**FINAL PROJECT**

1. PROBLEM STATEMENT:

The purpose of this project is to determine if the calling price on Craigslist of a used car in Charlotte can be predicted using the car’s age, mileage, color, title status, make, model, gas mileage, type (coup, sedan, SUV, truck, van), radio type (touch screen or not), and if being sold by an individual or dealership.

2. THE DATA:

The vehicle data was collected on [Charlotte Craigslist](https://charlotte.craigslist.org/search/cta?auto_bodytype=1&auto_bodytype=10&auto_bodytype=11&auto_bodytype=12&auto_bodytype=2&auto_bodytype=3&auto_bodytype=4&auto_bodytype=5&auto_bodytype=6&auto_bodytype=7&auto_bodytype=8&auto_bodytype=9&auto_cylinders=1&auto_cylinders=2&auto_cylinders=3&auto_cylinders=4&auto_cylinders=5&auto_cylinders=6&auto_cylinders=7&auto_cylinders=8&auto_fuel_type=1&auto_fuel_type=2&auto_fuel_type=3&auto_fuel_type=4&auto_fuel_type=6&auto_paint=1&auto_paint=10&auto_paint=11&auto_paint=2&auto_paint=20&auto_paint=3&auto_paint=4&auto_paint=5&auto_paint=6&auto_paint=7&auto_paint=8&auto_paint=9&auto_size=1&auto_size=2&auto_size=3&auto_size=4&auto_title_status=1&auto_title_status=2&auto_title_status=3&auto_title_status=4&auto_title_status=5&auto_title_status=6&auto_transmission=1&auto_transmission=2&auto_transmission=3&bundleDuplicates=1&condition=10&condition=20&condition=30&condition=40&condition=50&condition=60&hasPic=1&max_auto_miles=500000&max_auto_year=2020&max_price=100000&min_auto_miles=1&min_auto_year=1900&min_price=500). The following filters were used to ensure that the posting had the required test variables.

**A. Has image**

**B. Bundle Duplicates**

**C. Price: $500 - $100,000**

**D. Odometer: 1 mile – 500,000 miles**

**E. Condition:**

1. New
2. Like new
3. Excellent
4. Good
5. Fair
6. Salvage

**F. Cylinders:**

1. 3 cylinders
2. 4 cylinders
3. 5 cylinders
4. 6 cylinders
5. 10 cylinders
6. 12 cylinders
7. Other (to account for electric vehicles)

**G. Fuel:**

1. Gas
2. Diesel
3. Hybrid
4. Electric
5. Other

**H. Paint Color:**

1. Black
2. Blue
3. Brown
4. Green
5. Grey
6. Orange
7. Purple
8. Red
9. Silver
10. White
11. Yellow
12. Custom

**I. Size:**

1. Compact
2. Full-size
3. Mid-size
4. Sub-compact

**J. Title status:**

1. Clean
2. Salvage
3. Rebuilt
4. Parts only
5. Lien
6. Missing

**K. Transmission:**

1. Manual
2. Automatic
3. Other

**L. Type:**

1. Bus
2. Convertible
3. Coupe
4. Hatchback
5. Minivan
6. Offroad
7. Pickup
8. Sedan
9. Truck
10. SUV
11. Wagon
12. Van

To collect the data, I created a program in Visual Studio Code using Python to extract the data for the first two Craigslist Pages. I ran the program on April 23rd and 24th, then scrubbed the data for duplicate values. The data can be found here:



1. QUALITATIVE DATA:

**i. Makes:**

To reduce the number of variables in the model, models other than Ford, Chevy, Honda, Toyota, Dodge, GMC, Jeep, BMW, and Nissan were removed from the data set.

**ii. Paint Color:**

To reduce the number of variables used in the model, green, custom, yellow, and orange vehicles were removed from the data set.

**iii. Vehicle Type:**

To reduce the number of variables, “convertible” and” hatchbacks” were consolidated to the “coupe” vehicle type. “Pickup” type vehicles were consolidated into the “truck” vehicle type. “Mini-van” type vehicles were consolidated into the “van” category.

**iv. Title Status:**

To reduce the number of variables, the “lien” title status vehicle was removed from the dataset.

**v. Size:**

To reduce the number of extraneous variables, “sub-compact” vehicles were consolidated into the “compact” category.

**vi. Makes:**

The makes of the vehicles likely have an impact on the sale price of a vehicle, but there are too many makes in the dataset to make the qualitative variables useful.

1. QUANTITATIVE DATA:

**i. Age**:

The age was derived by subtracting the “year” of the vehicle from 2020. The years ranged from 1962 to 2019.

**ii. Number of Cylinders**:

The dataset includes 3, 4, 5, 6, 8, and 10-cylinder vehicles.

**iii. Odometer:**

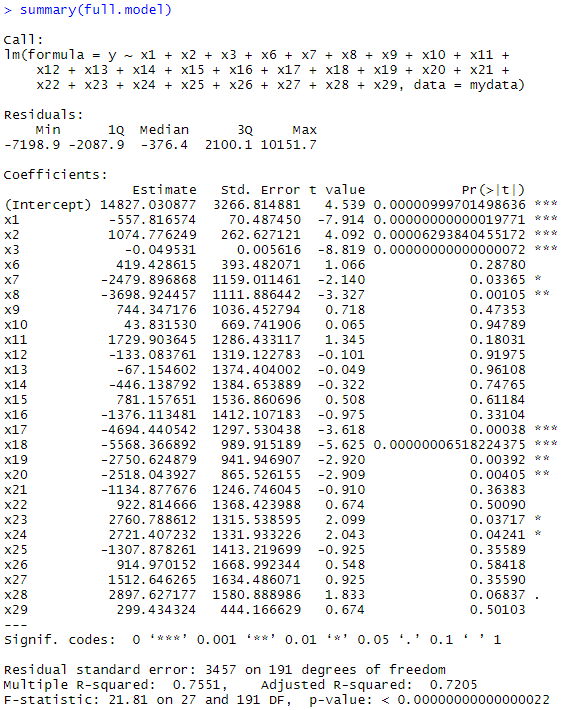
The odometer numbers were adjusted when the number of miles were written in “thousands”. For example, if a 1995 vehicle with a “fair” condition has 135 miles on the odometer, the seller likely meant that the vehicle had 135,000 miles. Without making these adjustments, the data would be skewed. The number of miles range from 283 to 350,000 miles.

1. Y-VALUES:

The purpose of the model is to predict the calling price of a used vehicle on Craigslist. There is a total of *n =*273 vehicles. The price

3. VARIABLE SELECTION:

***A. Assign Variables****:*

*****B. Full Linear Model****:*

***C. Stepwise Regression:***

*null.model = lm(y~1,data=mydata)*

*full.model = lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10*

*+x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21*

*+x22+x23+x24+x25+x26+x27+x28+x29, data = mydata)#Stepwise regression*

*step(null.model, scope = list(upper=full.model),data=mydata, direction="both")*

*Using the Stepwise regression method in RStudio in “both” directions, the best model using:*

*Three Important Variables:*

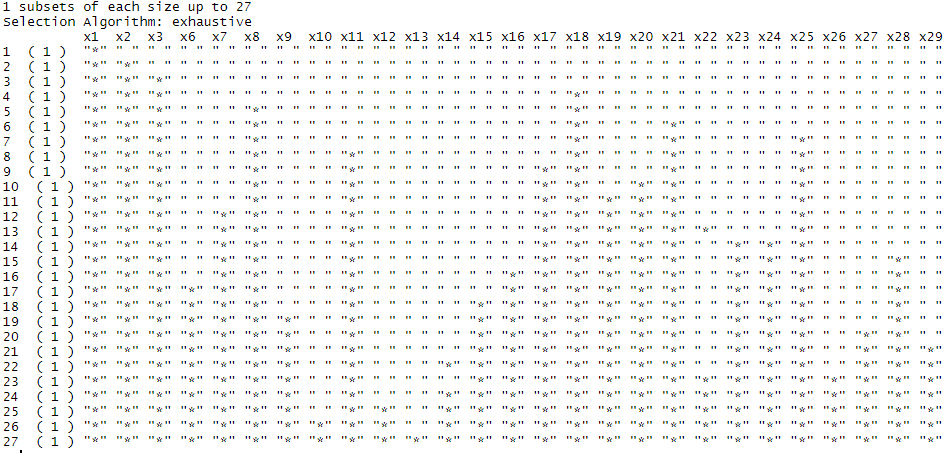
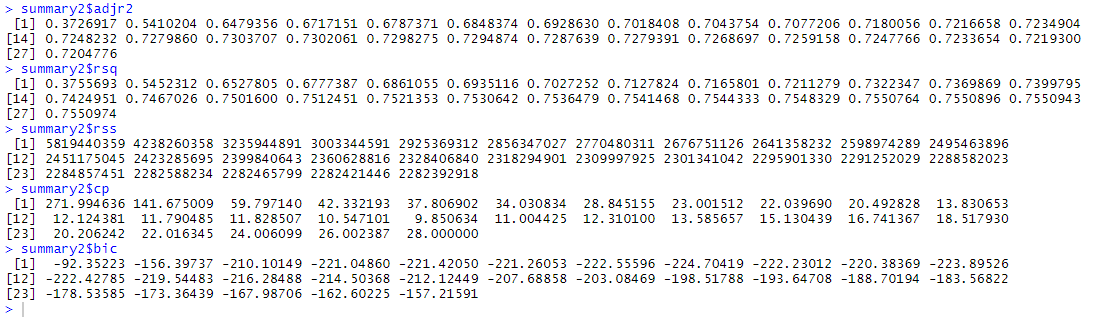
*Four Important Variables:*

***D. All-Possible-Regressions Selection Procedure:***

*full.model2 = regsubsets(y~* *x1+x2+x3 +x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21+x22*

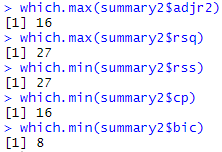
*+x23+x24+x25+x26+x27+x28+x29, mydata, nvmax = 27)*

