STOR 455 Homework #5

20 points - Due Thursday 03/30 at 12:00pm

Theory Part

1. True or False: For a linear model with 5 slopes, the sum of leverages is 5.

Your answer: False

2. True or False: The shrinkage in cross validation cannot be negative.

Your answer: True

Computing Part

Situation: Suppose that 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 that you get (the model) and how much it's been used. For this assignment you will investigate how the price might depend on the vehicle's year, state, and odometer reading. We focus on three states in this homework "CA", "NC", and "NY."

Data Source: To get a sample of vehicles, begin with the *UsedCars* csv file. The data was acquired by scraping Craigslist for vehicles for sale across the southeastern United States. For this assignment you will choose model of cars. Construct a subset of the *vehiclesSE* data for this model of vehicle. If your subset has cars with seemingly incorrect data (such as a price of $1, odometer reading of one million miles, year of 1900) you should remove those values from the data.

Directions: The code below should walk you through the process of selecting data from a particular model vehicle of your choice. The following R chunk begin with {r, eval=FALSE}. eval=FALSE makes this chunk not run when I knit the file. Before you knit this chunk, you should revert it to {r}.

```r
library(readr)
library(polynom)

vehicles_all <- read_csv("UsedCars.csv", show_col_types = FALSE)

vehicles_3States = subset(vehicles_all, State=="NY"|State=="NC"|State=="CA")

# Delete the ** below and enter your chosen model
ModelOfMyChoice = "3"


vehiclesSE= subset(vehicles_3States, Model==ModelOfMyChoice)
```

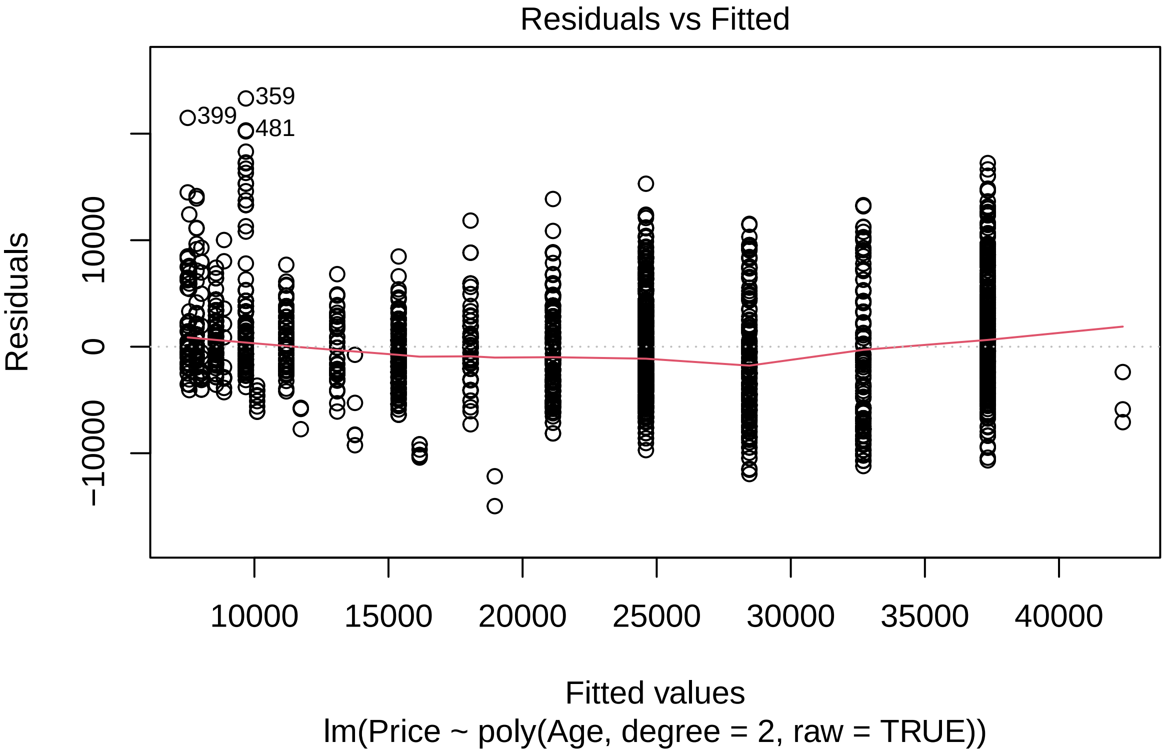
MODEL #1: Polynomial models

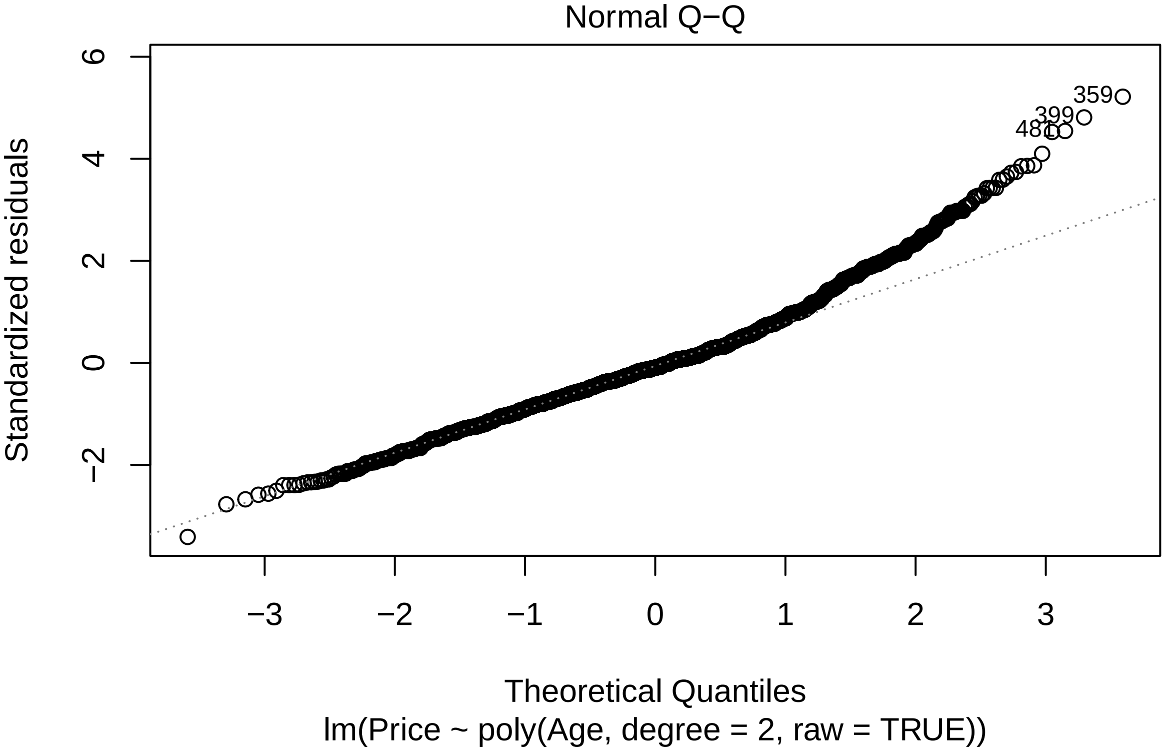
1. Construct a new variable called *Age* in the *vehiclesSE* dataframe. Since the vehicles were posted to Craigslist in 2021, define the *Age* of all vehicles to be their year subtracted from 2021.

