STOR 455 Homework #4

40 points - Due on 3/4 at 5:00pm

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

Data Source: To get a sample of vehicles, begin with the UsedCars CSV file. The data was acquired by scraping TrueCar.com for used vehicle listings on 9/24/2017 and contains more than 1.2 million used vehicles. For this assignment you should choose the same vehicle *Model* from North Carolina that you initially chose for homework #2.

Directions: The code below can again be used to select data from a particular *Model* of your choice from North Carolina. 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}.

```
library(readr)
UsedCars <- read_csv("UsedCars.csv")

## Rows: 1048575 Columns: 9

## -- Column specification ----------------------------------------
------------
## Delimiter: ","
## chr (5): City, State, Vin, Make, Model
## dbl (4): Id, Price, Year, Mileage

##
## i Use `spec()` to retrieve the full column specification for this da
ta.
## i Specify the column types or set `show_col_types = FALSE` to quiet
this message.

# Delete the *** below and enter the model from homework #2
ModelOfMyChoice = "EdgeSEL"
StateOfMyChoice = "NC"

# Takes a subset of your model vehicle from your state
MyVehicles = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMy
Choice)
```

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

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

```
mod4=lm(Price~Year+Mileage,data=MyVehicles)
summary(mod4)

##
## Call:
## lm(formula = Price ~ Year + Mileage, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7023.5 -1297.3  -231.1  1516.4  4944.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.353e+06  2.481e+05  -9.482   <2e-16 ***
## Year         1.181e+03  1.231e+02   9.592   <2e-16 ***
## Mileage     -8.971e-02  6.884e-03 -13.032   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2178 on 202 degrees of freedom
## Multiple R-squared:  0.8213, Adjusted R-squared:  0.8195
## F-statistic: 464.2 on 2 and 202 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.
   (1)For the predictor 'Year'

   H0: β1=0, Ha:β1≠0

   The null hypothesis is that the slope is 0, a horizontal line, a constant. The predictor year and the response price have no relationship.

   Since its P value is pretty small(<2e-16), we have evidence to show that price and year have linearity relationship.It is unlikely this would happen by chance if there is no relationship with the population.

   (2)For the predictor 'Mileage'

   H0: β2=0, Ha:β2≠0

   The null hypothesis is that the slope is 0, a horizontal line, a constant. The predictor Mileage and the response price have no relationship.

   Since its P value is pretty small(<2e-16), we have evidence to show that price and Mileage have linearity relationship.It is unlikely this would happen by chance if there is no relationship with the population.

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.

   H0: $\beta 1=\beta 2=0$, Ha:some $\beta i \neq 0$

   The null hypothesis is that all the slope are 0. All predictors (year and Mileage) have no relationship with the response price. The alternative hypothesis is that at least one slope is not zero.

   Since the overall P value is pretty small(<2e-16), we have evidence to show that price and predictors have linearity relationship. It is unlikely this would happen by chance if there is no relationship with the population.

```
source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scripts/anova455.R")
anova455(mod4)

## ANOVA Table
## Model: Price ~ Year + Mileage
##
##        Df     Sum Sq    Mean Sq F value     P(>F)
## Model   2 4402530766 2201265383  464.19 < 2.2e-16 ***
## Error 202  957922770    4742192
## Total 204 5360453537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Compute and interpret the variance inflation factor (VIF) for your predictors. Since both of the VIF of predictors are less than 5, there is no multicollinearity problem.

```
library(car)

## Loading required package: carData

vif(mod4)

##     Year  Mileage
## 2.097296 2.097296
```

5. Suppose that you are interested in purchasing a vehicle of this model that was four years old (in 2017) with 58K miles. Determine each of the following: a 90% confidence interval for the mean price at this Year and mileage, and a 90% prediction interval for the price of an individual vehicle at this Year and mileage. Write sentences that carefully interpret each of the intervals (in terms of vehicle prices)

(1) With 90% confidence I predict that the price of a four years old EdgeSEL with 58K miles sold in NC is between $14596.97 and $22249.95

(2) With 90% confidence I predict that the mean price of a four years old EdgeSEL with 58K miles sold in NC is between $18264.54 and $18582.38

```
newx=data.frame(Year=2013, Mileage=58000)
predict.lm(mod4, newx, interval="prediction",level=0.9)

##        fit      lwr      upr
## 1 18748.94 15140.26 22357.62

predict.lm(mod4, newx, interval="confidence",level=0.9)

##        fit      lwr      upr
## 1 18748.94 18477.15 19020.73
```

