STOR 455 Homework 4

40 points - Due on 10/13 at 5:00pm

Situation: Suppose that (again) you are interested in purchasing a used car. How much should you expect to pay? Obviously the price will depend on the type of car you get (the model) and how much it's been used. For this assignment you will investigate how the price might depend on the age and mileage, as well as the state where the car is purchased.

Data Source: To get a sample of cars, begin with the UsedCars CSV file. The data was acquired by scraping TrueCar.com for used car listings on 9/24/2017 and contains more than 1.2 million used cars. For this assignment you should choose the same car *Model* and *State* that you initially chose for homework #2. You should again add a variable called *Age* which is 2017-year (since the data was scraped in 2017).

Directions: The code below can again be used to select data from a particular *Model* and *State* of your choice. The R chunk below begins with {r, eval=FALSE}. eval=FALSE makes these chunks not run when I knit the file. Before you run this chunk, you should revert it to {r}.

```
setwd <- "C:/Users/adeve/Desktop"
UsedCars <- read.csv("UsedCars.csv")
source("https://raw.githubusercontent.com/JA-
McLean/STOR455/master/scripts/ShowSubsets.R")
library(readr)
library(car)

## Loading required package: carData

library(corrplot)

## corrplot 0.90 loaded

library(leaps)

# Delete the *** below and enter the model and state from homework #2
ModelOfMyChoice = "Accord"
StateOfMyChoice = "TX"

# Takes a subset of your model car from your state
MyCars = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoice)

# Add a new variable for the age of the cars.
MyCars$Age = 2017 - MyCars$Year
```

MODEL #4: Use Age and Miles as predictors for Price

1. Construct a model using two predictors (age and miles) with *Price* as the response variable and provide the summary output.

```
mod1 <- lm(Price~Age+Mileage, MyCars)
summary(mod1)

##
## Call:
## lm(formula = Price ~ Age + Mileage, data = MyCars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5702.4 -1750.4  -409.5  1161.0 15206.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.223e+04  1.077e+02  206.50   <2e-16 ***
## Age         -8.137e+02  3.702e+01  -21.98   <2e-16 ***
## Mileage     -5.226e-02  2.763e-03  -18.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2509 on 1374 degrees of freedom
## Multiple R-squared:  0.8018, Adjusted R-squared:  0.8015
## F-statistic:  2779 on 2 and 1374 DF,  p-value: < 2.2e-16
```

2. Assess the importance of each of the predictors in the regression model - be sure to indicate the specific value(s) from the summary output you are using to make the assessments. Include hypotheses and conclusions in context.

Based on the notes from chapter 3.4 in the textbook, "t-test for individual predictors show that the interaction term is or is not important to the model." If the term has a low p-value, it is most likely special and important to the regression model. Both Age and Mileage have extremely low t-tests, which shows that the probability of getting that exact result by chance would be extremely low.

3. Assess the overall effectiveness of this model (with a formal test). Again, be sure to include hypotheses and the specific value(s) you are using from the summary output to reach a conclusion.

ANOVA tells you if the model explains a signifigant amount of variability or if the variability is by chance. If the model is effective, then the sum of the squared error will be small and the MSModel will be realtively large compared to the MSE. The MSModel is the first row under the Sum Sq column. The MSE is calculated by: MSE = SSE/(n-2). Null Hypothesis: B1 = B2 = 0 (weak model) Alternate Hypothesis: Bi =/= 0 (Effective model)

```
anova(mod1)

## Analysis of Variance Table
##
```

```
## Response: Price
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## Age          1 3.2743e+10 3.2743e+10 5200.48 < 2.2e-16 ***
## Mileage      1 2.2528e+09 2.2528e+09  357.81 < 2.2e-16 ***
## Residuals 1374 8.6510e+09 6.2962e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The MSModel is very large: 3.2743e+10, or 32743326810. The MSE is much smaller compared to the MSModel: 6296213.

```
SSE <- anova(mod1)[3,2]
MSE <- SSE/(1374)
MSE

## [1] 6296213
```

Based on the analysis above, I conclude that this is most likely an effective model.

4. Compute and interpret the variance inflation factor (VIF) for your predictors. *VIF reflects the association between a predictor and all other predictors. We suspect multicollinearity if the VIF is greater than 5, or the R-squared is greater than 80%. Because the VIF of Age and Mileage are both under 5, we do not suspect any multicollinearity between the two predictors.*

```
vif(mod1)

##     Age Mileage
## 3.59231 3.59231
```

5. Suppose that you are interested in purchasing a car of this model that is four years old (in 2017) with 31K miles. Determine each of the following: a 90% confidence interval for the mean price at this age and mileage, and a 90% prediction interval for the price of an individual car at this age and mileage. Write sentences that carefully interpret each of the intervals (in terms of car prices)

We are 90% confident that the average price of any 4 year old texas honda accord with 31K miles is between $17156.56 and $17552.69. We are 90% confident that the average price of one specific 4 year old texas honda accord with 31K miles is between $12428.31 and $22280.94.

Choosing the single car must have a wider upper and lower bound because that is for one single car out of the whole batch, instead of the general sample population of cars.

```
confint(mod1, level = .90)

##                        5 %          95 %
## (Intercept)  2.205263e+04  2.240701e+04
## Age         -8.746787e+02 -7.528031e+02
## Mileage     -5.681339e-02 -4.771757e-02
```

```r
newx=data.frame(Age = 4, Mileage = 31000)
head(newx)
```

```
##   Age Mileage
## 1   4   31000
```

```r
predict.lm(mod1, newx, interval="confidence") # For all 4 year old texas accords
```

```
##        fit      lwr      upr
## 1 17354.63 17156.56 17552.69
```

```r
predict.lm(mod1, newx, interval="prediction") #For one specific car
```

```
##        fit      lwr      upr
## 1 17354.63 12428.31 22280.94
```

MODEL #5: Now Include a Categorical predictor

For this section you will combine both datasets used in Homework #2, as well as a third dataset. Each dataset from Homework #2 included cars from your specific *Model*, but from two different states. You should use the same code that you used in homework #2 to construct this second dataframe with cars from North Carolina, and a third dataframe with cars of your model from a third state of your choice. Then manipulate the code below to combine the three dataframes into one dataframe. Make sure to add the *Age* variable again to your dataframes for the additional states before binding them together. The R chunk below begins with {r, eval=FALSE}. eval=FALSE makes these chunks not run when I knit the file. Before you run this chunk, you should revert it to {r}.