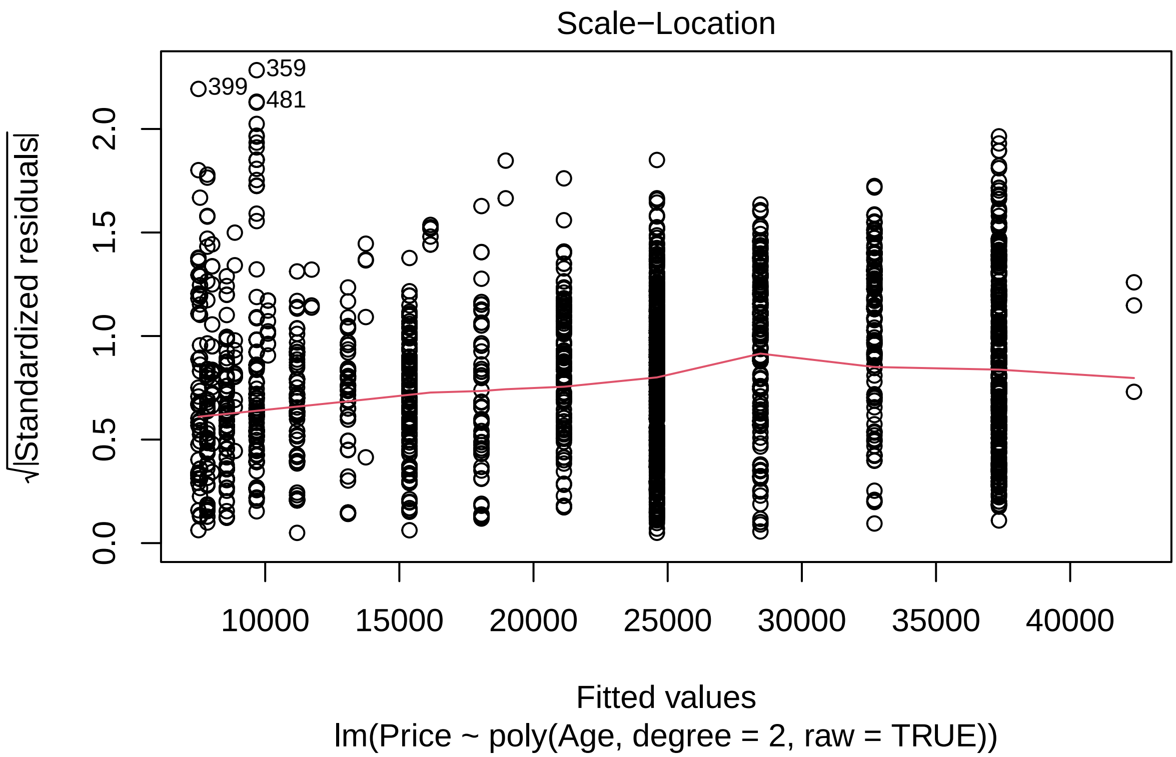
```r
vehiclesSE$Age = 2021 - vehiclesSE$Year
```

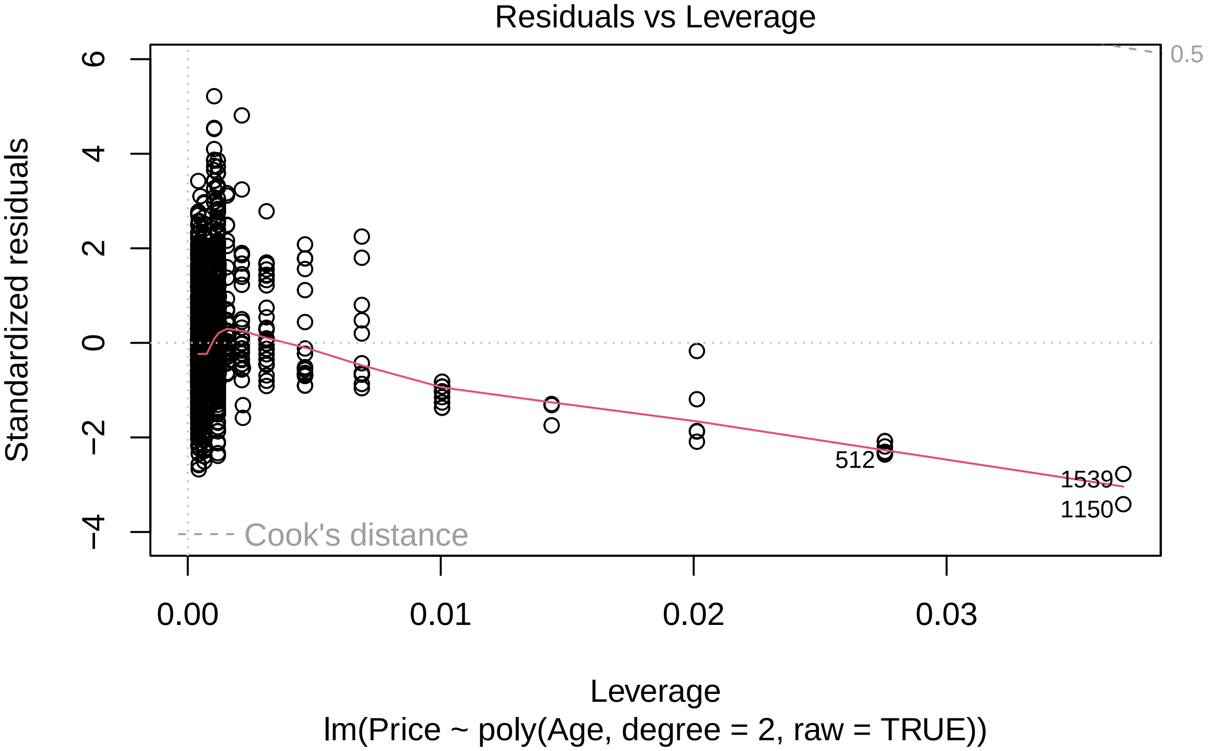
2. Fit a quadratic model using *Age* to predict *Price* and examine the residuals. Construct a scatterplot of the data with the quadratic fit included. You should discuss each of the conditions for the linear model.

```
QuadModel = lm(Price~poly(Age,degree=2, raw=TRUE), data=vehiclesSE)
plot(QuadModel)
```





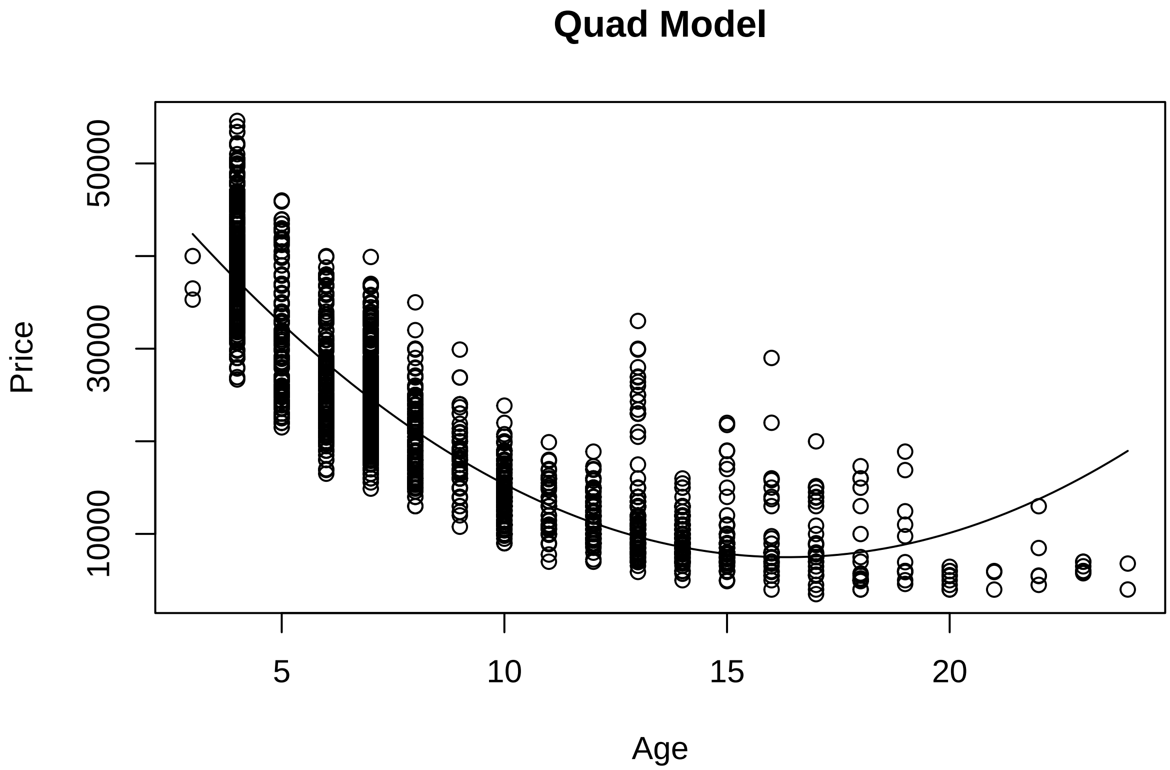




```
Quad_coef = summary(QuadModel)$coef[,1]

QuadCurve = as.function(polynomial(Quad_coef))

plot(Price~Age,main="Quad Model",data=vehiclesSE)
curve(QuadCurve, add = TRUE)
```