MODEL #5: Now Include a Categorical predictor

For this section you will combine both datasets used in Homework #2, as well as two new datasets. Each dataset from Homework #2 included vehicles 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 vehicles from the state of your choice, and a third and fourth dataframe with vehicles of your model from a third and fourth state (Choose either Arizona, Florida, or Ohio for the two additional states). Then manipulate the code below to combine the four dataframes into one dataframe. 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}.

```
State1 = MyVehicles
#fill in with the dataframe of cars of your model from state 2
StateOfMyChoice2 = "PA"
State2 = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoi
ce2)
#fill in with the dataframe of cars of your model from state 3
StateOfMyChoice3 = "AZ"
State3 = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoi
ce3)
#fill in with the dataframe of cars of your model from state 4
StateOfMyChoice4 = "FL"
State4 = subset(UsedCars, Model==ModelOfMyChoice & State==StateOfMyChoi
ce4)
# rbind combines the rows in one dataframe, assuming that the columns a
re the same.
CombinedStates = rbind(State1, State2, State3, State4)
```

6. Fit a multiple regression model using *Year*, *Mileage*, and *State* to predict the *Price* of the vehicle.

```
mod5=lm(Price~Year+Mileage+State,data=CombinedStates)
summary(mod5)

##
## Call:
## lm(formula = Price ~ Year + Mileage + State, data = CombinedStates)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6981.8 -1426.5  -291.5  1266.1 18492.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.442e+06  1.489e+05 -16.401  < 2e-16 ***
## Year         1.225e+03  7.387e+01  16.582  < 2e-16 ***
## Mileage     -8.141e-02  4.304e-03 -18.915  < 2e-16 ***
## StateFL     -1.081e+03  2.867e+02  -3.771 0.000178 ***
## StateNC      6.104e+01  2.877e+02   0.212 0.832024
## StatePA      2.914e+02  3.089e+02   0.943 0.345937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2254 on 638 degrees of freedom
## Multiple R-squared:  0.8029, Adjusted R-squared:  0.8014
## F-statistic: 519.9 on 5 and 638 DF,  p-value: < 2.2e-16
```

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

   Price=β0+β1Year+β2Mileage+β3StateFL+β4StateNC+β5StateNC+β6StatePA +ε

   H0: β3=β4=β5=β6=0 Ha:some βi≠0

   The null hypothesis is that the slopes of States are 0. The predictor State and the response price have no relationship.

   Since its P value is pretty small(9.584e-09), we have evidence to say that lines of mod4 and mod5 are different. We want to use the State term in our model because extra variability is explained by State.

```
mod4=lm(Price~Year+Mileage,data=CombinedStates)
anova(mod4, mod5)

## Analysis of Variance Table
##
## Model 1: Price ~ Year + Mileage
## Model 2: Price ~ Year + Mileage + State
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    641 3453013303
## 2    638 3242225952  3 210787351 13.826 9.584e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
mod5int=lm(Price ~ Year + State + Year*State + Mileage*State, data=Comb
inedStates)
summary(mod5int)

##
## Call:
## lm(formula = Price ~ Year + State + Year * State + Mileage *
##     State, data = CombinedStates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7023.5 -1419.8  -239.1  1203.4 18378.5
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.048e+06  4.008e+05  -5.108 4.31e-07 ***
## Year             1.029e+03  1.988e+02   5.176 3.05e-07 ***
## StateFL         -4.214e+05  4.734e+05  -0.890   0.3738
## StateNC         -3.053e+05  4.746e+05  -0.643   0.5204
## StatePA         -9.997e+05  5.303e+05  -1.885   0.0599 .
## Mileage         -9.086e-02  1.203e-02  -7.553 1.49e-13 ***
## Year:StateFL     2.082e+02  2.348e+02   0.886   0.3757
## Year:StateNC     1.516e+02  2.354e+02   0.644   0.5199
## Year:StatePA     4.964e+02  2.631e+02   1.887   0.0596 .
## StateFL:Mileage  2.379e-02  1.417e-02   1.679   0.0936 .
## StateNC:Mileage  1.149e-03  1.394e-02   0.082   0.9344
## StatePA:Mileage  6.399e-03  1.556e-02   0.411   0.6811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2231 on 632 degrees of freedom
## Multiple R-squared:  0.8088, Adjusted R-squared:  0.8055
## F-statistic: 243.1 on 11 and 632 DF,  p-value: < 2.2e-16
```

9. Perform a hypothesis test to determine the importance of the terms involving *State* in the model constructed in question 8. List your hypotheses, p-value, and conclusion.

   Price=β0+β1Year+…+βiMileage*State+ε

   H0: β3=β4=…=0 Ha: some βi≠0

   The null hypothesis is that the slopes of States are 0. The predictor State and the response price have no relationship.

