ISYE 6501, Week 10 HW

**Question 1**

The breast cancer data set breast-cancer-wisconsin.data.txt from http://archive.ics.uci.edu/ml/machinelearning-databases/breast-cancer-wisconsin/ (description at http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29 ) has missing values. 1. Use the mean/mode imputation method to impute values for the missing data.

2. Use regression to impute values for the missing data.

3. Use regression with perturbation to impute values for the missing data.

4. (Optional) Compare the results and quality of classification models (e.g., SVM, KNN) build using (1) the data sets from questions 1,2,3; (2) the data that remains after data points with missing values are removed; and (3) the data set when a binary variable is introduced to indicate missing values.

**Response –**

First, the ratio of missing values in column V7 is calculated.

Missing values ratio = 2.28%

Hence imputation can be performed

1. Mean/Mode Imputation

Mean of V7 without missing values = 3.54

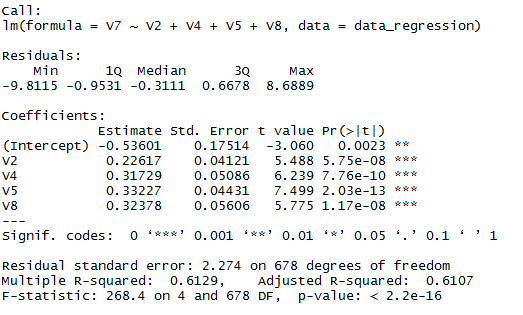
After rounding it to 4, the same value is assigned to all rows with “?” for V7

Mode of V7 without missing values = 1

The same value is assigned to all rows with “?” for V7

1. Regression

Response variable is not considered in regression. The remaining predictors are used to calculate V7 using step regression. V2, V4, V5 and V8 are the significant predictors that are used in the final regression model.



The missing values of V7 are calculated as

5 8 1 2 1 2 3 2 2 6 1 3 5 2 1 1

1. Regression with Perturbation

Perturbation is performed on V7 values calculated using regression model and applying normal distribution on them.

The upper and lower limits of values calculated are set to 10 and 1

The missing values of V7 are calculated as

4 8 1 5 2 1 4 3 3 5 4 3 4 1 3 1