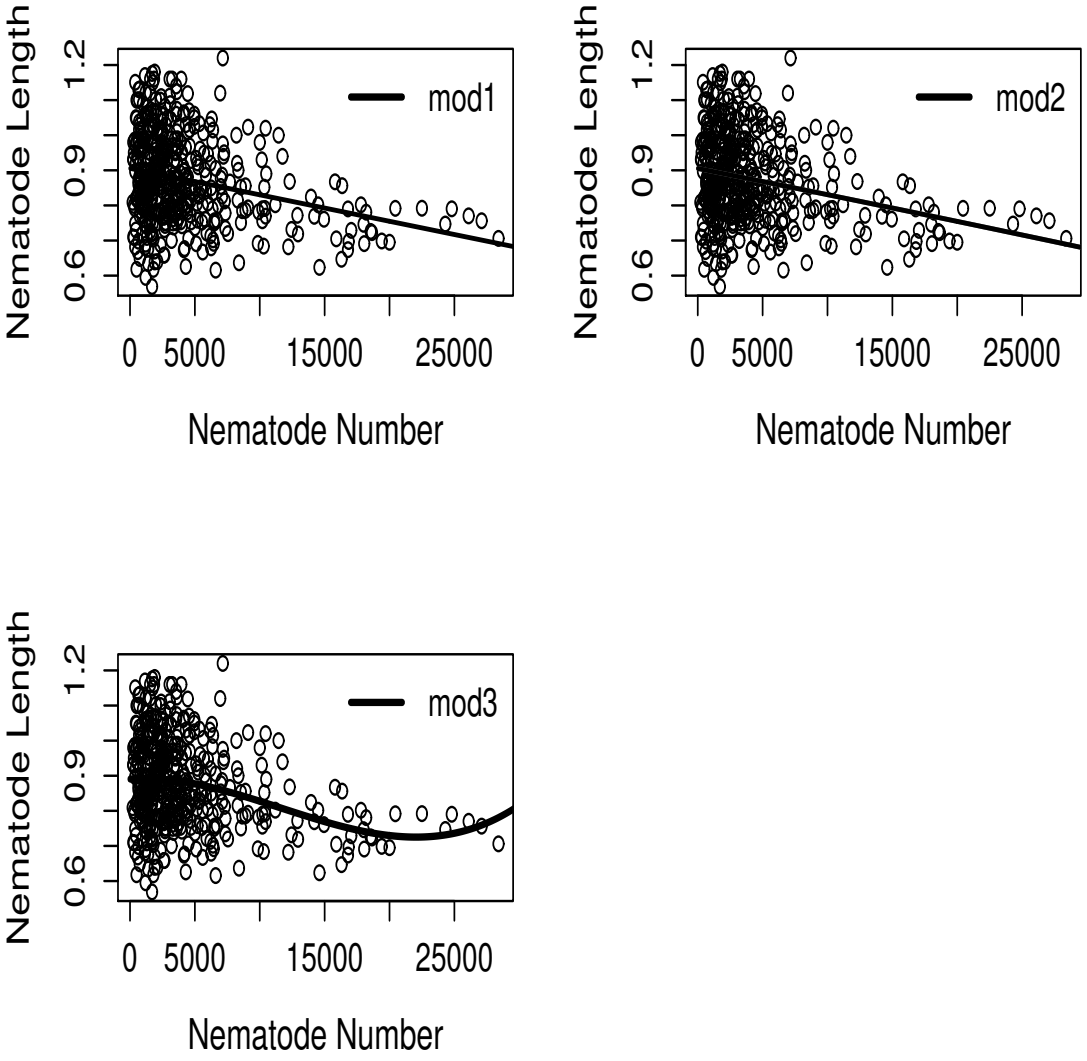
Looking at the residuals against fitted values, the residuals are scattered around zero but as nematode length increases, the amount of variability in the residuals increases and so we cannot assume the residuals have constant variance. From the Q-Q plot we can assume the residuals are normally distributed.





CONTINUED OVERLEAF/

(c) In order to assess the difference between the three models, the statistician plots each of the three fitted lines on a scatterplot of nematode number plotted against nematode length. Based on these fitted lines, which model would you choose? Give an explanation for your answer. **[2 MARKS]**



Based on a plot of each of the three model fits, we can see that do not gain anything by fitted a higher order polynomial function in mod2 or mod3 and therefore I would choose mod1 since it is the simplest of the three and we do not loose any information.

Markers Comments: Parts (a) and (b) are similar to questions seen in tutorials and

CONTINUED OVERLEAF/

lectures using a different data set. Part (c) has not been seen in examples.

Total: 60 MARKS

END OF QUESTION PAPER.