   Since its P value is pretty small(1.793e-09), we have evidence to say that lines of mod4 and mod5int are different. We want to use the State term in our model because extra variability is explained by State.

```
mod4=lm(Price~Year+Mileage,data=CombinedStates)
anova(mod4, mod5int)

## Analysis of Variance Table
##
```

```
## Model 1: Price ~ Year + Mileage
## Model 2: Price ~ Year + State + Year * State + Mileage * State
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    641 3453013303
## 2    632 3145145948  9 307867355 6.8738 1.793e-09 ***
## ---
## 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 vehicle" phenomenon where the predicted price is eventually negative as the line decreases for older vehicles. Let's see if adding one or more polynomial terms might help with this. For this section you should use the dataset with vehicles from four states that you used for model 5.

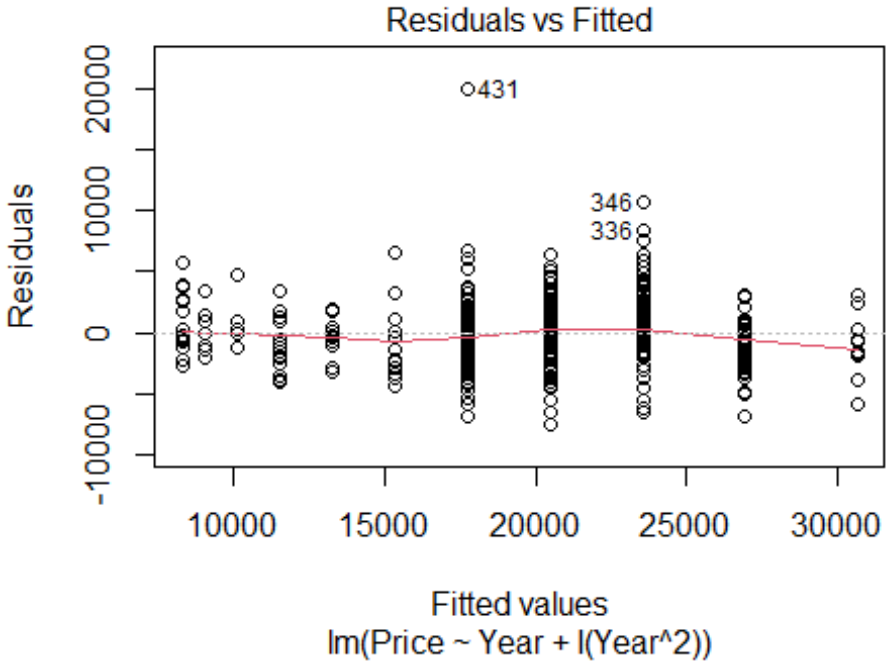