```r
# Delete the *** below and enter the model and state from homework #2
ModelOfMyChoice = "Accord"
StateOfMyChoice = "NC"

# Takes a subset of your model car from your state
MyCars.NC = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoice)

# Add a new variable for the age of the cars.
MyCars.NC$Age = 2017 - MyCars.NC$Year

# Delete the *** below and enter the model and state from homework #2
ModelOfMyChoice = "Accord"
StateOfMyChoice = "KY"

# Takes a subset of your model car from your state
MyCars.KY = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoice)

# Add a new variable for the age of the cars.
MyCars.KY$Age = 2017 - MyCars.KY$Year

State1 = MyCars
State2 = MyCars.NC #fill in with the name of your dataframe
```

```
State3 = MyCars.KY #fill in with the name of your dataframe

# rbind combines the rows in one dataframe, assuming that the columns are the
same.
CombinedStates = rbind(State1, State2, State3)
names(CombinedStates)

##  [1] "Id"      "Price"   "Year"    "Mileage" "City"    "State"   "Vin"
##  [8] "Make"    "Model"   "Age"
```

6. Fit a multiple regression model using *Age*, *Mileage*, and *State* to predict the *Price* of the car. *Need to create a variable for if you're a TX Car 0 or 1, or an NC car 0 or 1 KY's Cars are accounted by solving the linear model for 0 TX and 0 NC Mod2: Predicted Price = 2.147e+04 - 8.317e+02AGE - 4.804e-02MILEAGE + 7.626e+02NC + 5.864e+02TX*

```
CombinedStates$TX=(CombinedStates$State==2)*1
CombinedStates$NC=(CombinedStates$State==1)*1

mod2 = lm(Price~Age+Mileage+State, CombinedStates)
summary(mod2)

##
## Call:
## lm(formula = Price ~ Age + Mileage + State, data = CombinedStates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5635.1 -1756.6  -368.3  1233.2 15160.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.147e+04  2.016e+02 106.513  < 2e-16 ***
## Age         -8.317e+02  2.702e+01 -30.781  < 2e-16 ***
## Mileage     -4.804e-02  1.988e-03 -24.159  < 2e-16 ***
## StateNC      7.626e+02  2.129e+02   3.582 0.000348 ***
## StateTX      5.864e+02  2.046e+02   2.866 0.004198 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2468 on 2311 degrees of freedom
## Multiple R-squared:  0.8004, Adjusted R-squared:    0.8
## F-statistic:  2317 on 4 and 2311 DF,  p-value: < 2.2e-16
```

7. Perform a hypothesis test to determine the importance of *State* terms in the model constructed in question 6. List your hypotheses, p-value, and conclusion.

To find if state has a statistically signifigant effect on car prices, we look at the individual t-tests for each variable. We want that eatch have a low p-value because it tells us that your variable is not expected by random chance.

If we had a high p-value, then we may be able to get the variable results from a lot of differnet places, so it wouldn't be a special value. Since we see from the results above that all p-values are less than 0.05, cna can assue that the difference between states has a staistically signifigcant effect on our car prices.

The previous test assumes that we have a common slope. Is is reasonable to assume that we have a common slope?

Ho: B3 = 0 Ha: B3 =/= 0

Conclusion: We have evidence to reject the null hypothesis that B3 (State) = 0 because the p-value remains under 0.05.

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: Price
##             Df     Sum Sq    Mean Sq   F value    Pr(>F)
## Age          1 5.2849e+10 5.2849e+10 8675.1212 < 2.2e-16 ***
## Mileage      1 3.5258e+09 3.5258e+09  578.7555 < 2.2e-16 ***
## State        2 7.9240e+07 3.9620e+07    6.5036  0.001526 **
## Residuals 2311 1.4079e+10 6.0920e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8. Fit a multiple regression model using *Age*, *Mileage*, *State*, and the interactions between *Age* and *State*, and *Mileage* and *State* to predict the *Price* of the car.

```
mod3=lm(Price~Age+Mileage+factor(State)+Age*factor(State)+Mileage*factor(Stat
e),data=CombinedStates)
summary(mod3)
```

```
##
## Call:
## lm(formula = Price ~ Age + Mileage + factor(State) + Age * factor(State) +
##     Mileage * factor(State), data = CombinedStates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5702.4 -1753.4  -351.4  1227.6 15206.1
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.111e+04  3.149e+02  67.040  < 2e-16 ***
## Age                       -9.851e+02  1.188e+02  -8.290  < 2e-16 ***
## Mileage                   -2.976e-02  7.018e-03  -4.240 2.32e-05 ***
## factor(State)NC            9.482e+02  3.453e+02   2.746 0.006084 **
## factor(State)TX            1.116e+03  3.322e+02   3.359 0.000795 ***
## Age:factor(State)NC        1.701e+02  1.264e+02   1.345 0.178599
## Age:factor(State)TX        1.714e+02  1.242e+02   1.379 0.167996
## Mileage:factor(State)NC -1.653e-02  7.723e-03  -2.140 0.032469 *
```

```
## Mileage:factor(State)TX -2.251e-02  7.523e-03  -2.992 0.002803 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2462 on 2307 degrees of freedom
## Multiple R-squared:  0.8018, Adjusted R-squared:  0.8011
## F-statistic:  1167 on 8 and 2307 DF,  p-value: < 2.2e-16
```

```
anova(lm(Price~Age+Mileage, data=CombinedStates), mod3)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Age + Mileage
## Model 2: Price ~ Age + Mileage + factor(State) + Age * factor(State) +
##     Mileage * factor(State)
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1   2313 1.4158e+10
## 2   2307 1.3979e+10  6 178551630 4.9111 5.326e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9. Perform a hypothesis test to determine the importance of *State* terms in the model constructed in question 8. List your hypotheses, p-value, and conclusion. *To find if state has a statistically significant effect on car prices, we look at the individual t-tests for each variable. We want that eatch have a low p-value because it tells us that your variable is not expected by random chance.*

If we had a high p-value, then we may be able to get the variable results from a lot of differnet places, so it wouldn't be a special value. Since we see from the results above that all p-values are not less than 0.05, cna can assue that the difference between Age times factor(States) does not have a staistically signifigcant effect on our car prices.

The previous test assumes that we have a common slope. Is is reasonable to assume that we have a common slope?

