DATA 303/473 Assignment 1

Due 1159pm Friday 17 March 2023

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 (`tut1sol.Rmd`) used to create the Tutorial 1 Solutions is provided in the Tutorial Solutions section of Canvas 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"
date: "Due: 17 March 2023"
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
```
```

- **You will receive an email confirming your submission. Check the email to be sure it shows both the Rmd and PDF files have been submitted.**
- 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 1 Questions

Q1. (33 marks) Data on selected used car sales in India were obtained from the CarDekho website (https://www.cardekho.com/). The car sales occurred between 1983 and 2020. The data are available in the dataset `cardekho.csv` and include the variables:

- `price`: Selling price in thousand Rupees (Rs)
- `make`: Car make grouped into eight categories: `Ford`, `Honda`, `Hyundai`, `Mahindra`, `Maruti`, `Tata`, `Tpyota`, `Other`
- `kms`: Kilometres driven (x 1000)
- `fuel`: Fuel type: `Diesel` or `Petrol`
- `seller`: Seller type: `Dealer`, `Individual` or `Trustmark Dealer`
- `tx`: Transmission type: `Automatic` or `Diesel`
- `owner`: Current owner is: `First`, `Second` or `Third or above` owner
- `mileage`: Fuel economy in kilometres per litre (kmpl)
- `esize`: Engine size in cubic centimetres (CC)
- `power`: Maximum engine power in brake horse power (bhp)

We will analyse the data with `price` as the response variable and the rest of the variables listed above as predictors.

a. **(4 marks)** Use the `summary()` command to obtain a summary of the variables in the `cardekho.csv` dataset. Based on the results:
i) Identify any numerical variables (if any) that may have obviously incorrect values.
ii) Identify any variables (if any) that have missing observations.

Data cleaning is done to filter out some observations and a new dataset, `cardekho2.csv,` (available on Canvas) is created. This new dataset should be used to answer the rest of Question 1.

b. **(4 marks)** Read in the `cardekho2.csv` dataset and create a scatterplot matrix of all numerical variables in the dataset.
   i) Identify any predictors (if any) that have a non-linear relationship with the response variable `price`.
   ii) Will there be a need to apply a transformation to the response variable `price` when fitting a linear regression model? Explain your answer briefly.

c. **(3 marks)** Fit a linear model for `price` including all predictors with no transformations or interactions. Present a summary of the model in a table. Give an estimate of σ^2, the error variance.

d. **(2 marks)** List the predictor values for a car whose predicted price, $\widehat{E[Y|\mathbf{X}]}$, equals the intercept $\hat{\beta}_0 = -745843.41$ INR.

e. **(4 marks)** Based on the model fitted in part (c), give an interpretation of the coefficient for:
   i) `txManual`
   ii) `mileage`

f. **(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.

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

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

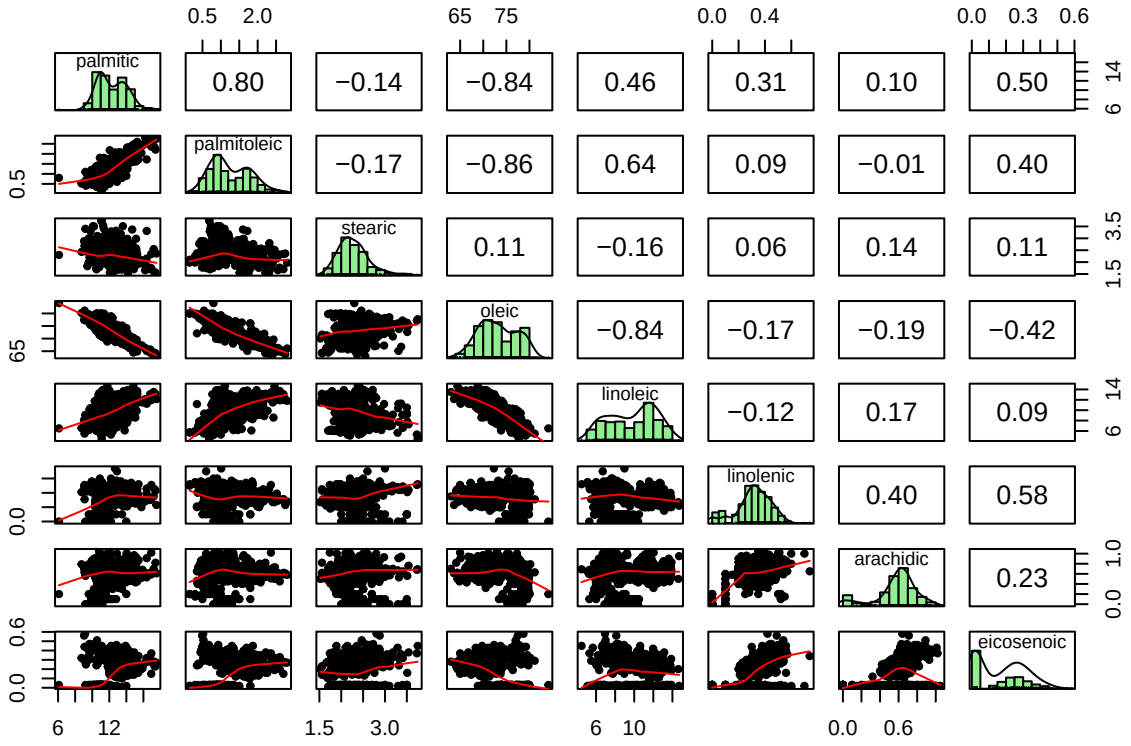
i. **(2 marks)** Use the VIF statistic to check whether or not there is evidence of severe multicollinearity among the predictors. Comment on the results.