10. Fit a quadratic model using *Year* to predict *Price* and examine the residuals. Construct a scatterplot of the data with the quadratic fit included. You should discuss each of the conditions for the linear model.

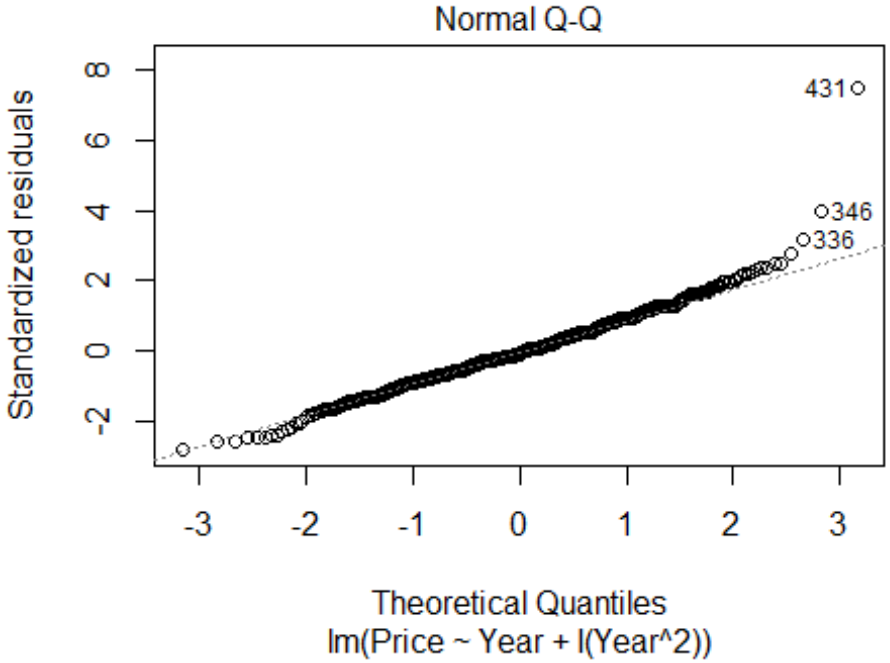
    (1)linearity is fine because the red line seems to be roughly horizontal at zero for the residuals.

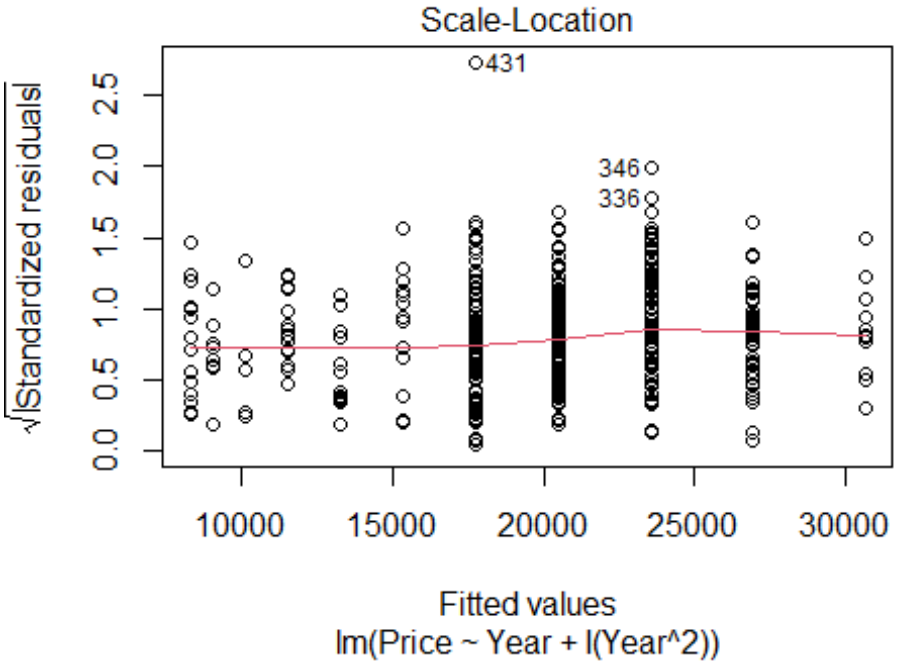
    (2)Constant variance seems to be roughly met. From the residuals vs fitted values plot we can see a similar spread above/below the red curve for all fitted values. The variability from the line does not follow any pattern as value of the predictor changes.

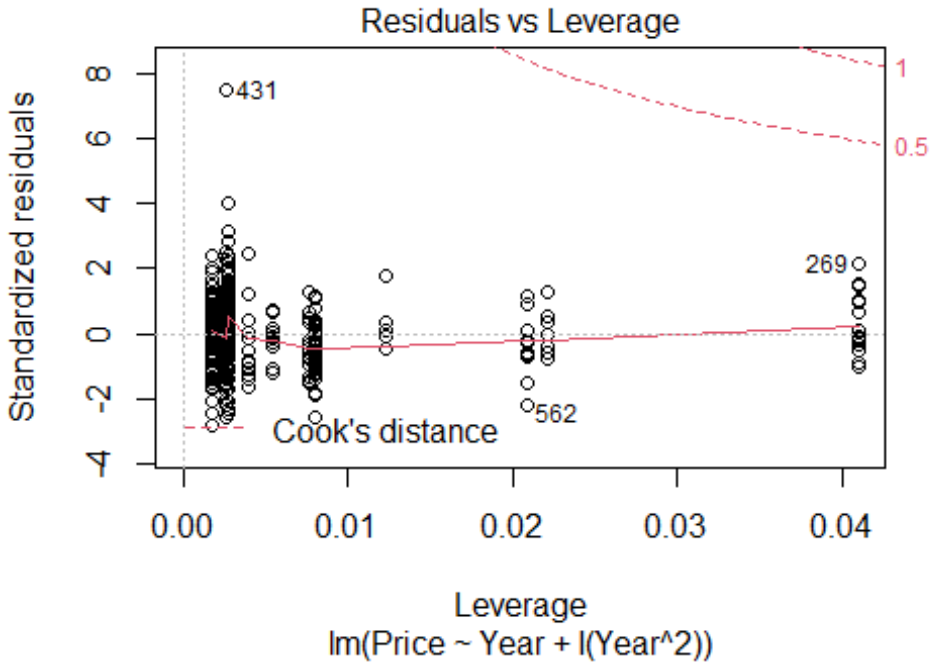
    (3)Normality of residuals seems to be roughly met since the data in the qqplot roughly fit the qqline. There are some deviations at each of the tails of the plot, showing some skewness of possible concern given the larger sample size of the data.