Ho: B4 = 0 Ha: B4 =/= 0

Conclusion: We have evidence to reject the null hypothesis that B3 (State) = 0 because the p-value remains under 0.05.

```
anova(mod2,mod3)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Age + Mileage + State
## Model 2: Price ~ Age + Mileage + factor(State) + Age * factor(State) +
##     Mileage * factor(State)
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1   2311 1.4079e+10
## 2   2307 1.3979e+10  4  99311635 4.0974 0.002597 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

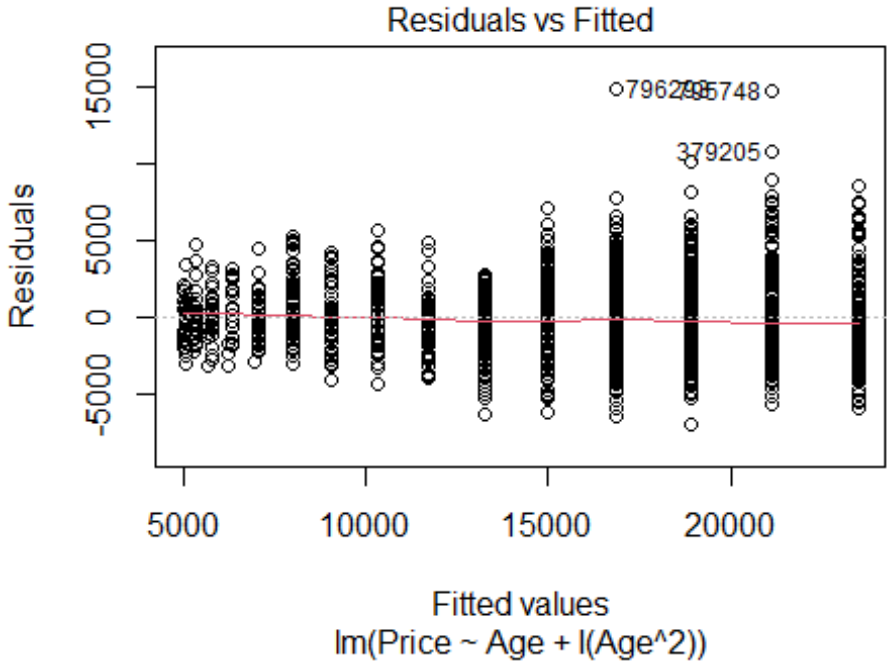
MODEL #6: Polynomial models

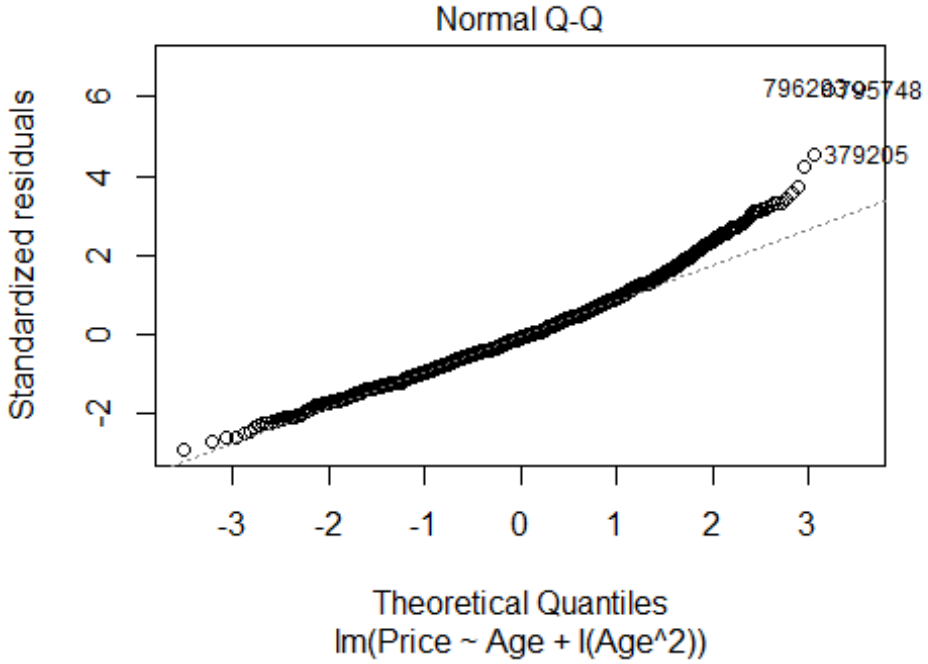
One of the drawbacks of the linear model in homework #2 was the "free car" phenomenon where the predicted price is eventually negative as the line decreases for older cars. Let's see if adding one or more polynomial terms might help with this. For this section you should use the dataset with cars from three states that you used for model 5.

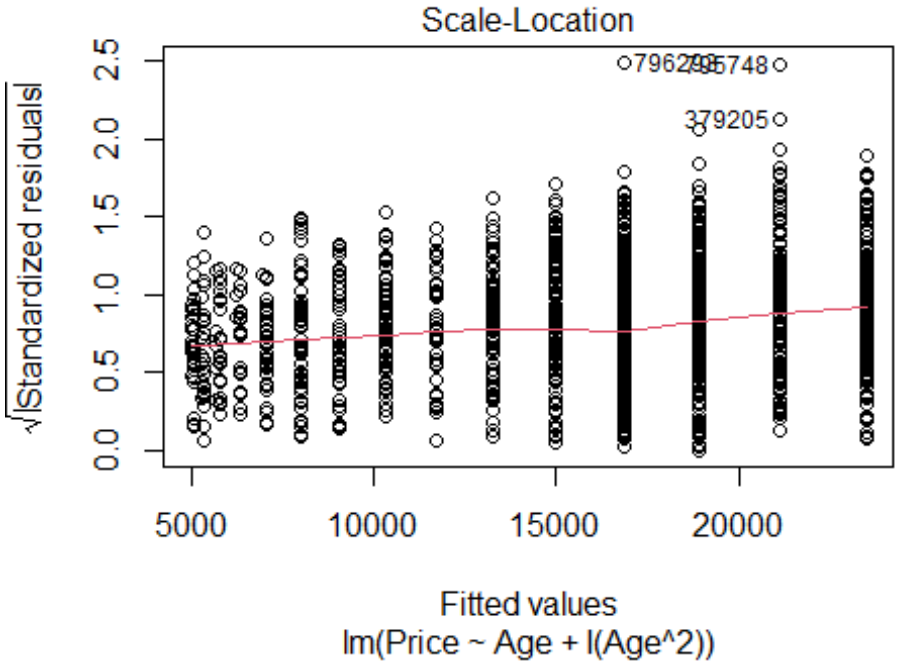
10. Fit a quadratic model using *Age* to predict *Price* and examine the residuals. Construct a scatterplot of the data with the quadratic fit included. You do not need to specifically cite all conditions for the linear model, but should discuss any issues that you see in the conditions.

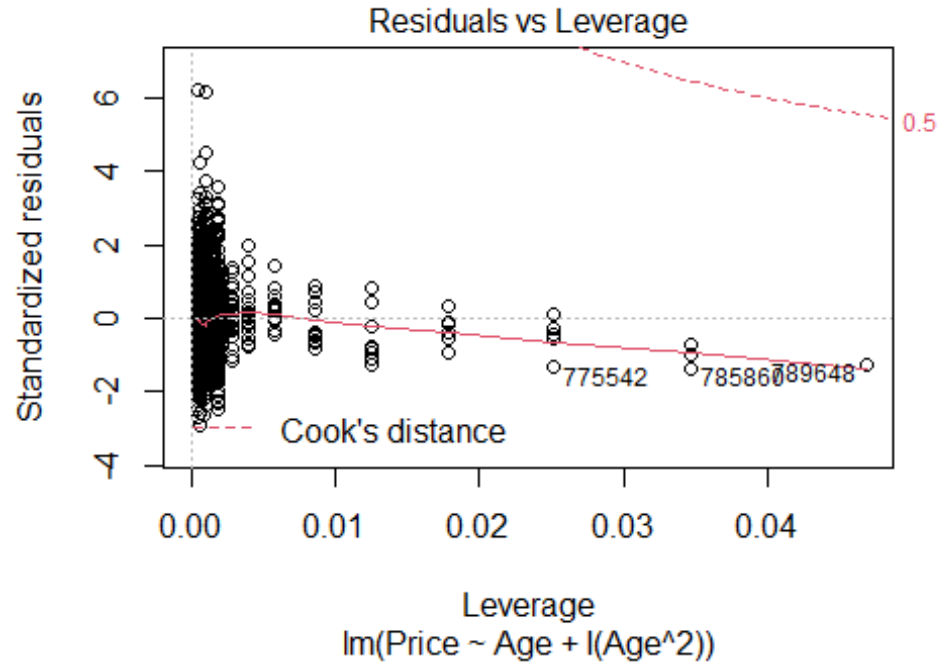
The biggest issue I see with the conditions are the residuals appear to be skewed slightly to the right. Other than that, all other conditions appear to be met.

