DASC 500 Homework 4

Each of the questions for this assignment involves the mtcars data set, provided as part of this assignment in Canvas. For many, mtcars is known as a "hello world" data set for students learning linear regression.

The data were extracted from the 1974 Motor Trend US magazine. The mtcars data includes one response variable mpg (fuel consumption) and 10 predictor variables that represent different aspects of automobile design and performance for 32 automobiles. If you're interested, R-Pubs by R-Studio posted a description of the variables in the mtcars data set.

Problem 1

Partition the mtcars data into a training set and a test set.

- Use a 70/30 split where 70% of the data is contained in the training set and 30% is contained in the test set.
- Verify the data is partitioned as intended.

As shown in the multiple linear regression lecture slides you can generate these data sets using the sample() and drop() functions that are included in a pandas.DataFrame object. When using this method you're able to include the random_state argument in the sample() function. By specifying a value for random_state you ensure that your random sample is repeatable. Using the same value for random_state makes it easier to replicate and compare results.

For this assignment, set random_state = 42.

Documentation for the pandas DataFrame.sample() statement is available at pandas.DataFrame.sample — pandas 1.3.3 documentation (pydata.org)

Problem 2

Using the training data generated in Problem 1, build 10 different simple linear regression models where mpg is the response variable, and the predictor is one of the other 10 variables.

Your answer to this problem should include a table that presents the results of the 10 simple linear regression models and contains the following columns:

- Predictor: The name of the predictor
- β_0: The constant term for each model
- β_1: The slope parameter for each model
- t-stat: The value of the t statistic for β_1
- LCL: The lower limit of the confidence interval for β_1
- UCL: The upper limit of the confidence interval for β_1
- R^2: The coefficient of determination

(Continued on next page)

Problem 3

Consider the results you obtained in Problem #2.

Which one of the 10 simple linear regression models do you consider to be the best?

Explain the rationale for your choice.

Problem 4

What is the equation for the simple linear regression model you selected in Problem #3?

What does it tell us about how the predictor variable affects mpg?

Problem 5

Using the training data generated in Problem 1, build and fit a single model that predicts mpg as a function of all 10 predictors; that is, a multiple linear regression model with an intercept term and 10 predictor variables.

Show a summary of the results from this model.

Problem 6

Compare the results you obtained in Problem #5 (the multiple linear regression) to those of Problem #2 (the 10 simple linear regression models).

Do the results of the larger model align with those seen in the 10 individual models?

In either case, explain your conclusion.

Extra Credit Problem (worth 20%)

A parsimonious model is one that achieves a suitable goodness of fit using as few explanatory variables as possible.

Using the training data set you generated in Problem #1, develop a parsimonious regression model of the mtcars data and compare it to the model that includes all 10 predictors.

Which model do you prefer and why?