Assignment 3 (due Nov 5) (6 points)

In this assignment you are given a small data set (train and test) similar with the one in Tutorial 4, where some attribute values are missing. The goal is to explore using regression to impute missing values. Furthermore, you explore whether missing value imputation helps in classification.

1- Classification with missing value samples removed: Load the training data set called *Assign3trainMissingValues.csv*. This training data set, however, contains some samples with missing values on the attribute *loan*. Remove those samples with missing values and train a logistic regression model called MClean. using the attributes *income*, *age* and *loan* to classify the label *default10yr*. Make sure that you pick the optimal cut-off from the training data. Apply MClean to test on the test set called *Assign3Test.csv*. Report the sensitivity and specificity on the test set.

2- Missing value imputation: Use regression to predict the missing *loan* values in *Assign3trainMissingValues.csv*. There are 3 possible regression models you can build. The first one uses *income* only; let us call this regression model RIncome. The second uses *age* only; let us call this RAge. The third one uses both *income* and *age*; let us call this RBoth. Apply RIncome to impute the missing values. Load the file *Assign3TrueValues* which gives the true values of the missing loans. Compute the total error of the imputed values. Similarly, apply RAge and RBoth to impute. Report the regression model that gives the smallest error of the imputed values.

3- Classification with imputed missing values: since you have the missing values imputed from the regression models, re-train the logistic regression model. Specifically, add the imputed samples to the samples with no missing values in *Assign3trainMissingValues.csv* and get models MIncome, MAge and MBoth corresponding to the three regression models of imputation. Apply the three logistic regression models to the test set *Assign3Test.csv*. Report the sensitivities and specificities on the test set.

4- **Hand in:** Apart from submitting your R script and reporting the sensitivities and specificities of the four models, comment on the following:
   a. The motivation of missing value imputation is to increase the number of training samples. Do any of the three regression models help?
   b. Does the regression model that gives the smallest error in imputation also give the best classifier?