Regression Modelling, December 2017
Answer Key

1. For each answers 1 mark is assigned for choosing the right answer and 1 more mark for providing the reasoning.

   (a) State whether the following three statements are TRUE or FALSE. Provide a one line explanation for your answer. **[6 MARKS]**

   i. FALSE: Cannot comment on causation **[2 MARKS]**

   ii. FALSE: Only non linear relationship **[2 MARKS]**

   iii. TRUE: OLS estimator **[2 MARKS]**

   (b) (iv) The model is pre determined. Need to know the data or plot the data to Estimate whether the association is linear or non-linear.

   (c) (ii)

   (d) (iv)

2. (a) After reviewing the residual plot of each of the datasets (a), (b), (c) and (d) identify the plots which indicates **[4 MARKS]**

   - (A) — None
   - (B) — Strong non-constant variance
   - (C) — Mild non-constant variance
   - (D) —Non-linearity

   (b)
   - (A) — None
   - (B) — Transformation — log or other
   - (C) — Transformation — log or other
   - (D) — include higher order term or other variables that have not been included

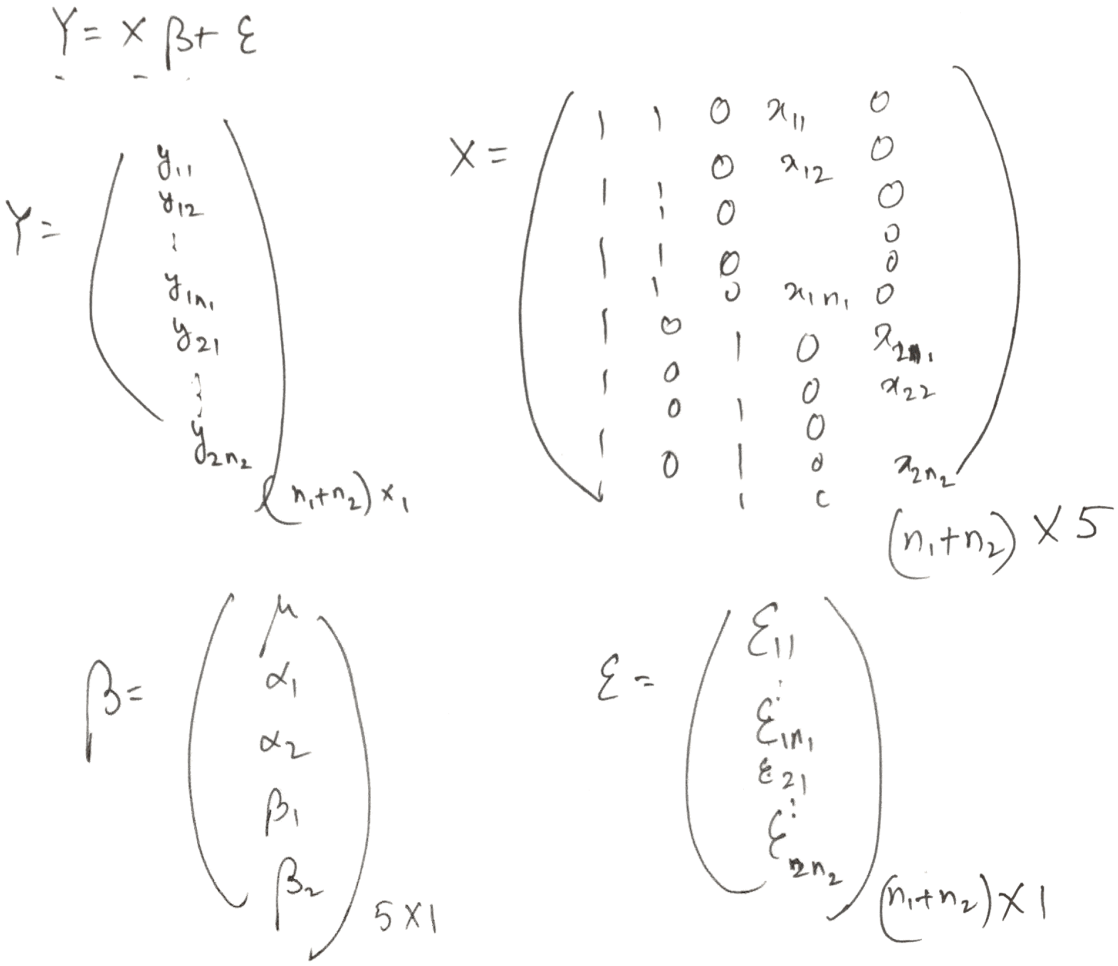
   (c) An influential point is an outlier (either in predictor/response or both) that greatly affects the slope of the regression line. One way to test the influence of an outlier is to compute the regression equation with and without the outlier. **[2 MARKS]**

