Untitled

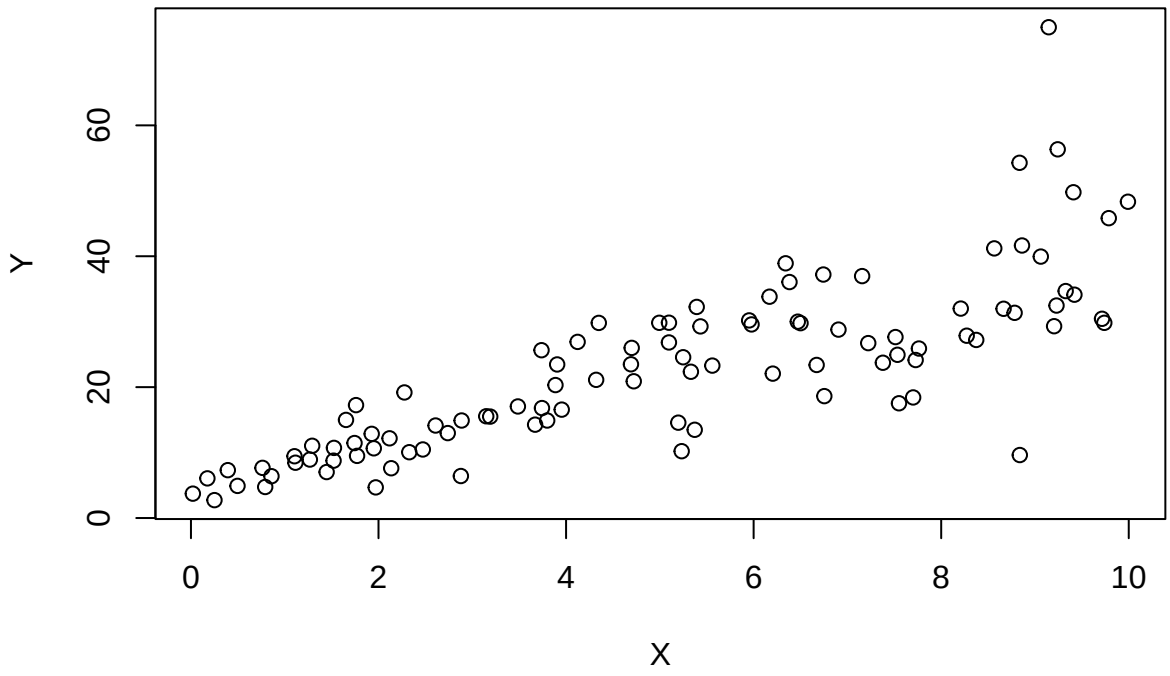
TILAK KUMAR BONALA

2022-11-13

#install.packages("mlbench")

```
#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
```

```
plot(Y~X)
```



#b) Construct a simple linear model of Y based on X.

```r
tab <- lm(Y ~ X)
summary(tab)
```

```
##
## 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
```

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

#c) How the Coefficient of Determination R^2, of the model above is related to the correlation

```r
#Coefficient of Determination= (Correlation Coefficient)2
cor(X,Y)^2
```

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

```r
summary(tab)
```

```
##
## 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
```

#(c)Ans: Coefficiemt of R^2 is equal to The square of relationship coefficient is same as coefficient of determination 65.17%

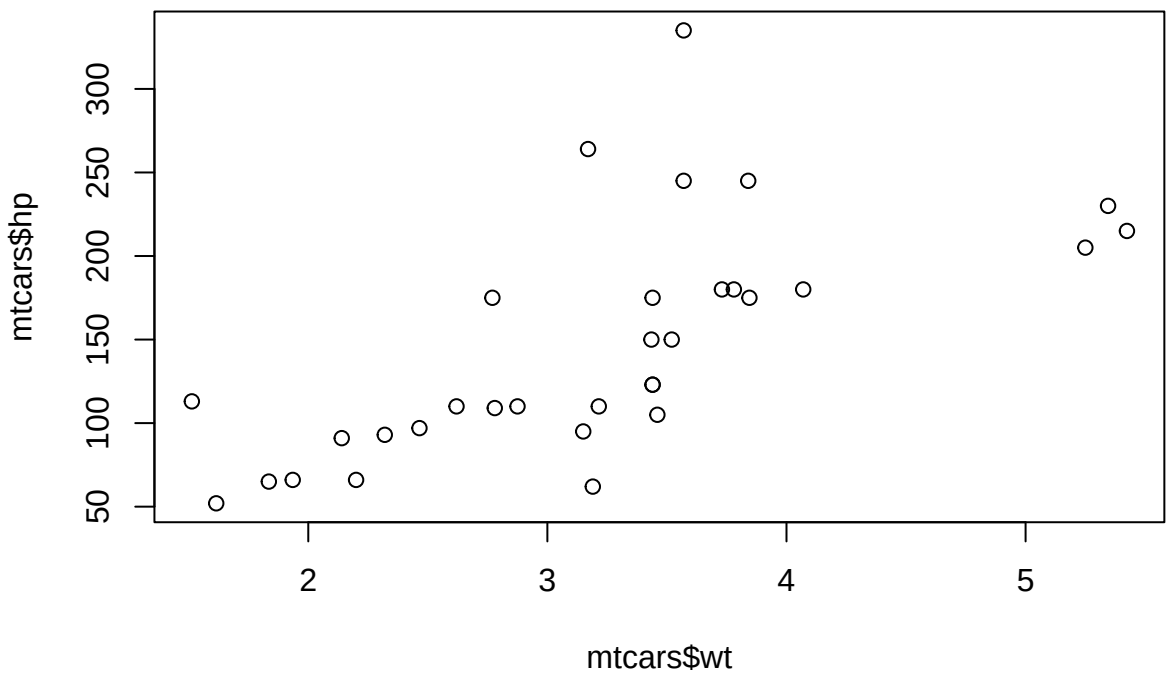
#2.We will use the 'mtcars' dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset. The description of the dataset can be found here.

```
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
```

