Assignment 1 - 50 marks

John O'Sullivan
ESILV-J - Data Statistics
Dorset College Dublin

Due by midnight on Sunday the 05th of April 2020

Instructions

- This assignment involves two tasks - Question 1 (25 marks) involves a simple linear regression, and Question 2 (25 marks) involves a multiple linear regression.
- You should submit your assignment to the 'Assignment 1' object on Moodle.
- You should submit two files:
  (i) a single .Rmd script file containing all of the commented code you used to obtain your answers
  (ii) the HTML (or pdf) file which you produced from the .Rmd script
- I have uploaded a template and some files to help you get started using RMarkdown. I will also share a video with you to help you do this.
- You may need to find some new functions in order to do some of these tasks. Remember to use R's search engine, as well as checking online.
- Make sure that your file is readable and has a neat presentation and clear flow. The HTML (or pdf) output file **must be a stand-alone document containing the answers to all questions and showing all necessary code and all necessary output**. Marks will only be given for clear and detailed answers.
- I advise you to first create an R script with all of your answers. When you are happy with this, convert it piece-by-piece into an .Rmd file.
- There are marks in this assignment for document presentation - your final document should be neat with a clear layout, showing a good use of RMarkdown to mix free-flowing text, code, and output.

Question 1 - Simple Linear Regression

All steps of the calculations for these question must be shown fully in order to get full marks. This is similar to what we have seen in the lectures, e.g. the 'long' way to find S_{xx}, S_{xy}, etc. You can use R to check your answers, but must write out the formula and show how the answers are found.

This question involves a **simple linear regression**. Access the dataset *salaries.csv* on Moodle. Read this into R. It contains the salaries and experience (in years) of lecturers in a particular college. Now answer the following questions.

Question 1.a

Explore the dataset. Describe both variables, and include any plots which you feel are necessary.

Produce a scatterplot of years of experience (x-axis) vs. salary (y-axis). Do you think linear regression is appropriate for this dataset? Explain your reasoning.

Question 1.b

Estimate the slope and the intercept using least-squares regression equations. That is, evaluate the best estimate of the intercept, β_0, and the best estimate of the slope, β_1, using the equations we saw in class. Interpret both of these values.

Question 1.c

An employee thinks that salary is independent of a lecturer's years of experience. Conduct a hypothesis test to test this claim. (Hint: this is testing if the slope β_1 equals 0 (and therefore $y = \beta_0$, a constant). Use a significance level of $\alpha = 0.05$ when doing this test.)

You may use R's **lm()** function to get the necessary information for this question. As with any hypothesis test, you must clearly show the following: state the null and alternative hypotheses; calculate the test statistic; then find the critical value/calculate the p-value; and finally, reach a conclusion.

Question 1.d

Calculate and interpret the R^2 value for the linear regression of Salary on Years of Experience. Show the steps for this calculation (though as before, you can check your answer from R's lm() function).

Is the model a useful one?

Question 1.e

A new lecturer has joined the college, and wants to know what they should expect to earn. They have 6.5 years of experience. They ask you to use your model to calculate:

1. a **point prediction** (that is, a single 'best guess' value) of the salary they should expect to earn in the college, which is denoted $E(y|x = 6.5) = \bar{y}_{6.5}$.
2. a 95% **confidence interval** for the mean salary of people with 6.5 years of experience
3. a 95% **prediction interval** for the salary of the new lecturer

Note: the **confidence interval** here refers to the expected average salary of all people with 6.5 years of experience; the **prediction interval** refers to a prediction for a single new observation - that is, one individual starting work in the college. You will need to do some research for these formula, and can use R to evaluate them. As before, code for all calculations must be shown (not just your final answer). You must include the steps needed to apply the formula explicitly, e.g.: $\bar{y}_{6.5} = \beta_0 + \beta_1(6.5)$, where these coefficients are used from Question 1.b or later parts. (You can use the predict() function here to check your answers, but must find the answers clearly using the relevant formula - these slides may help.)

Interpret your answers - how would you explain these three pieces of information (your three answers) to the new employee?

Question 1.f

A lecturer with 30 years of experience is about to join the company. They ask you to predict their salary using your model. Should you use your model to do this? If so, what is your point prediction of their salary?

Question 1.g

Produce a nice plot of the final model, suitable for inclusion in a work presentation.

This plot should show the underlying data, and the line of best fit. Labels should be neat and clear. If you can figure out how to do this, it would also be helpful to include confidence and prediction intervals for the line (you will need to research how to do this).

Question 2 - Multiple Linear Regression

There is a file called **medical-expenses.csv** on Moodle. It contains the following information on 1,338 people in Ireland: age, sex, BMI (body mass index), number of children, smoker, region, and annual medical expenses.

An insurance company wants to use this information to help it to decide on the premiums to charge new customers. They ask you to write a report where you present a model they can use to do this.

In your report, consider structuring it in the following way: Introduction, Data Exploration, Methodology, Results, Conclusion. You should apply some/all of the 6 steps we have seen in class for multiple linear regression, and clearly outline the decisions you make as you conduct your analysis.

Your report is intended for the insurance company workers to read, and should be neat and clear and concise.

Note that some of these variables are continuous and some are not. Also consider that you may want to include interactions in your model choice. As with the previous question, you may need to do some extra research online to help you.

As a final test of your analysis, the company asks you to predict the expected medical expenses for a new customer: a male aged 30 with 2 children, who is a smoker with a BMI of 32, and who lives in the southeast of Ireland.