On the residual plot there seems to be lots of variance and some outliers. On the residual line there seems to be lots of clusters surrounding it. The red line also looks like a parabola curve. All of these things make me come to the conclusion that there is good correlation between Age and Price. When looking through the linear conditions the variance looks about the same for all values of X, the relationship looks relatively linear, and for any value of X the Y's look normally distributed on the QQ plot.

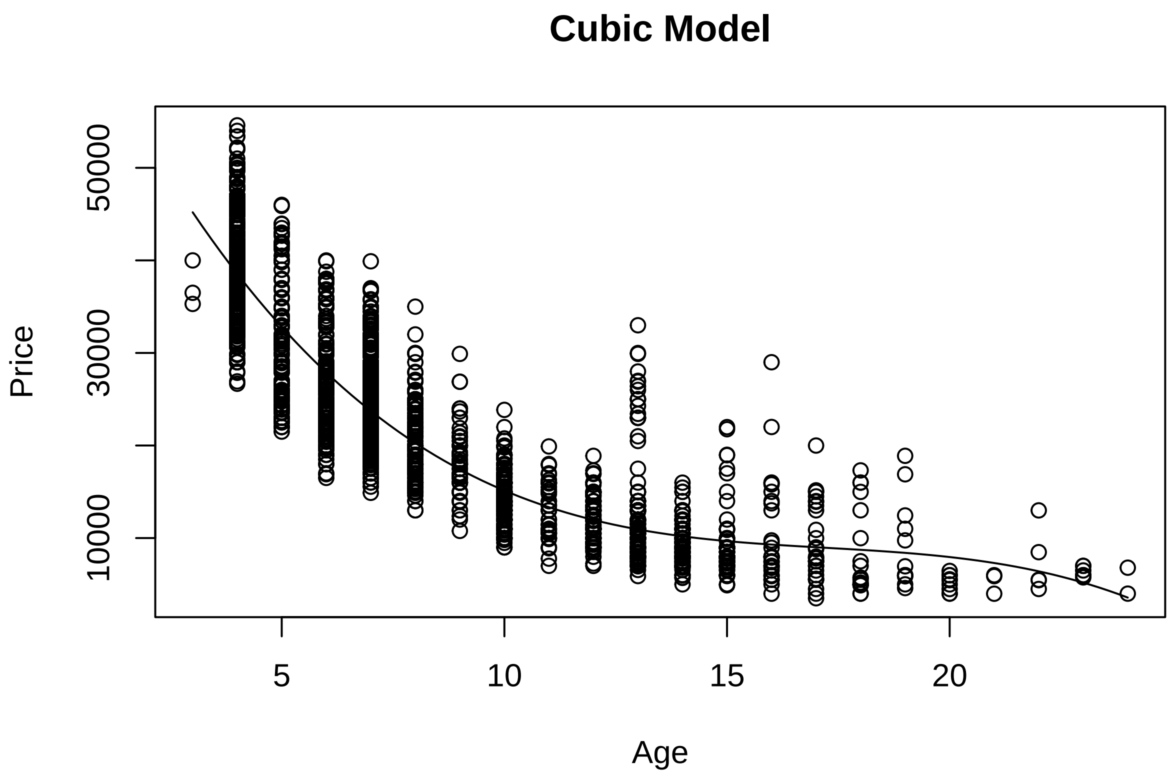
3. Would the fit improve significantly if you also included a cubic term? Does expanding your polynomial model to use a quartic term make significant improvements? Justify your answer.

```
CubicModel = lm(Price~poly(Age,degree=3, raw=TRUE), data=vehiclesSE)

Cubic_coef = summary(CubicModel)$coef[,1]

CubicCurve = as.function(polynomial(Cubic_coef))

plot(Price~Age,main="Cubic Model", data=vehiclesSE)
curve(CubicCurve, add = TRUE)
```

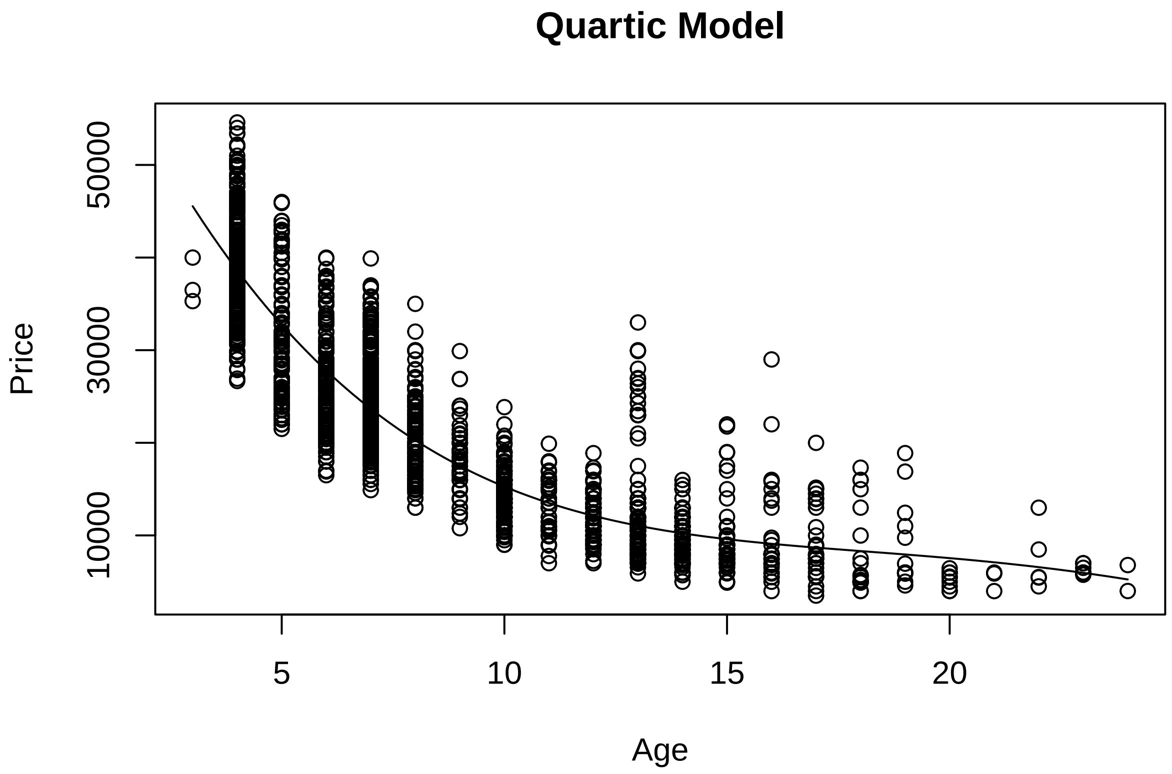


```
QuarticModel = lm(Price~poly(Age,degree=4, raw=TRUE), data=vehiclesSE)

Quartic_coef = summary(QuarticModel)$coef[,1]

QuarticCurve = as.function(polynomial(Quartic_coef))

plot(Price~Age,main="Quartic Model", data=vehiclesSE)
curve(QuarticCurve, add = TRUE)
```

