Homework4

Justin Markey

10/15/2021

# Import the data

data1 <- read.csv("../data/HW4 Data.csv", stringsAsFactors = TRUE)

# Q1

Optimal\_1 <- step(lm(Y~., data=data1), trace = FALSE)  
summary(Optimal\_1)$adj.r.square

## [1] 0.934498

Optimal\_1

##   
## Call:  
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = data1)  
##   
## Coefficients:  
## (Intercept) X1 X2 X3 X4   
## -7.31929 5.68268 -0.01801 7.27125 0.04125

# Q2

Optimal\_2 <- step(lm(Y~.^2,data=data1), trace = FALSE)  
summary(Optimal\_2)$adj.r.square

## [1] 0.9972923

Optimal\_2

##   
## Call:  
## lm(formula = Y ~ X1 + X3 + X4 + X5 + X1:X3 + X1:X4 + X3:X5, data = data1)  
##   
## Coefficients:  
## (Intercept) X1 X3 X4 X5 X1:X3   
## 0.062577 3.199262 -0.040836 0.032613 0.011882 2.404996   
## X1:X4 X3:X5   
## -0.008887 -0.007844

# Q3

## partition the data

rowindex<-sample(1:nrow(data1),2000)  
traindata<-data1[rowindex,]  
validdata<-data1[-rowindex,]

## train model

Optimal\_3<- step(lm(Y~., data=traindata), trace = FALSE)  
summary(Optimal\_3)$adj.r.square

## [1] 0.932661

Optimal\_3

##   
## Call:  
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = traindata)  
##   
## Coefficients:  
## (Intercept) X1 X2 X3 X4   
## -7.43138 5.70692 -0.01952 7.30004 0.04125

# Q4

## validate the models

RMSE\_train <- sqrt(mean((validdata$Y - predict(Optimal\_3,newdata=traindata))^2))  
RMSE\_valid <- sqrt(mean((validdata$Y - predict(Optimal\_3,newdata=validdata))^2))  
RMSE\_train

## [1] 13.52519

RMSE\_valid

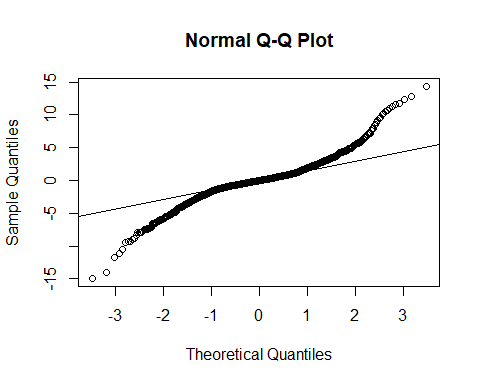
## [1] 2.386551

### Conclusion

Optimal\_3 does not overfit because the prediction error is low for the valid testing compared the to training testing. What this really means is that the underlying data for each set is different, and therefore it cannot be overfitted.

# Q5 normaility testing

qqnorm(Optimal\_3$residuals)  
qqline(Optimal\_3$residuals)



shapiro.test(Optimal\_3$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: Optimal\_3$residuals  
## W = 0.92199, p-value < 2.2e-16

### Conclusion

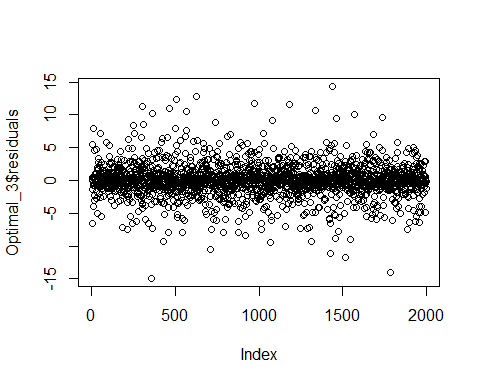
The graph shows that the residuals do not follow the qqline indicating some sort of non normal data pattern. This first impression is backed up by the fact that the shapiro test fails.

# Q6

library(car)

## Loading required package: carData

plot(Optimal\_3$residuals)



ncvTest(Optimal\_3)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 1.202093, Df = 1, p = 0.2729

### Conclusion

After looking at the plot of the residuals, we can see that the data contains high volitility. The residuals are very spuratic. The ncv test also confirms this volitility as the test fails.