```
mod6 = lm(Price~ Year+ I(Year^2), data=CombinedStates)
plot(mod6)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Price ~ Year + I(Year^2))
431
346
336
-10000
0
10000
20000
10000
15000
20000
25000
30000

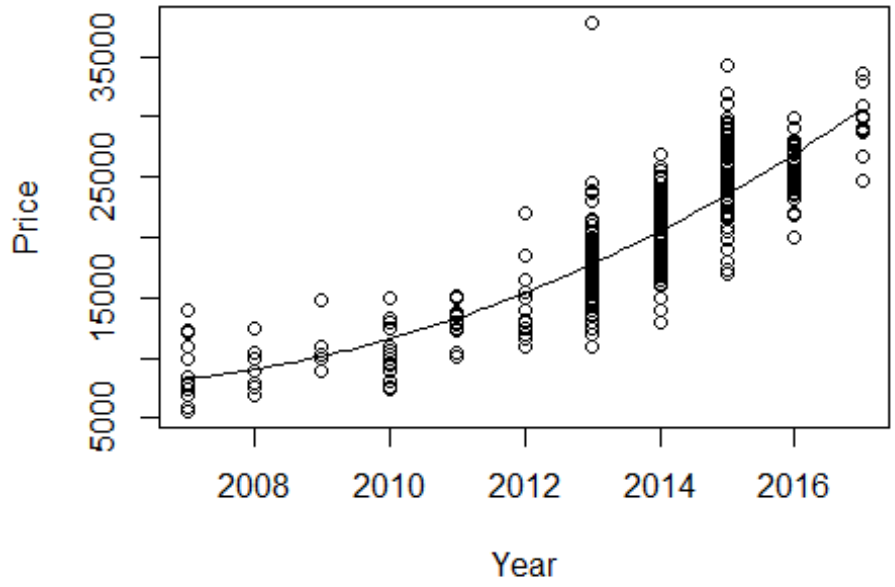

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ Year + I(Year^2))
431
346
336
-2
0
2
4
6
8
-3
-2
-1
0
1
2
3






```
plot(Price~ Year, data=CombinedStates)
B0_mod6 = summary(mod6)$coef[1,1]
B1_mod6 = summary(mod6)$coef[2,1]
```

```
B2_mod6 = summary(mod6)$coef[3,1]
curve(B0_mod6 + B1_mod6*x + B2_mod6*x^2, add=TRUE)
```



11. Perform a hypothesis test to determine if any of the coefficients in this model have nonzero coefficients. List your hypotheses, p-value, and conclusion.
Assuming that adding predictors can not significantly explain variability.
(1)From the first line of summary, we know that P value is small so compared with the null model, adding Year can significantly explain some variability.
(2) From the second line of summary, we know that P value is small so compared with the single linear model, adding Year^2 can significantly explain some variability.

```
anova(mod6)

## Analysis of Variance Table
##
## Response: Price
##             Df     Sum Sq    Mean Sq  F value    Pr(>F)
## Year         1 1.1216e+10 1.1216e+10 1560.587 < 2.2e-16 ***
## I(Year^2)    1 6.2931e+08 6.2931e+08   87.565 < 2.2e-16 ***
## Residuals  641 4.6067e+09 7.1868e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12. You are looking at a vehicle that was 4 years old (in 2017) 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 vehicle, and include an interpretive sentence in context.

With 95% confidence I predict that the price of a four years old EdgeSEL sold in the four states is between $12506 and $23048.51

```
newxqua=data.frame(Year=2013)
predict.lm(mod6, newxqua, interval="prediction",level=0.95)

##        fit   lwr      upr
## 1 17777.26 12506 23048.51
```

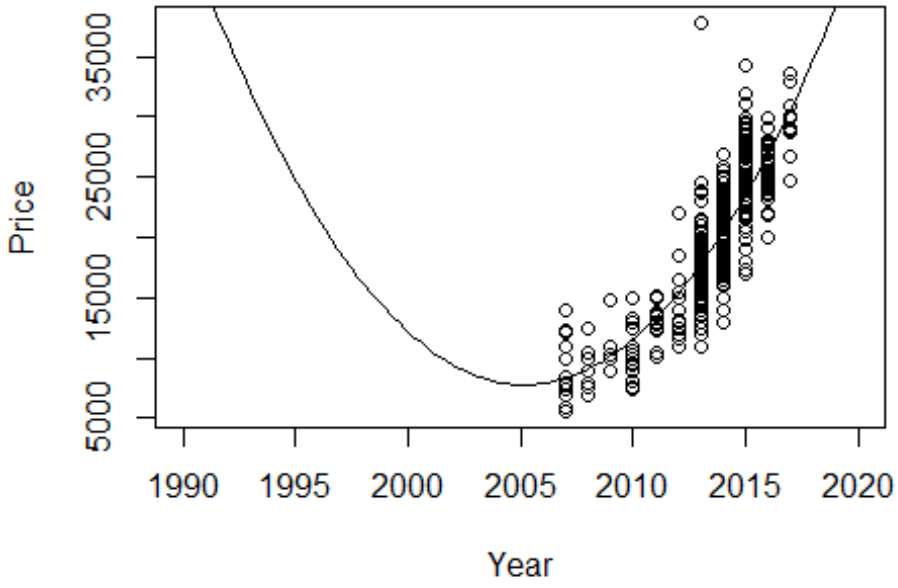
13. Does the quadratic model allow for some *Year* where a vehicle has a zero or negative predicted price? Justify your answer using a calculation or graph.

No it does not. By using the discriminant for quadratic equations, which is Δ=b^2-4ac, we know that β2>0 and Δ=-5097379<0, so there is no roots in this equation and the parabola is above abscissa.