```
# polymod <- lm(Price~poly(Age, degree = 2, raw=TRUE), CombinedStates); THis
is the same thing as below
polymod <- lm(Price~Age+I(Age^2), CombinedStates)
summary(polymod)

##
## Call:
## lm(formula = Price ~ Age + I(Age^2), data = CombinedStates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6917.0 -1532.5  -207.7  1349.2 14791.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23513.170    104.002  226.08   <2e-16 ***
## Age         -2461.304     40.948  -60.11   <2e-16 ***
## I(Age^2)       81.618      2.888   28.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2384 on 2313 degrees of freedom
## Multiple R-squared:  0.8136, Adjusted R-squared:  0.8135
## F-statistic:  5049 on 2 and 2313 DF,  p-value: < 2.2e-16

plot(polymod)
```

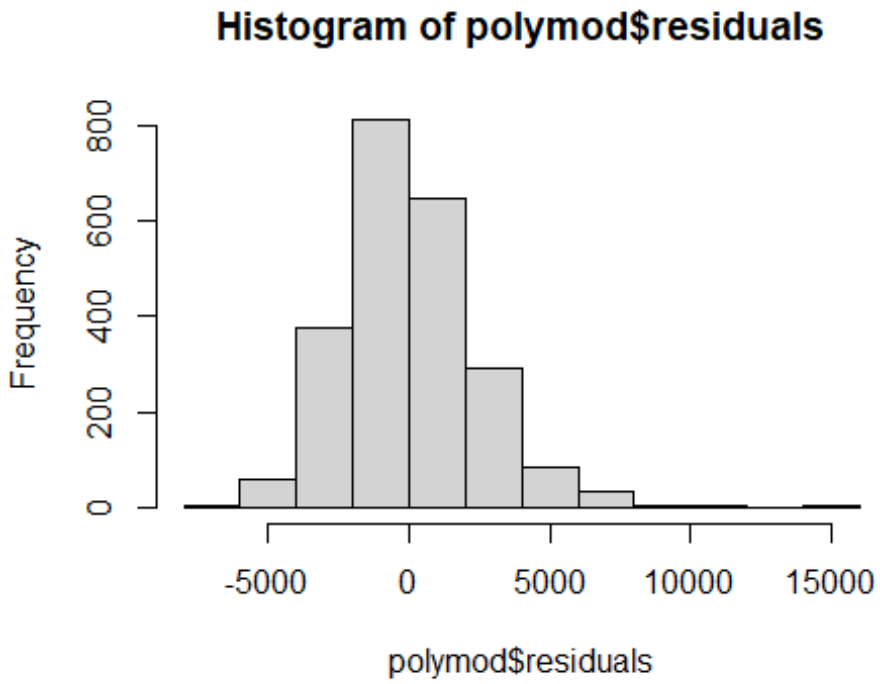

Residuals vs Fitted
Residuals
15000
5000
0
-5000
379205
5000
10000
15000
20000
Fitted values
lm(Price ~ Age + I(Age^2))


Normal Q-Q
Standardized residuals
6
4
2
0
-2
379205
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Price ~ Age + I(Age^2))






```r
hist(polymod$residuals)
```



11. Perform a hypothesis test to determine if this model is significant. List your hypotheses, p-value, and conclusion. *To find if age has a statistically significant effect on car prices, we look at the individual t-tests for each variable. We want that each have a low p-value because it tells us that your variable is not expected by random chance.*

If we had a high p-value, then we may be able to get the variable results from a lot of different places, so it wouldn't be a special value. Since we see from the results above that all p-values are less than 0.05, can assume that Age does have a statistically significant effect on our car prices.

Ho: B2 = 0 Ha: B2 =/= 0

Conclusion: We have evidence to reject the null hypothesis that B2 (Age, degree = 2) = 0 because the p-value remains under 0.05.

```
anova(polymod)
```

```
## Analysis of Variance Table
##
## Response: Price
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## Age          1 5.2849e+10 5.2849e+10 9299.65 < 2.2e-16 ***
## I(Age^2)     1 4.5391e+09 4.5391e+09  798.74 < 2.2e-16 ***
## Residuals 2313 1.3144e+10 5.6829e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12. You are looking at a 4-year-old car of your model and want to find an interval that is likely to contain its *Price* using your quadratic model. Construct an interval to predict the value of this car, and include an interpretive sentence in context.

We are 90% confident that the average price of any 4 year old honda accord in TX, NC, or KY is between $14854.42 to $15093.24. We are 90% confident that the average price of one specific 4 year old TX, NC, or KY honda accord is between $10297.55 to $19650.12.

Choosing the single car must have a wider upper and lower bound because that is for one single car out of the whole batch, instead of the general sample population of cars.

```
confint(polymod, level = .90)
```

```
##                      5 %        95 %
## (Intercept) 23342.03280 23684.30660
## Age         -2528.68383 -2393.92422
## I(Age^2)       76.86543    86.36958
```

```
newx=data.frame(Age = 4)
head(newx)
```

```
##   Age
## 1   4
```

```
predict.lm(polymod, newx, interval="confidence") # For all 4 year old TX, NC, and KY accords
```

```
##        fit      lwr      upr
## 1 14973.83 14854.42 15093.24
```

```
predict.lm(polymod, newx, interval="prediction") # For one specific car
```

```
##        fit      lwr      upr
## 1 14973.83 10297.55 19650.12
```

13. Does the quadratic model allow for some *Age* where a car has a zero or negative predicted price? Justify your answer using a calculation or graph.

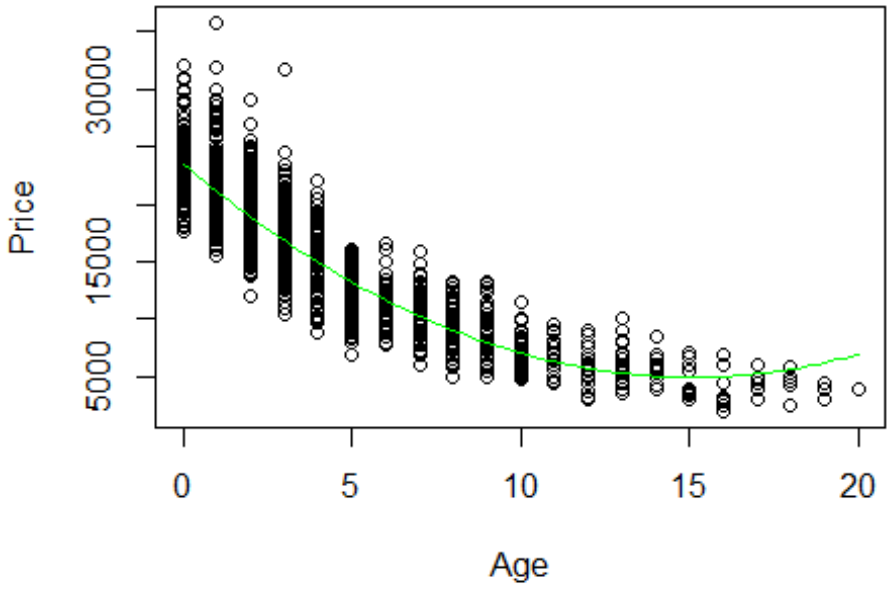
As shown on the graph, the quadratic model does not allow for some age where the car has a zero or negative predicted price. The minimum predicted price, based on the graph, appears to be about $5,000. The graph then begins to curve upwards before it reaches 0.

```
summary(polymod)
```

```
##
## Call:
## lm(formula = Price ~ Age + I(Age^2), data = CombinedStates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6917.0 -1532.5  -207.7  1349.2 14791.2
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23513.170    104.002  226.08   <2e-16 ***
## Age         -2461.304     40.948  -60.11   <2e-16 ***
## I(Age^2)       81.618      2.888   28.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2384 on 2313 degrees of freedom
## Multiple R-squared:  0.8136, Adjusted R-squared:  0.8135
## F-statistic:  5049 on 2 and 2313 DF,  p-value: < 2.2e-16
```

```
B0.1 = summary(polymod)$coefficients[1,1] # Intercept
B1.1 = summary(polymod)$coefficients[2,1] # Slope Age
B2.1 = summary(polymod)$coefficients[3,1] # Slope Age^2

plot(Price~Age, CombinedStates)
curve(B1.1*x+B2.1*I(x^2)+B0.1, col = "green", add=TRUE)
```



