DATA 303/473 Assignment 1

Nokuthaba Sibanda, 301111111

Due 1159pm Friday 15 March 2024

Instructions

- Prepare your assignment using Rmarkdown
- Submit your solutions in two files: an Rmarkdown file named e.g. `a1.Rmd` and the PDF file named `a1.pdf` that results from knitting the Rmd file. The Rmarkdown file (`assignment1.Rmd`) used to create these assignment questions is provided as an example for you to follow should you wish.
- The YAML header of your Rmarkdown file must contain your name and ID number in the author field, and should have the output format set to `pdf_document`. For example:

```
---
title: "DATA 303/473 Assignment 1"
author: "Nokuthaba Sibanda, 301111111"
output: pdf_document
---
```

- While you are developing your code you may find it easiest to have the output set to `html_document`, but change it to `pdf_document` when you submit.
- In your submission, embed any executable R code in code chunks, and make sure both the R code and the output is displayed correctly when you knit the document.
- If there are any R code errors, then the Rmarkdown file will not knit, and no output will be created at all. So if you can't get your code to work, but want to show your attempted code, then put `error=TRUE` in the header of the R code chunk that is failing.

```
‘‘‘{r, error=TRUE}
your imperfect R code
‘‘‘
```

- Title each question answer with its question numbers as `Q1.`, `Q2`,… instead of `1.`,`2.`,….
- Where you are asked to perform a hypothesis test, state the hypotheses being tested and give the test statistic, p-value and conclusion.

Assignment Questions

Q1. (28 marks) The noise generated by an aircraft is an efficiency and environmental matter for the aerospace industry. A vital component of the total airframe noise is the airfoil self-noise, caused by the interaction between an airfoil blade and the turbulence it produces. The airfoil self-noise dataset was obtained by NASA from a series of tests of airfoil blade sections conducted in an echoless wind-tunnel. The dataset, obtained from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/291/airfoil+self+noise) and available in the file `airfoil_self_noise.csv`, contains the following variables:

- `frequency`: Frequency, in Hertzs.
- `angle`: Angle of attack, in degrees.
- `chord_length`: Chord length, in metres.
- `speed`: Free-stream velocity, in metres per second.
- `displace`: Suction side displacement thickness, in meters.
- `decibels`: Scaled sound pressure level, in decibels.

We will carry out a regression analysis to investigate the relationship between the response variable `decibels` and the other variables in the dataset as predictors.

a. **(4 marks)** Carry out an exploratory data analysis (EDA). NOTE: The predictors `chord.length` and `speed` are numerical, but only have a few different values each (6 for `chord.length` and 4 for `speed`). Such variables are best treated as categorical variables during the analysis. List any key points of note from your EDA, including any considerations you might make during a regression analysis.

b. **(3 marks)** Fit a linear model to the data, including all predictors with no transformations or interactions. Present a summary of the model in a table and write the fitted model equation. Give an estimate of σ^2, the error variance.

c. **(3 marks)** Based on the fitted model results in part (b), give an interpretation of the coefficients for `angle` and `speed71.3`.

d. **(2 marks)** Does it make practical sense to interpret the intercept in this case? Justify your answer.

e. **(3 marks)** Obtain 95% confidence and prediction intervals for the last three observations in the dataset. Explain briefly why the prediction intervals are wider than the confidence intervals.

f. **(4 marks)** Use the `plot` function to carry out residual diagnostics for the model you fitted in part (b). Comment on what the residual plots indicate about regression assumptions or the existence of influential observations.

g. **(4 marks)** Test the assumptions of normality and constant variance in the errors. Do the results confirm the conclusions you reached in part (f) about these assumptions? In your response, include the hypotheses being tested in each test.

h. **(2 marks)** Use the VIF statistic to check whether or not there is evidence of severe multicollinearity among the predictors. Comment on the results.