```
B1_mod6^2-4*B2_mod6*B0_mod6

## [1] -5097379

plot(Price~ Year, data=CombinedStates, xlim=c(1990,2020))
curve(B0_mod6 + B1_mod6*x + B2_mod6*x^2, 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.

There is a multicollinearity issue so adding the cubic term does not create a better model. VIF test shows that there are aliased coefficients in the model.

```
mod6cub = lm(Price~ Year+I(Year^2)+I(Year^3), data=CombinedStates)
summary(mod6cub)

##
## Call:
## lm(formula = Price ~ Year + I(Year^2) + I(Year^3), data = CombinedSt
ates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7507.2 -1740.6  -252.8  1492.8 20017.7
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.591e+08  7.092e+07   9.293   <2e-16 ***
## Year        -6.574e+05  7.049e+04  -9.325   <2e-16 ***
## I(Year^2)    1.639e+02  1.752e+01   9.358   <2e-16 ***
## I(Year^3)           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2681 on 641 degrees of freedom
## Multiple R-squared:   0.72,  Adjusted R-squared:  0.7191
## F-statistic: 824.1 on 2 and 641 DF,  p-value: < 2.2e-16
```

MODEL #7: Complete second order model

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

15. Fit a complete second order model for predicting a used vehicle *Price* based on *Year* and *Mileage* and examine the residuals. You should discuss each of the conditions for the linear model.

    (1)linearity is fine because the red line seems to be roughly horizontal at zero for the residuals.

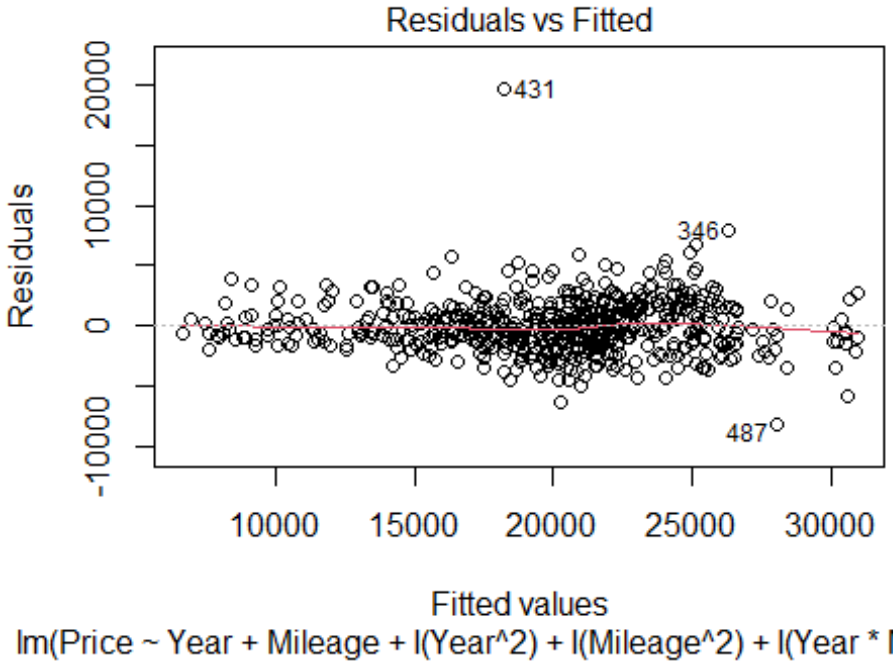
    (2)Constant variance seems to be roughly met. From the residuals vs fitted values plot we can see a similar spread above/below the red curve for all fitted values.