*summary2 = summary(full.model2)*



*plot(summary2$bic ,xlab = "Number of Variables ", ylab=" BIC"*

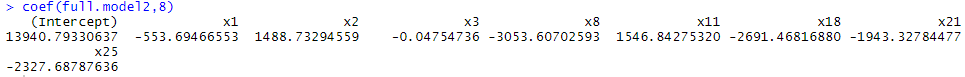
*,type="l")*



***E. Fitting the Model:***

*According to the criterion, the variables should be included in the group of most important predictors.*

*Using the coef() method in RStudio:*



*Fitted Model:*

4. MODEL SELECTION:

***A. Models:***

**Model9**: **Black**

**Model8**: **Black + Red**

**Model7**: **Black + Red + Light Green**

**Model6**: **Black + Red + Light Green + Light Blue**

**Model5**: **Black + Red + Light Green + Light Blue + Purple**

**Model4**: **Black + Red + Light Green + Light Blue + Purple + Orange**

**Model3**: **Black + Red + Light Green + Light Blue + Purple + Orange + Blue**

**Model2**: **Black + Red + Light Green + Light Blue + Purple + Orange + Blue + Gold**

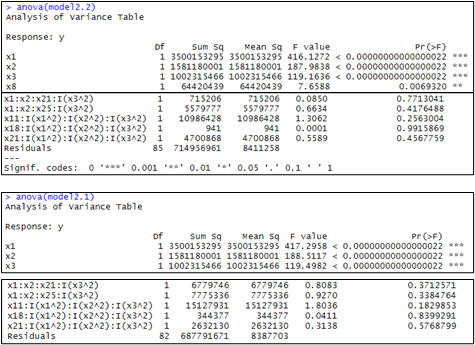
**Model1**: **Black + Red + Light Green + Light Blue + Purple + Orange + Blue + Gold + Pink**

**A screenshot of a cell phone

Description automatically generated**

***B. Partial F-Tests****:*

*Partial F-Test 1: Model2 vs. Model1*

*Test to see whether or not the product of one quadratic term and two individual terms is a useful predictor for a used vehicle on Craigslist.*

*714956961*

*687791671*

*8387703*

*Since the F-statistic is less than the F-critical value when , we fail to reject the null hypothesis and conclude that the product of one quadratic term and two individual terms are not statistically useful to the model.*

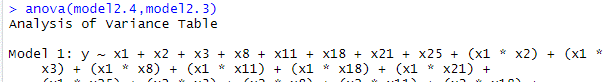
*Partial F-Test 2: Model3 vs. Model2*

 *Test to see whether or not the product of three quadratic terms and an individual term is a useful predictor for a used vehicle on Craigslist.*

*Since the F-statistic is less than the F-critical value when , we fail to reject the null hypothesis and conclude that the product of three quadratic terms and an individual term are not statistically useful to the model.*

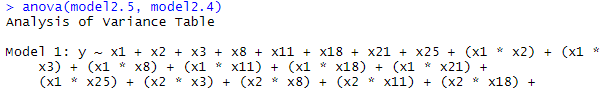
*.*

*Partial F-Test 3: Model4 vs. Model3*

 *Test to see whether or not the individual quadratic terms are useful predictors for a used vehicle on Craigslist.*

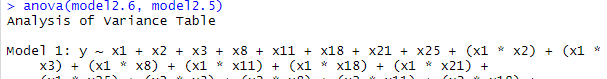
*Since the F-statistic is less than the F-critical value when , we fail to reject the null hypothesis and conclude that the individual quadratic terms are statistically useful to the model.*

*Partial F-Test 4: Model5 vs. Model4*

 *Test to see whether or not the product of one quadratic term and three individual terms is a useful predictor for a used vehicle on Craigslist.*

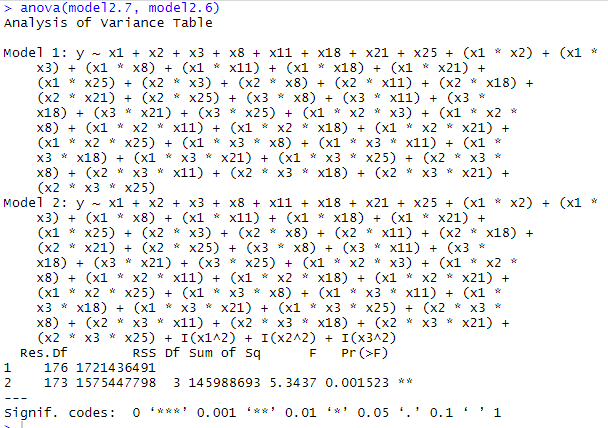
*Since the F-statistic is less than the F-critical value when , we fail to reject the null hypothesis and conclude that the product of one quadratic term and three individuals term is not statistically useful to the model.*

*Partial F-Test 5: Model6 vs. Model5*

 *Test to see whether or not the product of one quadratic term and one individual term is a useful predictor for a used vehicle on Craigslist.*

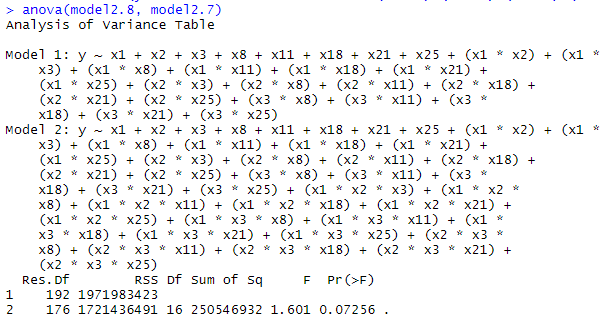
*Since the F-statistic is greater than the F-critical value when , we reject the null hypothesis and conclude that at least one product of one quadratic term and one individual term contributes to the model.*

*Partial F-Test 6: Model7 vs. Model6*

 *Test to see whether or not the product of one quadratic term and one individual term is a useful predictor for a used vehicle on Craigslist.*

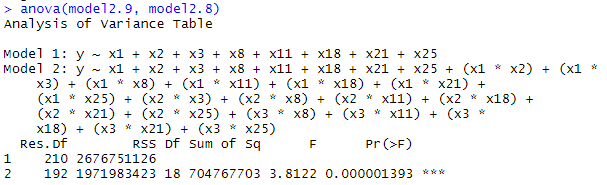
*Since the F-statistic is greater than the F-critical value when , we reject the null hypothesis and conclude that at least one product of one quadratic term and one individual term contributes to the model.*

*Partial F-Test 7: Model8 vs. Model7*

 *Test to see whether or not the product of three individual terms is a useful predictor for a used vehicle on Craigslist.*

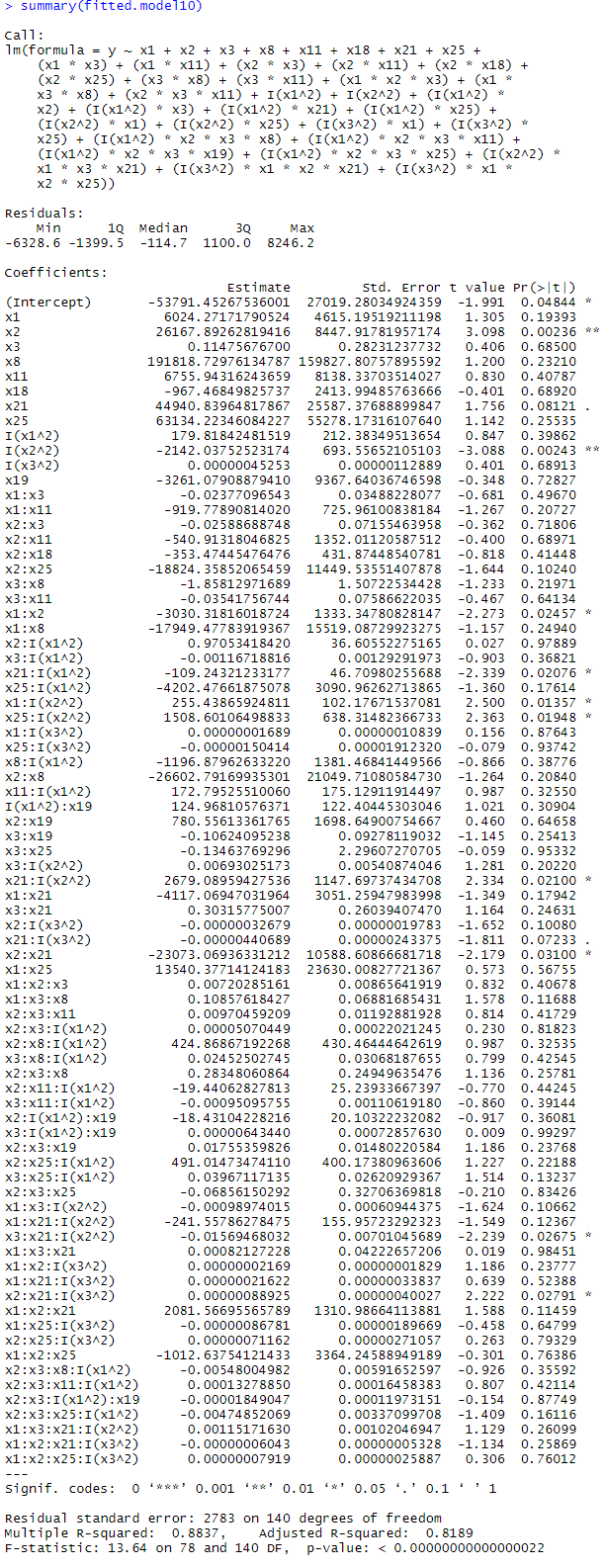
*Since the F-statistic is greater than the F-critical value when , we reject the null hypothesis and conclude that at least one product of three individual terms contributes to the model.*

*Partial F-Test 8: Model9 vs. Model8*

 *Test to see whether or not the product of two individual terms is a useful predictor for a used vehicle on Craigslist.*

