BA\_Assignment\_3

Tejaswini Yeruva

2022-11-13

##1. Running the code that is provided.

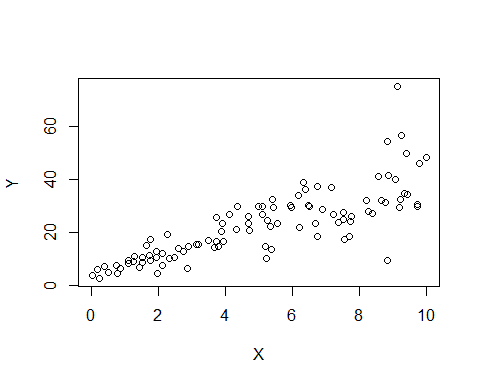
set.seed(2017)   
X=runif(100)\*10   
Y=X\*4+3.45   
Y=rnorm(100)\*0.29\*Y+Y

## a) Using plot function Y against X using the below command.

cor(X,Y)

## [1] 0.807291

plot(X,Y)



## Since the Plot shows the positive correlation, Linear model can fit Y based on X.

## b) Simple linear model Y based on X.

model<-lm(Y~X)  
summary(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

## The equation model is Y=3.6108\*X+4.4655.

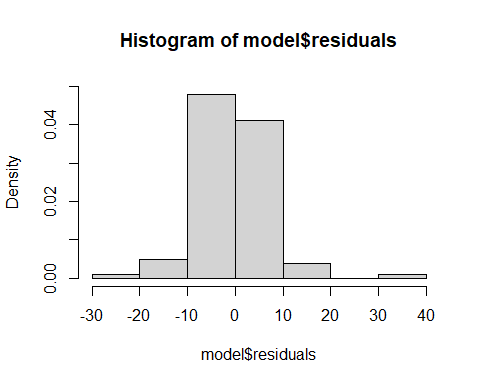
## Accuracy of the above linear model is 65.17%, above equation explains Y based on x.

## c) Coefficient of Determination

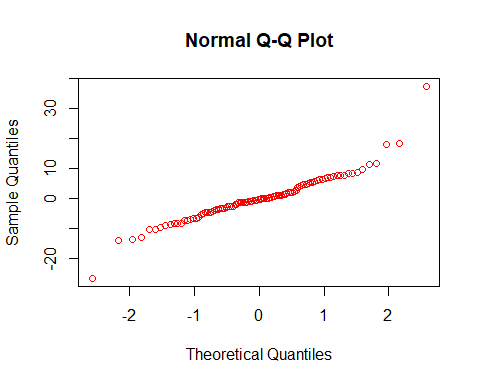
(cor(Y,X))^2

## [1] 0.6517187

## Coefficient of Determination= (Correlation Coefficient)^2  
## Multiple R-square can be determined by squaring of correlation.  
  
hist(model$residuals,freq = FALSE,ylim = c(0,0.05))



qqnorm(model$residuals,col="red")



## The above graph illustrates that the residuals are normally distrubuted, So the linear model is appropriate.

## 2) Using ‘mtcars’ dataset:

## a)

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

summary(lm(hp~wt,data=mtcars))

##   
## 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

summary(lm(hp~mpg,data=mtcars))

##   
## 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

## By using the above linear model we see that the Multiple R-squared, mpg has high r square value 60% compared to wt of car i.e 43.95%  
  
## Opinion made by Chris is right.

## b)

summary(model2<-lm(hp~cyl+mpg,data = mtcars))

##   
## 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

((model2$coefficients[2]\*4)+model2$coefficients[1])+(model2$coefficients[3]\*22)

## cyl   
## 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

## 3) Installing the required package:

library(mlbench)

## Warning: package 'mlbench' was built under R version 4.2.2

data(BostonHousing)

## a)

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

##   
## 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

## R-Square value is very low i.e 36% by this we can tell that it is not an accurate model.

## b1)

summary(hos1<-lm(medv~chas,data = BostonHousing))

##   
## 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

hos1$coefficients

## (Intercept) chas1   
## 22.093843 6.346157

(hos1$coefficients[2]\*0)+hos1$coefficients[1]

## chas1   
## 22.09384

(hos1$coefficients[2]\*1)+hos1$coefficients[1]

## chas1   
## 28.44

## From the above correlation coefficient, the house bound with Chas river is more expensive than the one not bound of 0 with value 4.3

## b2)

summary(hos2<-lm(medv~ptratio,data = BostonHousing))

##   
## Call:  
## lm(formula = medv ~ ptratio, data = BostonHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.8342 -4.8262 -0.6426 3.1571 31.2303   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 62.345 3.029 20.58 <2e-16 \*\*\*  
## ptratio -2.157 0.163 -13.23 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.931 on 504 degrees of freedom  
## Multiple R-squared: 0.2578, Adjusted R-squared: 0.2564   
## F-statistic: 175.1 on 1 and 504 DF, p-value: < 2.2e-16

(hos2$coefficients[2]\*15)+hos2$coefficients[1]

## ptratio   
## 29.987

(hos2$coefficients[2]\*18)+hos2$coefficients[1]

## ptratio   
## 23.51547

## From the above correlation coefficients, the coefficients are negative hence we can say that if the ptratio increases the housing price decreases.  
  
## The price of house which has ptratio of 15 is more expensive compared to price of house which has a ptratio of 18 by 6.471

## c)

summary(hos)

##   
## 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

## A low p-value i.e < 0.05 tells that we can reject the null hypothesis.  
## Hence from the model summary none of the independent variables are considerable.

## d)

anova(hos)

## 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

## Order of importance of the values by comparing p values:  
## 1) crim - Accounts for 15.08%   
## 2) ptratio - accounts for 11.02%  
## 3) zn - accounts for 8.32%  
## 4)chas - accounts for 1.56%  
  
## In total the model accounts for 64.01 and it can be improved.