BMIG 6201
Homework # 3: LDA, QDA and Subset Selection

Horacio Gómez-Acevedo, PhD.

February 26, 2021

1. Using the *Auto* dataset, provide a solution to the following questions

   (a) Create a binary variable, $mpg01$, that contains a 1 if mpg contains a value above its median, and a 0 if mpg containes a value below its median.

   (b) Explore the data graphically in order to investigate the associate between $mpg01$ and the other features using scatterplots and boxplots. Describe your findings.

   (c) Split the data into a training set and a test set.

   (d) **Fit a QDA and LDA on the training data in order to predict $mpg01$ using the variables that seemed most associated with $mpg01$. Calculate the test error of the models obtained.**

2. Using the *Concrete* data set

   `https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength`

   data set you will use a regression model to predict the strength.

   (a) Scale the **"condition*" variables (not the rest)** using a logarithmic scale.

   (b) Use the snipped provided of the **forward selection** to reduce the number of covariates. **You need to import the following:** `import statsmodels.formula.api as smf`.

   (c) Produce a python program that calculates the best subset selection.

   (d) Compare the best subset selection with the one obtained by the forward selection.