303-2_Assignment1_Template

January 10, 2022

1 STAT303-2: Assignment 1

1.1 Instructions:

a. You may talk to a friend, discuss the questions and potential directions for solving them. However, you need to write your own solutions and code separately, and not as a group activity.

b. Do not write your name on the assignment. (1 point)

c. Export your Jupyter notebook as a PDF file. If you get an error, make sure you have downloaded the MikTex software (for windows) or MacTex (for mac). Note that after installing MikTex/MacTex, you will need to check for updates, install the updates if needed, and re-start your system. Submit the PDF file. (1 point)

d. Please include each question (or question number) followed by code and your answer (if applicable). Write your code in the 'Code' cells and your answer in the 'Markdown' cells of the Jupyter notebook. Ensure that the solution is well-organized, clear, and concise (3 points)

1. It's easy enough to identify different sections of the homework assignment (e.g., if there are different sections of an assignment, they're clearly distinguishable by section headers or the like)

2. It's clear which code/markdown blocks correspond to which questions.
3. There aren't excessively long outputs of extraneous information (e.g., no printouts of entire data frames without good reason)

This assignment is due at 11:59pm on Wednesday, January 19th. Good luck!

Graded out of 42 points (max 47 points possible) – 37 points for code & answers, 5 points for anonymity and proper formatting

1.2 Part 1

Read the dataset petrol_consumption_train.csv. It contains the following five columns:

Petrol_tax: Petrol tax (cents per gallon)

Per_capita_income: Average income (dollars)

Paved_highways: Paved Highways (miles)

Prop_license: Proportion of population with driver's licenses

Petrol_consumption: Consumption of petrol (millions of gallons)

(20 points - max 25 possible)

a. Make a pairwise plot of all the variables in the dataset. Which variable seems to have the highest linear correlation with petrol_consumption? Let this variable be predictor P. (3 points for visualization, 1 point for answer)

Note: If you cannot figure out P by looking at the visualization, you may find the pairwise correlation coefficient to identify P.

b.Fit a simple linear regression model to predict petrol_consumption based on predictor P (identified in the previous part). Print the model summary. (2 points for code)

c. What is the increase in petrol consumption for an increase of 0.01 in P? (1 point for answer)

d. Does petrol consumption have a statistically significant relationship with the predictor P? (1 point for answer)

e. What is the R-squared? Interpret its value. (1 point for answer, 1 point for interpretation)

f. Estimate the petrol consumption for a state in which 50% of the population has a driver's license. What are the confidence and prediction intervals for your estimate? (2 points for code, 1.5 points for answer)

g. Estimate the petrol consumption for a state in which 10% of the population has a driver's license. Are you getting a reasonable estimate? Why or why not? (2 points for code, 1.5 points for answer)

h. Estimate the petrol consumption for the observations in petrol_consumption_train.csv. Find the RMSE. (2 points for code, 1 point for answer)

1.2.1 Bonus point questions: (5 points with no partial credit, will only give credit if all parts are answered correctly)

i. Fit a simple linear regression model to predict petrol_consumption based on predictor P, but without an intercept term.

j. Estimate the petrol consumption for the observations in petrol_consumption_test.csv using the model in developed in (i). Find the RMSE

k. The RMSE for the models in (i) and (b) are similar, which indicates that both models are almost equally good. However, the R-squared for the model in (j) is much higher than the R-squared for the model in (b). Why?

1.3 Part 2

A study was conducted on 97 men with prostate cancer who were due to receive a radical prostatectomy. The dataset prostate.csv contains data on 9 measurements made on these 97 men. The description of variables can be found here: https://rafalab.github.io/pages/649/prostate.html

(17 points total)

a. Fit a linear regression model with lpsa as the response and the other variables as the predictors. Write down the equation to predict lpsa based on the other eight variables. (2 points for code, 1 point for answer)

b. Is the relationship between lpsa and the predictor variables significant at 0.05? (1 point for answer)

c. Report the p-value for age. What do you conclude about the significance of this variable? (1 point for answer, 1 point for conclusion)

d. What is the 95% confidence interval for the coefficient of age? Can you conclude anything about its significance based on the confidence interval? (1 point for answer, 1 point for conclusion)

e. Fit a simple linear regression on lpsa against age. What is the p-value for age? (2 points for code, 1 point for answer)

f. Explain why this p-value in (e) is different from the p-value in (b). (3 points for answer)

g. Predict lpsa of a 65-year old man with lcavol = 1.35, lweight = 3.65, lbph = 0.1, svi = 0.22, lcp = -0.18, gleason = 6.75, and pgg45 = 25 and find 95% prediction intervals. (2 points for code, 1 points for answer)