STAT201 Assignment 3

Linear Regression with a Factor

Due by 3pm on Friday 25 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, training.csv, contains information from students who participated in a sports training programme, where strength training was added to reduce injuries. The time (in minutes) spent on strength training per week and injury score (a score of the number and severity of the injury) was collected from each student. The students either used a supervised strength training programme or did their strength training in an unsupervised, informal way.

a. Download the data file, training.csv, 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()`. The variables are Injury, Time and Formal. Formal is a factor with two level, yes (supervised training) or no (informal, unsupervised training).
b. Explore the data with a dot plot that shows how injuries and time spent in strength training are related (for students in formal and not formal training), `plot(Injury~Time, data=training, pch = 19, col = Formal).` With this code, the black dots are students with informal training and the red dots are students with formal training.
   Next, create a boxplot of the injury scores for the two levels of the variable "Formal", `boxplot(Injury~Formal, data=training).`
   Write a few sentences about what you see in both graphs. Explain why your box plot is suggesting that formal training is not as good as informal training for reducing injuries.
c. Create a linear regression model to predict injury from time spent training, using the lm function in R:
   `training.lm1<- lm(Injury ~ Time * Formal, data = training).`
   Use the summary and anova options in R, `summary(training.lm1)`, `anova(training.lm1)`, to print out your model. Explain, in a few sentences, what information is in the output. Can you reduce your model to a simpler model?
d. Look at the residual plots for your model and comment on each:

```
plot(training.lm1, which = 1)
plot(training.lm1, which = 2)
plot(training.lm1, which = 4)
```

e. Use your model, and looking at the dot-plot you made with injury and time for the formal and informal training, explain whether strength training reduces injury
f. Write out by hand (to show all your working) what you predict as an injury score for a student who spends 300 minutes per week training with informal training. Repeat this for a student who has formal training.

Question 2

The dataset, autompg.csv, has the miles per gallon (MPG) of 4, 6 or 8 cylinder cars, and the displacement in cubic inches (Disp).

a. As in the question above, download the data file. Look at the variables and check if cylinder is a factor or a numeric. If it is a factor you can use the code `as.factor(Cyl)`.
b. Use a suitable graph to look at miles per gallon with displacement. To show the different cylinder cars you can use the code below. With this simple code the colours of the dots, are the default, and are black for the 4 cylinder cars, red for the 6 cylinder cars and green for the 8 cylinder cars. There are tidier ways to plot this where you can specify the colours, you don't need to do this for this assignment but explore how to do this if you want. Comment on your graph.
```
plot(MPG~Disp, data = autompg, pch = 19, col = as.factor(Cyl))
```
c. Fit a linear regression model to predict MPG from displacement and the count of cylinders:
`cars.lm1<-lm(MPG~Disp*as.factor(Cyl), data = autompg)`, and print out your model with the summary and the anova option in R.
d. Look at the three residual plots, as in the previous question, for your model and comment on each.
e. Can you simplify your model? Explain your answer.
f. Using your dot-plot where you should be able to see the data for the 8, 6 and 4 cylinder cars, and the summary outputs of your model, explain what is the relationship between MPG and displacement, and the number of cylinders.