BA Assignment-3

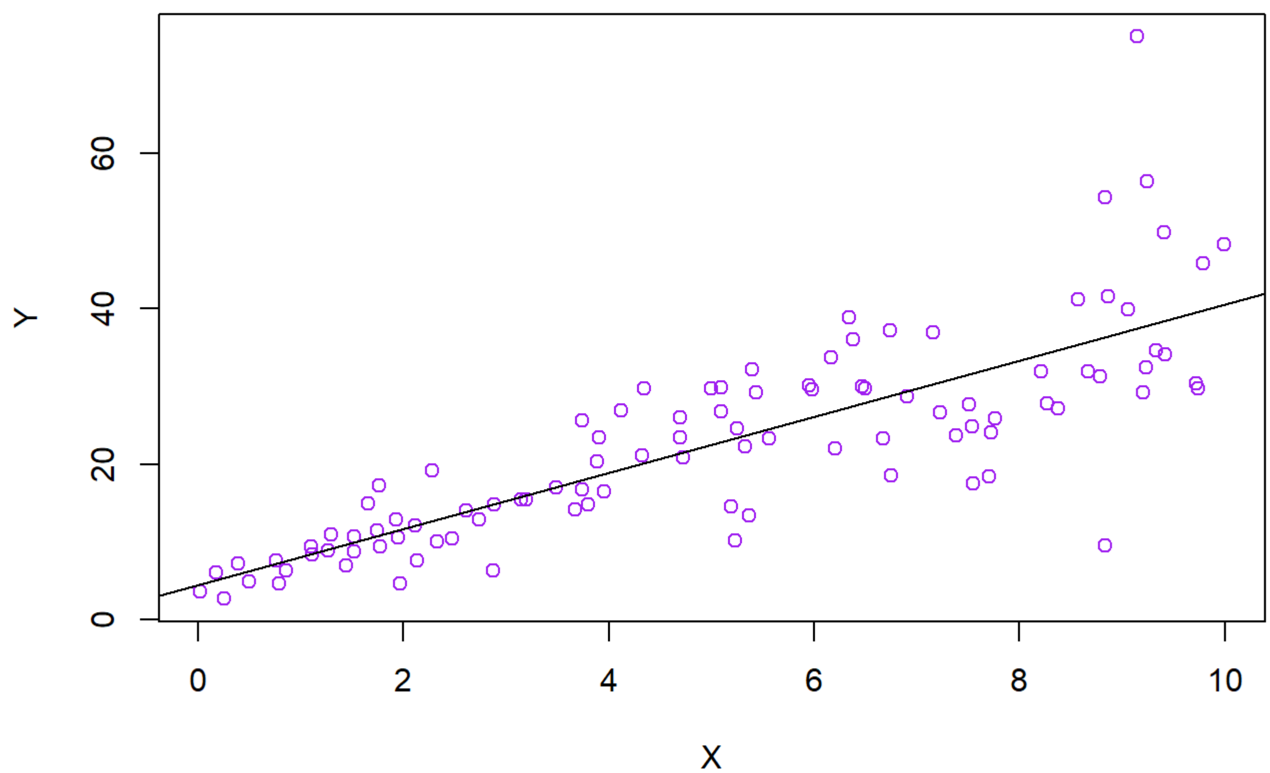
Nikitha Chigurupati

11/13/2022

```
#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 yo
u 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?

plot(Y~X,xlab='X',ylab='Y',col='purple')
abline(lsfit(X, Y),col = "black")
```



```
#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?

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

