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.

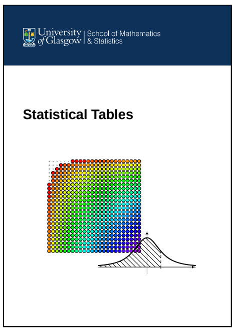
Regression Models (Level M)

This paper consists of 10 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:

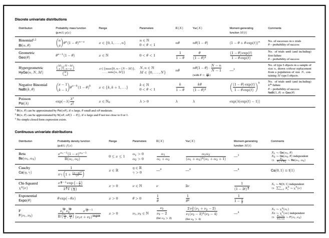
Statistical tables*



University of Glasgow | School of Mathematics & Statistics

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]**

   (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]**

   (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]**

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

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

2. Consider the following regression model

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

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

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

   (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]

   (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]

   (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 $S_{x_iy_i}$ 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]**

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)
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
```

CONTINUED OVERLEAF/

Summary statistics

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

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

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

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

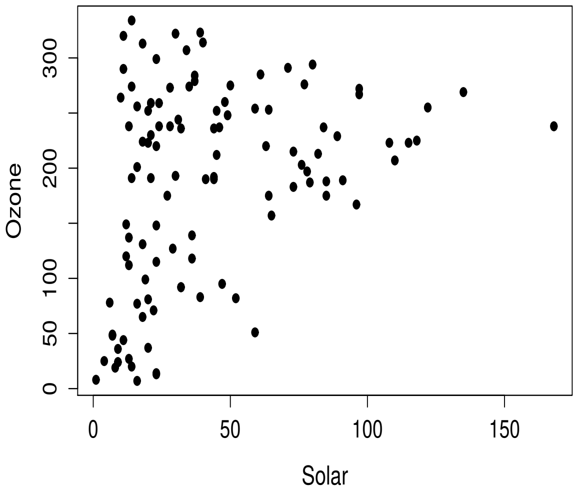
at any height. **[1 MARKS]**

(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]**

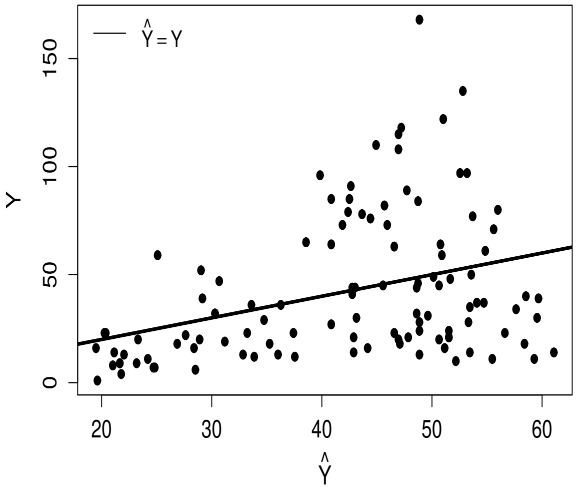
3. Data are available giving daily air quality measurements in Glasgow. A researcher is particuarly interested in the relationship between the mean Ozone concentration in parts per billion (Ozone) and solar radiation measured in Langleys (Solar). These data are plotted below.

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

CONTINUED OVERLEAF/



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]**

(c) Assuming a linear relationship is appropriate, calculate the correlation between

CONTINUED OVERLEAF/

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]

(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]

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

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

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.

CONTINUED OVERLEAF/

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, $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$, 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]

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

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

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

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.

CONTINUED OVERLEAF/

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

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.

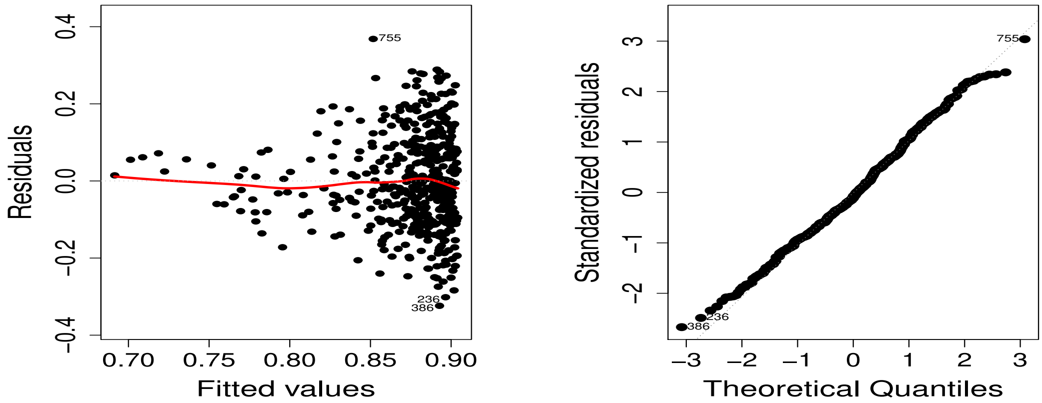
(a) The results of the fitted models are given below. Which model would you choose? Give an explanation for your answer. **[2 MARKS]**