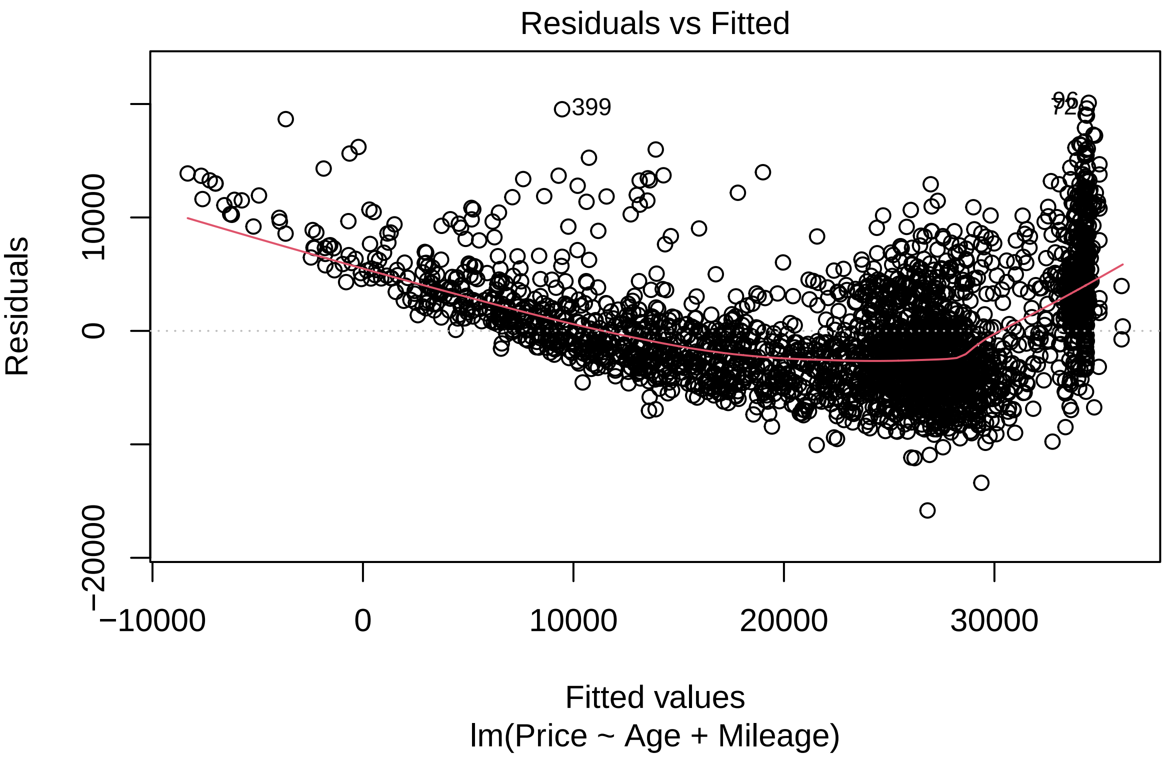


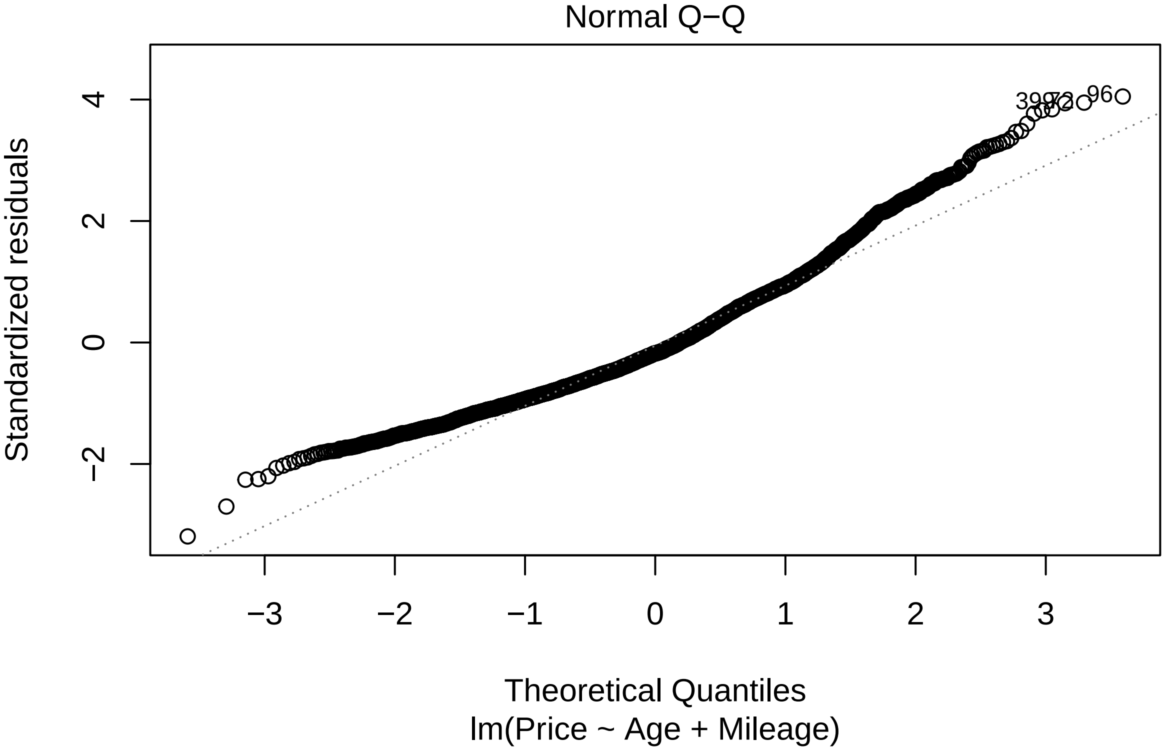
When looking at the the curves graphed onto the scatter plot they seem to be more accurate the higher the degree is. This is due to the higher degrees allowing for more curves in the graph making it more malleable to the data. Another note is that as the degrees get higher the difference becomes smaller and smaller. The jump between the quadratic and cubic is pretty big but the jump from cubic to quartic isn't very significant.