*Since the F-statistic is greater than the F-critical value when , we reject the null hypothesis and conclude that at least one product of two individual terms contributes to the model.*

*Fitting the Model:*



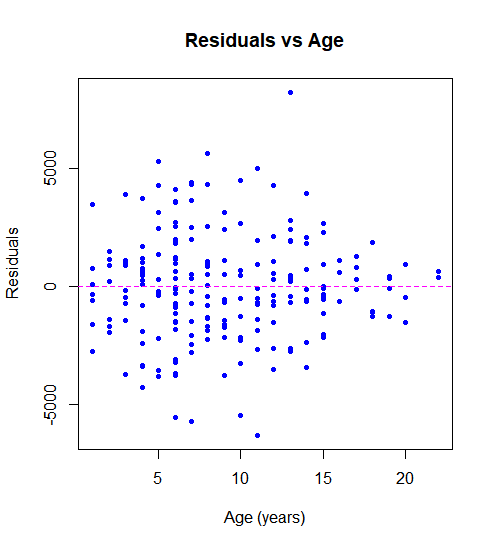
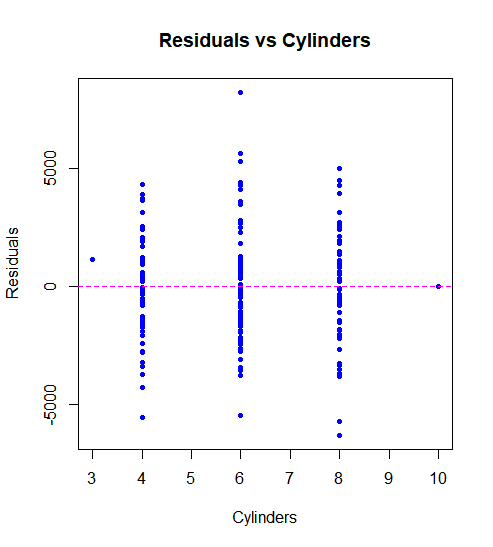
*For the models with useful predictors, each predictor was individually removed, tested, and then the p-value of the model was measured. If the p-value increased when a predictor was removed, then the predictor was kept. If the p-value decreased, then the predictor was not used.*

*Fitted model:*

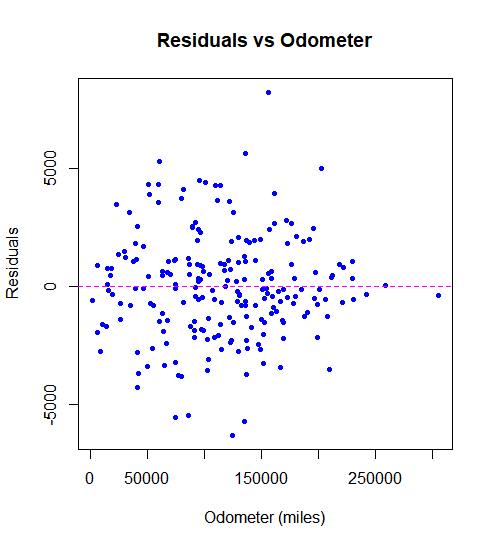
5. ASSUMPTION CHECK:

***A. Lack of Fit:***

*X1: Age X2: Cylinders*

*X3: Odometer*



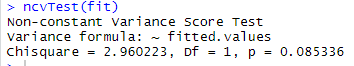
*There are no distinct patterns between the residuals and x1 and x3. At first glance, there appears to be a parabolic distribution*

*in x2. However, there are outliers*

*in the data to account for the 3 and 10-cylinder vehicles. Once those outliers are removed, x2 should not have the pattern.*

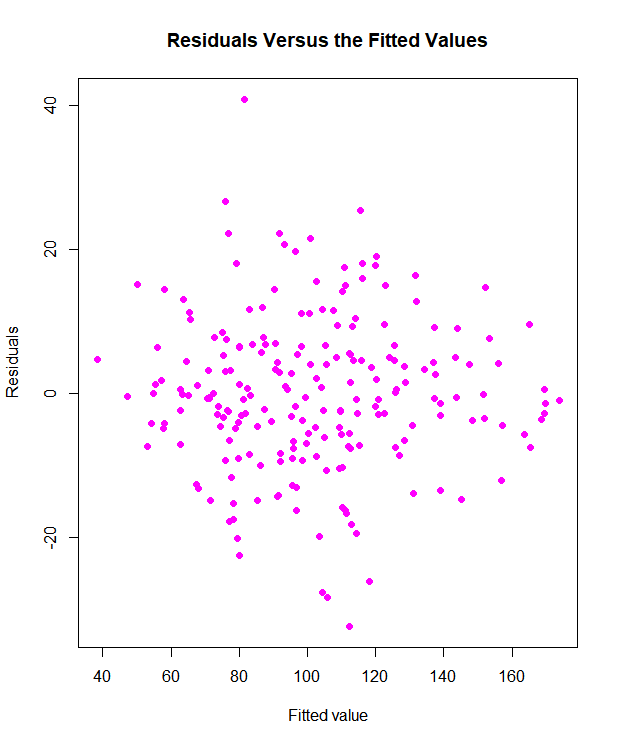
***B. Unequal Variance (Homoscedasticity):***

*ncvTest(fit)*



*p-value = 0.0853*

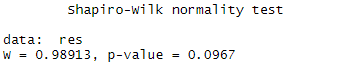
*Since the p-value is greater than alpha, we fail to reject the null hypothesis and conclude that the errors have a constant variance at* .



***C. Normality:***

***res = fit$residuals***

***shapiro.test(res)***



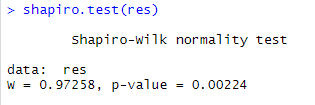
***qqnorm(fit$residuals)***

***qqline(fit$residuals)***

*Since the p-value is less than alpha, we reject the null hypothesis and conclude that the errors are not normally distributed at* .

*Solution:*

*Using the Box-Cox method, the value generated is 0.1919192.*



*New Fitted Model:*

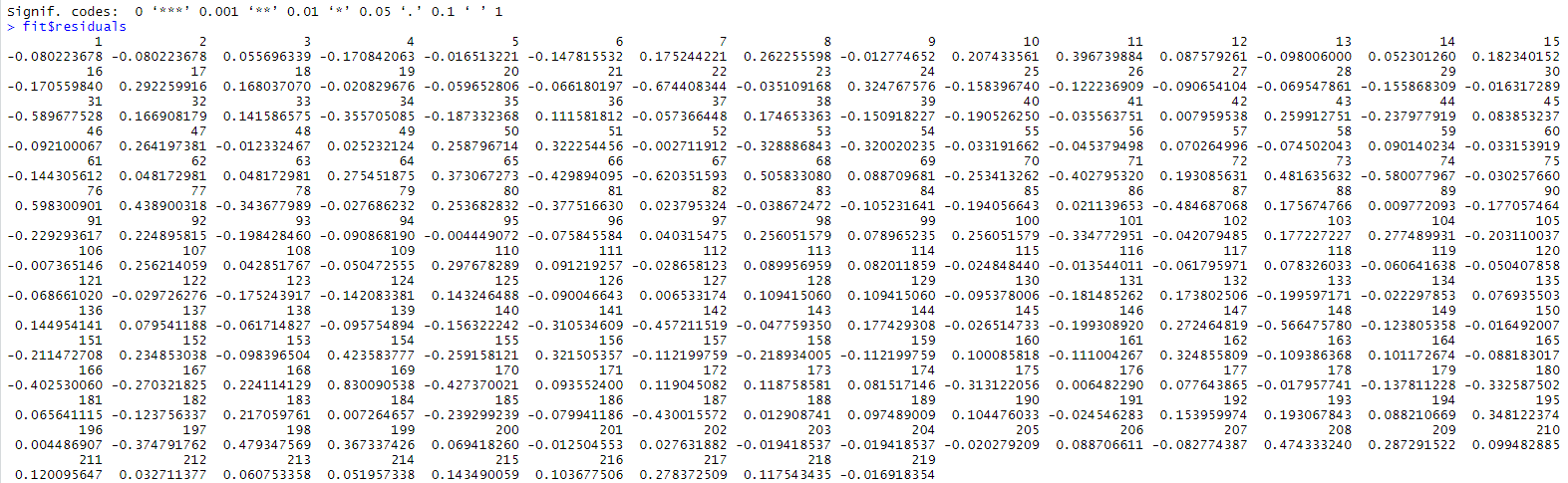
*fit = lm((y^0.1919192) ~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x3)+(x1\*x11)+(x2\*x3)+*

*(x2\*x11)+(x2\*x18)+(x2\*x25)+(x3\*x8)+(x3\*x11)+(x1\*x2\*x3)+(x1\*x3\*x8)+(x2\*x3\*x11)+I(x1^2)+I(x2^2)+(I(x1^2)\*x2)+(I(x1^2)\*x3)+(I(x1^2)\*x21)+(I(x1^2)\*x25)+(I(x2^2)\*x1)+(I(x2^2)\*x25)+(I(x3^2)\*x1)+(I(x3^2)\*x25)+(I(x1^2)\*x2\*x3\*x8)+(I(x1^2)\*x2\*x3\*x11)+(I(x1^2)\*x2\*x3\*x19)+(I(x1^2)\*x2\*x3\*x25)+(I(x2^2)\*x1\*x3\*x21)+(I(x3^2)\*x1\*x2\*x21)+(I(x3^2)\*x1\*x2\*x25))*

***D. Outliers:***

*Residuals*:

