PES University, Bengaluru
UE18CS312 - Data Analytics

Session: Aug – Dec 2020
Weeks 3-4 – Worksheet 2(a) (for Unit 2)

Compiled by: Ms. Richa and Ms. Mainaki Saraf
VII CSE, PES University RR Campus

1.0 Getting started

For this worksheet, we will be using the Boston housing dataset from the ISLR package in R. The dataset can also be found here:
https://www.kaggle.com/altavish/boston-housing-dataset
The description of the dataset can be found here:
https://www.kaggle.com/c/boston-housing

1. Read the dataset
2. Check for missing values and use an appropriate technique to fill it in.

1.1 Correlation

1. Plot a correlogram and check the correlation between the variables. Check out the different parameters and types of plot.
2. What inferences can you draw from the correlogram?
3. Try finding the different correlation (spearman rank, point biserial, phi coefficient) from the data by dividing them into ordinal or binary data based on a certain range. What can you see from the output?
   Hint: Explore **phi, biserial.cor** and the method parameter of **cor.test**

1.2 Simple Linear Regression (SLR)

1. Explore the **lm** package in R for this section.
2. Plot the best fit line for all the variables with the target variable **medv.** Play around with the variables and map them to the concepts studied in theory.
3. Plot the residual plots for all the variables as well. Which plot satisfies all the assumptions of linear regression?
   Hint: Use the **plot** function in R with appropriate parameters
4. Find the best predictor for the target variable **medv** using the following methods:
   a. Correlation: Which is better - high correlation or low correlation?
   b. Residual plots
5. Is there any variable (apart from the best predictor) that can be transformed to fit the SLR model while satisfying all assumptions? If yes, perform the valid transformation and try to fit the SLR model again.
   Hint: Plot a histogram to draw inferences

1.3 Multiple Linear Regression (MLR)

1. For this section also we will be using the **lm** package. Explore how we can use it for MLR models.
2. Using correlation and residual plots from Section 1.2, decide the variables you want to keep.
3. Is there any multicollinearity?
   Hint: Use VIF to verify
4. Analyse the residual plot to ensure that all the assumptions of MLR are fulfilled.
5. Plot the best fit line and compare it to SLR.

Part 2

2.0 Getting started

For this worksheet, we will be using the dataset and models from the previous worksheet (both SLR and MLR model).
The dataset can also be found here:
https://www.kaggle.com/altavish/boston-housing-dataset
The description of the dataset can be found here:
https://www.kaggle.com/c/boston-housing

2.1 R-square analysis

1. In section 1. 2 and 1.3 we calculated the best fit SLR and MLR models for the Boston Housing dataset. Calculate the R-square value of both the models and compare. What do you infer?
2. Calculate the adjusted R-square value too. Is there a large variation from R-square? If yes, why?

2.2 Q-Q plot analysis

1. Analyse the residuals from the two models above and check if they are normally distributed or not.
2. Plot the summary of both the models.

2.3 ANOVA

1. Find the one-way ANOVA of the dataset with the best predictor of SLR and another variable.

a. Print the summary of both and compare them. List down the differences.
b. Play around with the parameters and see how it influences the output.

2. Find the two-way ANOVA of the dataset for the variables you think are most significant based on correlation.
   Print the summary of two of them and compare the results.

2.4 SLR with Gradient Descent

1. Set the SSE as the loss function and using the concept of SLR with Gradient Descent find the optimal values of m and c.

a. Start with m = 0 and c = 0, learning rate = 0.01, 0.001
b. Run it for 250, 500 and 1000 epochs, check the R-square at each point.
c. Draw inferences from the R-square value and reason out as to why we get those values.