MSiA-400 Everything Starts with Data
Lab Exercise #2

Due Date: Thursday, November 15, 12 pm

EXERCISE INSTRUCTIONS: Please submit one report file that includes: short answer, related code and print for each problem if necessary. Push your answers to Github.

The redwine data includes 11 input (explanatory or independent) variables and 1 output (response or dependent) variable for regression analysis. The 11 explanatory variables include:

> fixed acidity(FA), volatile acidity(VA), citric acid(CA), residual sugar(RS),
> chlorides(CH), free sulfur dioxide(FS), total sulfur dioxide(SD), density(DE),
> pH(PH), sulphates(SU), and alcohol(AL).

The output variable is quality(QA) (score between 0 and 10).

Please find the attached data file *redwine.txt* which contains missing values in attributes SD and RS. The summary of the data set is following.

Name of the data set	redwine
Number of columns	12 (11 explanatory variables and 1 response variable)
Number of observations	1599
Number of missing values	39 (22 in RS and 17 in SD)

Answer the following questions.

Problem 1

Recall that RS and SD have missing values. Calculate the averages of RS and SD by ignoring the missing values.

Problem 2

After correlation analysis, Mr. Klabjan observed that there exists a significant correlation between SD and FS. Create vectors of SD.obs and FS.obs by omitting observations with missing values in SD. Build (simple) linear regression model to estimate SD.obs using FS.obs. That is, SD.obs is used as response variable and FS.obs is used as explanatory variable for the regression analysis. Print out the coefficients of the regression model.
Hint: If you save the output from `lm` function to `ABC`, then the coefficients of the regression model can be obtained by `coefficients(ABC)`.

Problem 3

Create a vector (of length 17) of estimated SD values using the regression model in Problem 2 and FS values of the observations with missing SD values. Impute missing values of SD using the created vector. Print out the average of SD after the imputation.

Problem 4

Mr. Klabjan decided RS is not significantly correlated to other attributes. Impute missing values of RS using the average value imputation method from the lab. Print out the average of RS after the imputation.

Problem 5

We have imputed all missing values in the data set. Build multiple linear regression model for the new data set and save it as winemodel. Print out the coefficients of the regression model.
Hint 1: built multiple linear regression by winemodel = lm(redwine$QA∼redwine$FA+...+redwine$AL)

Problem 6

Print out the summary of the model. Pick one attribute that is least likely to be related to QA based on p-values.

Problem 7

Perform 5-fold cross validation for the model you just built. Print out the average error rate.

Problem 8

Mr. Klabjan is informed that the attribute picked in Problem 6 actually contains outliers. Calculate the average μ and standard deviation σ of the selected attribute. Create a new data set after removing observations that is outside of the range $[\mu - 3\sigma, \mu + 3\sigma]$ and name the data set as redwine2. Print out the dimension of redwine2 to know how many observations are removed.

Problem 9

Build regression model winemodel2 using the new data set from Problem 8 and print out the summary. Compare this model with the model obtained in Problem 6 and decide which one is better. Pick 5 attributes that is most likely to be related to QA based on p-values.