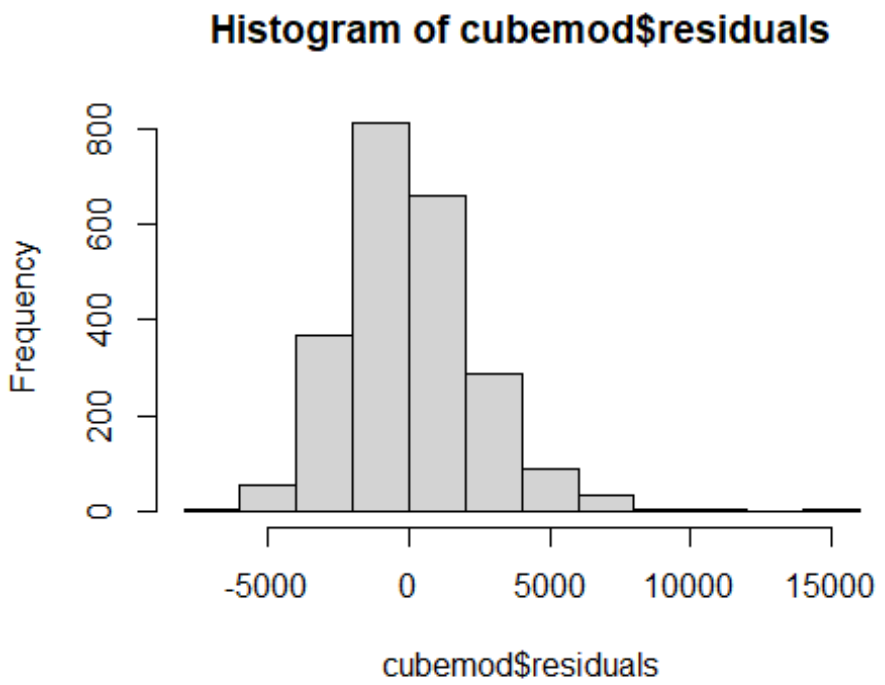
14. Would the fit improve significantly if you also included a cubic term? Does expanding your polynomial model to use a quartic term make significant improvements? Justify your answer.

Cubemod appears to fit the graph much better than the previous models. There is a slight skew in the residuals, but other than that the other graphs associated look good. I will do a nested f-test further down to further test if cubemod is truely better than quarticmod

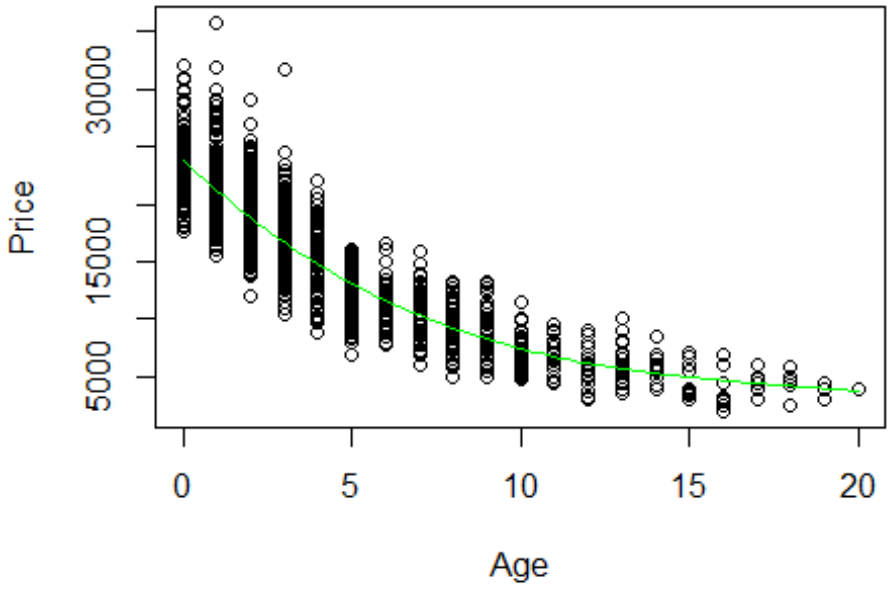
```r
cubemod <- lm(Price~Age+I(Age^2)+I(Age^3), CombinedStates)
summary(cubemod)
```

```
##
## Call:
## lm(formula = Price ~ Age + I(Age^2) + I(Age^3), data = CombinedStates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6840.3 -1561.4  -229.1  1295.4 14952.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23880.2147   128.1698 186.317  < 2e-16 ***
## Age         -2790.0804    79.0340 -35.302  < 2e-16 ***
## I(Age^2)      140.1596    12.3958  11.307  < 2e-16 ***
## I(Age^3)       -2.5502     0.5253  -4.855 1.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2372 on 2312 degrees of freedom
## Multiple R-squared:  0.8155, Adjusted R-squared:  0.8153
## F-statistic:  3407 on 3 and 2312 DF,  p-value: < 2.2e-16
```

```r
# plot(cubemod), I commented this out for the sake of the knitted file
hist(cubemod$residuals)
```



```
B0.1.cube = summary(cubemod)$coefficients[1,1] # Intercept
B1.1.cube = summary(cubemod)$coefficients[2,1] # Slope Age
B2.1.cube = summary(cubemod)$coefficients[3,1] # Slope Age^2
B3.1.cube = summary(cubemod)$coefficients[4,1] # Slope Age^3

plot(Price~Age, CombinedStates)
curve(B1.1.cube*x+B2.1.cube*I(x^2)+B3.1.cube*I(x^3)+B0.1.cube, col = "green",
add=TRUE)
```

