Assignment – Regression Analytics

Purpose

This assignment integrates your knowledge from the practice activities, readings, and narrated presentations within this Module.

Learning Outcomes

- Identify opportunities for capturing value in various domains which can be solved using regression analysis. (CLO 2)
- Explain assumptions, terminology, variables, and the output of linear regression analysis. (CLOs 1, 6)
- Evaluate the role of coefficient of determination in regression models as it applies to business metrics. (CLOs 4, 6)
- Evaluate the fitness of a linear regression model as it applies to business metrics. (CLOs 4, 6)
- Identify various input features (e.g. numerical and categorical) into a regression model and evaluate the model output. (CLOs 3, 6)
- Implement and execute linear regression models using R. (CLO 5)

Instructions

Please answer all questions from 1 to 3. You should use R to solve the questions and include the screen shots in your submission. The Golden questions are optional and carries additional marks. This means that you will not lose marks if you do not answer that question. Your solution should include the calculation steps and your conclusion that is clearly expressed.

Questions

1. Run the following code in R-studio to create two variables X and Y.

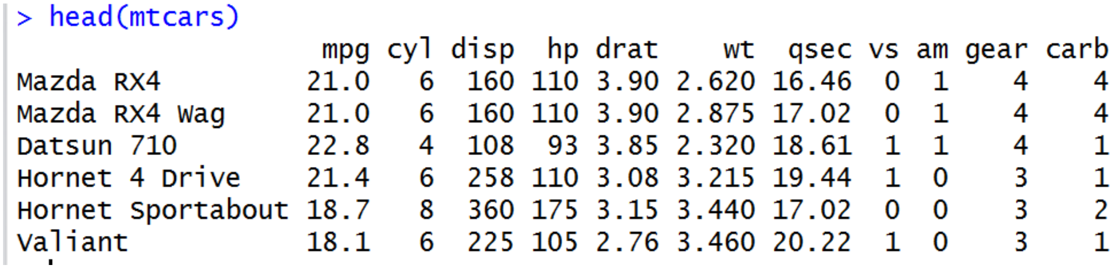
   ```
   set.seed(2017)

   X=runif(100)*10

   Y=X*4+3.45

   Y=rnorm(100)*0.29*Y+Y
   ```

   a) Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X? (8% of total points)

   b) Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model? (8% of total points)

c) How the Coefficient of Determination, R^2, of the model above is related to the correlation coefficient of X and Y? (8% of total points)

2. We will use the ‘mtcars’ dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset. The description of the dataset can be found here.



```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

a) James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question. (17% of total points)

b) Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp). Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22? (17% of total points)

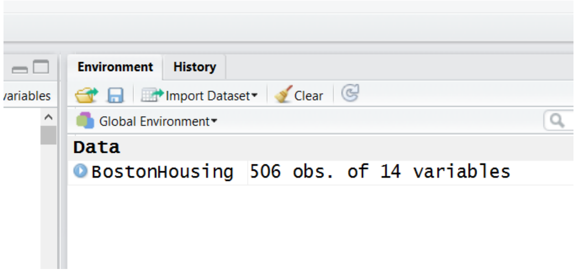
3. For this question, we are going to use BostonHousing dataset. The dataset is in ‘mlbench’ package, so we first need to instal the package, call the library and the load the dataset using the following commands

install.packages('mlbench')

library(mlbench)

data(BostonHousing)

You should have a dataframe with the name of BostonHousing in your Global environment now.



```
Data
BostonHousing  506 obs. of 14 variables
```

The dataset contains information about houses in different parts of Boston. Details of the dataset is explained here. Note the dataset is old, hence low house prices!

a) Build a model to estimate the median value of owner-occupied homes (medv)based on the following variables: crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River(chas). Is this an accurate model? (Hint check R^2) (8% of total points)

b) Use the estimated coefficient to answer these questions?

   I. Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much? (8% of total points)

   II. Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15 and in the other one is 18. Which one is more expensive and by how much? **(Golden Question: 4% extra)**

c) Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer. (8% of total points)

d) Use the anova analysis and determine the order of importance of these four variables. (18% of total points)

General Submission Instructions:

All work must be your own. Copying other people's work or from the Internet is a form of plagiarism and will be prosecuted as such.

You may submit a Microsoft Word (.doc/.docx) document as an attachment using the Canvas Assignment tool, or you may copy and paste your answer into the provided box within the Assignment tool. If you attach a document for your assignment, be sure to include your name in the text of the document and in the name of the document.

- You can only submit once, so make sure you are completely finished before submitting and that you attach the correct word .doc/.docx file.
- Submissions sent by email will NOT be accepted.

Due dates are listed in the Assignment Schedule document.