```
##
## 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
```

```
#Y=4.4655+3.6108*X
#Accuracy is 0.6517 or 65%
```

```
#c) How the Coefficient of Determination, R2, of the model above is related to the correlation c
oefficient of X and Y? (5 marks)

cor(X,Y)^2
```

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

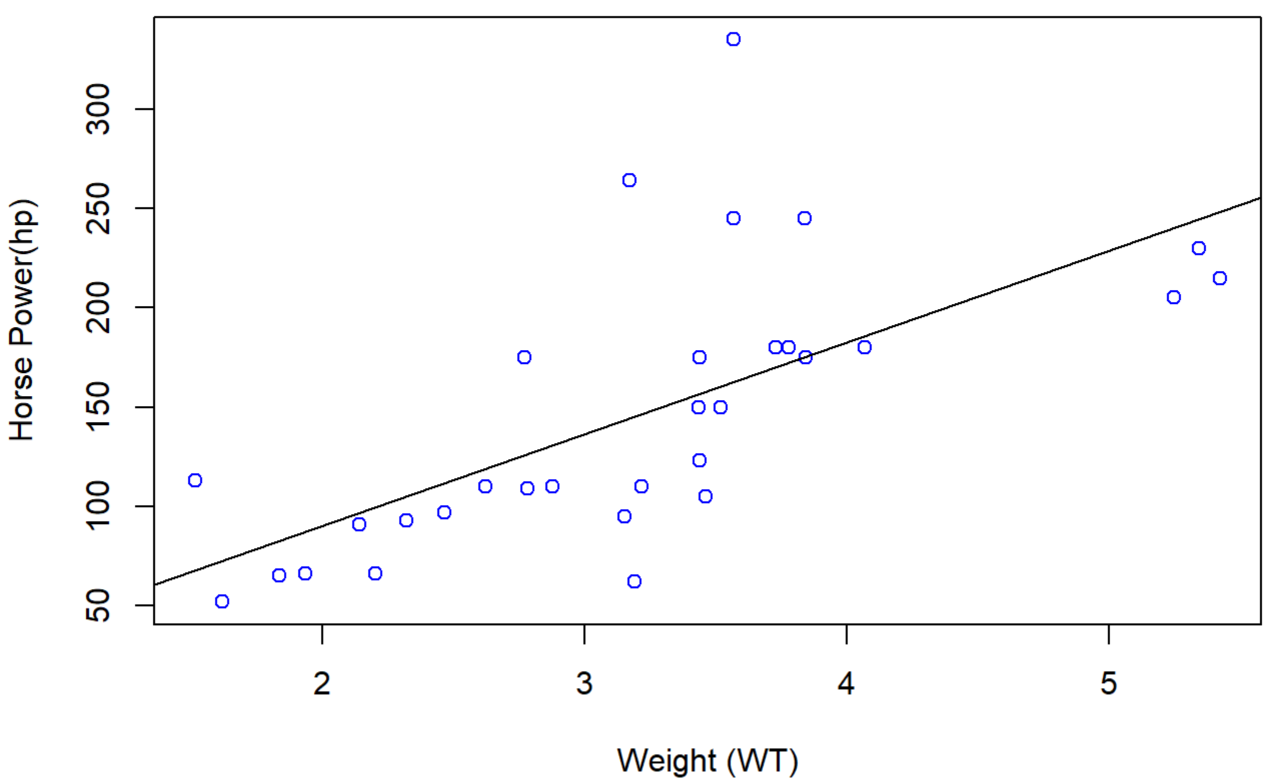
```
#2.We will use the ‘mtcars’ dataset for this question.The dataset is already included in your R
 distribution. The dataset shows some of the characteristics of different cars. The following sh
ows few samples (i.e.the first 6 rows) of the dataset.The description of the dataset can be foun
d here.

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 Pow
er (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 da
ta to answer the question.