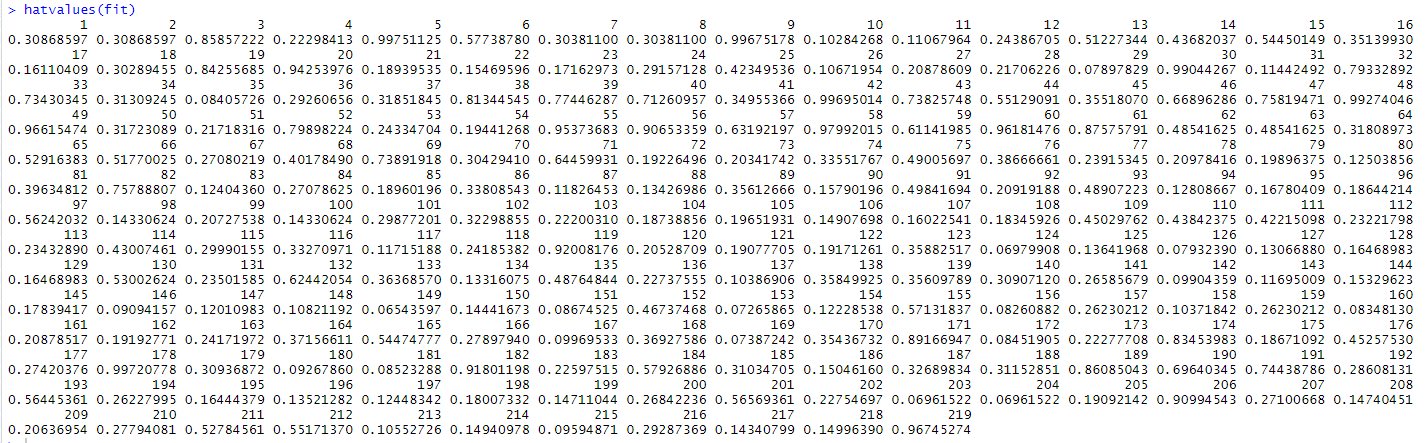
*fit$residuals*



*None of the residuals were higher than three standard deviations (0.8432).*

*Influence value* :

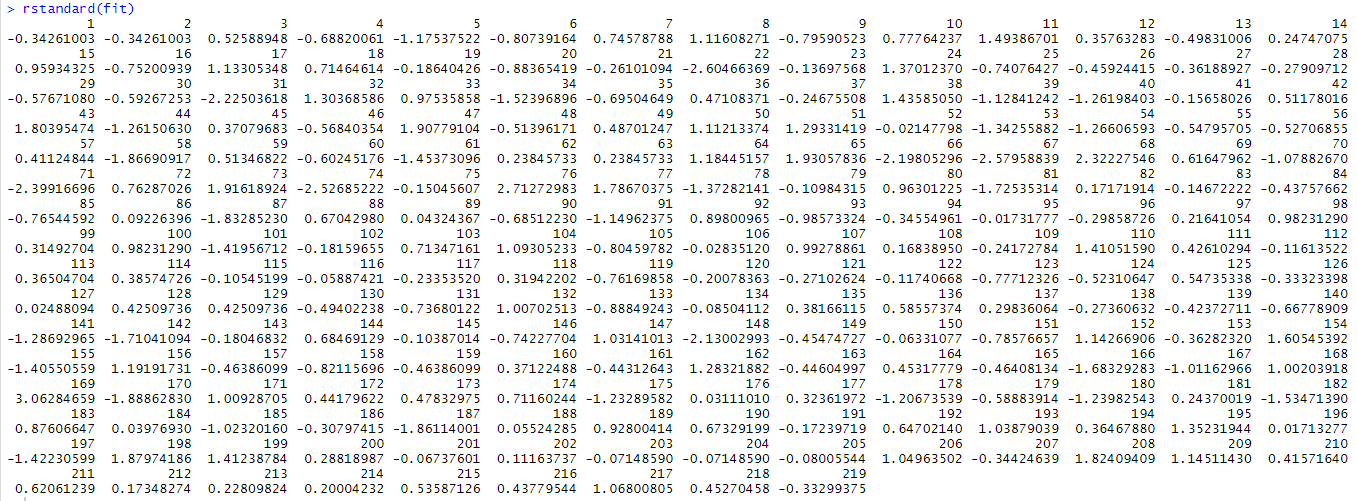
*Test value = ==*



*158 observations had a higher value than 0.1735. Values with a hat value of 0.8 or higher were removed from the dataset:*

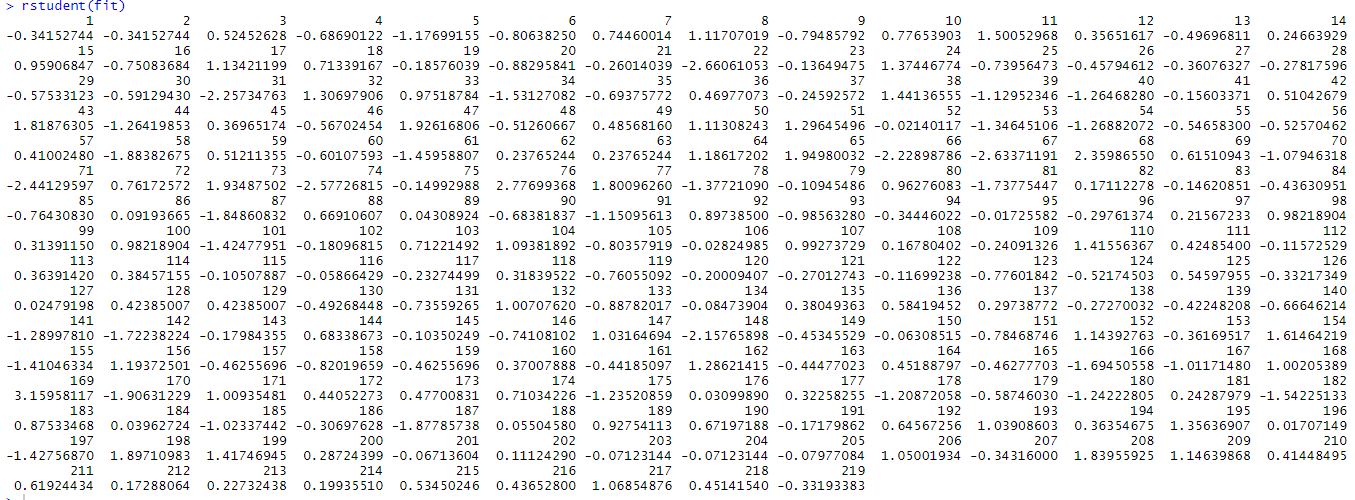
*Observations removed: 9, 19, 30, 38, 48, 49, 55, 56, 60, 119, 174, and 206.*

*Standardized Residuals*:



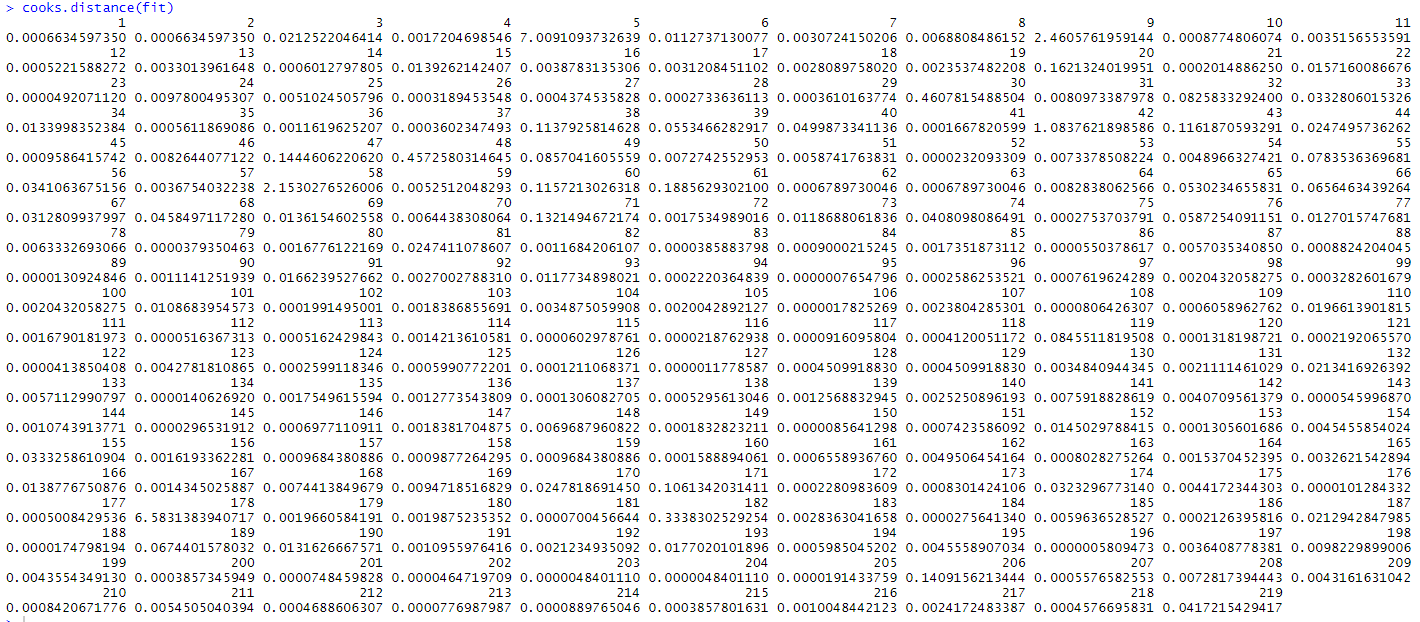
*Observation 169 has a RStandard value (3.0628) higher than 3. This value was removed from the dataset*

*Stundentized Residuals*:



*Observation 169 has a RStudent value (3.1596) higher than 3. This value was removed from the dataset*

*Cooks Distance*:



*Observations 42 (1.0838), 58 (2.1530), and 178 (6.5831) have values greater than 1. These values were removed from the dataset.*

*Dfbetas*:

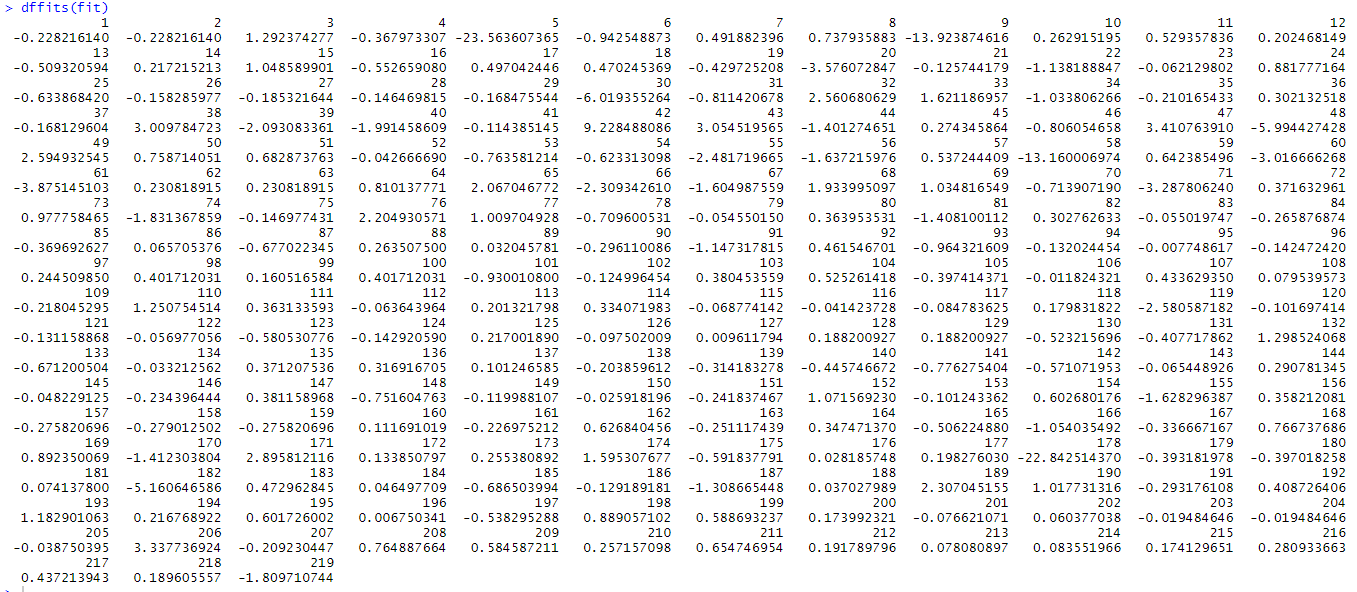
*Test value =*

*There are 10,950 Dfbeta values for this model – too many to display in this paper. I will provide an excel sheet with the Dfvalues which have been highlighted if they exceeded 0.1351. 100 observations exceeded this threshold. In order to avoid excess bias in the data and save time combing over the data, I’ve removed those observations that exceeded the 0.1351 threshold in 15 or more predictors:*

*Observations removed: 5, 20, 32, 34, 43, 47, 58, 61, 65, 66, 68, 71, 74, 76, 81, 91, 93, 110, 119, 132, 155, 162, 166, 170, 171, 178, 182, 187, 189, and 193.*

*Dffits*:

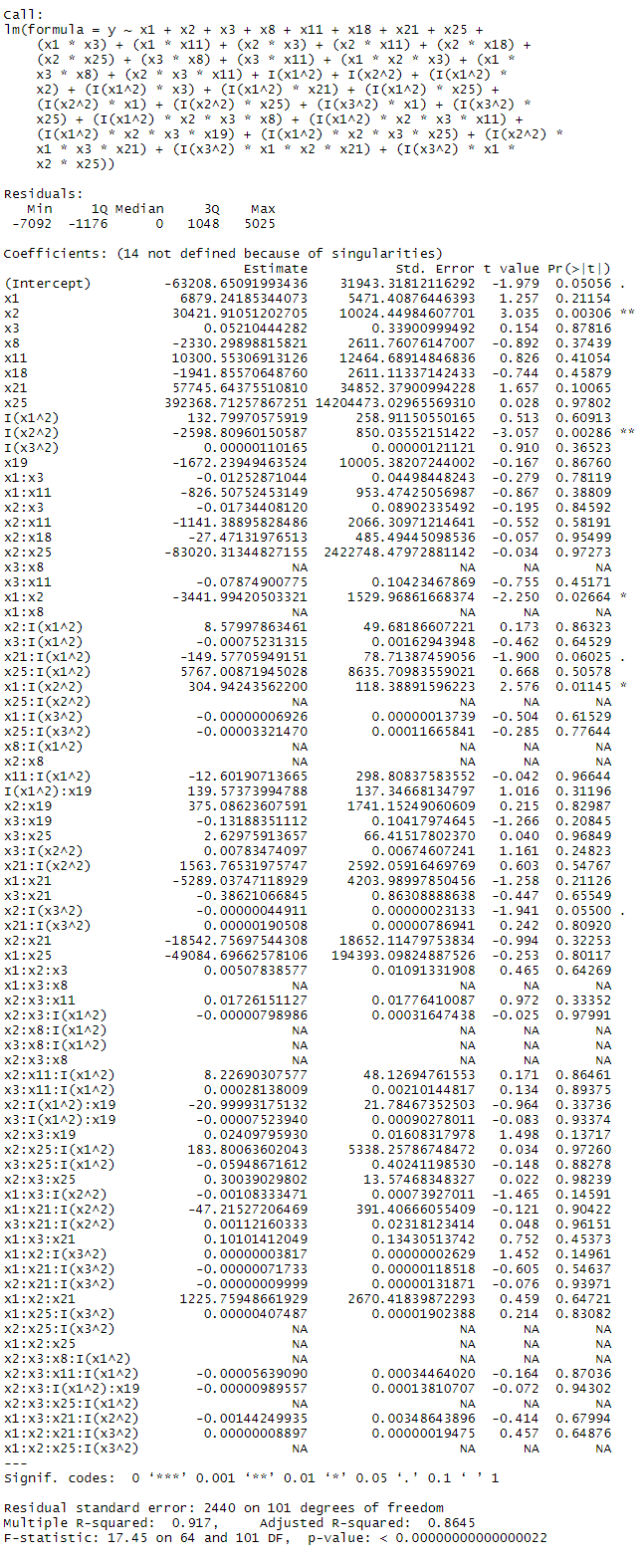
*Test value =*



*There were 77 observations which exceeded the 0.6340 threshold. Many of the 77 observation values were identified as influencers in the Dfbetas. The following observations were removed from the dataset:*

*Observations removed: 3, 39, 46, 51, 73, 78, 87, 168, and 219.*

***E. Global F-Test****:*

*Test to see whether or not the model is statistically useful for predicting the calling price of a used vehicle on Craigslist. The Box-Cox normalizing transformation value was removed to provide an accurate standard error value.*

*Since the p-value is less than , we reject the null hypothesis and conclude that the model is statistically useful for predicting the calling price of a used vehicle on Craigslist (y).*

6. CONCLUSION:

Provide adjusted R squared, model standard error and explain is this model good to use.

***Standard Error of Model =2440***

*From the value, 91% of the sample variation can be explained by the model.*

*The model can predict the value of a used car on Charlotte Craigslist to within 2s = $2,440 of its true value.*

*The value and Standard Error of the Model are sufficient for predicting the estimated price of a used car on Charlotte Craigslist.*

7. CHECK:

*Checking the accuracy using the current data:*

set.seed(100)

trainingRowIndex=sample(1:nrow(mydata),0.8\*nrow(mydata))

trainingData = mydata[trainingRowIndex, ]

testData = mydata[-trainingRowIndex, ]

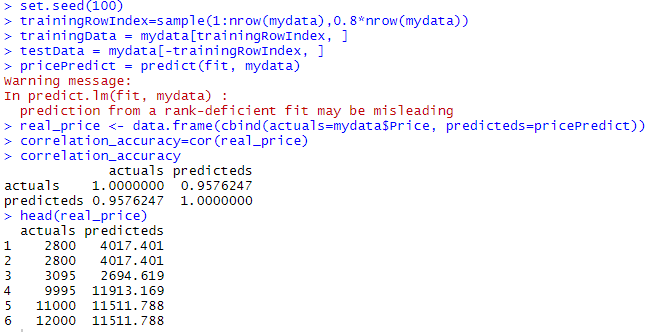
pricePredict = predict(fit, mydata)

real\_price <- data.frame(cbind(actuals=mydata$Price, predicteds=pricePredict))

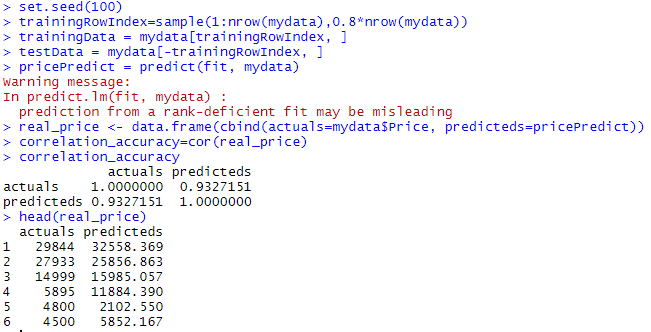
correlation\_accuracy=cor(real\_price)

correlation\_accuracy

head(real\_price)



Checking the accuracy using the new data collected on April 29th:



***The predicted price is close to the original price.***

7. R Code:

#install.packages("leaps")

library(leaps)

#install.packages("readxl")

library("readxl")