   (d) Cook's Distance **[1 MARKS]**

END OF QUESTION PAPER.

3. (a) See handwritten solution

   (b) See handwritten solution



   (c) Col2+Col3=Col1 **[2 MARKS]**
   If X is not of full rank a unique $(X^TX)^{-1}$ does not exist. **[2 MARKS]**

   Possible models **[2 MARKS]**

   Model 2: $Y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \quad \epsilon_{ij}$
   One can also use the grand mean and the indicators of group 1 or 2.

4. (a) There doesn't appear to be a substantial relationship between minute ventilation (Vent) and percentage of oxygen (O2).

END OF QUESTION PAPER.

The relationship between minute ventilation (Vent) and percentage of carbon dioxide (CO2) appears to be curved and with increasing error variance.

(b) The plot between percentage of oxygen (O2) and percentage of carbon dioxide (CO2) is the classical appearance of a scatter plot for the experimental conditions. The plot suggests that there is no correlation at all between the two variables. You should be able to observe from the plot the 4 levels of O2 and the 5 levels of CO2 that make up the $54 = 20$ experimental conditions.

(c)

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) + \epsilon_i$$

where

where:

y_i is percentage of minute ventilation of nestling bank swallow i x_{i1} is percentage of oxygen exposed to nestling bank swallow i x_{i2} is percentage of carbon dioxide exposed to nestling bank swallow i

and the independent error terms ϵ_i follow a normal distribution with mean 0 and equal variance σ^2.

(d)
- Multiple R-squared: 0.2682 **[2 MARKS]**
- Adjusted R-squared: 0.2557 **[2 MARKS]**
- Interpretation: Only 26.82% of the variation in minute ventilation is explained by taking into account the percentages of oxygen and carbon dioxide. **[2 MARKS]**

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

(i) Oxygen $H_0 : \beta_1 = 0 \quad vs \quad \beta_1 \neq 0$ Not rejected as p-value 0.408. Not a significant predictor **[2.5 MARKS]**

(ii) Co2 $H_0 : \beta_2 = 0 \quad vs \quad \beta_2 \neq 0$ rejected as p-value $\leq .05$. Significant predictor **[2.5 MARKS]**

(f) Need to state the general formula

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t\left(n-p; \frac{1+c}{2}\right)\sqrt{\frac{RSS}{n-p}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

and also get the t-value from the t-table

and then calculate b=c(1,15,5)

$$161.4647 \pm t(117; 0.975)\sqrt{\frac{2897566}{117}\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b}}$$

END OF QUESTION PAPER.

$$= 161.4647 \pm 1.98\sqrt{\frac{2897566}{117}0.01023148}$$

95% confidence interval (129.9396, 192.9898) **[3 MARKS]**

(g) Need to state the general formula

$$\mathbf{b}^T\hat{\boldsymbol{\beta}} \pm t\left(n-p; \frac{1+c}{2}\right)\sqrt{\frac{RSS}{n-p}(1+\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}$$

and then calculate

$$161.4647 \pm t(117; 0.975)\sqrt{\frac{2897566}{117}(1+\mathbf{b}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{b})}$$

$$= 161.4647 \pm 1.98\sqrt{\frac{2897566}{117}(1+0.01023148)}$$

95% prediction intervals (-151.79 474.7194)

[3 MARKS]

END OF QUESTION PAPER.