Assignment 3

Harish Kumar Uddandi

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(e1071)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
library(dplyr)
```

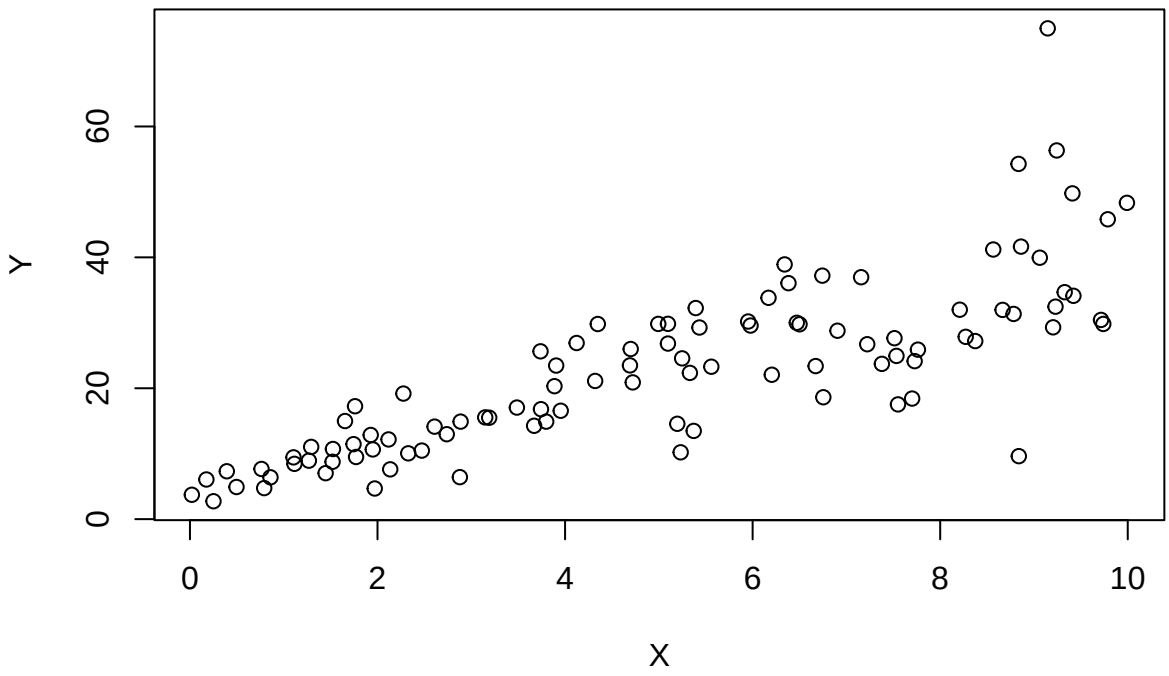
Q1) Run the following code in R-studio to create two variables X and Y.

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?

Answer:

```
library(tidyverse)
library(ggplot2)
set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y
plot(X,Y)
```



Yes, it appears that a linear model might be fit.

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?

```
accuracy <- lm(Y ~ X)
summary(accuracy)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.755  -3.846  -0.387   4.318  37.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4655     1.5537   2.874  0.00497 **
## X             3.6108     0.2666  13.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.756 on 98 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6482
## F-statistic: 183.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

Here, the formula is Y=3.61*X+4.46. R2 in this instance is 0.65. This indicates that 65% of the variability is explained by the model.

c)

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

```
cor(X,Y)^2
```

```
## [1] 0.6517187
```

Answer: We are reiterating the equality of the two values. so that means Coefficient of Determination = (Correlation Coefficient)^2

2. We will use the 'mtcars' 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.

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

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.

```
summary(lm(mtcars$hp~mtcars$wt))
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$wt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.430 -33.596 -13.587   7.913 172.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.821     32.325  -0.056    0.955
## mtcars$wt     46.160      9.625   4.796 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 52.44 on 30 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4151
## F-statistic:    23 on 1 and 30 DF,  p-value: 4.146e-05
```

R^2 of the model based on the weight is 0.43

```
summary(lm(mtcars$hp~mtcars$mpg))
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mtcars$mpg     -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

R^2 of the model based on mpg is 0.60 Therefore, it is more accurate model.So, Chris is right here.

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).

```
summary(lm(mtcars$hp~mtcars$cyl+mtcars$mpg))
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$cyl + mtcars$mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067     86.093   0.628  0.53492
## mtcars$cyl    23.979      7.346   3.264  0.00281 **
## mtcars$mpg    -2.775      2.177  -1.275  0.21253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.22 on 29 degrees of freedom
## Multiple R-squared:  0.7093, Adjusted R-squared:  0.6892
## F-statistic: 35.37 on 2 and 29 DF,  p-value: 1.663e-08
```

The Equation is hp=23.979cyl-2.775mpg+54.067 So therefore, For a car with 4 cyl and mpg=22, hp=23.979$4$-2.775$22$+54.067=88.93

The same can be done by using,

```
Model= lm(hp~cyl+mpg, data=mtcars)
predict(Model, newdata=data.frame(cyl=4, mpg=22))
```

```
##        1
## 88.93618
```

3. For this question, we are going to use BostonHousing dataset. The dataset is in 'mlbench' package, so we first need to install the package, call the library and the load the dataset using the following commands

install.packages('mlbench')

```
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.2.2
```

```
data(BostonHousing)
```

a) Build a model to estimate the median value of owner-occupied homes (medv)based on thefollowing 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 R^2)

```
set.seed(123)
Modelestimate<-lm(medv~crim+zn+ptratio+chas,data = BostonHousing)
summary(Modelestimate)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.91868    3.23497  15.431  < 2e-16 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## chas1        4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

Here R^2 is 0.359, which is not very impressive. Therefore, the model is not precise enough.

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?

The coefficient chas1 is 4.5839 In other words, if the variable is 1 rather than 0, we will add 4.5839 to the estimate of price. That is $4,583.9 since the price is expressed in $1000(The median value of owner-occupied homes). So the house bounding the River is $4,583.9 more expensive.

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?

The ptratio coefficient is 1.493. Therefore, there will be a 1.493 unit fall in the price of the neighborhood's properties, or 1493 in thousands of dollars, for every unit rise in the pupil-teacher ratio (i.e., fewer teachers and more crowded classrooms in schools).

(18*1493= 26874) (15*1493=22395) Substracting these we get $4479

Therefore, the home with a pupil-teacher ratio of 18 is less expensive ($4479) than the nearby home with a pupil-teacher ratio of 15. A 15 student-teacher ratio is therefore more expensive.

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.

Answer:

The p-values of all coefficients are very small.Therefore, all four variables are statistically significant and are related to the home price.

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

```
anova(Modelestimate)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## crim        1  6440.8  6440.8 118.007 < 2.2e-16 ***
## zn          1  3554.3  3554.3  65.122 5.253e-15 ***
## ptratio     1  4709.5  4709.5  86.287 < 2.2e-16 ***
## chas        1   667.2   667.2  12.224 0.0005137 ***
## Residuals 501 27344.5    54.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: Variables with higher Sum Square are more important.(Here crim variable) The order of significance in this model 1. crim 2. ptratio 3. Zn and 4. chas