MATH3823 / MATH5824M Computer Practical

For the week 26 February - 1 March

General information This R computer practical is not included in the module assessment but you are strongly encouraged to submit the key numerical answers as requested on the Microsoft Form accessed using the link "*Microsoft Form for for submission of key numerical answers*" in the "*Week 5 Computer Practical*" folder in Minerva. Any answers submitted before **2pm on Friday 1st March** will be graded automatically with feedback sent by email at the start of the following week. Please note that it is my intention to offer a similar **partial check** in advance of the final submission deadline of the Assessed Coursework and hence making use of the option for this non-assessed work will be good practice.

The aim of this R Computer Practical is to give you practice using R as preparation for the Assessment which takes place in a few weeks. Please note, however, that currently in lectures we have mainly discussed normal linear models but that the Assessment will also consider generalized linear models and hence will require more time to complete. The instructions below refer to a number $\texttt{N} \in (00, 01, \ldots, 99)$. This number represents the last two digits of your student ID. For example, if your student ID is 200123456, then $\texttt{N} = 56$ whereas if your ID is 200678900, then $\texttt{N} = 00$ – note that this is *double zero*. For these two student ID, the file names mentioned below would be *EngineEmissions-56.csv* and *EngineEmissions-00.csv* respectively.

When reporting your answers, but not for intermediate calculations, you should round values to 2 decimal places if RStudio returns more decimal places. As examples, a returned answer of 5 should be returned as 5; 5.1 should be entered as 5.1; 5.13 should be entered as 5.13; but 5.138 should be entered as 5.14, etc. The RStudio command `round` can be used, for example `round(5.138, 2)`.

Description of the problem Exhaust gases from road transport have a major impact on air quality and hence on health. For example, in the UK over 20% of total carbon dioxide emissions come from cars, vans and lorries. Currently, petrol cars are tested annually for three pollutants: hydrocarbons (HC), carbon monoxide (CO) and nitrogen oxides (NOX). These measurements are recorded in grams per minute (g/min). The testing is controlled by the Ministry of Transport (MOT) with cars being awarded a test PASS if the level of HC is below 0.6 g/min, CO is below 6 g/min and NOX is below 1.2 g/min.

There is a general belief that a car engine that is working well will have low readings on all tests and that badly maintained cars will emit high levels of all pollutants. Over the years governments have imposed strict regulations on car manufacturers to reduce the emission levels of new cars but over 20% of the cars on UK road are, however, more than 13 years old.

This computer practical investigates the relationships between emission levels of different pollutants and the prediction of an MOT Pass/Fail.

A sample of data in the file *EngineEmissions-*`N`*.csv* contains gas exhaust emission measurements, in grams per minute (g/min), collected from a local MOT testing centre. You can find this file in the usual online data folder.

Start by calculating the correlation matrix between the variables and comment on any important features.

Next, consider the relationship between the levels of HC and NOX. The gas test for NOX requires more equipment and is more expensive than the HC test and hence it is of interest to see if the NOX level can be predicted from the HC level. It is believed that a linear regression model will be appropriate, with NOX as the response variable and HC as the explanatory variable. Use the correlation coefficient and a scatter plot to comment on the relationship. Then, fit the linear regression to the data. Describe the key steps in your analysis, including parameter estimates, and qualitatively comment on goodness of fit of the fitted model

relative to the data using an appropriate graph. Also, perform a residual analysis to check that the linear model is sufficient, or not, and support your comments with appropriate graphs. Use the fitted model to predict NOX values when the HC is 0.4 g/min and 0.8 g/min. Do you think that NOX levels can be reliably predicted from the HC level? Which of these two predictions do you think is most reliable?

Finally, do you think that it is worthwhile to also include the carbon monoxide (CO) level in the prediction of NOX? Use ANOVA testing to support your choice of best model. Again, you should perform a residual analysis to check that the new linear model is sufficient, or not, and support your comments with appropriate graphs.

End of Practical