Assignment2_Regression

Ram

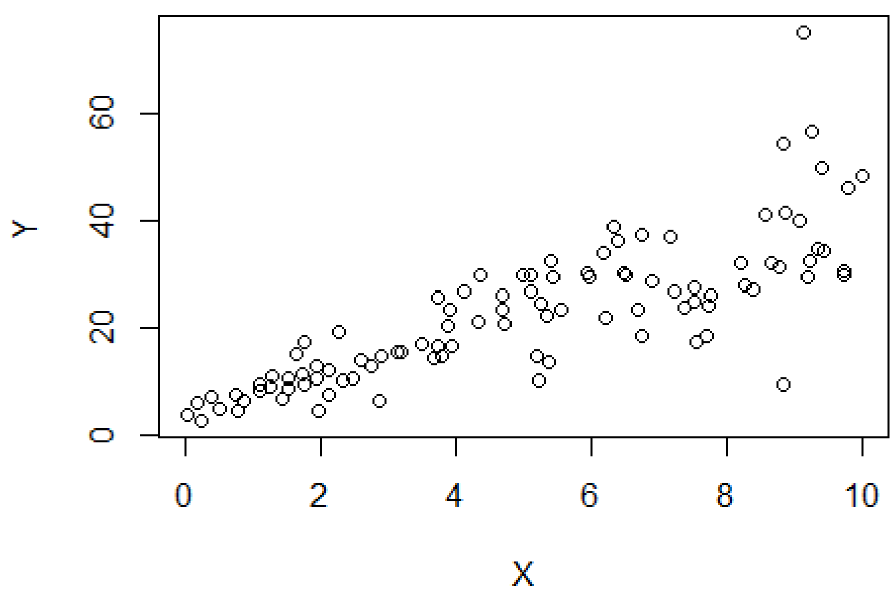
11/16/2021

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?

```
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.

## [1] 0.807291
```

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?

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

##
## 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

LinearModel$coefficients

## (Intercept)           X
##    4.465490    3.610759
```

#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

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

## [1] 0.6517187
```

2. We will use the â€˜mtcarsâ€™ 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.

```
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

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

##
## Call:
## lm(formula = mtcars$hp ~ mtcars$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
## 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

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

##
## Call:
## lm(formula = mtcars$hp ~ mtcars$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 ***
## 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
```

#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?

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

##
## Call:
## lm(formula = mtcars$hp ~ mtcars$mpg + mtcars$cyl, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067     86.093   0.628  0.53492
## mtcars$mpg    -2.775      2.177  -1.275  0.21253
## mtcars$cyl    23.979      7.346   3.264  0.00281 **
## ---
## 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

LinearModel4$coefficients

## (Intercept)  mtcars$mpg  mtcars$cyl
##   54.066600   -2.774769   23.978626
```

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

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

## (Intercept)
##    88.93618
```

#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 â€˜mlbenchâ€™ package, so we first need to instal the package, call the library and the load the dataset using the following commands

```
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)

```
HousingModel<-
lm(BostonHousing$medv~BostonHousing$crim+BostonHousing$zn+BostonHousing$ptrat
io+BostonHousing$chas,data=BostonHousing)
summary(HousingModel)

##
## 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
```

#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?

```
HousingModel1<-lm(medv~chas,data = BostonHousing)
summary(HousingModel1)

##
## Call:
```

```
## lm(formula = medv ~ chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.094  -5.894  -1.417   2.856  27.906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.0938     0.4176  52.902  < 2e-16 ***
## chas1         6.3462     1.5880   3.996 7.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.064 on 504 degrees of freedom
## Multiple R-squared:  0.03072,    Adjusted R-squared:  0.02879
## F-statistic: 15.97 on 1 and 504 DF,  p-value: 7.391e-05
```

```
HousingModel1$coefficients
```

```
## (Intercept)       chas1
##   22.093843    6.346157
```

#For chas = 0 and #For chas = 1

```
withoutchas <-
((HousingModel1$coefficients[2]*0)+HousingModel1$coefficients[1])
withoutchas
```

```
##    chas1
## 22.09384
```

```
withchas <- ((HousingModel1$coefficients[2]*1)+HousingModel1$coefficients[1])
withchas
```

```
## chas1
## 28.44
```

#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?

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

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$ptratio, data =
BostonHousing)
```

```
##
## Coefficients:
##           (Intercept)  BostonHousing$ptratio
##                62.345                 -2.157
```

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

```
ptratio15 <-
((HousingModel2$coefficients[2]*15)+HousingModel2$coefficients[1])
ptratio15

## BostonHousing$ptratio
##                29.987

ptratio18 <-
((HousingModel2$coefficients[2]*18)+HousingModel2$coefficients[1])
ptratio18

## BostonHousing$ptratio
##              23.51547
```

#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.

```
summary(HousingModel)

##
## 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 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.

```
anova(HousingModel)

## 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
```

#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