Title: Assignment #5: (*Regression Modeling on Different Datasets*)

Purpose: *This assignment is to study regression models in R and how do they work on different datasets and how to validate the model built using the dataset.*

Dataset(s): *The dataset is obtained from the UCI Machine Learning Repository. The link for dataset can be found here:*

https://archive.ics.uci.edu/ml/datasets/Auto+MPG

https://archive.ics.uci.edu/ml/datasets/Student+Performance

Approach:

For Dataset -1 – Auto MPG – Predict MPG for the dataset

- *The dataset is loaded.*
- *The dataset is quite tricky, it is separated by both tabs and spaces. I did a work around and merged the data frame obtained to separated based on both the whitespaces and tabs.*
- *It consists of a 8 observations. There were NA's in the dataset which were omitted and the dataset was of '392x8' dimensions.*
- *The carName had not much significance with the model prediction. Hence it was removed.*
- *The model was built using the remaining dimensions and I observed :*
    1. *There was no collinear dimension.*
    2. *I couldn't find any other dimension which had no significance.*
    3. *None of the dataset was non-linear with the model built.*
- *The model was built using a training dataset of about 68% and the square root of MSE obtained was 13% of the average value of MPG*
- *The LOOCV model was built and the K-Fold CV model was built and the observation showed the dataset provided the similar observation for prediction roughly about 14% of the average value of MPG.*

For Dataset 2 – Student Performance – Predict the Student Grade (G3):

- *First the dataset is loaded. Checked for NA's if any (this dataset had none).*
- *It had significantly large dimensions consisting of 33 features to determine the prediction variable.*
- *The dataset had about 646 values and the linear model was built on top of it.*
- *When the linear model was built, the following observations were made:*
    1. *I couldn't find any collinear features for predicted vector.*
    2. *There were no features which I can find as not so significant.*

- *The model built using training dataset of about 68% and the square root of MSE obtained was about 9% of the average value of G3.*
- *The LOOCV and K-Fold CV model was built and a similar observation for the prediction was made.*

Summary:

It is observed that the prediction result accurately depends on finding the features which are significant and produce more accurate results for dataset/model. The collinear features don't provide much significance. The significant variables can be more accurately obtained by also doing PCA to remove features which are similar among them.

Appendix :

This includes the R code that I have done to obtain the analysis for regression models:

```
library(readr)
auto_mpg <- read_delim("~/Downloads/Education/Spring 17/R for Data
Scientists/Assignments/auto-mpg.data-original",
                "\t", escape_double = FALSE, col_names = FALSE,
                trim_ws = TRUE)
View(auto_mpg)
head(auto_mpg)
summary(auto_mpg)
dim(auto_mpg)
colnames(auto_mpg) = c
("MPG","Cylinder","Displacement","HorsePower","Weight","Acc","ModelYr","Origin","CarNa
me")
myData <- auto_mpg
myData <- na.omit(myData)
dim(auto_mpg)
dim(myData)
lr.model.1 <- lm(MPG~.,data=myData)
summary(lr.model.1)

#try the model using train and test
myData = myData[,-8]
train_data = sample(392,266) #68%
lr.model.2 <- lm(MPG~., data=myData, subset=train_data)
summary(lr.model.2)
mean(lr.model.2$residuals^2) #12.47473
summary(myData$MPG)  #0.66% of average MPG, so that seems to be okay

mean( (MPG - predict(lr.model.2, myData) ) [-train_data]^2) #9.9745
```

```
#try doing with LOOCV
library(boot)
lr.model.boot = glm(MPG~., data=myData)
summary(lr.model.boot)

cv.error.boot = cv.glm(myData,lr.model.boot)

cv.error.boot$delta #12.08526

cv.error.kcv <- cv.glm(myData,lr.model.boot,K=10)

cv.error.kcv$delta #12.20015 - 14.8%
```

```
library(readr)
student_por <- read_delim("~/Downloads/Education/Spring 17/R for Data
Scientists/Assignments/student/student-por.csv", ";", escape_double = FALSE, trim_ws =
TRUE)

myData_grade <- as.data.frame(student_por)

lr.model.1 <- lm(myData_grade$G3~.,data=myData_grade)

summary(lr.model.1)

#train and test
train_data = sample(649,441) #68%

lr.model.2 <- lm(G3~., data=myData_grade, subset=train_data)
summary(lr.model.2)

mean(lr.model.2$residuals^2) #1.26 - about 9% of data falls

#prediction
mean( (myData_grade$G3 - predict(lr.model.2, myData_grade) ) [-train_data]^2) #2.0593

#try with LOOCV
lr.model.boot = glm(G3~., data=myData_grade)
summary(lr.model.boot)

cv.error.boot = cv.glm(myData_grade,lr.model.boot)

cv.error.boot$delta #1.686186
```

```
cv.error.kcv <- cv.glm(myData_grade,lr.model.boot,K=10)

cv.error.kcv$delta #1.675 - 9%
```