CMTH 642 Lab 11

Please first of all install the needed packages with the install.packages() function. And then, call the following libraries.

```
library(RCurl) # getURL
library(MASS) # stepwise regression
library(leaps) # all subsets regression
```

Dataset

In our examples, we will use the a computer dataset which contains information about price, speed, hd, ram, screen, cd, multi, premium, ads, and trend. The purpose of this lab is finding the best com

We first download the dataset as follows:

```
u <- getURL("http://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/Computers.csv")
c_prices <- read.csv(text = u)
```

Q1) Split the dataset to 70% of training and 30% of test sets. We want to make sure that the training set and the test set do not have any common data points.

Q2) Multiple Linear Regression Algorithm

- **a)** Create lm model based on train set. Use multiple linear regression model to predict the 'price' variable based on 'ram', 'screen', 'speed', 'hd' and 'ads' as independent variables.
- **b)** Use predict() function on the test set
- **c)** Calculate error (prediction price – test price) in predictions and show the histogram of error
- **d)** Calculate mean square error (mse) and find the percentage of cases with less than 25% error.

Q3) Use simple linear regression model by using 'ram' as an independent variable. Compare the results with the multiple linear regression.

a) **Create lm model based on train set.**

b) **Use predict() function on the test set**

c) **Calculate error (prediction price – test price) in predictions and show the histogram of error**

d) **Calculate mean square error (mse) and find the percentage of cases with less than 25% error.**

e) **Compare the results with the multiple linear regression.**

Q4) Forward and Backward selection algorithm

a) **Forward**: Start with 'null', which means none of the indepenent variables are selected. You will come up with a selection of independent variables between 'null' and 'full'. 'full' means all the independent variables are included. You will end up using all the variables. Set 'trace=TRUE' to see all the steps.

b) **Backward**: We can also use 'backward' elimination, which will start with 'full'.

Q5) Variable selection using automatic methods

The R package leaps has a function regsubsets that can be used for best subsets, forward selection and backwards elimination depending on which approach is considered most appropriate for the application under consideration.

a) Use regsubsets() to see the best combination of the 6 attributes

b) What are the best 4 attributes to predict computer price based on this analysis?

Q6) Price Prediction using k Nearest Neighbor Regression

a) Suppose we have the following new record, and we want to predict its price based on kNN regression method and all the previous computer price records available.

c(7000,**0**,32,90,8,15,'no','no','yes',200,2)

For now, its price is set as 0, so, the algorithm should predict the price. Use **knn.reg()** for prediction of the price. (use k=7 and algorithm="kd_tree")