#install.packages("car")

library(car)

options("scipen"= 100)

mydata = read\_excel("C:/Users/Matt/Desktop/UNCC/UNCC Spring 20/STAT 2223-002 Pramesh Subedi - Elements of Statistics II/STAT Final Project/Final Project/CraigslistCarData.xlsx")

y = mydata$Price

#Quantitative Variables

x1 = mydata$Age #in years

x2 = mydata$Cylinders

x3 = mydata$Odometer

x4 = mydata$Min\_Price

x5 = mydata$Max\_Price

x6 = mydata$Rating

#Qualitative Variables

#Title Status

x7 = ifelse(mydata$`Title\_Status` == "rebuilt", 1, 0) #Clean is base

x8 = ifelse(mydata$`Title\_Status` == "salvage", 1, 0) #Clean is base

#Size

x9 = ifelse(mydata$Size == "compact", 1, 0) #Full-size is base

x10 = ifelse(mydata$Size == "mid-size", 1, 0) #Full-size is base

#Color

x11 = ifelse(mydata$Color == "white", 1, 0) #Brown is base

x12 = ifelse(mydata$Color == "silver", 1, 0) #Brown is base

x13 = ifelse(mydata$Color == "black", 1, 0) #Brown is base

x14 = ifelse(mydata$Color == "grey", 1, 0) #Brown is base

x15 = ifelse(mydata$Color == "red", 1, 0) #Brown is base

x16 = ifelse(mydata$Color == "blue", 1, 0) #Brown is base

#Vehicle Type

x17 = ifelse(mydata$Type == "coupe", 1, 0) #Truck is base

x18 = ifelse(mydata$Type == "sedan", 1, 0) #Truck is base

x19 = ifelse(mydata$Type == "SUV", 1, 0) #Truck is base

x20 = ifelse(mydata$Type == "van", 1, 0) #Truck is base

#Vehicle Make

x21 = ifelse(mydata$Make == "Ford", 1, 0) #Nissan is base

x22 = ifelse(mydata$Make == "Chevy", 1, 0) #Nissan is base

x23 = ifelse(mydata$Make == "Honda", 1, 0) #Nissan is base

x24 = ifelse(mydata$Make == "Toyota", 1, 0) #Nissan is base

x25 = ifelse(mydata$Make == "Dodge", 1, 0) #Nissan is base

x26 = ifelse(mydata$Make == "GMC", 1, 0) #Nissan is base

x27 = ifelse(mydata$Make == "Jeep", 1, 0) #Nissan is base

x28 = ifelse(mydata$Make == "BMW", 1, 0) #Nissan is base

x29 = mydata$`Num Doors`

null.model = lm(y~1,data=mydata)

full.model = lm(y~x1+x2+x3+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26+x27+x28+x29, data=mydata)summary(full.model)

step(null.model, scope = list(upper=full.model), data=mydata, direction="both")

full.model2 = regsubsets(y~x1+x2+x3+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21+x22+x23+x24+x25+x26+x27+x28+x29, data=mydata, nvmax = 27)

summary2 = summary(full.model2)

summary2

summary2$adjr2

summary2$rsq

summary2$rss

summary2$cp

summary2$bic

which.max(summary2$adjr2)

which.max(summary2$rsq)

which.min(summary2$rss)

which.min(summary2$cp)

which.min(summary2$bic)

full.model.y.hat=full.model$fitted.values

full.model.res=full.model$residuals

coef(full.model2,8)

plot(x1,y, xlab="Age (Years)", ylab="Price", main="Price vs Age", col="6", pch=19)

plot(x2,y, xlab="Cylinders", ylab="Price", main="Price vs Cylinders", col="6", pch=19)

plot(x3,y, xlab="Odometer (Miles)", ylab="Price", main="Price vs Odometer", col="6", pch=19)

length(y)

fitted.model2 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25, data=mydata)

full.model2 = regsubsets(y~x1+x2+x3+x8+x18+x21+x25, data = mydata)

full.model2 = lm((y^.5)~x1+x2+x3+x8+x18+x21+x25, data = mydata)

model2.1 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x2)+(x1\*x3)+(x1\*x8)+(x1\*x11)+(x1\*x18)+(x1\*x21)+(x1\*x25)+(x2\*x3)+(x2\*x8)+(x2\*x11)+(x2\*x18)+(x2\*x21)+(x2\*x25)+(x3\*x8)+(x3\*x11)+(x3\*x18)+(x3\*x21)+(x3\*x25)+(x1\*x2\*x3)+(x1\*x2\*x8)+(x1\*x2\*x11)+(x1\*x2\*x18)+(x1\*x2\*x21)+(x1\*x2\*x25)+(x1\*x3\*x8)+(x1\*x3\*x11)+(x1\*x3\*x18)+(x1\*x3\*x21)+(x1\*x3\*x25)+(x2\*x3\*x8)+(x2\*x3\*x11)+(x2\*x3\*x18)+(x2\*x3\*x21)+(x2\*x3\*x25)+I(x1^2)+I(x2^2)+I(x3^2)+(I(x1^2)\*x2)+(I(x1^2)\*x3)+(I(x1^2)\*x8)+(I(x1^2)\*x11)+(I(x1^2)\*x18)+(I(x1^2)\*x21)+(I(x1^2)\*x25)+(I(x2^2)\*x1)+(I(x2^2)\*x3)+(I(x2^2)\*x8)+(I(x2^2)\*x11)+(I(x2^2)\*x18)+(I(x2^2)\*x21)+(I(x2^2)\*x25)+(I(x3^2)\*x1)+(I(x3^2)\*x2)+(I(x3^2)\*x8)+(I(x3^2)\*x11)+(I(x3^2)\*x18)+(I(x3^2)\*x21)+(I(x3^2)\*x25)+(I(x1^2)\*x2\*x3\*x8)+(I(x1^2)\*x2\*x3\*x11)+(I(x1^2)\*x2\*x3\*x18)+(I(x1^2)\*x2\*x3\*x19)+(I(x1^2)\*x2\*x3\*x21)+(I(x1^2)\*x2\*x3\*x25)+(I(x2^2)\*x1\*x3\*x8)+(I(x2^2)\*x1\*x3\*x11)+(I(x2^2)\*x1\*x3\*x18)+(I(x2^2)\*x1\*x3\*x21)+(I(x2^2)\*x1\*x3\*x25)+(I(x3^2)\*x1\*x2\*x8)+(I(x3^2)\*x1\*x2\*x11)+(I(x3^2)\*x1\*x2\*x18)+(I(x3^2)\*x1\*x2\*x21)+(I(x3^2)\*x1\*x2\*x25)+(I(x1^2)\*x2\*x3)+(I(x1^2)\*x2\*x8)+(I(x1^2)\*x2\*x11)+(I(x1^2)\*x2\*x18)+(I(x1^2)\*x2\*x21)+(I(x1^2)\*x2\*x25)+(I(x1^2)\*x3\*x8)+(I(x1^2)\*x3\*x11)+(I(x1^2)\*x3\*x18)+(I(x1^2)\*x3\*x21)+(I(x1^2)\*x3\*x25)+(I(x2^2)\*x1\*x3)+(I(x2^2)\*x1\*x8)+(I(x2^2)\*x1\*x11)+(I(x2^2)\*x1\*x18)+(I(x2^2)\*x1\*x21)+(I(x2^2)\*x1\*x25)+(I(x2^2)\*x3\*x8)+(I(x2^2)\*x3\*x11)+(I(x2^2)\*x3\*x18)+(I(x2^2)\*x3\*x21)+(I(x2^2)\*x3\*x25)+(I(x3^2)\*x1\*x3)+(I(x3^2)\*x1\*x8)+(I(x3^2)\*x1\*x11)+(I(x3^2)\*x1\*x18)+(I(x3^2)\*x1\*x21)+(I(x3^2)\*x1\*x25)+(I(x3^2)\*x2\*x8)+(I(x3^2)\*x2\*x11)+(I(x3^2)\*x2\*x18)+(I(x3^2)\*x2\*x21)+(I(x3^2)\*x2\*x25)+(I(x1^2)\*I(x2^2)\*I(x3^2)\*x8)+(I(x1^2)\*I(x2^2)\*I(x3^2)\*x11)+(I(x1^2)\*I(x2^2)\*I(x3^2)\*x18)+(I(x1^2)\*I(x2^2)\*I(x3^2)\*x21)+(I(x1^2)\*I(x2^2)\*I(x3^2)\*x25)+(I(x1^2)\*I(x2^2)\*x3)+(I(x1^2)\*I(x2^2)\*x8)+(I(x1^2)\*I(x2^2)\*x11)+(I(x1^2)\*I(x2^2)\*x18)+(I(x1^2)\*I(x2^2)\*x21)+(I(x1^2)\*I(x2^2)\*x25)+(I(x1^2)\*I(x3^2)\*x2)+(I(x1^2)\*I(x3^2)\*x8)+(I(x1^2)\*I(x3^2)\*x11)+(I(x1^2)\*I(x3^2)\*x18)+(I(x1^2)\*I(x3^2)\*x21)+(I(x1^2)\*I(x3^2)\*x25)+(I(x2^2)\*I(x3^2)\*x1)+(I(x2^2)\*I(x3^2)\*x8)+(I(x2^2)\*I(x3^2)\*x11)+(I(x2^2)\*I(x3^2)\*x18)+ (I(x2^2)\*I(x3^2)\*x21)+(I(x2^2)\*I(x3^2)\*x25))

anova(model2.2, model2.1)

