University of Glasgow

December 2017
1 hour 30 mins

EXAMINATION FOR THE DEGREE OF MASTERS (SCIENCE)

Regression Modelling

"Hand calculators with simple basic functions (log, exp, square root, etc.) may be used in examinations. No calculator which can store or display text or graphics may be used, and any student found using such will be reported to the Clerk of Senate".

NOTE: Candidates should attempt all 4 questions. Questions are not equally weighted.

1. Answer the following questions:

   (a) State whether the following three statements are TRUE or FALSE. Provide a one line explanation for your answer.

      i. When using simple linear regression analysis, if there is a strong correlation between the independent and dependent variable, then we can conclude that an increase in the value of the independent variable causes an increase in the value of the dependent variable.

      [2 MARKS]

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

      iii. If we fit the model $Y = \alpha + \epsilon$ to a set of data, we will always get $\hat{\alpha} = \bar{Y}$ and $\hat{Y} = \bar{Y}$. **[2 MARKS]**

CONTINUED OVERLEAF/

For part (b)-(f) choose the correct answers from the four options. Note that in some cases more than one answer may be correct

(b) Which of the following can **NOT** be answered from a regression equation? **[2 MARKS]**

i. Predict the value of y at a particular value of x.

ii. Estimate the slope between y and x.

iii. Estimate whether the linear association is positive or negative.

iv. Estimate whether the association is linear or non-linear.

(c) Based on 1988 census data for the 50 States in the United States, the correlation between the number of churches per State and the number of violent crimes per State was 0.85. We can conclude that **[2 MARKS]**

i. There is a causal relationship between the number of churches and the number of violent crimes committed in a city.

ii. The correlation is partly spurious because of the confounding variable of population size: both number of churches and number of violent crimes are related to the population size.

iii. Since the data comes from a census, or nearly complete enumeration of the United States, there must be a causal relationship between the number of churches and the number of violent crimes.

iv. The relationship is not causal because only correlations of +1 or 1 show causal relationships.

(d) A researcher reports that the correlation between two quantitative variables is $r = 0.8$. Which of the following statements is correct? **[2 MARKS]**

i. The average value of y changes by 0.8 when x is increased by 1.