#LINEAR MODELS OF HORSE POWER(HP) AND WEIGHT (WT):
plot(mtcars$hp~mtcars$wt,xlab='Weight (WT)',ylab='Horse Power(hp)',col='blue')
abline(lsfit(mtcars$wt,mtcars$hp),col = "black")
```

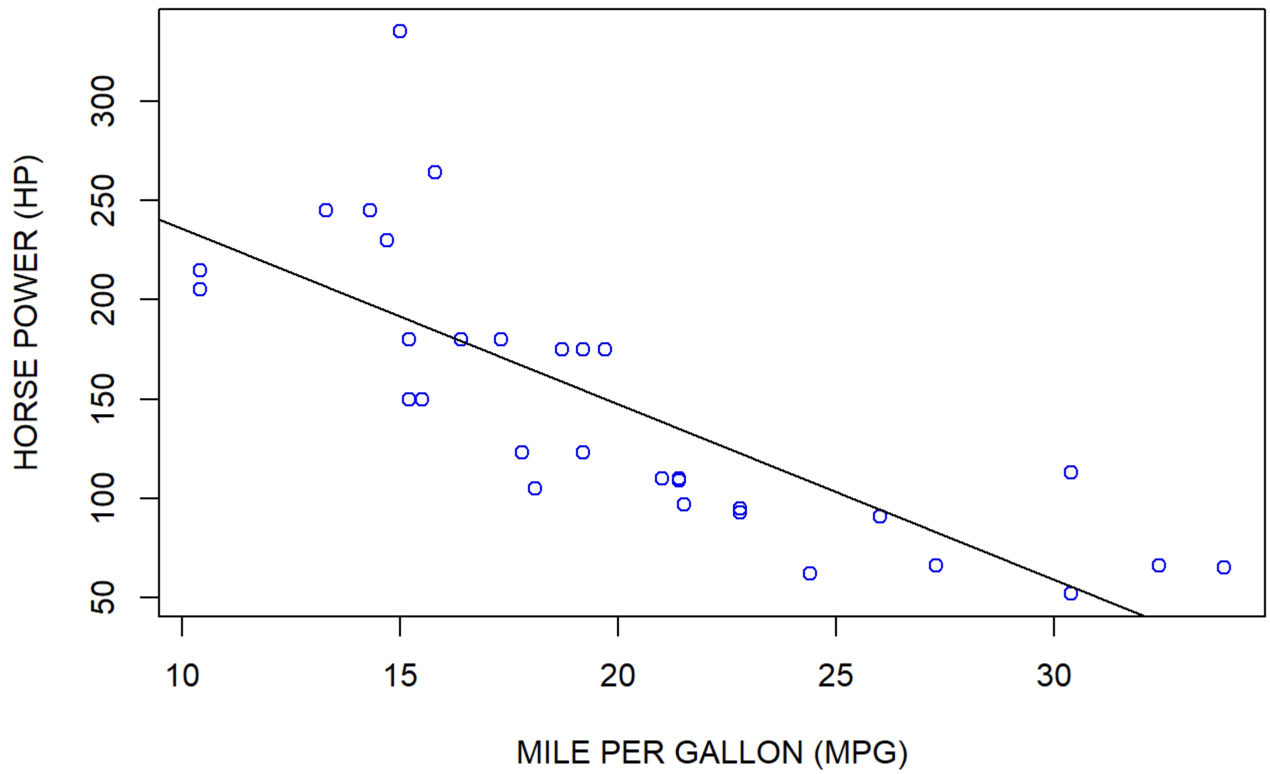



```
Model_HP_WT<-lm(formula =hp~wt, data = mtcars )
summary(Model_HP_WT)
```

```
##
## 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
```

```
#Accuracy of Model_HP_WT is 0.4339 or 43.39%


#LINEAR MODELS OF HORSE POWER(HP) AND MILE PER GALLON (MPG):
plot(mtcars$hp~mtcars$mpg,xlab='MILE PER GALLON (MPG)',ylab='HORSE POWER (HP)',col='blue')
abline(lsfit(mtcars$mpg, mtcars$hp),col = "black")
```




```
Model_HP_MPG<-lm(formula =hp~mpg, data = mtcars )
summary(Model_HP_MPG)
```

```
##
## 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
```

```
#Accuracy of the model_HP_MPG is 0.6024 OR 60.24%


#CONCLUSION: Mile Per Gallon (MPG) is a better estimator of the HORSE POWER (HP)
```

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

```
##
## 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
```

```
Estimated_HP<-predict(Model_cyl_mpg,data.frame(cyl=4,mpg=22))
Estimated_HP
```

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

```
#3.For this question, we are going to use BostonHousing dataset. The dataset is in ‘mlbench’ pac
kage, so we first need to install the package, call the library and load the dataset using the f
ollowing commands

library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.1.3
```

```
data(BostonHousing)