model2.2 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x2)+(x1\*x3)+(x1\*x8)+(x1\*x11)+(x1\*x18)+(x1\*x21)+(x1\*x25)+(x2\*x3)+(x2\*x8)+(x2\*x11)+(x2\*x18)+(x2\*x21)+(x2\*x25)+(x3\*x8)+(x3\*x11)+(x3\*x18)+(x3\*x21)+(x3\*x25)+(x1\*x2\*x3)+(x1\*x2\*x8)+(x1\*x2\*x11)+(x1\*x2\*x18)+(x1\*x2\*x21)+(x1\*x2\*x25)+(x1\*x3\*x8)+(x1\*x3\*x11)+(x1\*x3\*x18)+(x1\*x3\*x21)+(x1\*x3\*x25)+(x2\*x3\*x8)+(x2\*x3\*x11)+(x2\*x3\*x18)+(x2\*x3\*x21)+(x2\*x3\*x25)+I(x1^2)+I(x2^2)+I(x3^2)+(I(x1^2)\*x2)+(I(x1^2)\*x3)+(I(x1^2)\*x8)+(I(x1^2)\*x11)+(I(x1^2)\*x18)+(I(x1^2)\*x21)+(I(x1^2)\*x25)+(I(x2^2)\*x1)+(I(x2^2)\*x3)+(I(x2^2)\*x8)+(I(x2^2)\*x11)+(I(x2^2)\*x18)+(I(x2^2)\*x21)+(I(x2^2)\*x25)+(I(x3^2)\*x1)+(I(x3^2)\*x2)+(I(x3^2)\*x8)+(I(x3^2)\*x11)+(I(x3^2)\*x18)+(I(x3^2)\*x21)+(I(x3^2)\*x25)+(I(x1^2)\*x2\*x3\*x8)+(I(x1^2)\*x2\*x3\*x11)+(I(x1^2)\*x2\*x3\*x18)+(I(x1^2)\*x2\*x3\*x19)+(I(x1^2)\*x2\*x3\*x21)+(I(x1^2)\*x2\*x3\*x25)+(I(x2^2)\*x1\*x3\*x8)+(I(x2^2)\*x1\*x3\*x11)+(I(x2^2)\*x1\*x3\*x18)+(I(x2^2)\*x1\*x3\*x21)+(I(x2^2)\*x1\*x3\*x25)+(I(x3^2)\*x1\*x2\*x8)+(I(x3^2)\*x1\*x2\*x11)+(I(x3^2)\*x1\*x2\*x18)+(I(x3^2)\*x1\*x2\*x21)+(I(x3^2)\*x1\*x2\*x25)+(I(x1^2)\*x2\*x3)+(I(x1^2)\*x2\*x8)+(I(x1^2)\*x2\*x11)+(I(x1^2)\*x2\*x18)+(I(x1^2)\*x2\*x21)+(I(x1^2)\*x2\*x25)+(I(x1^2)\*x3\*x8)+(I(x1^2)\*x3\*x11)+(I(x1^2)\*x3\*x18)+(I(x1^2)\*x3\*x21)+(I(x1^2)\*x3\*x25)+(I(x2^2)\*x1\*x3)+(I(x2^2)\*x1\*x8)+(I(x2^2)\*x1\*x11)+(I(x2^2)\*x1\*x18)+(I(x2^2)\*x1\*x21)+(I(x2^2)\*x1\*x25)+(I(x2^2)\*x3\*x8)+(I(x2^2)\*x3\*x11)+(I(x2^2)\*x3\*x18)+(I(x2^2)\*x3\*x21)+(I(x2^2)\*x3\*x25)+(I(x3^2)\*x1\*x3)+(I(x3^2)\*x1\*x8)+(I(x3^2)\*x1\*x11)+(I(x3^2)\*x1\*x18)+(I(x3^2)\*x1\*x21)+(I(x3^2)\*x1\*x25)+(I(x3^2)\*x2\*x8)+(I(x3^2)\*x2\*x11)+(I(x3^2)\*x2\*x18)+(I(x3^2)\*x2\*x21)+(I(x3^2)\*x2\*x25)+(I(x1^2)\*I(x2^2)\*I(x3^2)\*x8)+(I(x1^2)\*I(x2^2)\*I(x3^2)\*x11)+(I(x1^2)\*I(x2^2)\*I(x3^2)\*x18)+(I(x1^2)\*I(x2^2)\*I(x3^2)\*x21)+(I(x1^2)\*I(x2^2)\*I(x3^2)\*x25))

anova(model2.3, model2.2)

model2.3 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x2)+(x1\*x3)+(x1\*x8)+(x1\*x11)+(x1\*x18)+(x1\*x21)+(x1\*x25)+(x2\*x3)+(x2\*x8)+(x2\*x11)+(x2\*x18)+(x2\*x21)+(x2\*x25)+(x3\*x8)+(x3\*x11)+(x3\*x18)+(x3\*x21)+(x3\*x25)+(x1\*x2\*x3)+(x1\*x2\*x8)+(x1\*x2\*x11)+(x1\*x2\*x18)+(x1\*x2\*x21)+(x1\*x2\*x25)+(x1\*x3\*x8)+(x1\*x3\*x11)+(x1\*x3\*x18)+(x1\*x3\*x21)+(x1\*x3\*x25)+(x2\*x3\*x8)+(x2\*x3\*x11)+(x2\*x3\*x18)+(x2\*x3\*x21)+(x2\*x3\*x25)+I(x1^2)+I(x2^2)+I(x3^2)+(I(x1^2)\*x2)+(I(x1^2)\*x3)+(I(x1^2)\*x8)+(I(x1^2)\*x11)+(I(x1^2)\*x18)+(I(x1^2)\*x21)+(I(x1^2)\*x25)+(I(x2^2)\*x1)+(I(x2^2)\*x3)+(I(x2^2)\*x8)+(I(x2^2)\*x11)+(I(x2^2)\*x18)+(I(x2^2)\*x21)+(I(x2^2)\*x25)+(I(x3^2)\*x1)+(I(x3^2)\*x2)+(I(x3^2)\*x8)+(I(x3^2)\*x11)+(I(x3^2)\*x18)+(I(x3^2)\*x21)+(I(x3^2)\*x25)+(I(x1^2)\*x2\*x3\*x8)+(I(x1^2)\*x2\*x3\*x11)+(I(x1^2)\*x2\*x3\*x18)+(I(x1^2)\*x2\*x3\*x19)+(I(x1^2)\*x2\*x3\*x21)+(I(x1^2)\*x2\*x3\*x25)+(I(x2^2)\*x1\*x3\*x8)+(I(x2^2)\*x1\*x3\*x11)+(I(x2^2)\*x1\*x3\*x18)+(I(x2^2)\*x1\*x3\*x21)+(I(x2^2)\*x1\*x3\*x25)+(I(x3^2)\*x1\*x2\*x8)+(I(x3^2)\*x1\*x2\*x11)+(I(x3^2)\*x1\*x2\*x18)+(I(x3^2)\*x1\*x2\*x21)+(I(x3^2)\*x1\*x2\*x25)+(I(x1^2)\*x2\*x3)+(I(x1^2)\*x2\*x8)+(I(x1^2)\*x2\*x11)+(I(x1^2)\*x2\*x18)+(I(x1^2)\*x2\*x21)+(I(x1^2)\*x2\*x25)+(I(x1^2)\*x3\*x8)+(I(x1^2)\*x3\*x11)+(I(x1^2)\*x3\*x18)+(I(x1^2)\*x3\*x21)+(I(x1^2)\*x3\*x25)+(I(x2^2)\*x1\*x3)+(I(x2^2)\*x1\*x8)+(I(x2^2)\*x1\*x11)+(I(x2^2)\*x1\*x18)+(I(x2^2)\*x1\*x21)+(I(x2^2)\*x1\*x25)+(I(x2^2)\*x3\*x8)+(I(x2^2)\*x3\*x11)+(I(x2^2)\*x3\*x18)+(I(x2^2)\*x3\*x21)+(I(x2^2)\*x3\*x25)+(I(x3^2)\*x1\*x3)+(I(x3^2)\*x1\*x8)+(I(x3^2)\*x1\*x11)+(I(x3^2)\*x1\*x18)+(I(x3^2)\*x1\*x21)+(I(x3^2)\*x1\*x25)+(I(x3^2)\*x2\*x8)+(I(x3^2)\*x2\*x11)+(I(x3^2)\*x2\*x18)+(I(x3^2)\*x2\*x21)+(I(x3^2)\*x2\*x25))

anova(model2.4, model2.3)

model2.4 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x2)+(x1\*x3)+(x1\*x8)+(x1\*x11)+(x1\*x18)+(x1\*x21)+(x1\*x25)+(x2\*x3)+(x2\*x8)+(x2\*x11)+(x2\*x18)+(x2\*x21)+(x2\*x25)+(x3\*x8)+(x3\*x11)+(x3\*x18)+(x3\*x21)+(x3\*x25)+(x1\*x2\*x3)+(x1\*x2\*x8)+(x1\*x2\*x11)+(x1\*x2\*x18)+(x1\*x2\*x21)+(x1\*x2\*x25)+(x1\*x3\*x8)+(x1\*x3\*x11)+(x1\*x3\*x18)+(x1\*x3\*x21)+(x1\*x3\*x25)+(x2\*x3\*x8)+(x2\*x3\*x11)+(x2\*x3\*x18)+(x2\*x3\*x21)+(x2\*x3\*x25)+I(x1^2)+I(x2^2)+I(x3^2)+(I(x1^2)\*x2)+(I(x1^2)\*x3)+(I(x1^2)\*x8)+(I(x1^2)\*x11)+(I(x1^2)\*x18)+(I(x1^2)\*x21)+(I(x1^2)\*x25)+(I(x2^2)\*x1)+(I(x2^2)\*x3)+(I(x2^2)\*x8)+(I(x2^2)\*x11)+(I(x2^2)\*x18)+(I(x2^2)\*x21)+(I(x2^2)\*x25)+(I(x3^2)\*x1)+(I(x3^2)\*x2)+(I(x3^2)\*x8)+(I(x3^2)\*x11)+(I(x3^2)\*x18)+(I(x3^2)\*x21)+(I(x3^2)\*x25)+(I(x1^2)\*x2\*x3\*x8)+(I(x1^2)\*x2\*x3\*x11)+(I(x1^2)\*x2\*x3\*x18)+(I(x1^2)\*x2\*x3\*x19)+(I(x1^2)\*x2\*x3\*x21)+(I(x1^2)\*x2\*x3\*x25)+(I(x2^2)\*x1\*x3\*x8)+(I(x2^2)\*x1\*x3\*x11)+(I(x2^2)\*x1\*x3\*x18)+(I(x2^2)\*x1\*x3\*x21)+(I(x2^2)\*x1\*x3\*x25)+(I(x3^2)\*x1\*x2\*x8)+(I(x3^2)\*x1\*x2\*x11)+(I(x3^2)\*x1\*x2\*x18)+(I(x3^2)\*x1\*x2\*x21)+(I(x3^2)\*x1\*x2\*x25))

anova(model2.5, model2.4)

