Tutorial and Laboratories

Week 13

The purpose of this week's tutorial and lab exercises is to understand the data modelling and how to build a linear regression model between the explanatory and predictor variables.

The objectives of this tutorial and lab are to:

1- Understand the linear regression analysis.
2- Practice on using R to build single and multiple linear regression model and to be able to evaluate these models.

Linear Regression:

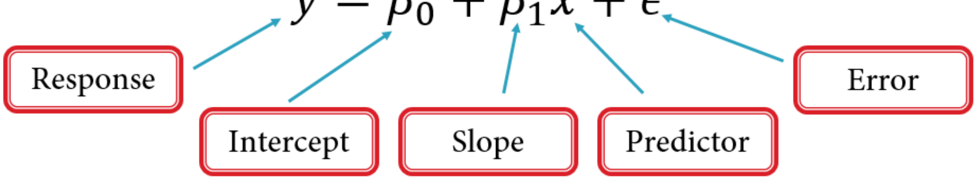
Linear regression is a linear approximation of the causal relationship between two or more variables. The aim of using the linear regression is to predict the parameters that explain the relationship between predictors (*i.e.,* explanatory) variables and predicted (*i.e.,* response) variable.

To estimate a relationship between variables, the following steps are used:

1. Get the sample data.
2. Design a model that works for that sample.
3. Make predictions for the whole population.
4. Measure performance

The linear regression can be estimated by finding the parameters (B_0 and B_1) the following formula:

$$y = \beta_0 + \beta_1 x + \epsilon$$



Exercise 1

1- Download the data from this link (advertising data under week 12)
2- Read the data in R-studio and get familiar with the included variables.
3- Remove the un-named data variable(s) if any!
4- Split the data into training and testing sets, where the training is 60% and the testing is 40% of the data.

5- Build a model that describes the changes of the sales with the changes in the money spent in advertising with the TV. This is meant to build a single linear regression model.
6- Build another model that would describe the changes of the sales with the changes in the money spent on all the variables. This means building a multi-linear regression model.
7- Test both models on the test set.
8- Compute the residual mean square error (RMSE) to evaluate the output of the two models.

Exercise 2

1- Download the data from this link[1](Housing data from week 13)
2- Read the data and get familiar with the included variables.
3- Check the distribution of the price variable visually. Is it normally distributed?
4- Check the correlation matrix between the variables, numerically and graphically.
5- Check the outliers of the variables visually using boxplot.
6- Split the data into training and testing sets, where the training is 70% and the testing is 30% of the data.
7- Build a model that describes the changes of the prices with that variable that produces the highest correlation with the price variable.
8- Can you print and interpret the model parameters?
9- Visualise the model!
10- Test the model on the testing data and then evaluate the model performance using Residual Mean Square Error (RMSE) metric.
11- Build another model that would describes the changes of the prices with the changes in all the variables. This means building a multi-linear regression model.
12- Test the second model on the testing data and evaluate its performance using RMSE metric.
13- Visualise the distributions of the residuals of the two models.

Answering questions:

Please use this time to ask your tutor about any part or any exercises, which you feel that it was not clear to you.

[1] Please note that the original data has been downloaded from Kaggle (https://www.kaggle.com/aariyan101/usa-housingcsv/)