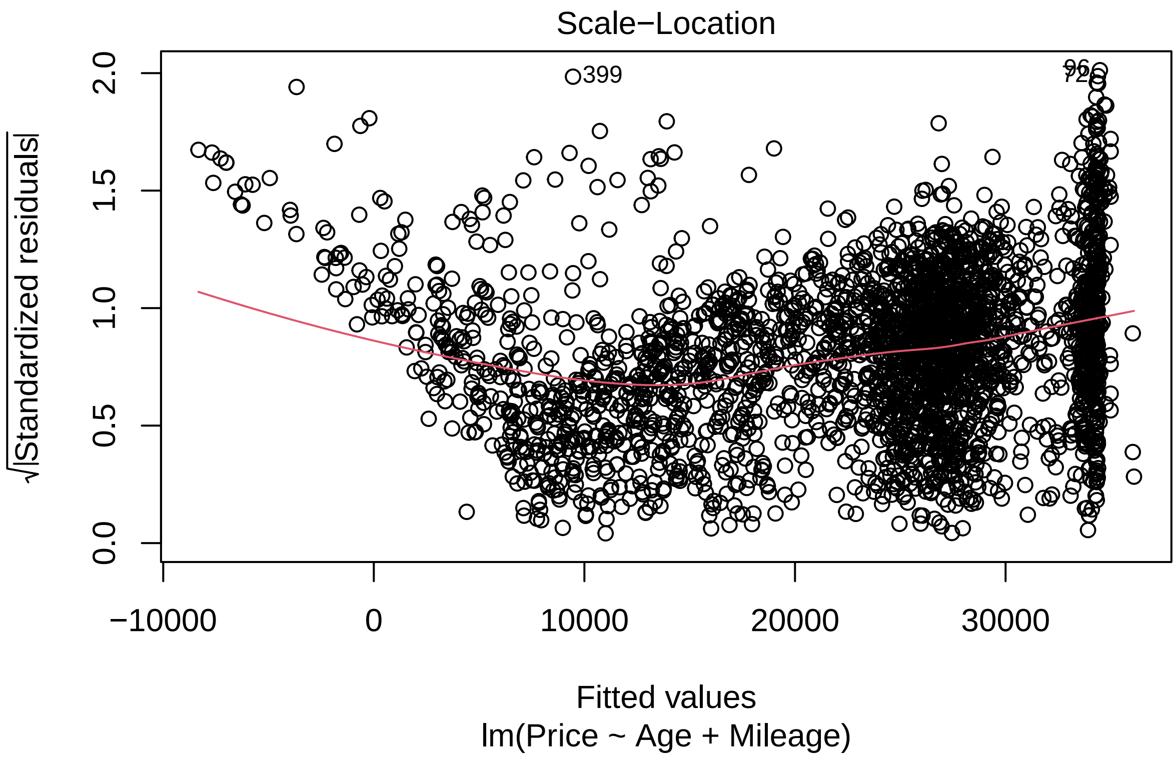
MODEL #2: Complete second order model

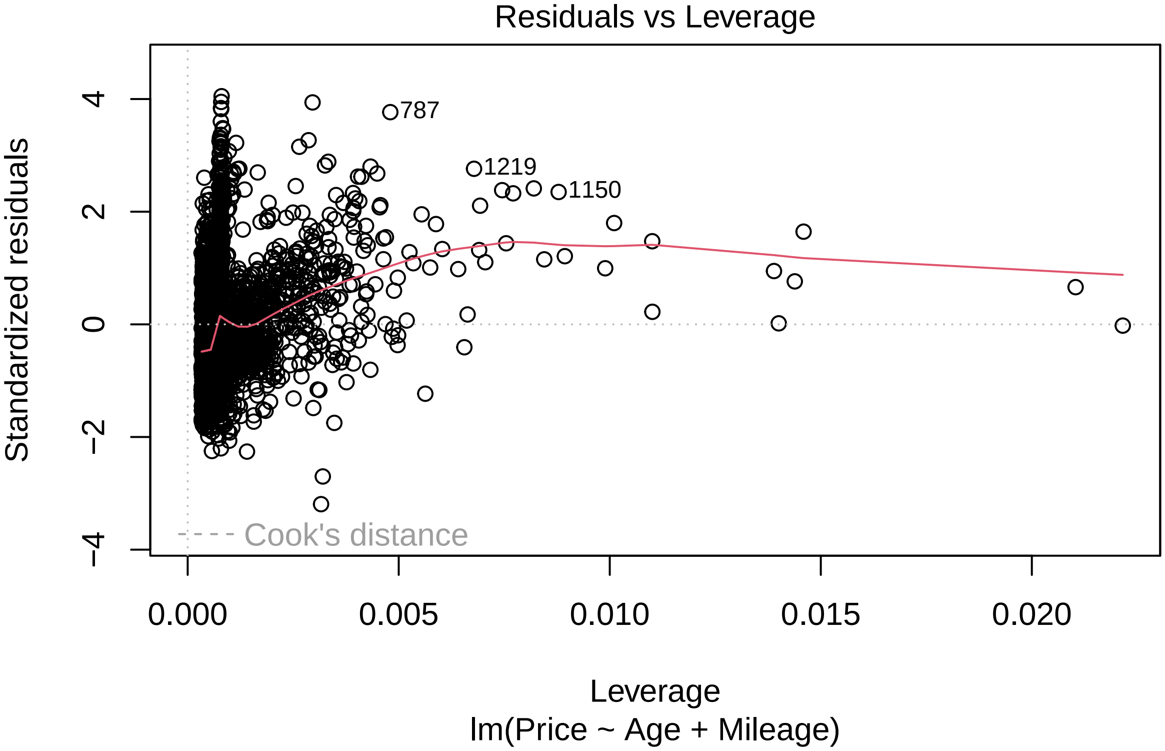
4. Fit a complete second order model for predicting a used vehicle *Price* based on *Age* and *Mileage* and examine the residuals. You should discuss each of the conditions for the linear model.

```
Model = lm(Price~Age+Mileage, data = vehiclesSE)
plot(Model)
```


Residuals vs Fitted
Residuals
−20000
0
10000
−10000
0
10000
20000
30000
399
96
72
Fitted values
lm(Price ~ Age + Mileage)


Normal Q−Q
Standardized residuals
−2
0
2
4
−3
−2
−1
0
1
2
3
399
72
96
Theoretical Quantiles
lm(Price ~ Age + Mileage)