j. **(4 marks)** Based on a global usefulness test, is it worth going on to further analyse and interpret a model of `price` against each of the predictors? Carry out the test, give the conclusion and justify your answer.

Q2. (7 marks) Data were collected on the fatty acid composition of 572 olive oil samples from Italy. There was interest in investigating the relationship between palmitic acid, and some of the other constituents: linoleic acid, stearic acid and oleic acid. The data are available in the file `olive.csv` and were read into R and analysed as follows:

```
olive<-read.csv("olive.csv", header=T)
str(olive)
```

```
## 'data.frame':    572 obs. of  10 variables:
##  $ region     : chr  "Southern Italy" "Southern Italy" "Southern Italy" "Southern Italy" ...
##  $ area       : chr  "North-Apulia" "North-Apulia" "North-Apulia" "North-Apulia" ...
##  $ palmitic   : num  10.75 10.88 9.11 9.66 10.51 ...
##  $ palmitoleic: num  0.75 0.73 0.54 0.57 0.67 0.49 0.66 0.61 0.6 0.55 ...
##  $ stearic    : num  2.26 2.24 2.46 2.4 2.59 2.68 2.64 2.35 2.39 2.13 ...
##  $ oleic      : num  78.2 77.1 81.1 79.5 77.7 ...
##  $ linoleic   : num  6.72 7.81 5.49 6.19 6.72 6.78 6.18 7.34 7.09 6.33 ...
##  $ linolenic  : num  0.36 0.31 0.31 0.5 0.5 0.51 0.49 0.39 0.46 0.26 ...
##  $ arachidic  : num  0.6 0.61 0.63 0.78 0.8 0.7 0.56 0.64 0.83 0.52 ...
##  $ eicosenoic : num  0.29 0.29 0.29 0.35 0.46 0.44 0.29 0.35 0.33 0.3 ...
```



```
fit2<-lm(palmitic ~  linoleic + stearic, data=olive)
summary(fit2)
```

```
##
## Call:
## lm(formula = palmitic ~ linoleic + stearic, data = olive)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.193 -1.238  0.069  1.078  4.519
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.15814    0.51779  19.618   <2e-16 ***
## linoleic     0.30855    0.02625  11.755   <2e-16 ***
## stearic     -0.37847    0.17344  -2.182   0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.493 on 569 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.216
## F-statistic: 79.67 on 2 and 569 DF,  p-value: < 2.2e-16
```

```
fit3<-lm(palmitic ~  linoleic + stearic + oleic, data=olive)
summary(fit3)
```

```
##
## Call:
## lm(formula = palmitic ~ linoleic + stearic + oleic, data = olive)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25377 -0.20139 -0.00748  0.19258  1.53568
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.70728    0.61290  115.36   <2e-16 ***
## linoleic    -0.67472    0.01149  -58.75   <2e-16 ***
## stearic     -0.80706    0.04020  -20.08   <2e-16 ***
## oleic       -0.68283    0.00678 -100.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.344 on 568 degrees of freedom
## Multiple R-squared:  0.9586, Adjusted R-squared:  0.9584
## F-statistic:  4381 on 3 and 568 DF,  p-value: < 2.2e-16
```

a. **(2 marks)** In the `fit2` model, the coefficient for `linoleic` is positive, while it is negative in the `fit3` model. Making reference to the scaterplot matrix, give a possible explanation for the change in sign for the linoleic acid coefficient.

b. **(2 marks)** Use the model in `fit3` to obtain 95% confidence and prediction intervals of `palmitic` for an olive oil sample with:

   - `linoleic`= 0.3
   - `stearic` = 2.2
   - `oleic`= 73.0

c. **(1 mark)** If all regression assumptions hold, what other condition would have to be met for the prediction in part (b) to be valid.

d. **(2 marks)** Some of the olive oil samples originated from the same region of Italy. What regression assumption is violated in this case? Explain your answer briefly.

Assignment total: 40 marks