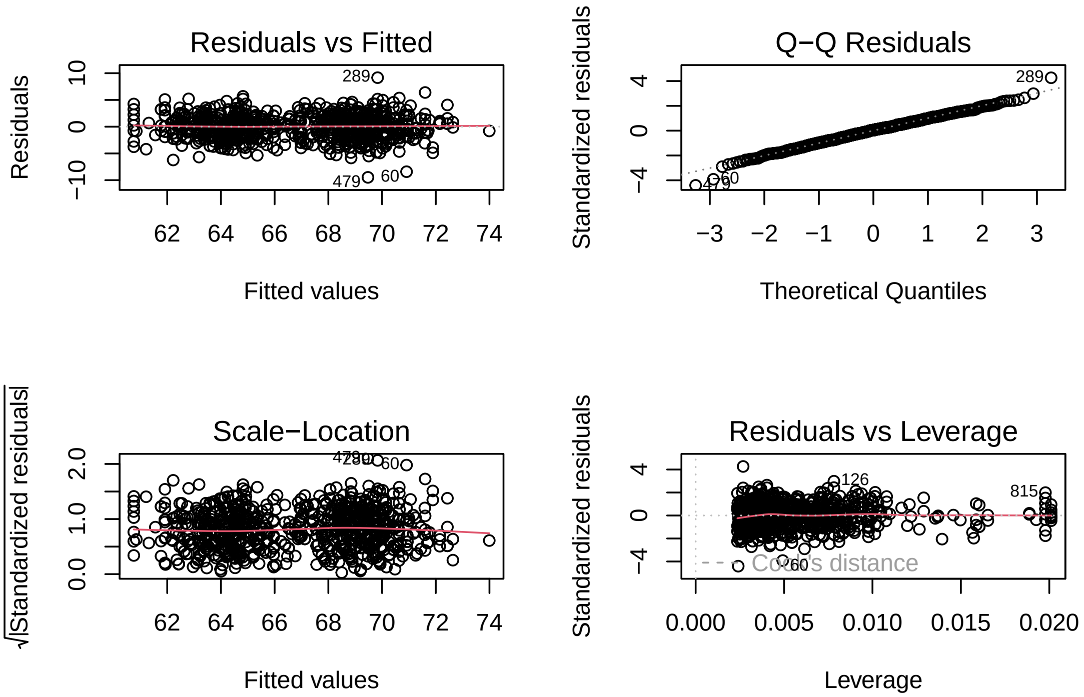
i. **(3 marks)** Based on a global usefulness test, is it worth going on to further analyse and interpret a model of `decibels` against each of the predictors? Carry out the test, give the conclusion and justify your answer.

Q2. (12 marks) Francis Galton's 1866 dataset (cleaned) lists individual observations on height for 899 children. Galton coined the term "regression" following his study of how children's heights related to heights of their parents. The data are available in the file `galton.csv` and contain the following variables:

- `familyID`: Family ID
- `father`: Height of father
- `mother`: Height of mother
- `gender`: gender of child
- `height`: Height of child
- `kids`: Number of childre in family
- `midparent`: Mid-parent height calculated as (‘father + 1.08*mother)/2
- `adltchld`: `height` if `gender`=M, otherwise 1.08*`height` if `gender`= F

All heights are measured in inches.

a. **(3 marks)** Read the data into R and fit a linear model for `height` with the variables `father`, `mother`, `gender`, `kids` and `midparent` as predictors. Provide a summary of the fitted model. You will notice that estimates for `midparent` are listed as `NA`. Why might this be the case and what regression problem does this point to?
b. **(2 marks)** What action might you take to resolve the problem identified in part (a)?
c. **(2 marks)** Based on the model fitted in part (a) give an interpretation of the coefficient for `genderM`.
d. **(2 marks)** Determine the number of families in the dataset.
e. **(3 marks)** The problem in part (a) is resolved and a new linear model is fitted.No observations are excluded. The plots below are obtained to investigate regression assumptions for this new model. Based on your answer in part (d) and the plots below, do the data meet all the regression assumptions? Explain your answer briefly.



Assignment total: 40 marks