**UCS410: PROBABILITY AND STATISTICS**

**Laboratory Assignment – 7**

Q1) Simulate normal distribution values. Imagine a population in which the average height is 1.70 m with a standard deviation of 0.1. Use rnorm to simulate the height of 1000 people and save it in an object called heights.

a) Plot the density of the simulated values.
b) Generate 10000 values with the same parameters and plot the respective density function on top of the previous plot in red to differentiate it.

This plot will show you how much a sample with 10000 simulations approximate to the real normal distribution.
c) Find the 90% interval of a population with mean = 1.70 and standard deviation = .1 between 0.05 and 0.95.
d) Calculate the qvalue corresponding to every percentile in standard normal distribution.
e) Calculte the pvalues corresponding to z values ranging from 0 to 1 at an interval of 0.05.

Q2) Download the Auto.csv data set from LMS. Based on it program the following problems in R.

a) Calculate simple (linear) correlation between car price and its fuel economy (measured in miles per gallon, or mpg)
b) Create a correlation matrix by selecting each pair of columns from dataset one by one and calculate correlation between selected pairs. Fill the values in matrix named as correlation matrix.
c) Create a new dataframe, auto_num, that contains only columns with numeric values from the auto dataframe. You can do this using the Filter function. Use the cor function to create a matrix of correlation coefficients for variables in the auto_num dataframe.

d) Use the corrgram function from the corrgram package to create a default correlogram to visualize correlations between variables in the auto dataframe.
e) Create a new dataframe, auto_subset, by subsetting the auto dataframe to include only the Price, MPG, Hroom, and Rseat variables. Use the new dataframe to create a correlogram that (1) shows correlation coefficients on the lower panel, and (2) shows scatter plots (points) on the upper panel.
f) Analyze the correlation values to understand the association between pair of column datasets.

## Laboratory Assignment – 8

**Q1** Implement the linear regression on a regression dataset to be downloaded from LMS using the concept of training and testing in order to understand the accuracy of results using the following metric.

**a)** Correlation between predicted and actual value on testing part of data.
**b)** Accuracy metric
**c)** Visualization of best fit line.

Q2 Execution of the following 3 R commands will give us the data {(x(i), y(i), z(i), i = 1, 2, …,100}.

```
x<-rpois(100, 50)
y<-rpois(100, 100)
z<-rpois(100, 150)
```

Using this data:

a) Fit the linear regression model of the form $z = a + b.x + c.y$ using,
b) Fit the 3 models of the form $y = a + b.x$, $y = a + b.x + c.x^2$, and $y = a.b^x$ to this data using
c) Find the coefficient of determination, with the help of formula, for the three models and decide for the best model.