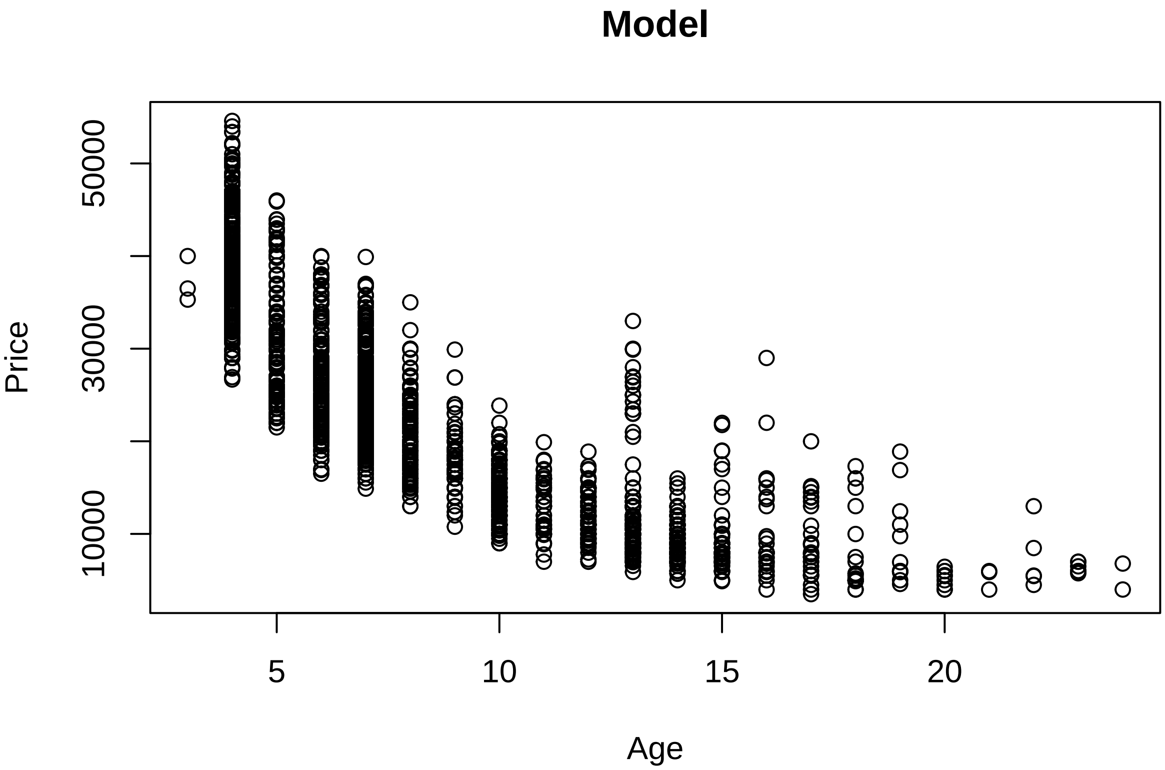




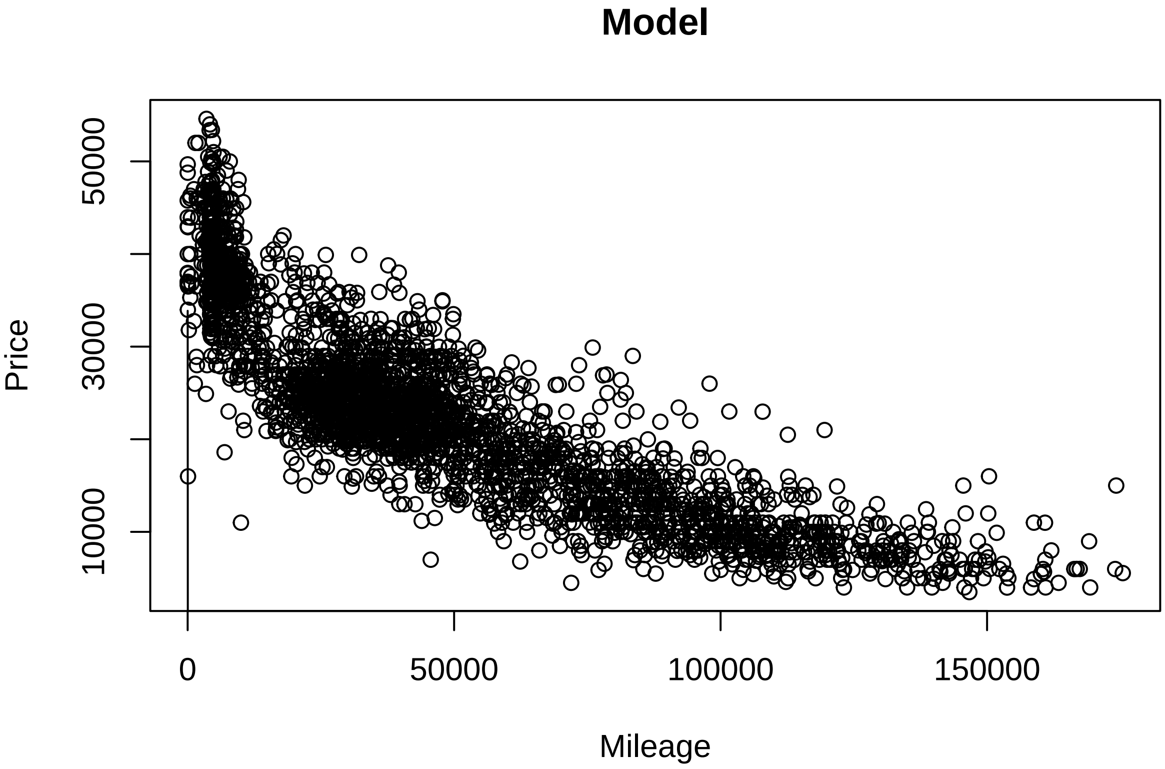
```
Mod_coef = summary(Model)$coef[,1]

ModCurve = as.function(polynomial(Mod_coef))

plot(Price~Age+Mileage, main="Model",data=vehiclesSE)
```



```
curve(ModCurve, add = TRUE)
```



When looking at the residuals vs fitted graph they are severely clustered to the right of the graph. On the left side there isn't much variance but on the right side this is lots of variance throughout. When looking through the linear conditions the variance is not the same for all values of X, the relationship doesn't look super linear, but for any value of X the Y's do look normally distributed as on the QQ plot the line is relatively straight.

Through these conditions we can conclude that this model does not meet the conditions for a linear model.

5. Perform a hypothesis test to determine the importance of just the second order terms (quadratic and interaction) in the model constructed in question 4. List your hypotheses, p-value, and conclusion.

```
anova(QuadModel)
```

```
## Analysis of Variance Table
##
## Response: Price
##                                     Df     Sum Sq    Mean Sq F value    Pr(>F)
## poly(Age, degree = 2, raw = TRUE)    2 2.6812e+11 1.3406e+11  6708.1 < 2.2e-16
## Residuals                         3037 6.0693e+10 1.9985e+07
##
## poly(Age, degree = 2, raw = TRUE) ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An F-test was used to see if there was a relationship between Price and Age^2. The Ho states that p = 0 and there is no correlation between Price and Age^2. The Ha states that p is != 0 and there is a correlation between Price and Age^2. The p-value for Age^2 is 2.2e^-16 which is incredibly small and less than 0.05. From this we can reject the null hypothesis and can assume that there is a correlation between Price and Age^2.

6. Perform a hypothesis test to determine the importance of just the terms that involve *Mileage* in the model constructed in question 4. List your hypotheses, p-value, and conclusion.

```
anova(Model)
```

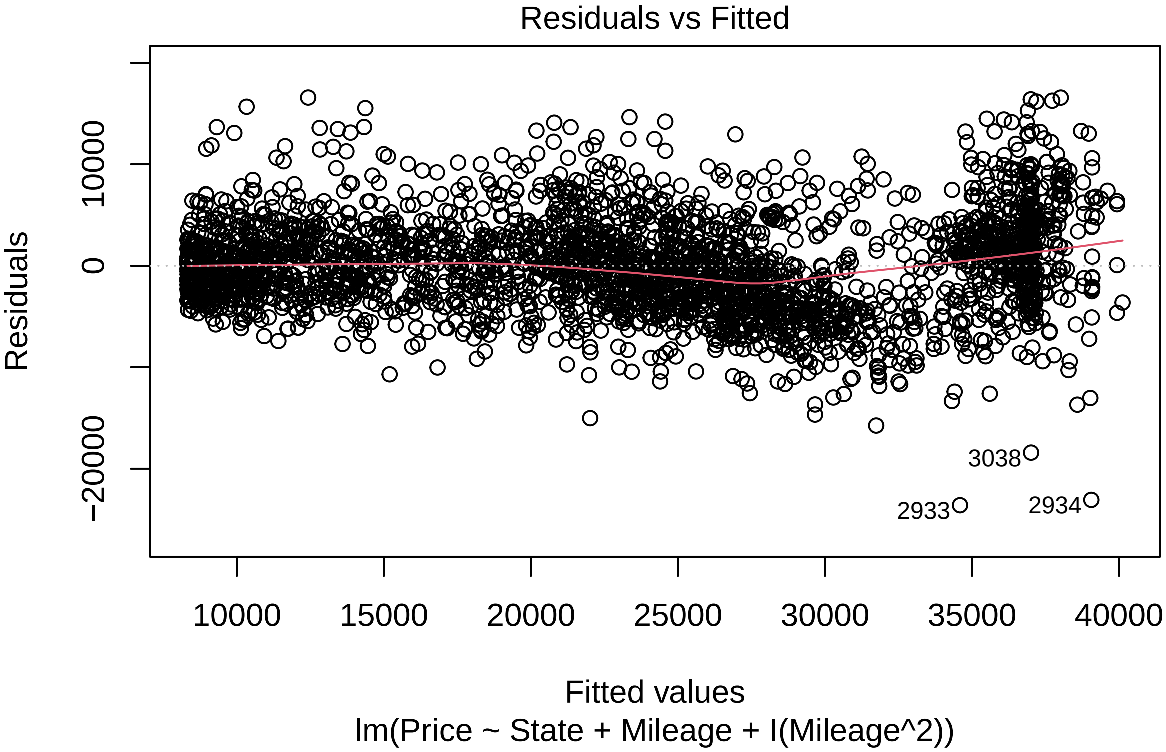
```
## Analysis of Variance Table
##
## Response: Price
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## Age          1 2.3088e+11 2.3088e+11 9360.77 < 2.2e-16 ***
## Mileage      1 2.3026e+10 2.3026e+10  933.56 < 2.2e-16 ***
## Residuals 3037 7.4907e+10 2.4665e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