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


boston <- lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn + BostonHousin
g$ptratio + BostonHousing$chas, data = BostonHousing)

summary(boston)
```

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +
##     BostonHousing$ptratio + BostonHousing$chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           49.91868    3.23497  15.431  < 2e-16 ***
## BostonHousing$crim    -0.26018    0.04015  -6.480 2.20e-10 ***
## BostonHousing$zn       0.07073    0.01548   4.570 6.14e-06 ***
## BostonHousing$ptratio -1.49367    0.17144  -8.712  < 2e-16 ***
## BostonHousing$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
```

```
#A house near the Chas River is $4584 more expensive than a house not near the river,according t
o the predicted coefficient.
```

```r
#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 ot
her does not. Which one is more expensive and by how much?

Boston_1 <- lm(formula = BostonHousing$medv ~ BostonHousing$chas, data = BostonHousing)

#using the coefficients, the value of both the houses can be calculated
House_1 <- Boston_1$coefficients[1] + Boston_1$coefficients[2]*0
House_2 <- Boston_1$coefficients[1] + Boston_1$coefficients[2]*1

print(paste('House with chas and more expensive by ', House_2 - House_1))
```

```
## [1] "House with chas and more expensive by  6.34615711252662"
```

```r
#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 ho
w much?

Boston_2 <- lm(formula = BostonHousing$medv ~ BostonHousing$ptratio , data = BostonHousing)
Boston_2
```

```
## 
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$ptratio, data = BostonHousing)
## 
## Coefficients:
##           (Intercept)  BostonHousing$ptratio  
##                62.345                 -2.157
```

```r
# coefficients can be used to find the values of both houses.
House_3 <- Boston_2$coefficients[1] + Boston_2$coefficients[2] * 15

House_4 <- Boston_2$coefficients[1] + Boston_2$coefficients[2] * 18

print(paste('The house in which the pupil-teacher ratio of the two houses is 15 and is more expe
nsive by ', House_3 - House_4))
```

```
## [1] "The house in which the pupil-teacher ratio of the two houses is 15 and is more expensive
by  6.47152588818295"
```

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

summary(BostonHousing)
```

```
##       crim                zn             indus       chas         nox        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   0:471   Min.   :0.3850  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1: 35   1st Qu.:0.4490  
##  Median : 0.25651   Median :  0.00   Median : 9.69           Median :0.5380  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14           Mean   :0.5547  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10           3rd Qu.:0.6240  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74           Max.   :0.8710  
##        rm             age              dis              rad        
##  Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000  
##  1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000  
##  Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000  
##  Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549  
##  3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000  
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000  
##       tax           ptratio            b              lstat      
##  Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73  
##  1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95  
##  Median :330.0   Median :19.05   Median :391.44   Median :11.36  
##  Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65  
##  3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95  
##  Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97  
##       medv      
##  Min.   : 5.00  
##  1st Qu.:17.02  
##  Median :21.20  
##  Mean   :22.53  
##  3rd Qu.:25.00  
##  Max.   :50.00
```

```
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
#It can be concluded that none of the variables are statistically important as the P values of t
he model are less than 0.05.
```

```
#d)Use the anova analysis and determine the order of importance of these four variables.(5 mark
s)
anova(boston)
```

```
## Analysis of Variance Table
## 
## Response: BostonHousing$medv
##                        Df  Sum Sq Mean Sq F value    Pr(>F)    
## BostonHousing$crim      1  6440.8  6440.8 118.007 < 2.2e-16 ***
## BostonHousing$zn        1  3554.3  3554.3  65.122 5.253e-15 ***
## BostonHousing$ptratio   1  4709.5  4709.5  86.287 < 2.2e-16 ***
## BostonHousing$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
```

```
#We can see that the variety (sum squared) defined with the aid of using the crim variable is drastically better than different variables. We should bet this as including the crim, drastically stepped forward the model. Still we will see that a huge part of the variety is unexplained, this is proven with the aid of using residuals.
#The order of significance is crim, ptratio,zn, chas
```