BA Assignment 3

Pavan Chaitanya

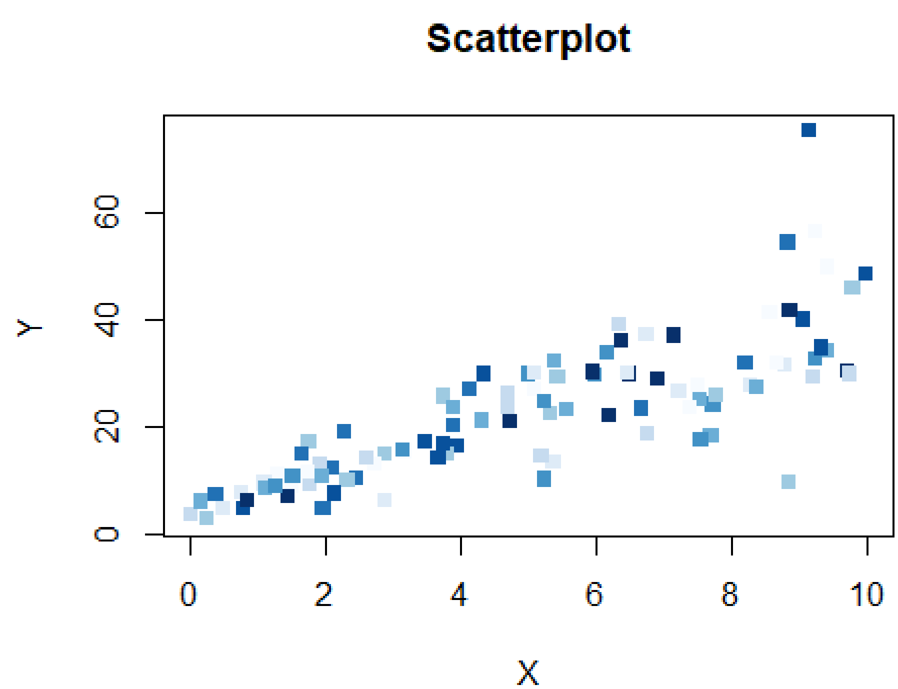
2022-11-09

Question 1: 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?

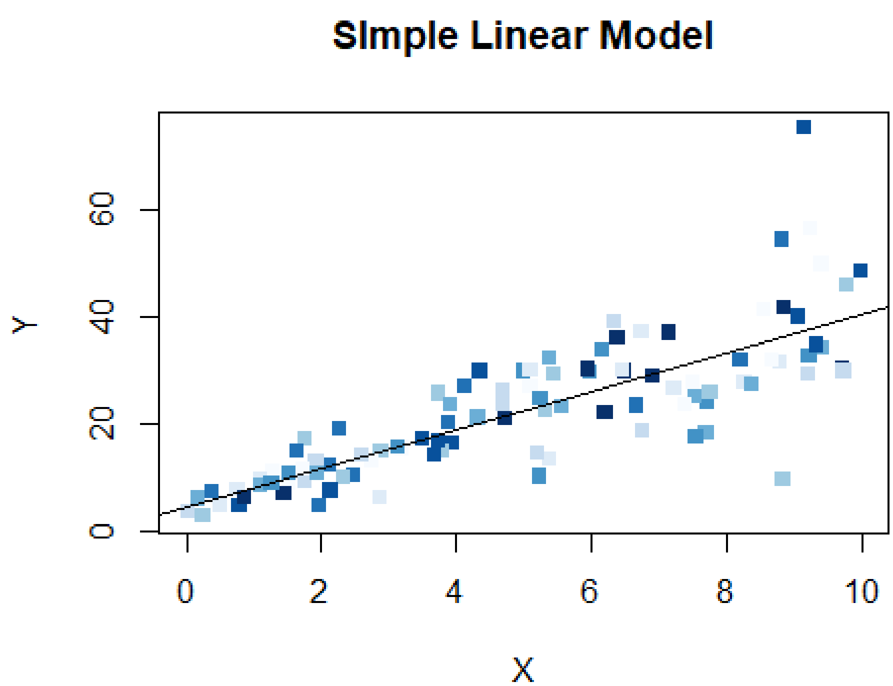
```
# scatterplot
plot(X,Y, main="Scatterplot",xlab = "X",ylab = "Y",col =blues9,pch=15 )
```



```
# Yes, based on the scatterplot we can fit a linear model to explain Y based on X.
```

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?

```
plot(X,Y,xlim=c(0, 10),xlab="X", ylab="Y", main="SImple Linear Model",col=blues9,pch=15)
abline(lsfit(X, Y),col = "black")
```




```
# The Equation is Y = B0 + B1X + E .
# The error term that the regression model was unable to explain is represented by the regression coefficient B0, which also represents the intercept, B1, which also represents the slope.
# The model's R square accuracy is 65%.
```

c) How the Coefficient of Determination, R^2, of the model above is related to the correlation coefficient of X and Y?

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

## [1] 0.6517187
```

```
# Coefficient of determination R^2 is equal (r)^2, that is, Correlation
Coefficient squared. R^2 or coefficient of determination shows percentage
variation in y that is explained by the independent variable x. R^2 is
usually between 0 and 1. It is obtained by getting the square value of the
Coefficient of correlation, “r” value.
```

Question 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. The description of the dataset can be found here.