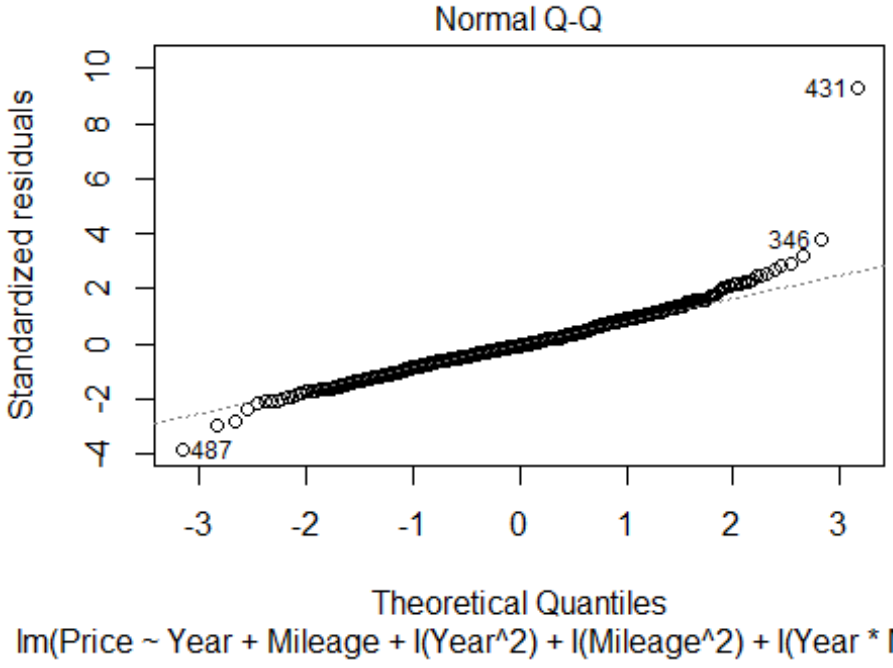
    (3)Normality of residuals seems to be roughly met since the data in the qqplot roughly fit the qqline. There are some deviations at each of the tails of the plot, showing some skewness of possible concern given the larger sample size of the data.

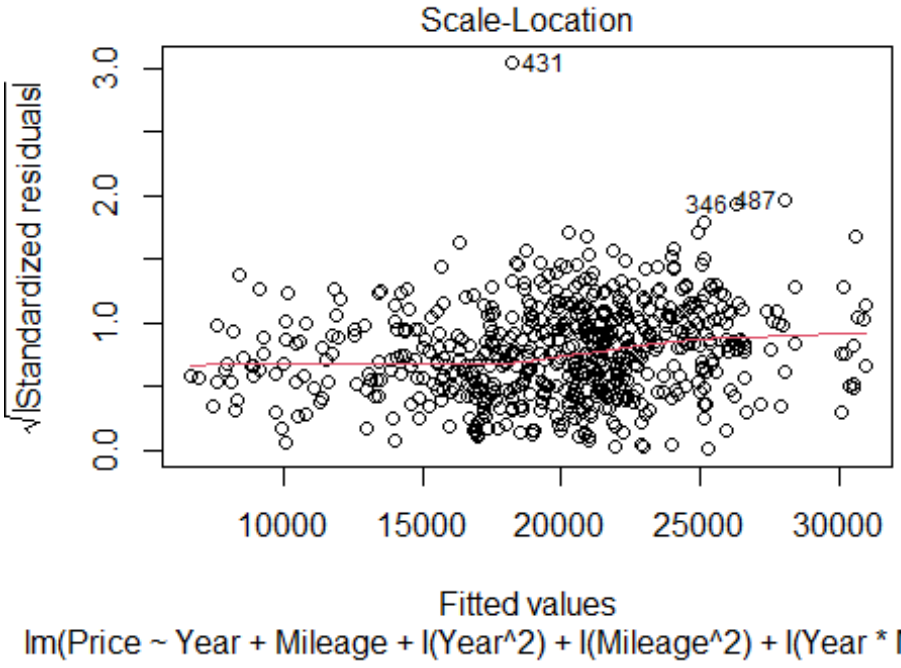
```
mod7=lm(Price~Year+Mileage+I(Year^2)+I(Mileage^2)+I(Year*Mileage),data=
CombinedStates)
summary(mod7)
```

