Assessment 3: Model Building, Variable Screening and Residual Analysis

Material from Chapters 5, 6, 7 & 8
Due 11 th April   Total marks: 100

Refer to the **Assessment Criteria and Guidelines** in the Assessment block of the Moodle site. Five marks will be awarded for the following:

- Clear expression including correct use of terminology and notation

- Presentation of figures and tables including relevant R output.

- Clearly and concisely annotated R code.

Question 1 [45 marks]

At the start of a second wave of the corona virus in a particular country, epidemiologists attempted to predict when the numbers of new cases would begin to decline due to lockdown procedures. The data is in file `virus.txt`. Data was available for the number of covid-19 cases (*cases*) recorded per week (*week*) in 5 cities.

(a) Plot *cases* against *week*. What does the plot suggest? [5 marks]

(b) Determine the order of the polynomial model required to fit the data. With reference to relevant outputs, justify each step in the process. [15 marks]

(c) For the final model in (b), check the model assumptions, and also identify any potential outliers or influential points. [15 marks]

(d) Plot the fitted values on a scatter plot of the data. Include a plot of the 95% confidence bands on your graph [10 marks]

Question 2 [50 marks]

A gas turbine is a combustion engine that can convert natural gas or other liquid fuels to mechanical energy which is then turned into electrical energy. Gas turbines are often used to power aircraft, trains, ships, electrical generators, pumps etc. Several measurements are used to analyse gas turbines performance including

- heat rate (`HEATRATE`) which indicates the amount of fuel required to generate one unit of electricity. Heat rate is measured in kilojoules per kilowatt per hour, and generally a lower heat rate is ideal.

- exhaust gas temperature (`EXH`) which generally stays between 400 - 700 ^0C

- inlet temperature (`INLET`) which can go up to 1600^0C

- air mass flow rate (`AIRFLOW`) which is a volumetric flow at standard conditions translates to a specific mass flow rate, in kilograms per second

A researcher collected these data on 67 gas turbines (`GASTURBINE.txt`) and tried to model `HEATRATE` as a function of three predictors `EXH`, `INLET` and `AIRFLOW`.

(a) Produce and interpret a pairs plot. What does the plot suggest as an appropriate model? Explain your response. [10 marks]

(b) Fit a *main effects* model. Produce relevant outputs that will allow you to check the four indicators of multicollinearity. Summarise your findings. [15 marks]

(c) Run *forward stepwise* model selection with the "upper" model containing all possible interactions. Include all relevant outputs and the summary table for the final model.

[15 marks]

(d) Write a concise (one to two paragraphs), informative conclusion based on your analysis and results. [10 marks]

Note: checking model assumptions is not required for this question.