Assignment II – Regression Analytics

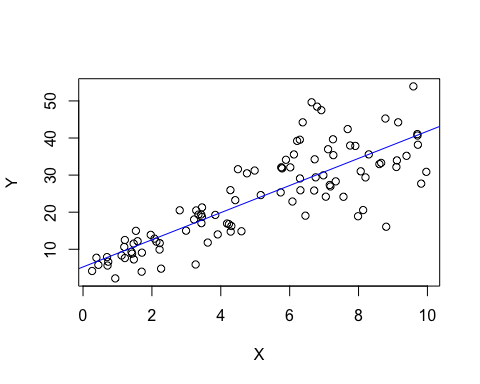
Khushboo Yadav

11/3/2020

#### Q1.1. Run the following code in R-studio to create two variables X and 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? (5 Marks)

set.seed(2000)  
  
X=runif(100)\*10  
Y=X\*4+3.45  
Y=rnorm(100)\*0.29\*Y+Y  
  
#Plot Y against X  
plot(X,Y)  
abline(lsfit(X,Y),col="blue")

 Answer:-

Based on the graph ,we can see that there is positive linear relationship between X and Y.Therefore, I think we can fit 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? (5 Marks)

#Constructing a simple linear model  
Simple\_Linear\_Model<-lm(Y~X)  
  
summary(Simple\_Linear\_Model)

##   
## Call:  
## lm(formula = Y ~ X)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.3668 -3.9507 -0.6603 4.1766 20.1204   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.236 1.468 3.568 0.000559 \*\*\*  
## X 3.659 0.251 14.573 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.054 on 98 degrees of freedom  
## Multiple R-squared: 0.6843, Adjusted R-squared: 0.681   
## F-statistic: 212.4 on 1 and 98 DF, p-value: < 2.2e-16

#Equation to explain Y based on X  
  
#Y^ = βo + β1X (βo=sample intercept,β1=the sample slopes)  
  
Y\_cap=5.236+3.658\*X #(saving y^ value)

Answer:-

Equation that explains Y based on X is :

Y^ = βo + β1X (βo=sample intercept,β1=the sample slopes).Therefore equation will be Y^=5.236+3.658\*X.

Accuracy of the model: Multiple R-squared and Adjusted R-squared is 68% of the model.

Which means the model explains 68% variability of the target (response) variable. Accuracy of the model is 68%.

#### c) How the Coefficient of Determination, R 2 , of the model above is related to the correlation coefficient of X and Y? (5 marks)

Answer :-

Since for this model , we have only one independent variable and 1 dependent variable , it is a simple linear regression model.

Therefore ,

Coefficient of determination = (correlation coefficient)^2

0.6843 = 0.6843

Since we know the value of the R-squared i.e 0.6843 , the value of the calculated squared correlation coefficient should be same.

We can also confirm the value of the correlation coefficient below :

#Assigning variable to the squared correlation coefficient value  
Correlation\_coefficient\_squared<-round(cor(X,Y)^2,digits = 4)  
  
#print the value  
print(Correlation\_coefficient\_squared) #0.684

## [1] 0.6843

Therefore ,Coefficient of determination = (correlation coefficient)^2

#### 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.

#### 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. (10 marks)

library('ISLR')  
data(mtcars)  
  
#Viewing the data  
#view(mtcars)  
  
#Creating a simple linear model to analyze James thinking  
James\_model<- lm(mtcars$hp~mtcars$wt,mtcars)  
summary(James\_model)

##   
## 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

#Creating a simple linear model to analyze Chris thinking  
Chris\_model <- lm(mtcars$hp~mtcars$mpg,mtcars)  
summary(Chris\_model)

##   
## 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

Answer:-

As per the model analysis for James and Chris, we can see that Multiple R-squared value and adjusted R-squared value is higher for Chris.This means that Chris model is more accurate than James.

Therefore, Chris model is better.

#### 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 cylinder and mpg of 22? (10 mark)

# Linear model for mtcars ,horsepower as dependent variable and cylinders(cyl), mile per gallon (mpg) as independent variable  
New\_model\_mtcars<-lm(mtcars$hp~cyl+mpg,mtcars)  
summary(New\_model\_mtcars)

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

#Predicting the model based with cylinder = 4 , mpg=22  
pre\_mtcars<- predict(New\_model\_mtcars,data.frame(cyl=c(4),mpg=c(22)))  
print(pre\_mtcars)

## 1   
## 88.93618

Answer:- The estimated horse power with cylinder =4,mpg=22 will be 88.93618.

