Assignment 5

Alka Santosh Naik

2023-10-03

#Alka Santosh Naik  
rm(list=ls())  
library(rio)  
library(moments)  
cardata=import("6304 Module 5 Assignment Data.xlsx")  
colnames(cardata)=tolower(make.names(colnames(cardata)))  
set.seed(16999752)  
sample.data=cardata[sample(which(cardata$model.year<=76),100),]  
sample.data$cylinders=as.factor(sample.data$cylinders)  
sample.data$model.year=as.factor(sample.data$model.year)  
attach(sample.data)

#1 question  
str(sample.data)

## 'data.frame': 100 obs. of 9 variables:  
## $ mpg : num 31 15.5 16.5 30 14 16.5 21 24 25 15 ...  
## $ cylinders : Factor w/ 3 levels "4","6","8": 1 3 3 1 3 2 1 1 1 2 ...  
## $ cubic.inches : num 76 304 350 79 400 168 140 119 90 250 ...  
## $ horsepower : num 52 120 180 70 175 120 72 97 71 72 ...  
## $ weight : num 1649 3962 4380 2074 4385 ...  
## $ model.year : Factor w/ 7 levels "70","71","72",..: 5 7 7 2 3 7 4 6 6 6 ...  
## $ country.of.origin: chr "Japan" "USA" "USA" "Europe" ...  
## $ auto.make : chr "toyota" "amc" "cadillac" "peugeot" ...  
## $ auto.model : chr "corona" "matador" "seville" "304" ...

Analysis: str of sample.data

#2 question  
mean(sample.data$mpg)

## [1] 19.965

mean(sample.data$cubic.inches)

## [1] 214

Analysis: mean of MPG and Cubic inches

#3 question  
cardata.out=lm(mpg~cubic.inches+horsepower+weight,data=sample.data)  
summary(cardata.out)

##   
## Call:  
## lm(formula = mpg ~ cubic.inches + horsepower + weight, data = sample.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.9447 -1.6543 0.0207 1.7867 5.4684   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.0154116 1.1823541 31.307 < 2e-16 \*\*\*  
## cubic.inches -0.0137750 0.0064762 -2.127 0.036 \*   
## horsepower -0.0126335 0.0134169 -0.942 0.349   
## weight -0.0040330 0.0006835 -5.901 5.42e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.457 on 96 degrees of freedom  
## Multiple R-squared: 0.847, Adjusted R-squared: 0.8422   
## F-statistic: 177.2 on 3 and 96 DF, p-value: < 2.2e-16

Analysis: From the P-Values we can say intercept, cubic inches and weight are significant and horsepower shows no significance as P-Value is greater than 0.05.

B0=here the intercept cannot be interpreted and shouldn’t be reported

B1=for every 100 increase in cubic inches the MPG decreases by 1.3

B2=for every 100 increase in horsepower the MPG decreases by 1.2

B3=for every 1000 increase in weights (pounds) the MPG decreases by 4

#4 question  
confint(cardata.out)

## 2.5 % 97.5 %  
## (Intercept) 34.668457307 39.3623658795  
## cubic.inches -0.026630179 -0.0009198015  
## horsepower -0.039265897 0.0139989675  
## weight -0.005389686 -0.0026763510

Analysis:

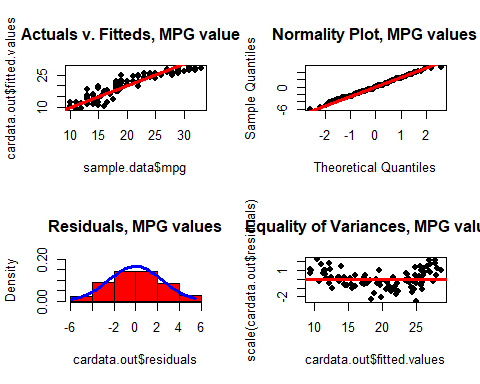
B0= we are 95% confident that the intercept is between 34.6 to 39.3

B1=for every 10000 increase in cubic inches we are 95% confident that MPG will decrease between the values -266 to -9 (higher width)

B2= for every 100 increases in horsepower we are 95% confident that MPG will decrease or increase between the values -3.9 to 1.3 (moderate width)