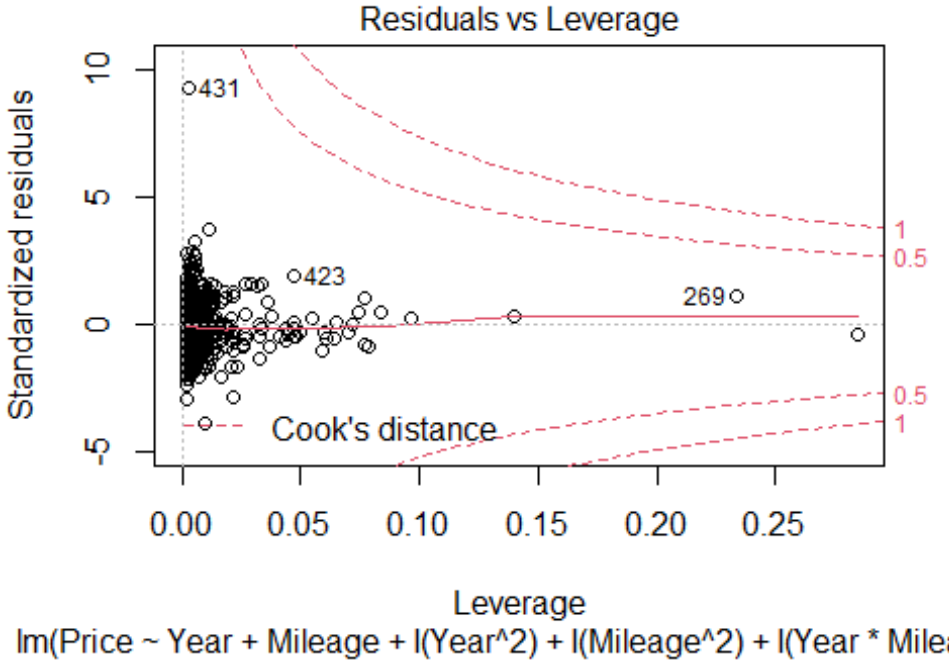
```
##
## Call:
## lm(formula = Price ~ Year + Mileage + I(Year^2) + I(Mileage^2) +
##     I(Year * Mileage), data = CombinedStates)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8079  -1230   -142   1142  19589
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.349e+08  1.162e+08   2.882  0.00409 **
## Year              -3.343e+05  1.153e+05  -2.900  0.00386 **
## Mileage            6.475e+00  6.685e+00   0.969  0.33310
## I(Year^2)          8.346e+01  2.861e+01   2.918  0.00365 **
## I(Mileage^2)       1.922e-07  1.204e-07   1.596  0.11092
## I(Year * Mileage) -3.268e-03  3.314e-03  -0.986  0.32444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2117 on 638 degrees of freedom
## Multiple R-squared:  0.8263, Adjusted R-squared:  0.8249
## F-statistic: 606.8 on 5 and 638 DF,  p-value: < 2.2e-16

plot(mod7)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Price ~ Year + Mileage + I(Year^2) + I(Mileage^2) + I(Year * Milea
431
346
487
20000
10000
0
-10000
10000
15000
20000
25000
30000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ Year + Mileage + I(Year^2) + I(Mileage^2) + I(Year * Milea
431
346
487
10
8
6
4
2
0
-2
-4
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
431
346
487
0.0
1.0
2.0
3.0
10000
15000
20000
25000
30000
Fitted values
lm(Price ~ Year + Mileage + I(Year^2) + I(Mileage^2) + I(Year * Milea


Residuals vs Leverage
Standardized residuals
431
423
269
Cook's distance
1
0.5
0.5
1
-5
0
5
10
0.00
0.05
0.10
0.15
0.20
0.25
Leverage
lm(Price ~ Year + Mileage + I(Year^2) + I(Mileage^2) + I(Year * Milea

16. Perform a hypothesis test to determine if any of the coefficients in this model have nonzero coefficients. List your hypotheses, p-value, and conclusion.
    Assuming all of those coefficients are zero while the alternative is that in those coefficients at least one of them is non zero. P value is big(0.3244) so we should not use the full second order.

```
mod7_reduced1=lm(Price~Year+Mileage+I(Year^2)+I(Mileage^2),data=Combine
dStates)
anova(mod7_reduced1,mod7)

## Analysis of Variance Table
##
## Model 1: Price ~ Year + Mileage + I(Year^2) + I(Mileage^2)
## Model 2: Price ~ Year + Mileage + I(Year^2) + I(Mileage^2) + I(Year
*
##     Mileage)
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    639 2862728210
## 2    638 2858371420  1   4356790 0.9725 0.3244
```

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.
    Assuming that the second order terms can not significantly explain extra variability. P value is small so it makes sense to keep those second order terms.

```
mod7_reduced2=lm(Price~Year+Mileage,data=CombinedStates)
anova(mod7_reduced2,mod7)

## Analysis of Variance Table
##
## Model 1: Price ~ Year + Mileage
## Model 2: Price ~ Year + Mileage + I(Year^2) + I(Mileage^2) + I(Year
*
##     Mileage)
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    641 3453013303
## 2    638 2858371420  3 594641883 44.242 < 2.2e-16 ***
## ---
## 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.
    Assuming that the year term can not significantly explain more variability. P value is small so it makes sense to keep the year term.

```
mod7_reduced3=lm(Price~Mileage+I(Mileage^2),data=CombinedStates)
anova(mod7_reduced3,mod7)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Mileage + I(Mileage^2)
## Model 2: Price ~ Year + Mileage + I(Year^2) + I(Mileage^2) + I(Year
*
##     Mileage)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    641 4653510267
## 2    638 2858371420  3 1795138847 133.56 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```