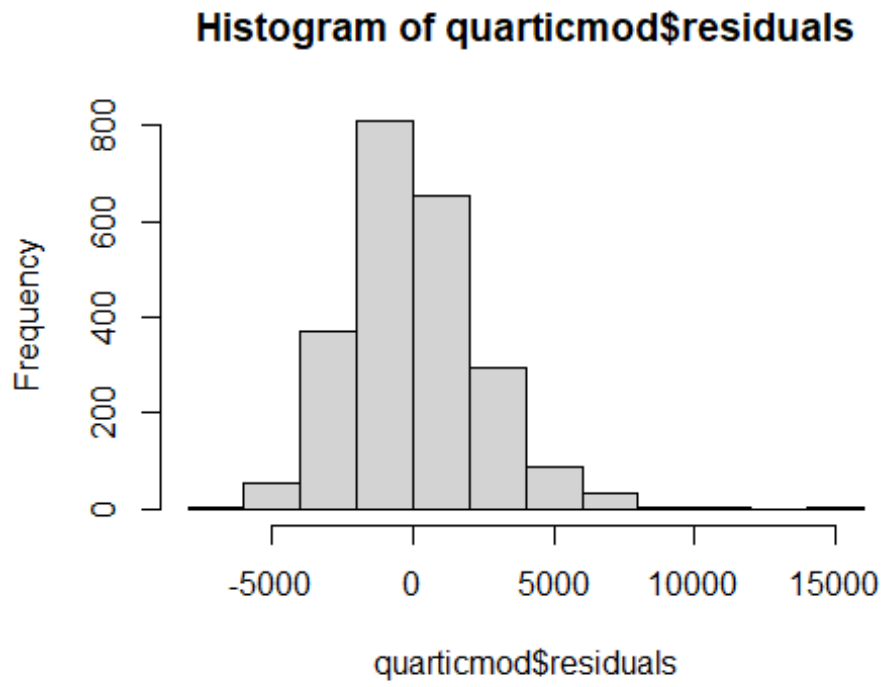


Based on the summary output of quarticmod, adding the quarticmod makes Age^3 and Age^4 no longer statistically significant in the model. This decreases the effectiveness of the model, and makes quarticmod a less desirable model to use

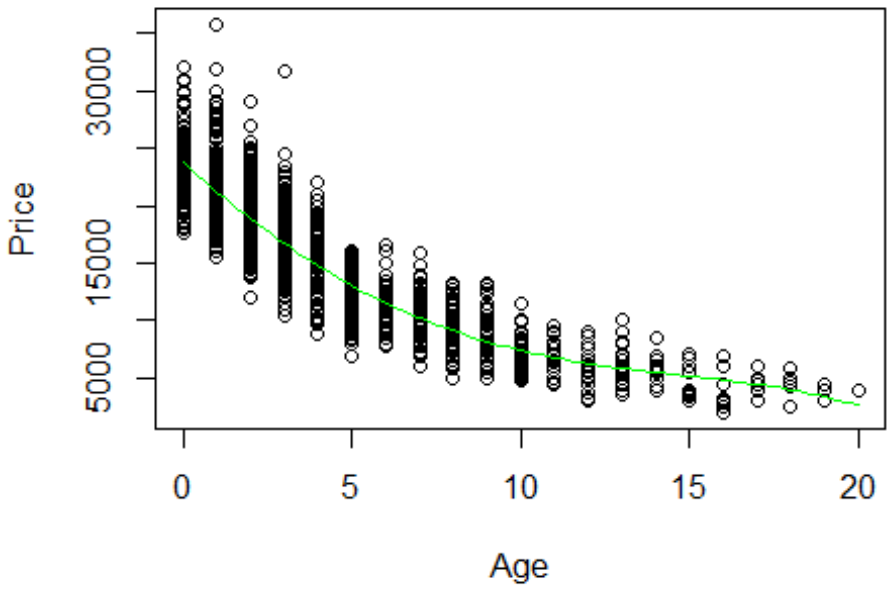
```
quarticmod <- lm(Price~Age+I(Age^2)+I(Age^3)+I(Age^4), CombinedStates)
summary(quarticmod)

##
## Call:
## lm(formula = Price ~ Age + I(Age^2) + I(Age^3) + I(Age^4), data =
CombinedStates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6880.1 -1572.4  -238.1  1272.6 14921.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 23796.6327   142.6004 166.876  < 2e-16 ***
## Age         -2659.2814   125.8052 -21.138  < 2e-16 ***
## I(Age^2)       97.3734    34.3364   2.836  0.00461 **
## I(Age^3)        1.8344     3.3233   0.552  0.58101
## I(Age^4)       -0.1352     0.1012  -1.336  0.18163
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2372 on 2311 degrees of freedom
## Multiple R-squared:  0.8157, Adjusted R-squared:  0.8153
## F-statistic:  2556 on 4 and 2311 DF,  p-value: < 2.2e-16
```

```
# plot(quarticmod) I commented this out for the sake of the knitted file.
hist(quarticmod$residuals)
```




```
B0.1.quartic = summary(quarticmod)$coefficients[1,1] # Intercept
B1.1.quartic = summary(quarticmod)$coefficients[2,1] # Slope Age
B2.1.quartic = summary(quarticmod)$coefficients[3,1] # Slope Age^2
B3.1.quartic = summary(quarticmod)$coefficients[4,1] # Slope Age^3
B4.1.quartic = summary(quarticmod)$coefficients[5,1] # Slope Age^4

plot(Price~Age, CombinedStates)
curve(B1.1.quartic*x+B2.1.quartic*I(x^2)+B3.1.quartic*I(x^3)+B4.1.quartic*I(x
^4)+B0.1.quartic, col = "green", add=TRUE)
```




```
anova(cubemod, quarticmod)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Age + I(Age^2) + I(Age^3)
## Model 2: Price ~ Age + I(Age^2) + I(Age^3) + I(Age^4)
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1   2312 1.3012e+10
## 2   2311 1.3002e+10  1  10044385 1.7853 0.1816
```

Based on the nested f test above, the quarticmod is not better for the data than the cubic mod.

MODEL #7: Complete second order model

For this section you should again use the dataset with cars from three states that you used for models 5 and 6.

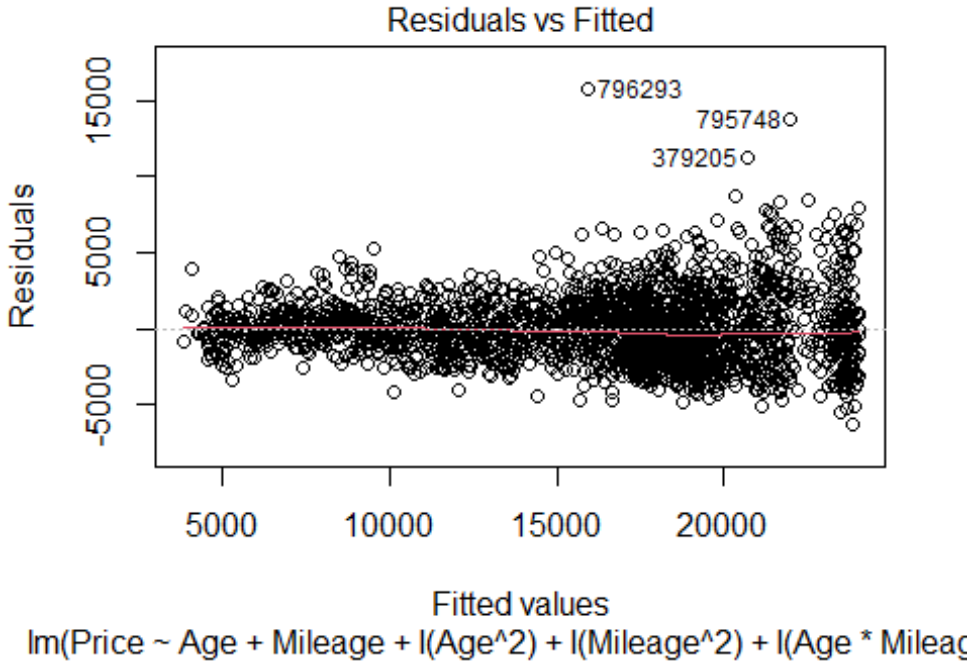
15. Fit a complete second order model for predicting a used car *Price* based on *Age* and *Mileage* and examine the residuals. You do not need to specifically cite all conditions for the linear model, but should discuss any issues that you see in the conditions.