B3= for every 1000 increase in weight we are 95% confident that MPG will decrease between the values -5.3 to -2.6 (moderate width)

#5 question  
par(mfrow=c(2,2))  
# Linearity  
plot(sample.data$mpg,cardata.out$fitted.values,  
 pch=19,main="Actuals v. Fitteds, MPG values")  
abline(0,1,col="red",lwd=3)  
# Normality  
qqnorm(cardata.out$residuals,pch=19,  
 main="Normality Plot, MPG values")  
qqline(cardata.out$residuals,lwd=3,col="red")  
hist(cardata.out$residuals,col="red",ylim=c(0,0.20),  
 main="Residuals, MPG values",  
 probability=TRUE)  
curve(dnorm(x,mean(cardata.out$residuals),  
 sd(cardata.out$residuals)),  
 from=min(cardata.out$residuals),  
 to=max(cardata.out$residuals),  
 lwd=3,col="blue",add=TRUE)  
# Equality of Variances  
plot(cardata.out$fitted.values,  
 scale(cardata.out$residuals),  
 pch=19,main="Equality of Variances, MPG values")  
abline(0,0,lwd=3,col="red")



par(mfrow=c(1,1))

Analysis: the model do not follow linearity , as from the graph “Actuals v. Fitteds, MPG values” we see a curve in the data plotted.

The model do not follow equality of variances, as from the graph” Equality of Variances, MPG values” we see pattern in the residuals

The mode follows normality which can be determined by observing the graphs “Normality Plot, MPG values” and “Residuals, MPG values”

#6 question  
cardata2.out=lm(mpg~horsepower+weight+cubic.inches+I(horsepower^2)+I(weight^2)+I(horsepower\*cubic.inches),data=sample.data)  
summary(cardata2.out)

##   
## Call:  
## lm(formula = mpg ~ horsepower + weight + cubic.inches + I(horsepower^2) +   
## I(weight^2) + I(horsepower \* cubic.inches), data = sample.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0197 -1.2134 0.1503 1.1967 4.2769   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.233e+01 3.203e+00 16.340 < 2e-16 \*\*\*  
## horsepower 3.639e-02 4.059e-02 0.897 0.372224   
## weight -1.317e-02 2.805e-03 -4.696 9.13e-06 \*\*\*  
## cubic.inches -6.022e-02 1.747e-02 -3.446 0.000855 \*\*\*  
## I(horsepower^2) -6.243e-04 2.125e-04 -2.938 0.004165 \*\*   
## I(weight^2) 1.462e-06 3.703e-07 3.948 0.000153 \*\*\*  
## I(horsepower \* cubic.inches) 3.933e-04 1.273e-04 3.089 0.002651 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.868 on 93 degrees of freedom  
## Multiple R-squared: 0.9143, Adjusted R-squared: 0.9088   
## F-statistic: 165.3 on 6 and 93 DF, p-value: < 2.2e-16

Analysis: Yes adding horsepower^2, weight^2 and horsepower\*cubic.inches definitely increased the model performance

As we see the R square value is increased to 91.4% and the residual standard error is decreased to 1.8.

#7 question  
revdata=sample.data  
revdata$cylinders=relevel(revdata$cylinders,"8")  
cardata3.out=lm(mpg~cubic.inches+horsepower+weight+cylinders,data=revdata)  
summary(cardata3.out)

##   
## Call:  
## lm(formula = mpg ~ cubic.inches + horsepower + weight + cylinders,   
## data = revdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.3849 -1.3668 -0.0062 1.5591 4.7272   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.4446744 3.2458874 11.228 < 2e-16 \*\*\*  
## cubic.inches -0.0057354 0.0084539 -0.678 0.4992   
## horsepower -0.0324949 0.0140860 -2.307 0.0233 \*   
## weight -0.0037134 0.0006764 -5.490 3.4e-07 \*\*\*  
## cylinders4 0.9799822 1.9768067 0.496 0.6212   
## cylinders6 -1.6552232 1.2271846 -1.349 0.1806   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.331 on 94 degrees of freedom  
## Multiple R-squared: 0.8651, Adjusted R-squared: 0.858   
## F-statistic: 120.6 on 5 and 94 DF, p-value: < 2.2e-16

Analysis: None of the Cylinders factor levels appear to be significant.