---
title: "Assignment2_Regression"
author: "Ram"
date: "11/16/2021"
output: word_document
---
1)Run the following code in R-studio to create two variables X and Y.

```{r}
set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y
```

a) Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X?

```{r}
plot(X,Y) # The Plot describes we can fit the linear Model.
cor(X,Y) # since the correlation is positive, we can fit the linear
model.
```

b) Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model?

```{r}
LinearModel<-lm(Y~X)
summary(LinearModel)
LinearModel$coefficients
```
#Y=3.6108X+4.4655 is the equation of the model
#The accuracy of above linear model is 65.17%, Varability y can be explained by x.

c) How the Coefficient of Determination,R2 of the model above is related to the correlation coefficient of X and Y?

#The square of correlation coefficient is same as coefficient of determination, 0.6517
#Coefficient of Determination = (Correlation Coefficient)^2

```{r}
Coefficient_Determination<-(cor(Y,X))^2
Coefficient_Determination
```

2. We will use the Ã¢â‚¬ËœmtcarsÃ¢â‚¬â„¢ dataset for this question. The dataset is already included in your R
distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset. The description of the dataset can be found here

a) James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question.

```{r}
head(mtcars)

LinearModel2<- lm(mtcars$hp~mtcars$wt,data = mtcars)#hp is HorsePower and
wt is weight
summary(LinearModel2)

LinearModel3<- lm(mtcars$hp~mtcars$mpg,data = mtcars)#hp=horsepower,
mpg=Mile per Gallon
summary(LinearModel3)
```

#chris is correct by seeing the multiple r-squared value, mpg got high r square value 0.6024 compared to wt of car 0.4339.

b) Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp). Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22?

```{r}
LinearModel4<-lm(mtcars$hp~mtcars$mpg+mtcars$cyl,data = mtcars)
summary(LinearModel4)
LinearModel4$coefficients
```

To check the Estimated HP of a car with 4 cylinder and mpg of 22:

```{r}
Predict_Model <- (LinearModel4$coefficients[1] +
(LinearModel4$coefficients[2]*22) + (LinearModel4$coefficients[3]*4))
Predict_Model
```

#Estimated Horse Power of a car with 4 cylinder and mpg of 22 is 88.93618

3)For this question, we are going to use BostonHousing dataset. The dataset is in Ã¢â‚¬ËœmlbenchÃ¢â‚¬â„¢ package, so we first need to instal the package, call the library and the load the dataset using the following commands

```{r}
library(mlbench)
data(BostonHousing)
```

a) Build a model to estimate the median value of owner-occupied homes (medv)based on the following variables: crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River(chas). Is this an accurate model? (Hint check R2)

```{r}
HousingModel<-
lm(BostonHousing$medv~BostonHousing$crim+BostonHousing$zn+BostonHousing$p
tratio+BostonHousing$chas,data=BostonHousing)
summary(HousingModel)

```

#The Model Accuracy is 0.3599, and is not Accurate model.

b) Use the estimated coefficient to answer these questions?
I). Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much?

```{r}
HousingModel1<-lm(medv~chas,data = BostonHousing)
summary(HousingModel1)
HousingModel1$coefficients
```

#For chas = 0 and #For chas = 1
```{r}
withoutchas <-
((HousingModel1$coefficients[2]*0)+HousingModel1$coefficients[1])
withoutchas
withchas <-
((HousingModel1$coefficients[2]*1)+HousingModel1$coefficients[1])
withchas
```

#By using the correlation cofficients the house with chas is more expensive than house without chas. Differnece is 6.34616

II). Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15 and in the other one is 18. Which one is more expensive and by how much?

```{r}
HousingModel2<-lm(BostonHousing$medv~BostonHousing$ptratio,data =
BostonHousing)
HousingModel2
```

#By using the correlation coffecients, the coffecients are negative. We can say that the ptratio increases as the housing price decreases.

```{r}
ptratio15 <-
((HousingModel2$coefficients[2]*15)+HousingModel2$coefficients[1])
ptratio15
ptratio18 <-
((HousingModel2$coefficients[2]*18)+HousingModel2$coefficients[1])
ptratio18
```

````
```
````

#The price of house whice has ptratio of 15 is high compared to price of house which has a ptratio of 18 by difference of 6.47153

c) Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer.

```{r}
summary(HousingModel)
```

#A low p-value is low (< 0.05) we you can reject the null hypothesis.Looking at the p-values we can say that none of the independent varialbes are statistically significant.

d) Use the anova analysis and determine the order of importance of these four variables.

```{r}
anova(HousingModel)
```

#By comparing p values,below is the order from high to low

1) crim = 6440.8
2) ptratio = 4709.5
3) zn = 3554.3
4) chas = 667.2
```