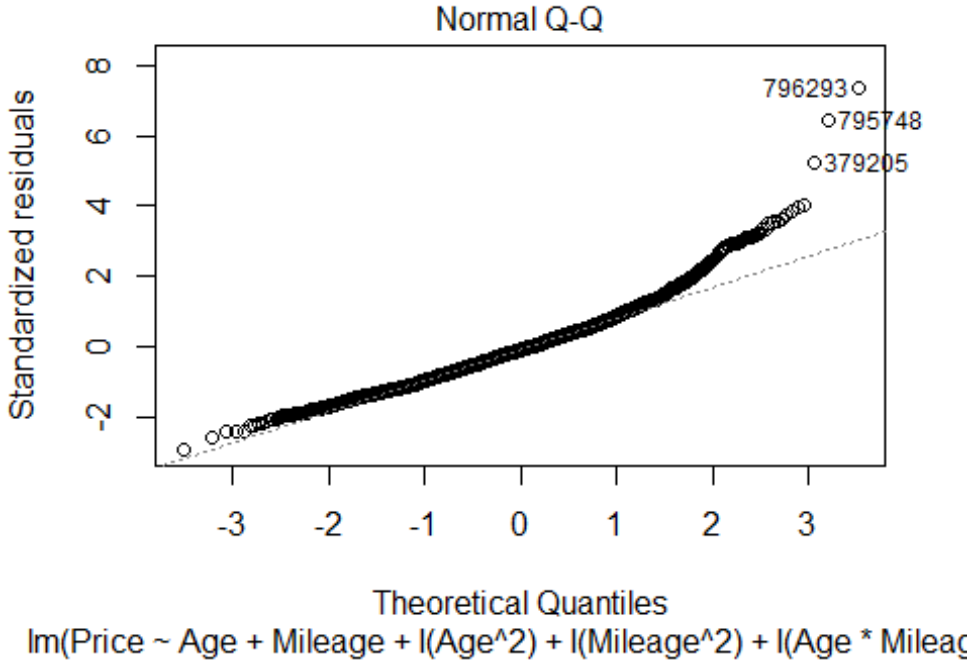
The conditions for linear appear to be met. The residuals appear to be skewed to the right. Additionally, the cook's distance plot shows that there are some points that have very high leverage, but there does not appear to be any points outside of the 0.5 and 1 cook's distance plot.

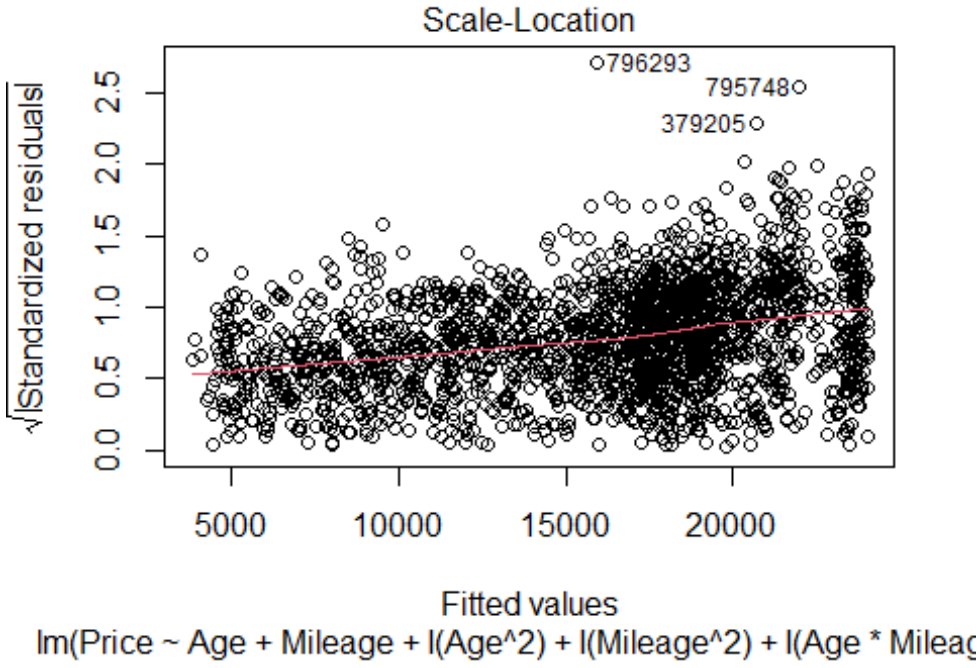
```r
secondordermod <- lm(Price~Age+Mileage+I(Age^2)+I(Mileage^2)+I(Age*Mileage),
CombinedStates)
summary(secondordermod)
```

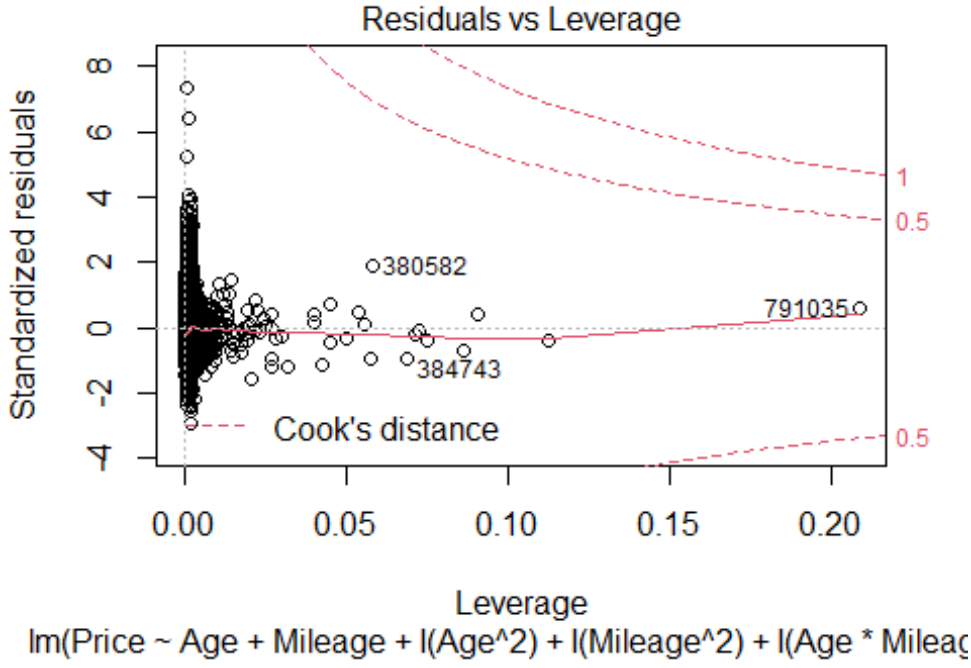
```
##
## Call:
## lm(formula = Price ~ Age + Mileage + I(Age^2) + I(Mileage^2) +
##     I(Age * Mileage), data = CombinedStates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6237.9 -1444.5  -157.5  1113.4 15747.4
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.401e+04  1.016e+02 236.407  < 2e-16 ***
## Age              -1.603e+03  5.599e+01 -28.625  < 2e-16 ***
## Mileage          -6.801e-02  3.944e-03 -17.243  < 2e-16 ***
## I(Age^2)          3.646e+01  5.595e+00   6.517 8.79e-11 ***
## I(Mileage^2)      5.166e-08  2.874e-08   1.798  0.07238 .
## I(Age * Mileage)  2.647e-03  7.082e-04   3.738  0.00019 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2149 on 2310 degrees of freedom
## Multiple R-squared:  0.8488, Adjusted R-squared:  0.8485
## F-statistic:  2594 on 5 and 2310 DF,  p-value: < 2.2e-16
```

```r
plot(secondordermod)
```