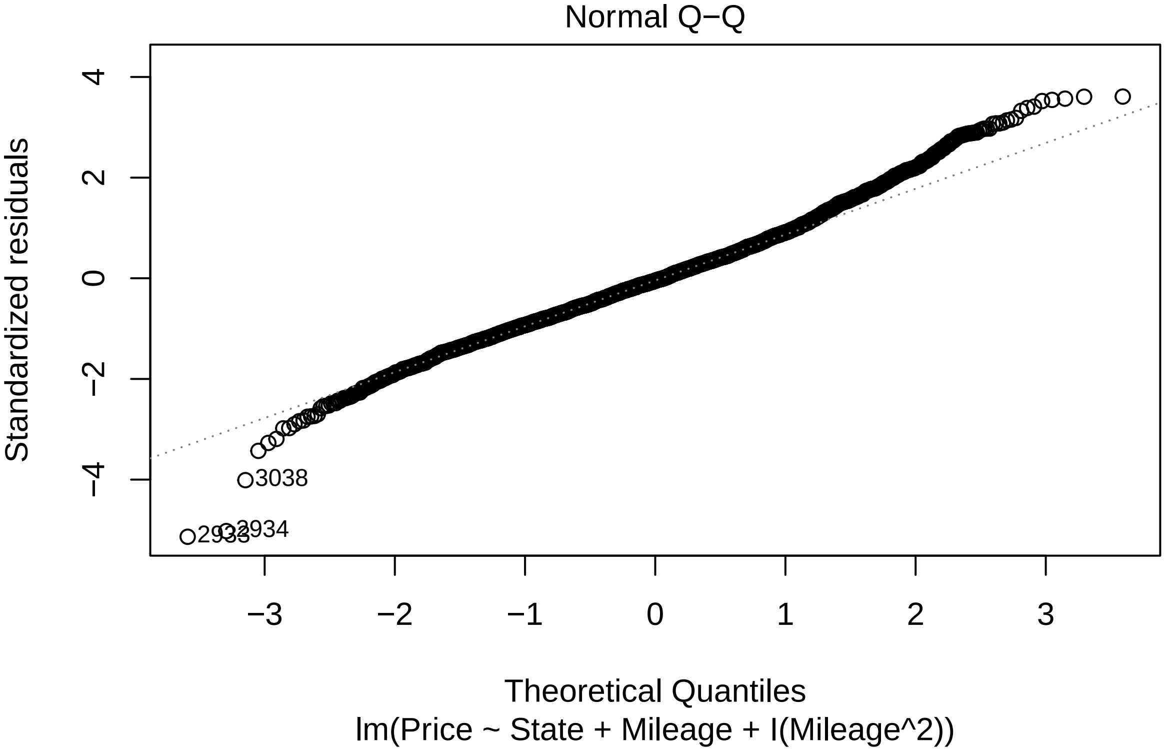
An F-test was used to see if there was a relationship between Price and Mileage. The Ho states that p = 0 and there is no correlation between Price and Mileage. The Ha states that p is != 0 and there is a correlation between Price and Mileage. The p-value for Mileage is 2.2e^-16 which is incredibly small and less than 0.05. From this we can reject the null hypothesis and can assume that there is a correlation between Price and Mileage.

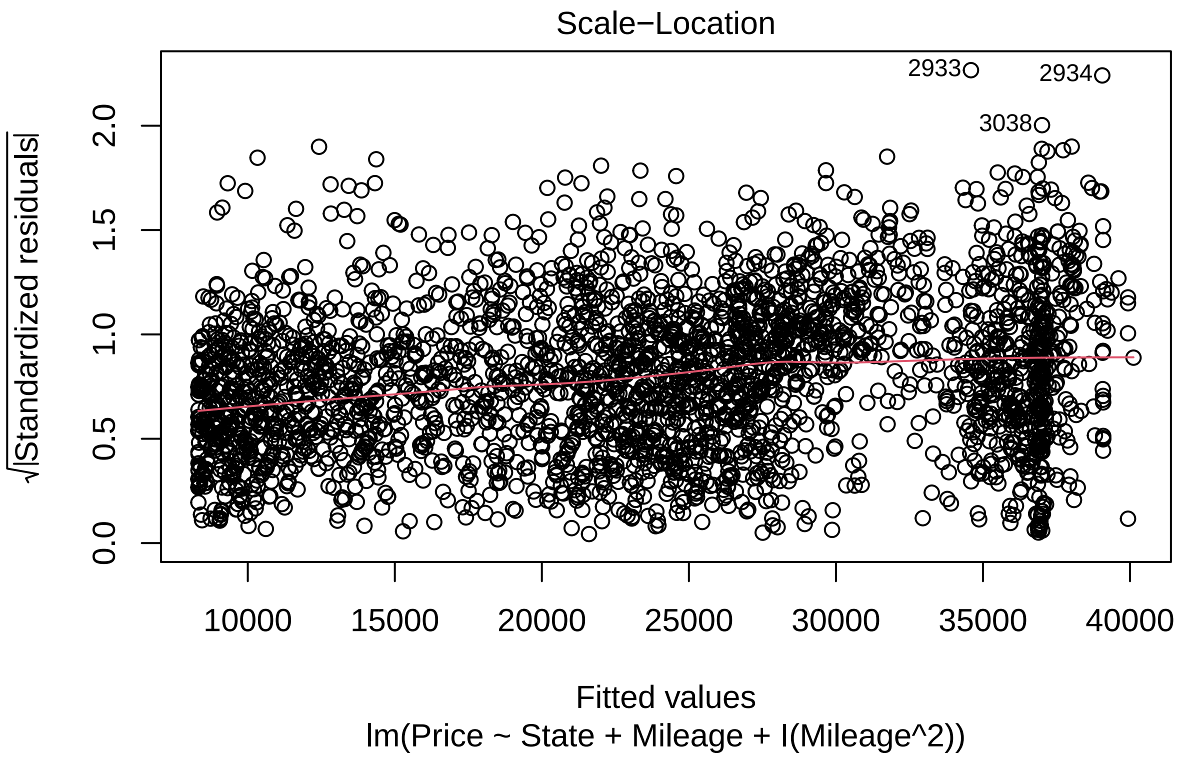
MODEL #3: Polynomial model with a categorical predictor