```
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.

```
# James opinion about the HorsePower (hp) of cars
model <- lm(hp ~ wt, data = mtcars)
summary(model)

##
## Call:
## lm(formula = hp ~ wt, data = mtcars)
##
## 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
## 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
```

```
# Chris' opinion about the Horse Power (hp) of cars
model <- lm(hp ~ mpg, data = mtcars)
summary(model)

##
## Call:
## lm(formula = hp ~ mpg, data = mtcars)
##
## 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 ***
## 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

# The linear model below demonstrates that Chris is correct since fuel
economy (MPG) accounts for 60% of the variation in horsepower, and James'
perspective is irrelevant because vehicle weight (wt) only accounts for 43%
of that variation.

# Consequently, mpg is a more accurate indicator of a car's horsepower.
```

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?

```
model <- lm(hp ~ cyl + mpg, data = mtcars)
summary(model)

##
## Call:
## lm(formula = hp ~ cyl + mpg, data = mtcars)
##
```

```
## 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   
## cyl           23.979      7.346   3.264  0.00281 **
## 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

predict(model, data.frame(cyl=4, mpg=22))

##        1 
## 88.93618

# The estimated horsepower = 89
```

Quuestion 3: For this question, we are going to use BostonHousing dataset. The dataset is in 'mlbench' package, so we first need to instal 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 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)

```
library('mlbench')
data(BostonHousing)
head(BostonHousing)

##      crim zn indus chas   nox    rm  age    dis rad tax ptratio      b
lstat
```

```
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

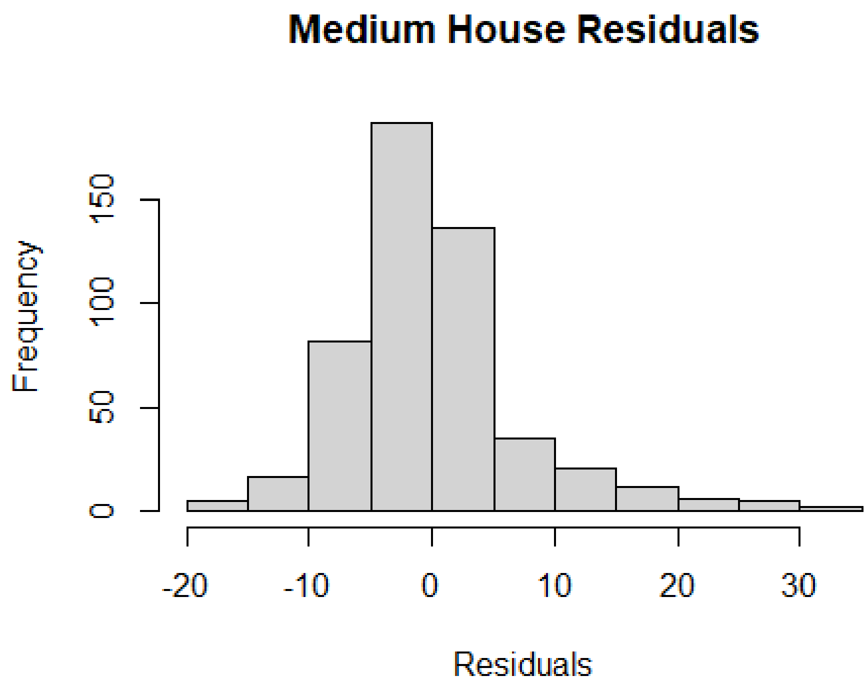
```
model <- lm(medv~crim+zn+ptratio+chas, data=BostonHousing)
summary(model)
```

```
##
## 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
```

```
# R2 = 36% for the coefficient of determination. Based on the above
variables, this is a shaky estimate of the median owner-occupied home value
(medv). This model's precision cannot be trusted.

# Based on the summary output lets plot the histogram to analyze the
```

```
assumptions of the regression Model.
hist(model$residuals,main = "Medium House Residuals",xlab = "Residuals",ylab
= "Frequency")
```



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?

```
summary(model)

##
## 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

# According to estimated coefficients, the price of the house along the Chas
River will be higher since it will rise by $4584 in comparison to any other
house.
```

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?

```
comparision <- 1494 *3
comparision

## [1] 4482

# If the coefficient of pupil to teacher ratio = -1.49367 then there will be
a decrease of approximately $1,494 to every unit change in the ptratio.
Therefore, if the pupil-teacher ratio is raised by 3 units (yielding pupil-
teacher ratio of 15 and 18 for the two houses). The estimated values
indicates that the pupil-teacher ratio of 18 will be less expensive compared
to that of pupil-teacher ratio of 15 ($1,494 *3) it'll be $4,482.
```

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.

```
summary(model)

##
## 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

# Given that each of the four variables' p-values is less than or equal to
0.05, they are all statistically significant.
```

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

```
print(anova(model))

## 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

# Using the sum square, the order of importance will be:
# 1. Crim = 6440.8
# 2. Ptratio= 4709.5
# 3. Zn = 3554.3
# 4. Chas = 667.2
```