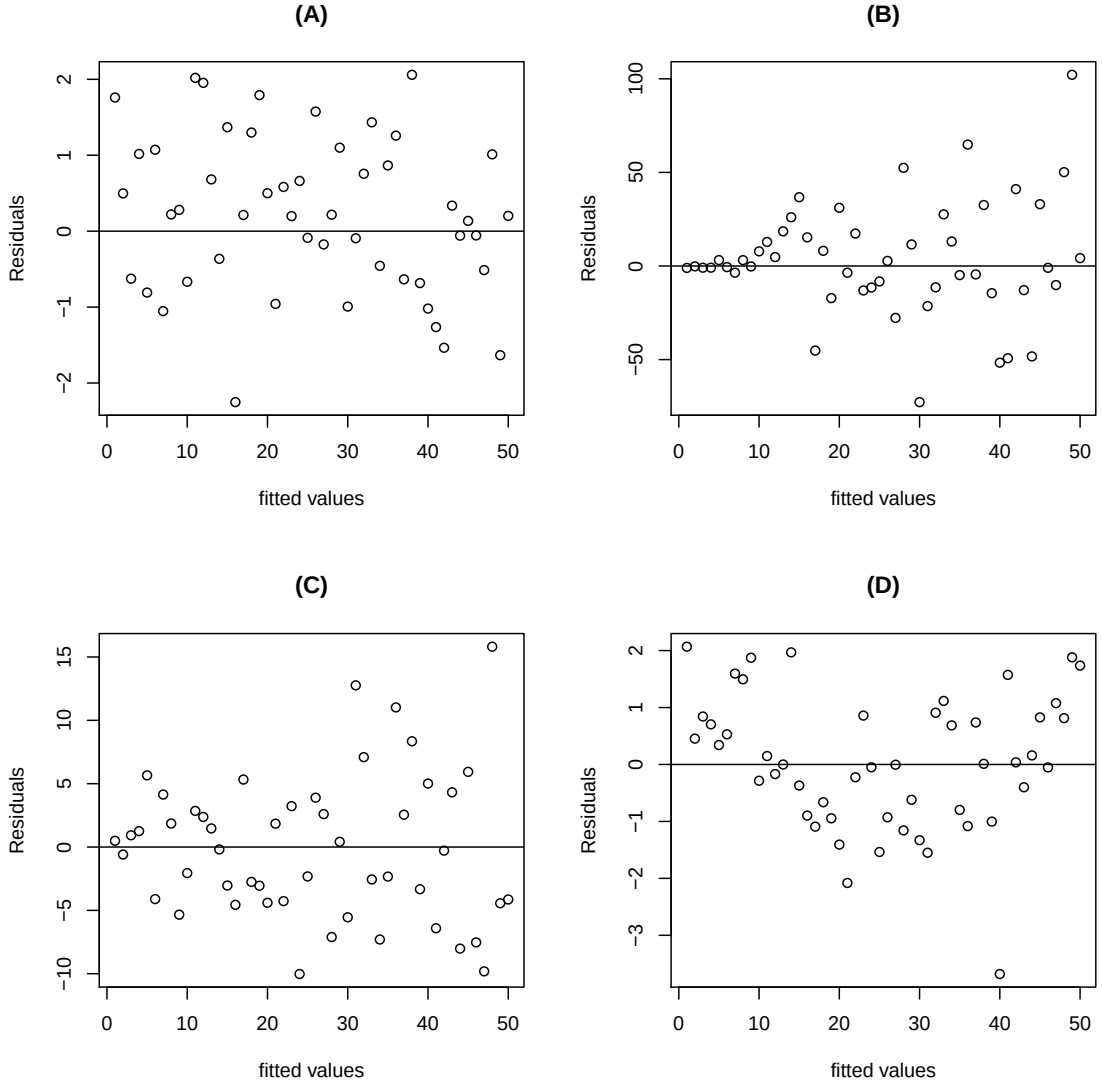
ii. The average value of x changes by 0.8 when y is increased by 1.

iii. The explanatory variable (x) explains 0.8 of the variation in the response variable (y)

iv. The explanatory variable (x) explains $0.8^2 = 0.64$ of the variation in the response variable (y).

CONTINUED OVERLEAF/

2. The following figure displays the residual plots after fitting the linear model $y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1:50$ on four different datasets.



(a) After reviewing the residual plot of each of the datasets (A), (B), (C) and (D) identify the plots which indicate

- Non-linearity
- Strong non-constant variance
- Mild non-constant variance

[4 MARKS]

(b) In each case suggest how you would overcome the violation of linear model assumptions.

[3 MARKS]

CONTINUED OVERLEAF/

(c) What is an influential point? **[2 MARKS]**

(d) Mention a technique that can diagnose a influential point. **[1 MARKS]**

3. Consider the following model:

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

   $Y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \quad \epsilon_{ij}$'s independent

   (a) For the above model, write the model in vector-matrix form $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, identifying clearly the elements of $\mathbf{Y}$, $\mathbf{X}$ and $\boldsymbol{\beta}$.
   [5 MARKS]

   (b) State the dimensions of the vector/matrix $\mathbf{Y}$, $\mathbf{X}$ and $\boldsymbol{\beta}$. **[3 MARKS]**

   (c) Show that matrix $\mathbf{X}$ is not of full rank. State why this is a potential problem in estimating the parameters. Suggest a re-parameterisation of the model which will overcome this problem. **[6 MARKS]**