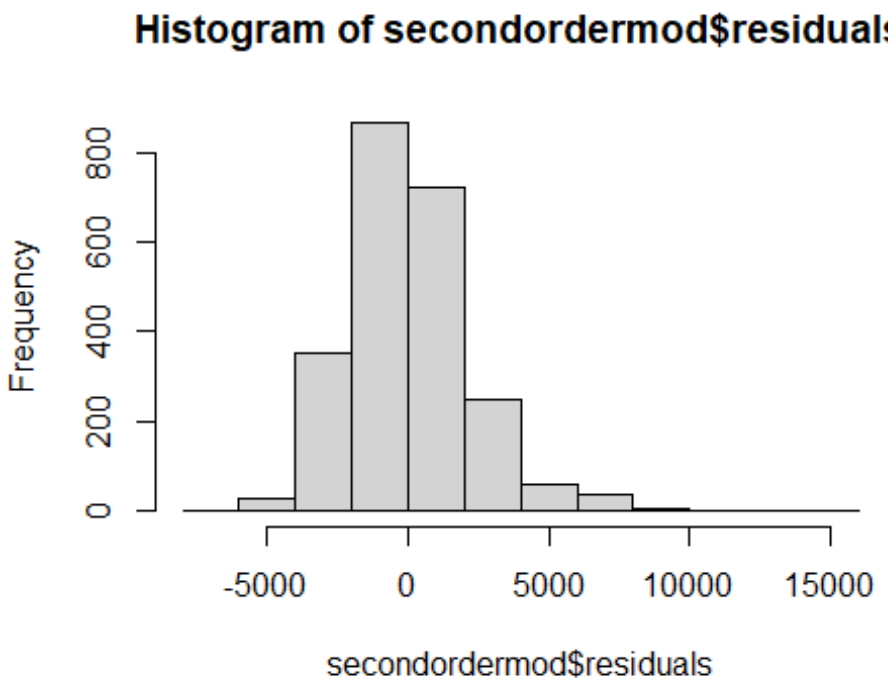

Residuals vs Fitted
796293
795748
379205
Residuals
15000
5000
-5000
5000
10000
15000
20000
Fitted values
lm(Price ~ Age + Mileage + I(Age^2) + I(Mileage^2) + I(Age * Mileag


Normal Q-Q
796293
795748
379205
Standardized residuals
8
6
4
2
0
-2
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Price ~ Age + Mileage + I(Age^2) + I(Mileage^2) + I(Age * Mileag






```r
hist(secondordermod$residuals)
```



16. Perform a hypothesis test to determine if the model constructed in question 15 is significant. List your hypotheses, p-value, and conclusion. *To find if the model in question is significant, we look at the individual t-tests for each variable. We want that each have a low p-value because it tells us that your variable is not expected by random chance.*

If we had a high p-value, then we may be able to get the variable results from a lot of different places, so it wouldn't be a special value. Since we see from the results above in the summary and below in the anova that all p-values are less than 0.05, can assume that the model is significant.

Ho: Bi = 0 Ha: Bi =/= 0

Conclusion: We have evidence to reject the null hypothesis that Bi = 0 because the p-value remains under 0.05.

```
anova(secondordermod)
```

```
## Analysis of Variance Table
##
## Response: Price
##                    Df     Sum Sq    Mean Sq   F value  Pr(>F)
## Age                 1 5.2849e+10 5.2849e+10 11448.858 < 2e-16 ***
## Mileage             1 3.5258e+09 3.5258e+09   763.804 < 2e-16 ***
## I(Age^2)            1 3.1005e+09 3.1005e+09   671.685 < 2e-16 ***
## I(Mileage^2)        1 3.2966e+08 3.2966e+08    71.415 < 2e-16 ***
## I(Age * Mileage)    1 6.4500e+07 6.4500e+07    13.973 0.00019 ***
```

```
## Residuals        2310 1.0663e+10 4.6161e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA tells you if the model explains a signifigant amount of variability or if the variability is by chance. If the model is effective, then the sum of the squared error will be small and the MSModel will be realtively large compared to the MSE. The MSModel is the first row under the Sum Sq column. The MSE is calculated by: MSE = SSE/(n-2). Null Hypothesis: B1 = B2 = 0 (weak model) Alternate Hypothesis: Bi =/= 0 (Effective model)

The MSModel is very large: 3100543131. The MSE is much smaller compared to the MSModel: 2256582.

```
SSE.second <- anova(secondordermod)[3,2]
SSE.second
```

```
## [1] 3100543131
```

```
MSE.second <- SSE.second/(1374)
MSE.second
```

```
## [1] 2256582
```

Based on the analysis above, I conclude that this is most likely an effective model.

17. Perform a hypothesis test to determine the importance of just the *second order terms* (quadratic and interaction) in the model constructed in question 15. List your hypotheses, p-value, and conclusion.

Ho: Bi = 0 Ha: Bi =/= 0

Conclusion: We do not have evidence to reject the null hypothesis that Bi = 0 because the p-value of I(AgexMileage) is greater than 0.05.

```
secondordermod.second.terms <- lm(Price~I(Age^2)+I(Mileage^2)+I(Age*Mileage),
CombinedStates)
anova(secondordermod.second.terms)
```

```
## Analysis of Variance Table
##
## Response: Price
##                    Df     Sum Sq    Mean Sq   F value Pr(>F)
## I(Age^2)            1 3.6855e+10 3.6855e+10 2985.1919 <2e-16 ***
## I(Mileage^2)        1 5.1328e+09 5.1328e+09  415.7444 <2e-16 ***
## I(Age * Mileage)    1 5.5266e+04 5.5266e+04    0.0045 0.9467
## Residuals        2312 2.8544e+10 1.2346e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

18. Perform a hypothesis test to determine the importance of just the terms that involve *Mileage* in the model constructed in question 15. List your hypotheses, p-value, and conclusion.

Ho: Bi = 0 Ha: Bi =/= 0

Conclusion: We have evidence to reject the null hypothesis that Bi = 0 because the p-value remains under 0.05.

```
secondordermod.mile <- lm(Price~Mileage+I(Mileage^2)+I(Age*Mileage),
CombinedStates)
anova(secondordermod.mile)
```

```
## Analysis of Variance Table
##
## Response: Price
##                    Df     Sum Sq    Mean Sq F value    Pr(>F)
## Mileage             1 5.0584e+10 5.0584e+10 7858.11 < 2.2e-16 ***
## I(Mileage^2)        1 3.1605e+09 3.1605e+09  490.98 < 2.2e-16 ***
## I(Age * Mileage)    1 1.9051e+09 1.9051e+09  295.95 < 2.2e-16 ***
## Residuals        2312 1.4883e+10 6.4372e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```