#### 

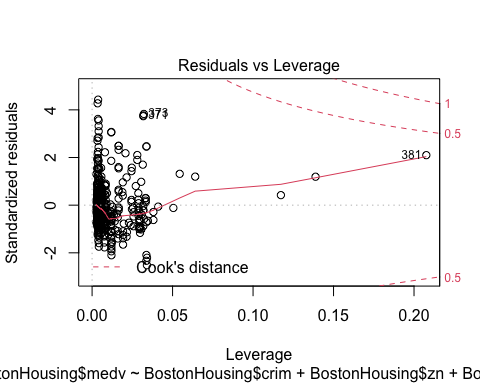
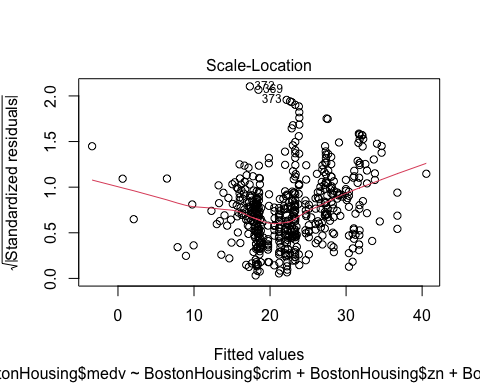
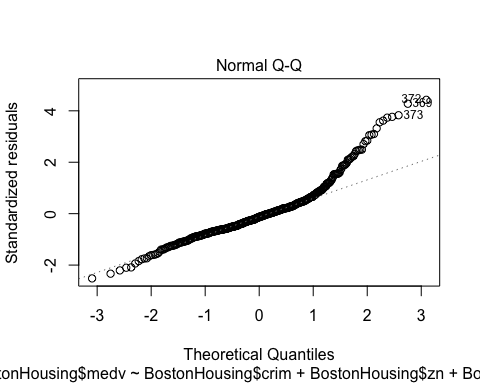
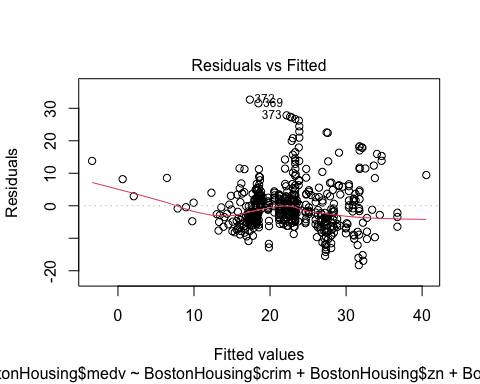
#### Question 3. For this question, we are going to use BostonHousing dataset. The dataset is in ‘mlbench’ package,so we first need to install the package, call the library and the load the dataset using the following commands

#### 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 R 2) (5 marks)

#install.packages('mlbench')  
library(mlbench)  
  
data(BostonHousing)  
  
#Building a linear model with median value of owner-occupied homes as dependent variable and crime rate,proportion of residential land zone, pupil-teacher ratio and weather tract bounds Chas river as independent values  
  
Boston\_model<-lm(BostonHousing$medv~BostonHousing$crim+BostonHousing$zn+BostonHousing$ptratio+BostonHousing$chas)  
  
summary(Boston\_model)

##   
## Call:  
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +   
## BostonHousing$ptratio + BostonHousing$chas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.282 -4.505 -0.986 2.650 32.656   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 49.91868 3.23497 15.431 < 2e-16 \*\*\*  
## BostonHousing$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

plot(Boston\_model)

 Answer:-

Since R-squared value is 0.3599, the Accuracy is 35.99 percent. It means that almost 40% variability of the target (response) variable can be explained by the model.

Considering in this model only 4 variables have been included, the model is appropriate. However, I believe using more number of variables in the model can certainly improve the accuracy of the 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 Riverand the other does not. Which one is more expensive and by how much? (5 marks)

Answer:-

Charles river location is a significant contributor and as per the above estimates we can see that presence of Charles River leads to impact the housing cost by 4583.93 dollars.

4.58393\*1000 = 4583.93

It means that It’s possible that the neighborhoods near the Charles river are more desirable.

#### 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? (Golden Question: 10 extra marks if you answer)

Answer:-

As per the above Coefficient estimates Pupil-teacher ratio increment by unit leads to -1.49367 which means it leads to decrements in the price of the housing.

4.48101\*1000(dollars in 1000 units) = 4481.01

Therefore,pupil-teacher ratio when it is 15 ,the housing will be more expensive and by 4481.01 dollars.

People are paying more money to send their kids to the schools where there are fewer students for every teacher.(considered to be better education)

#### 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.(5 mark)

Answer:-

All of the variables are statistically important .

Since the p value for all of the 4 variables are less than 5%.we reject the null hypothesis that β = 0.

A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct.

#### d) Use the anova analysis and determine the order of importance of these four variables. (10 marks)

# Anova Analysis on the Boston Model  
anova(Boston\_model)

## 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

Answer:-

We can see that the variability (sum squared) explained by the ‘crim’ variable is significantly higher than that of ptratio ,zn,chas.

The order of importance would be as follows:

1. crim (sum squared= 6440.8) ,
2. ptratio(sum squared =4709.5),
3. zn(sum squared =3554.3) ,
4. chas(sum squared=667.2)

Also, we can see that a large portion of the variability is unexplained, that is shown by residuals which is 27344.5.