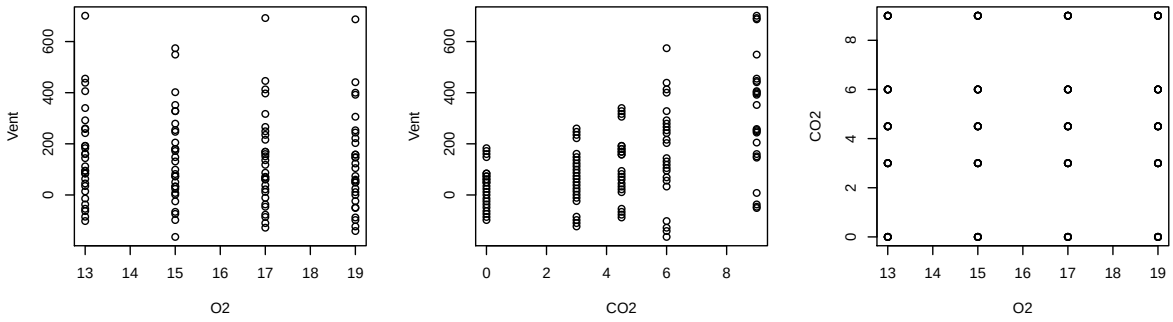
4. Some mammals burrow into the ground to live. Scientists have found that the quality of the air in these burrows is not as good as the air aboveground. In fact, some mammals change the way that they breathe in order to accommodate living in the poor air quality conditions underground.

   Some researchers (Colby, et al, 1987) wanted to find out if nestling bank swallows, which live in underground burrows, also alter how they breathe. The researchers conducted a randomized experiment on n = 120 nestling bank swallows. In an underground burrow, they varied the percentage of oxygen at four different levels and the percentage of carbon dioxide at five different levels. Under each of the resulting $5 \times 4 = 20$ experimental conditions, the researchers observed the total volume of air breathed per minute for each of 6 nestling bank swallows. In this way, they obtained the following data on the n = 120 nestling bank swallows:

   - Response (y): percentage increase in "minute ventilation," (`Vent`), i.e., total volume of air breathed per minute.
   - Potential predictor (x1): percentage of oxygen (`O2`) in the air the baby birds breathe.
   - Potential predictor (x2): percentage of carbon dioxide (`CO2`) in the air the baby birds breathe.

CONTINUED OVERLEAF/

Here's a scatter plot matrix of the resulting data obtained by the researchers:



(a) Comment on the relationship between the response and the two predictors. **[3 MARKS]**

(b) Comment on the relationship between the two predictors and idenitify if there are any issues with multicolinearity between them. **[2 MARKS]**

(c) Now we fit a linear model of the Response on the two possible predictors in `R` issuing the command

```
model <- lm(Vent ~ O2 + CO2)
```

Write the model corresponding to the `R` code in standard linear model notation. **[2 MARKS]**

(d) The `R` output from `summary` and `anova` of the model are given below

```
lm(formula = Vent ~ O2 + CO2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   85.901    106.006   0.810    0.419
O2            -5.330      6.425  -0.830    0.408
CO2           31.103      4.789   6.495  2.1e-09 ***

---------------------------------------------------------
Analysis of Variance Table

Response: Vent
           Df  Sum Sq Mean Sq F value    Pr(>F)
O2          1   17045   17045  0.6883    0.4084
CO2         1 1044773 1044773 42.1866 2.104e-09 ***
Residuals 117 2897566   24766
```

CONTINUED OVERLEAF/

(e) Calculate the R^2 and the R^2_{adj} for the above model and comment on them. **[6 MARKS]**

(f) Now we want test the following two research questions:

(i) Is oxygen related to minute ventilation, after taking into account carbon dioxide?

(ii) Is carbon dioxide related to minute ventilation, after taking into account oxygen?

Using the notation that you have used in (c), state the null and alternative hypothesis to address each of the above research questions and test your hypothesis. **[5 MARKS]**

(g) The following is the calculated $\mathbf{X}^T\mathbf{X})^{-1}$ for the above model

$$\begin{pmatrix} 0.45375 & -0.02667 & -0.00417 \\ -0.02667 & 0.00167 & 0 \\ -0.00417 & 0 & 0.00093 \end{pmatrix}$$

Stating the general formula, calculate the 95% confidence interval of the mean minute ventilation of nestling bank swallows whose breathing air is comprised of 15 units of oxygen and 5 units of carbon dioxide? **[3 MARKS]**

(h) Stating the general formula, calculate the 95% prediction interval for the minute ventilation of a nestling bank swallows whose breathing air is comprised of 15 units of oxygen and 5 units of carbon dioxide? **[3 MARKS]**

Total: 60

END OF QUESTION PAPER.