#Building a model based on James estimation:

```
plot(mtcars$hp~mtcars$wt)
```



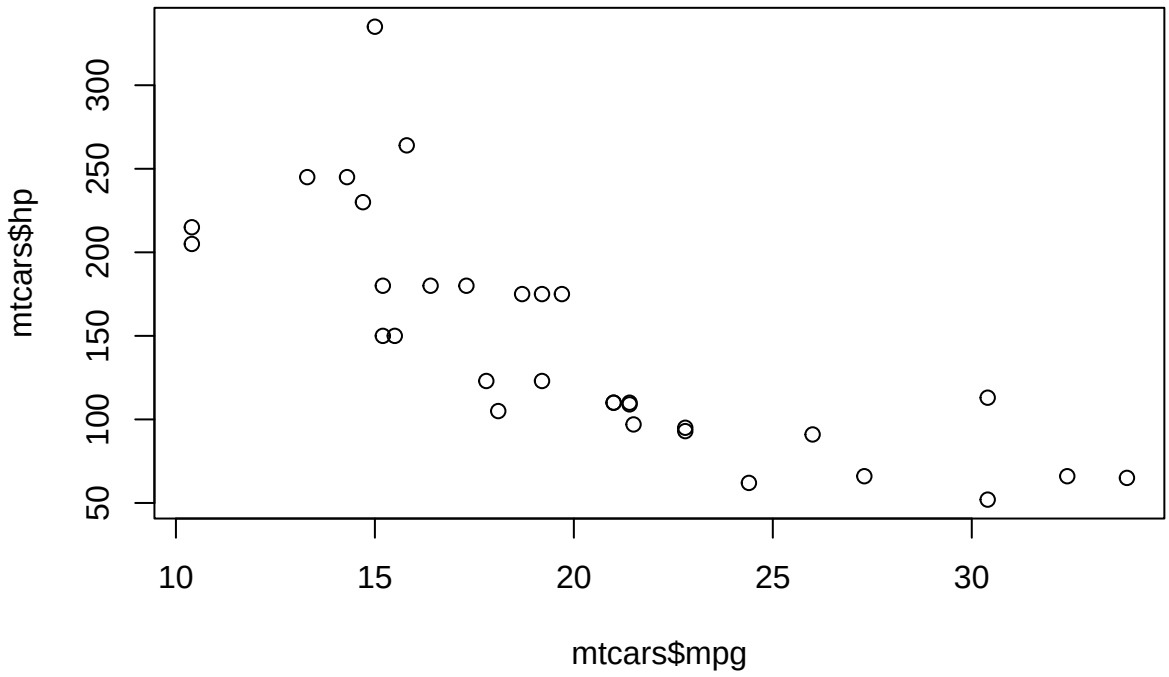

```
james_model<-lm(formula =hp~wt, data = mtcars )
summary(james_model)
```

```
##
```

```
## 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
```

#Building a model based on Chris estimation:

```
plot(mtcars$hp~mtcars$mpg)
```




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

#Conclusion:According to the results we get to know that chris has accuracy rate of 60.24 where as james as an accuracy rate of 43.39,as the accuracy rate is high for chris, he's estimation is correct.'

#2.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?

```
Model2<-lm(hp~cyl+mpg,data = mtcars)
summary(Model2)
```

```
##
## 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(Model2,data.frame(cyl=4,mpg=22))
estimated_HP
```

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

```
predict(Model2,data.frame(cyl=4,mpg=22),interval = "prediction",level = 0.85)
```

```
##        fit      lwr      upr
## 1 88.93618 28.53849 149.3339
```

#therefore, estimated horse power is 88.93618.

#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?

```
set.seed(123)
Model3<-lm(medv~crim+zn+ptratio+chas,data = BostonHousing)
summary(Model3)
```

```
##
## 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
```

#as the multiple-R-squared value is 0.3599 which is 36% nearly, it is not good model.

#b) Use the estimated coefficient to answer these questions?

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 is 4.58393, indicates that the houses that bounds by the Chas river are 4.58393 times more expensive than the houses which do not bounds by the river.

#Moreover, the values of chas river are 1 or 0 which means the houses which bound by the river are assigned a value of 1, orelse 0. So for the houses which do not bound by the river are going to have 0 times change in their value.

#(c) Finding which of the variables are statistically important:**

#All the variables including crime rate,proportion of residential land zoned for lots over 25,000 sq.ft,the local pupil-teacher ratio,the tract bounds Chas River are statistically important as all of them has very low P value

#(d)Determining the order of importance of the 4 variables using ANOVA analysis:**

```
anova_lm<-anova(Model3)
anova_lm
```

```
## 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
```

The importance of variables can be determined by their Sum of Squares value. Higher the Sum of squares, the more important is the variable in estimating the value of a dependent variable

#Order of importance of variables:

#crime rate by town=6440.8

#pupil-teacher ratio by town=3554.3

residential land zoned for lots over 25,000 sq.ft.=4709.5

#Charles River dummy variable=667.2