model2.5 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x2)+(x1\*x3)+(x1\*x8)+(x1\*x11)+(x1\*x18)+(x1\*x21)+(x1\*x25)+(x2\*x3)+(x2\*x8)+(x2\*x11)+(x2\*x18)+(x2\*x21)+(x2\*x25)+(x3\*x8)+(x3\*x11)+(x3\*x18)+(x3\*x21)+(x3\*x25)+(x1\*x2\*x3)+(x1\*x2\*x8)+(x1\*x2\*x11)+(x1\*x2\*x18)+(x1\*x2\*x21)+(x1\*x2\*x25)+(x1\*x3\*x8)+(x1\*x3\*x11)+(x1\*x3\*x18)+(x1\*x3\*x21)+(x1\*x3\*x25)+(x2\*x3\*x8)+(x2\*x3\*x11)+(x2\*x3\*x18)+(x2\*x3\*x21)+(x2\*x3\*x25)+I(x1^2)+I(x2^2)+I(x3^2)+(I(x1^2)\*x2)+(I(x1^2)\*x3)+(I(x1^2)\*x8)+(I(x1^2)\*x11)+(I(x1^2)\*x18)+(I(x1^2)\*x21)+(I(x1^2)\*x25)+(I(x2^2)\*x1)+(I(x2^2)\*x3)+(I(x2^2)\*x8)+(I(x2^2)\*x11)+(I(x2^2)\*x18)+(I(x2^2)\*x21)+(I(x2^2)\*x25)+(I(x3^2)\*x1)+(I(x3^2)\*x2)+(I(x3^2)\*x8)+(I(x3^2)\*x11)+(I(x3^2)\*x18)+(I(x3^2)\*x21)+(I(x3^2)\*x25))

anova(model2.6, model2.5)

model2.6 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x2)+(x1\*x3)+(x1\*x8)+(x1\*x11)+(x1\*x18)+(x1\*x21)+(x1\*x25)+(x2\*x3)+(x2\*x8)+(x2\*x11)+(x2\*x18)+(x2\*x21)+(x2\*x25)+(x3\*x8)+(x3\*x11)+(x3\*x18)+(x3\*x21)+(x3\*x25)+(x1\*x2\*x3)+(x1\*x2\*x8)+(x1\*x2\*x11)+(x1\*x2\*x18)+(x1\*x2\*x21)+(x1\*x2\*x25)+(x1\*x3\*x8)+(x1\*x3\*x11)+(x1\*x3\*x18)+(x1\*x3\*x21)+(x1\*x3\*x25)+(x2\*x3\*x8)+(x2\*x3\*x11)+(x2\*x3\*x18)+(x2\*x3\*x21)+(x2\*x3\*x25)+I(x1^2)+I(x2^2)+I(x3^2))

anova(model2.7, model2.6)

model2.7 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x2)+(x1\*x3)+(x1\*x8)+(x1\*x11)+(x1\*x18)+(x1\*x21)+(x1\*x25)+(x2\*x3)+(x2\*x8)+(x2\*x11)+(x2\*x18)+(x2\*x21)+(x2\*x25)+(x3\*x8)+(x3\*x11)+(x3\*x18)+(x3\*x21)+(x3\*x25)+(x1\*x2\*x3)+(x1\*x2\*x8)+(x1\*x2\*x11)+(x1\*x2\*x18)+(x1\*x2\*x21)+(x1\*x2\*x25)+(x1\*x3\*x8)+(x1\*x3\*x11)+(x1\*x3\*x18)+(x1\*x3\*x21)+(x1\*x3\*x25)+(x2\*x3\*x8)+(x2\*x3\*x18)+(x2\*x3\*x21)+(x2\*x3\*x25))

anova(model2.8, model2.7)

model2.8 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x2)+(x1\*x3)+(x1\*x8)+(x1\*x11)+(x1\*x18)+(x1\*x21)+(x1\*x25)+(x2\*x3)+(x2\*x8)+(x2\*x11)+(x2\*x18)+(x2\*x21)+(x2\*x25)+(x3\*x8)+(x3\*x11)+(x3\*x18)+(x3\*x21)+(x3\*x25))

model2.8 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x3)+(x1\*x11)+(x2\*x3)+(x2\*x11)+(x2\*x18)+(x2\*x25)+(x3\*x8)+(x3\*x11))

anova(model2.9, model2.8)

model2.9 = lm(y~x1+x2+x3+x8+x11+x18+x21+x25)

fit = lm(y~x1+x2+x3+x8+x11+x18+x21+x25+(x1\*x3)+(x1\*x11)+(x2\*x3)+(x2\*x11)+(x2\*x18)+(x2\*x25)+(x3\*x8)+(x3\*x11)+(x1\*x2\*x3)+(x1\*x3\*x8)+(x2\*x3\*x11)+I(x1^2)+I(x2^2)+(I(x1^2)\*x2)+(I(x1^2)\*x3)+(I(x1^2)\*x21)+(I(x1^2)\*x25)+(I(x2^2)\*x1)+(I(x2^2)\*x25)+(I(x3^2)\*x1)+(I(x3^2)\*x25)+(I(x1^2)\*x2\*x3\*x8)+(I(x1^2)\*x2\*x3\*x11)+(I(x1^2)\*x2\*x3\*x19)+(I(x1^2)\*x2\*x3\*x25)+(I(x2^2)\*x1\*x3\*x21)+(I(x3^2)\*x1\*x2\*x21)+(I(x3^2)\*x1\*x2\*x25))

summary(fit)

plot(x1,fit$residuals, ylab="Residuals", xlab="Age (years)", col="blue", pch=20, main="Residuals vs Age")

plot(x2,fit$residuals, ylab="Residuals", xlab="Cylinders", col="blue", pch=20, main="Residuals vs Cylinders")

plot(x3,fit$residuals, ylab="Residuals", xlab="Odometer (miles)", col="blue", pch=20, main="Residuals vs Odometer")

abline(0,0,lty=2, col="6")

ncvTest(fit)

plot(y,res, main="Residuals Versus PRICE", ylab="Residuals", xlab="PRICE", pch=19, col="6")

y.hat = fit$fitted.values

res = fit$residuals

plot(y.hat,res, main="Residuals Versus the Fitted Values", ylab="Residuals", xlab="Fitted value", pch=19, col="6")

shapiro.test(res)

qqnorm(fit$residuals)

qqline(fit$residuals)

anova(fit)

View(fit$residuals)

hatvalues(fit)

View(hatvalues(fit))

View(stdres(fit))

View(rstandard(fit))

rstandard(fit)

View(rstudent(fit))

rstudent(fit)

View(cooks.distance(fit))

cooks.distance(fit)

View(fit)

View(dfbetas(fit))

View(dffits(fit))

dffits(fit)

set.seed(100)

trainingRowIndex=sample(1:nrow(mydata),0.8\*nrow(mydata))

trainingData = mydata[trainingRowIndex, ]

testData = mydata[-trainingRowIndex, ]

pricePredict = predict(fit, mydata)

real\_price <- data.frame(cbind(actuals=mydata$Price, predicteds=pricePredict))

correlation\_accuracy=cor(real\_price)

correlation\_accuracy

head(real\_price)

mydata = read\_excel("C:/Users/Matt/Desktop/UNCC/UNCC Spring 20/STAT 2223-002 Pramesh Subedi - Elements of Statistics II/STAT Final Project/Final Project/CraigslistCarDataV2.xlsx")

y = mydata$Price

#Quantitative Variables

x1 = mydata$Age #in years

x2 = mydata$Cylinders

x3 = mydata$Odometer

x4 = mydata$Min\_Price

x5 = mydata$Max\_Price

x6 = mydata$Rating

#Qualitative Variables

#Title Status

x7 = ifelse(mydata$`Title\_Status` == "rebuilt", 1, 0) #Clean is base

x8 = ifelse(mydata$`Title\_Status` == "salvage", 1, 0) #Clean is base

#Size

x9 = ifelse(mydata$Size == "compact", 1, 0) #Full-size is base

x10 = ifelse(mydata$Size == "mid-size", 1, 0) #Full-size is base

#Color

x11 = ifelse(mydata$Color == "white", 1, 0) #Brown is base

x12 = ifelse(mydata$Color == "silver", 1, 0) #Brown is base

x13 = ifelse(mydata$Color == "black", 1, 0) #Brown is base

x14 = ifelse(mydata$Color == "grey", 1, 0) #Brown is base

x15 = ifelse(mydata$Color == "red", 1, 0) #Brown is base

x16 = ifelse(mydata$Color == "blue", 1, 0) #Brown is base

#Vehicle Type

x17 = ifelse(mydata$Type == "coupe", 1, 0) #Truck is base

x18 = ifelse(mydata$Type == "sedan", 1, 0) #Truck is base

x19 = ifelse(mydata$Type == "SUV", 1, 0) #Truck is base

x20 = ifelse(mydata$Type == "van", 1, 0) #Truck is base

#Vehicle Make

x21 = ifelse(mydata$Make == "Ford", 1, 0) #Nissan is base

x22 = ifelse(mydata$Make == "Chevy", 1, 0) #Nissan is base

x23 = ifelse(mydata$Make == "Honda", 1, 0) #Nissan is base

x24 = ifelse(mydata$Make == "Toyota", 1, 0) #Nissan is base

x25 = ifelse(mydata$Make == "Dodge", 1, 0) #Nissan is base

x26 = ifelse(mydata$Make == "GMC", 1, 0) #Nissan is base

x27 = ifelse(mydata$Make == "Jeep", 1, 0) #Nissan is base

x28 = ifelse(mydata$Make == "BMW", 1, 0) #Nissan is base

x29 = mydata$`Num Doors`

set.seed(100)

trainingRowIndex=sample(1:nrow(mydata),0.8\*nrow(mydata))

trainingData = mydata[trainingRowIndex, ]

testData = mydata[-trainingRowIndex, ]

pricePredict = predict(fit, mydata)

real\_price <- data.frame(cbind(actuals=mydata$Price, predicteds=pricePredict))

correlation\_accuracy=cor(real\_price)

correlation\_accuracy

head(real\_price)