```
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/

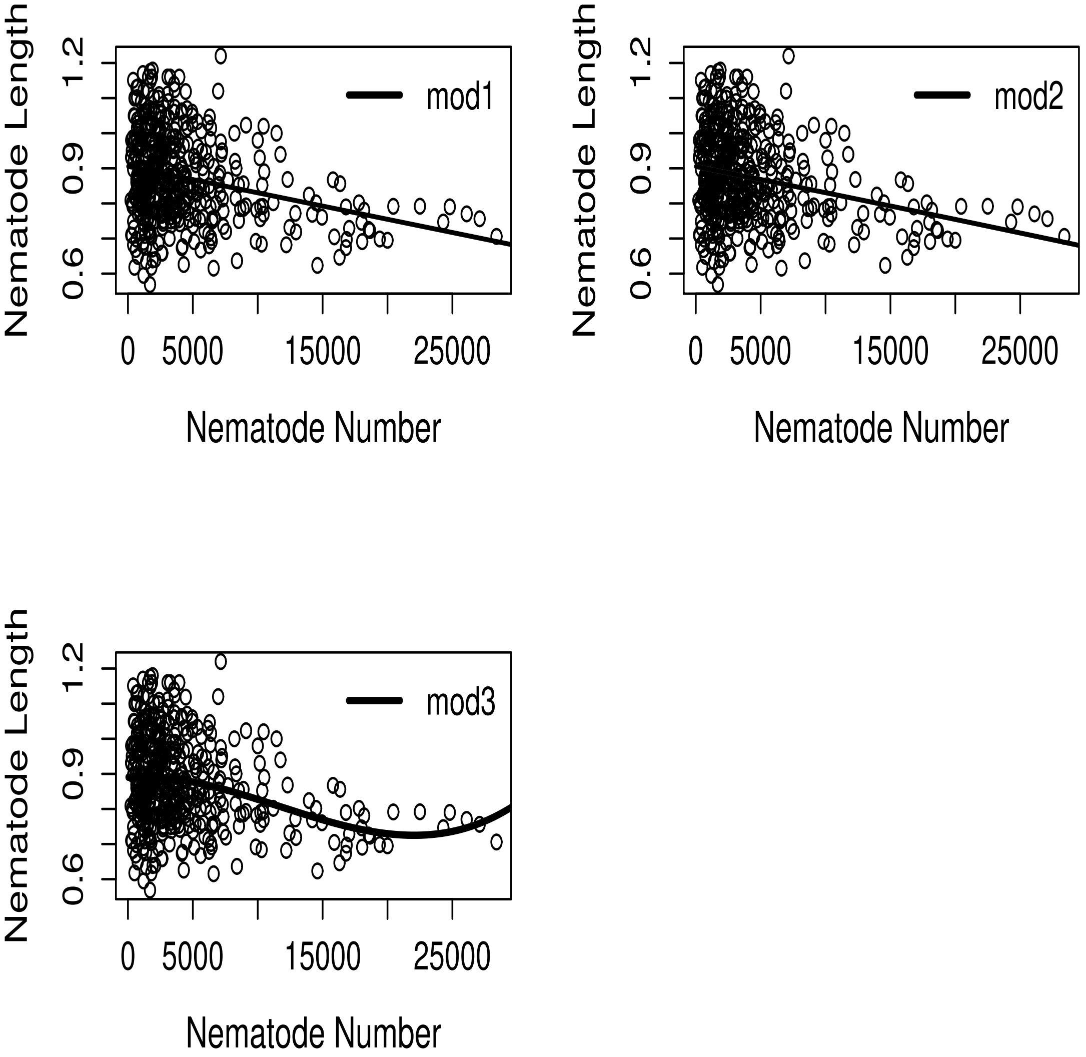
```
poly(number, 3)2   -0.005498   0.121222  -0.045   0.9638
poly(number, 3)3    0.217898   0.101222   1.798   0.0229
```

(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]**



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]**



Total: 60 MARKS

END OF QUESTION PAPER.