STAT201 Assignment 2

Linear and Quadratic Regression
Due by 3pm on Friday 18 March 2022

Your assignment is the answers to the questions below. You will need to copy the relevant output and graphs from R and put them into a Word document (or another editor). You can copy your graphs in R using export, and copy to clipboard.

Write your assignment so it can be read easily. You need to include your R code in your assignment, but do not put it in the main part of your assignment. Instead put your R code at the end as an appendix. You can save your R code from the script screen.

Submit your completed assignment by uploading it on the Learn webpage where you downloaded this question sheet. It must be uploaded before the due date.

You can upload more than one version of your assignment, and the most recent version is the one that will be marked.

Question 1

The dataset, poverty.csv, contains information from States in USA. There are two variables, one for a deprivation score of poverty in the State (Dep), and another for the percentage of people who have experienced crime (Crime).

a. Download the data file from Learn into a folder, and then, in R set the working directory to this. Import the dataset into R. Check the data has read in correctly by printing out the top 6 rows using `head()`, the bottom 6 rows using `tail()`.
b. Explore the data with a suitable graph that shows how crime and deprivation are related (refer to assignment 1). Also create a histogram of crime, to check if it has an approximate normal distribution. Write a few sentences about what you see in both graphs.
c. Create a linear regression model to predict crime from poverty, using the lm function in R:
   `poverty.lm1<-lm(Crime ~ Dep, data = poverty)`
d. Use the summary option in R, `summary(poverty.lm1)`, to print out your model. Describe the relationship between deprivation score of poverty and the percentage of people who have experienced crime.
e. Look at the residual plots for your model and comment on each:
   ```
   plot(poverty.lm1, which = 1)
   plot(poverty.lm1, which = 2)
   plot(poverty.lm1, which = 4)
   ```
f. Use your model to predict, by hand, the percentage of people who have experienced crime in States with a deprivation score of 10. Show your workings.
g. Use the predict function in R to predict percentage of people who have experienced crime States with a deprivation scores of 10, 15, and 20, and show the prediction intervals:

   ```
   pred<-data.frame(Dep=c(10,15,20))
   predict(poverty.lm1, pred, interval = “prediction”)
   ```

Question 2

The dataset, airquality.csv, has data from a study over time (measured in days) of air pollution levels.

a. As in the question above, download the data file. Check the data has read in correctly by printing out the top 6 rows using `head()`, the bottom 6 rows using `tail()`. Use a suitable graph to look at pollution levels over the days. Also create a histogram of the pollution measurements. Write a few sentences about what you see in both graphs.

b. Fit a linear regression model to predict pollution from time: `air.lm1<-lm(Pollution~Time, data = airquality)`, and print out your model with the summary option in R. Describe the relationship between pollution levels and time.

c. Look at the three residual plots, as in the previous question, for your model and comment on each. You do not need to remove any data points and instead just mention any concerns with the Cook's Distance graph.

d. Fit a quadratic model: `air.lm2<-lm(Pollution~Time+I(Time^2), data = airquality)` and print out your model with the summary option in R.

e. Does the summary output suggest you can reduce the quadratic model to the simpler linear model? Explain your answer.

f. Look at the three residual plots for your quadratic model and comment on each and compare these with your linear model. You do not need to remove any data points and instead just mention any concerns with the Cook's Distance graph.

g. Create a plot with the data points, and your linear and quadratic models. Looking at this graph, and the summary outputs of your linear and quadratic models, explain which of the two is your preferred model and why. Some R-code for this final graph is below:

```
x<-airquality$Time

plot(Pollution~Time, data = airquality)

lines(airquality$Time[x], fitted(air.lm1)[x], lwd=2)

lines(airquality$Time[x], fitted(air.lm2)[x], lwd=2, col="blue")
```