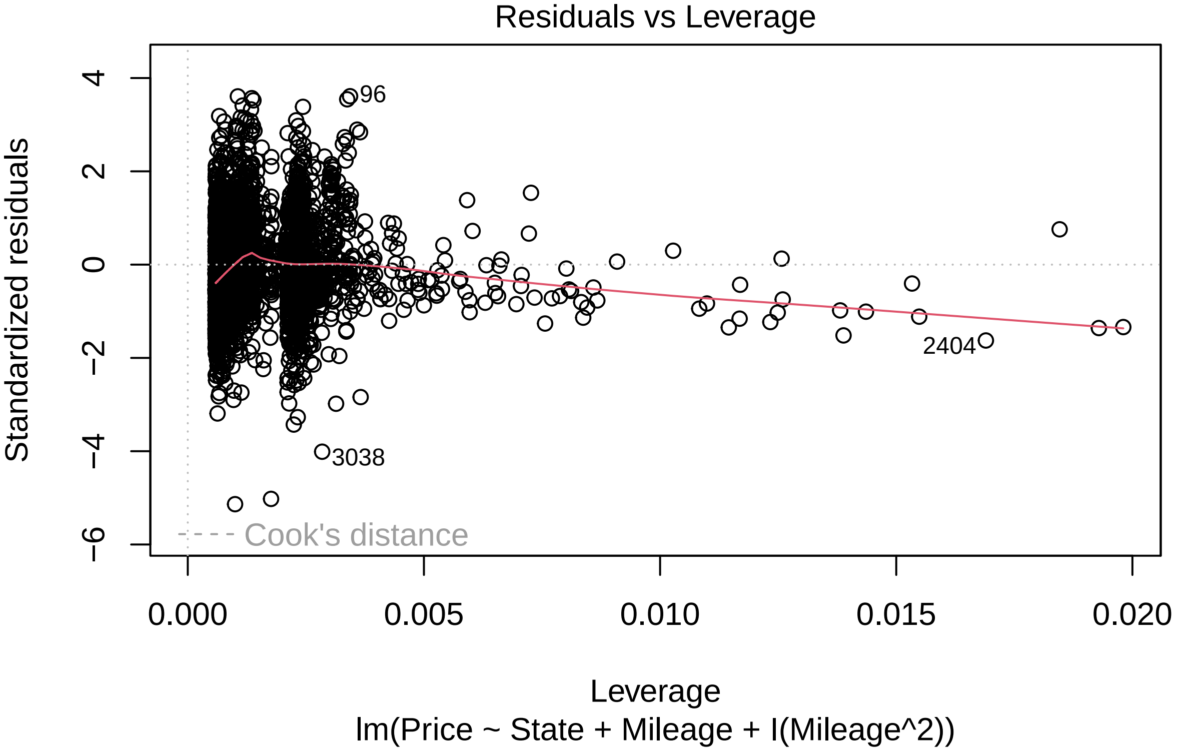
7. Fit a quadratic model regressing *Price* on *Mileage*, while considering *State* and its interactions with each term in the polynomial. Examine the residuals. You should discuss each of the conditions for the linear model.

```
QuadMileModel = lm(Price~State+Mileage+I(Mileage^2),data=vehiclesSE)
plot(QuadMileModel)
```


Residuals vs Fitted
Residuals
10000
0
−20000
10000
15000
20000
25000
30000
35000
40000
3038
2933
2934
Fitted values
lm(Price ~ State + Mileage + I(Mileage^2))


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
3038
2933
2934
Theoretical Quantiles
lm(Price ~ State + Mileage + I(Mileage^2))

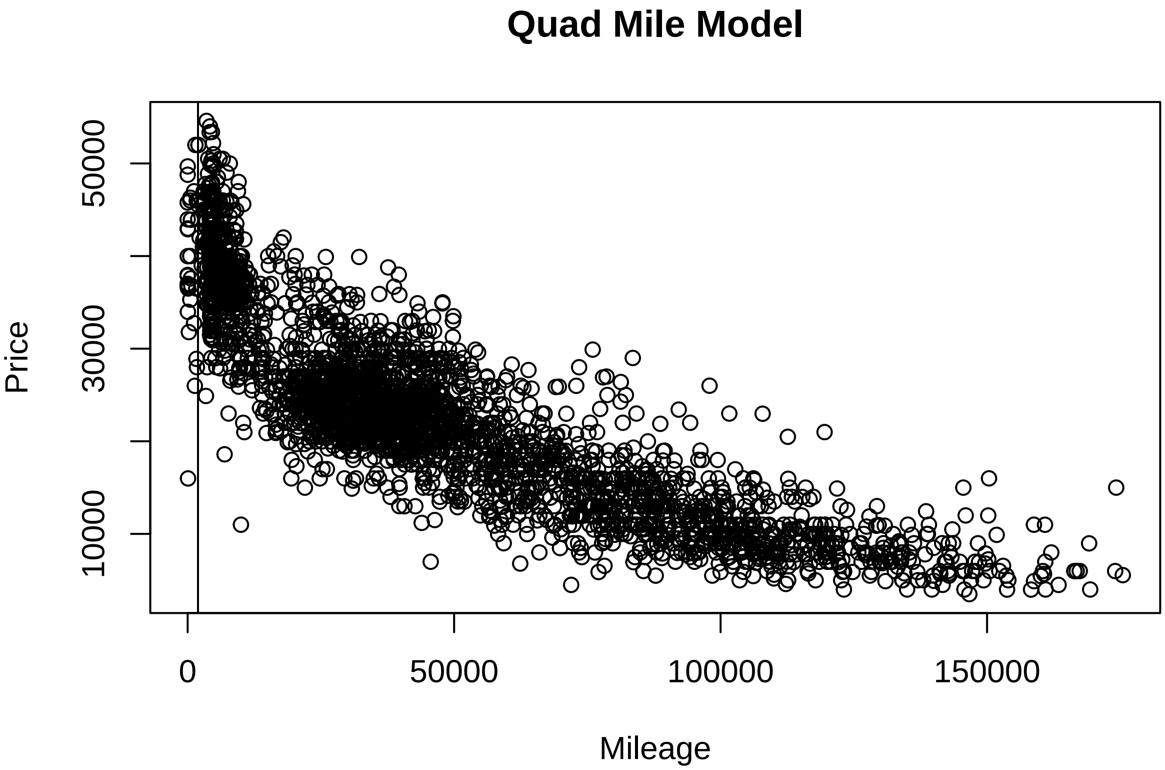




```r
QuadMile_coef = summary(QuadMileModel)$coef[,1]

QuadMileCurve = as.function(polynomial(QuadMile_coef))

plot(Price~Mileage,main="Quad Mile Model",data=vehiclesSE)
curve(QuadMileCurve, add = TRUE)
```



When looking at the residuals vs fitted graph they are spread pretty evenly and heavily clustered around the center throughout. When looking through the linear conditions the variance is relatively the same for all values of X, the relationship look decently linear, and for any value of X the Y's do look normally distributed as on the QQ plot the line is relatively straight. Through these conditions we can conclude that this model does meet the conditions for a linear model.

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

```
anova(QuadMileModel)
```

```
## Analysis of Variance Table
##
## Response: Price
##                 Df     Sum Sq    Mean Sq  F value    Pr(>F)
## State            2 4.6596e+09 2.3298e+09   110.24 < 2.2e-16 ***
## Mileage          1 2.3617e+11 2.3617e+11 11174.92 < 2.2e-16 ***
## I(Mileage^2)     1 2.3836e+10 2.3836e+10  1127.84 < 2.2e-16 ***
## Residuals     3035 6.4143e+10 2.1134e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An F-test was used to see if there was a relationship between Price and State. The Ho states that p = 0 and there is no correlation between Price and State. The Ha states that p is != 0 and there is a correlation between Price and State. The p-value for State is 2.2e^-16 which is incredibly small and less than 0.05. From this we can reject the null hypothesis and can assume that there is a correlation between Price and State.