STOR 455 Homework 2

40 points - Due Wednesday 9/15 at 5:00pm

Situation: Suppose that 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.

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 will choose a car *Model* for which there are at least 100 of that model listed for sale in a state of your choice (that is not North Carolina). After constructing a subset of the UsedCars data under these conditions, check to make sure that there is a reasonable amount of variability in the years and mileages for your car (ie, all of your cars are not from the same year). The model that you choose should have cars ranging over at least 5 years. You should add a variable called *Age* which is 2017-year (since the data was scraped in 2017).

Directions: The code below should walk you through the process of selecting data from a particular model and state of your choice. Each of the following two R chunks begin with {r, eval=FALSE}. eval=FALSE makes these chunks not run when I knit the file. Before you knit these chunks, you should revert them to {r}.

```r
library(readr)

# This line will only run if the UsedCars.csv is stored in the same directory as this notebook!
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 data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Delete the ** below and enter the two letter abbreviation for the state of your choice.
StateOfMyChoice = "TX"
```

```r
# Creates a dataframe with the number of each model for sale in your state
Cars = as.data.frame(table(UsedCars$Model[UsedCars$State==StateOfMyChoice]))

# Renames the variables
names(Cars)[1] = "Model"
names(Cars)[2] = "Count"

# Restricts the data to only models with at least 100 for sale
# Before submitting, comment this out so that it doesn't print while knitting
# Enough_Cars = subset(Cars, Count>=100)
# Enough_Cars

# Delete the ** below and enter the model that you chose from the Enough_Cars data.
ModelOfMyChoice = "Accord"

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

# Check to make sure that the cars span at least 5 years.
range(MyCars$Year)
```

```
## [1] 1998 2017
```

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

MODEL #1: Use Age as a predictor for Price

1. Calculate the least squares regression line that best fits your data. Interpret (in context) what the slope estimate tells you about prices and ages of your used car model. Explain why the sign (positive/negative) makes sense.

The least squares regression model for a linear regressions model for price on the age of Texas Honda Accords is yhat = -1408.64x + 21571.35. The negative slope makes sense because as the car's age increases, the price of the car is going to decrease.

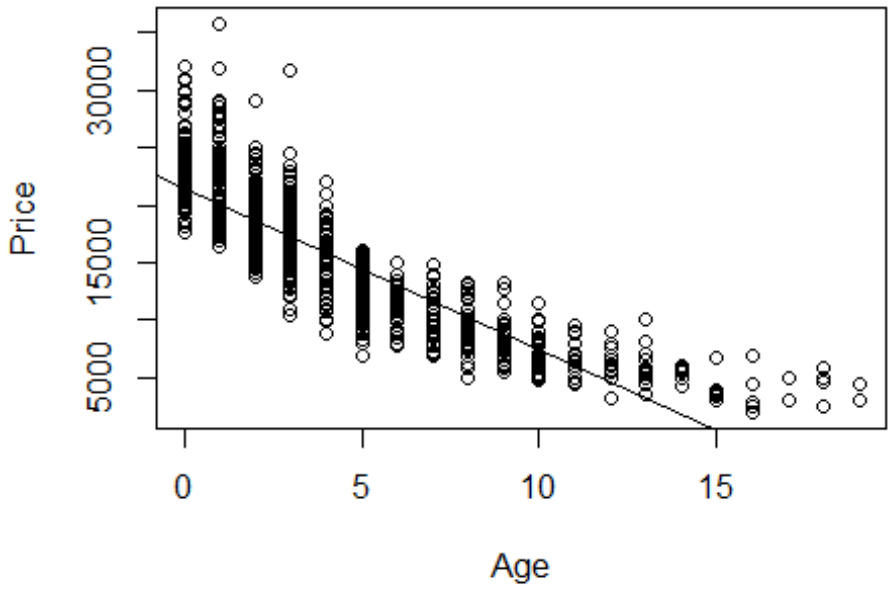
```r
mod1 <- lm(Price~Age, MyCars)
summary(lm(Price~Age, MyCars))
```

```
##
## Call:
## lm(formula = Price ~ Age, data = MyCars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7533.1 -1820.8  -376.3  1423.6 15619.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21571.35     114.33  188.68   <2e-16 ***
```

```
## Age         -1408.64      21.92  -64.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2816 on 1375 degrees of freedom
## Multiple R-squared:  0.7502, Adjusted R-squared:   0.75
## F-statistic:  4129 on 1 and 1375 DF,  p-value: < 2.2e-16
```

2. Produce a scatterplot of the relationship with the regression line on it.

```
plot(Price~Age, MyCars)
abline(mod1)
```

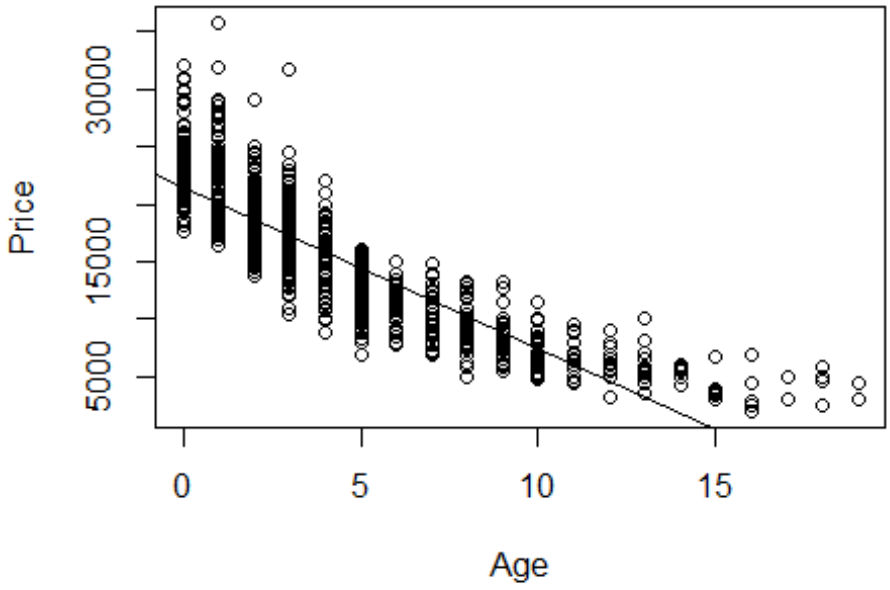


3. Produce appropriate residual plots and comment on how well your data appear to fit the conditions for a linear model. Don't worry about doing transformations at this point if there are problems with the conditions.

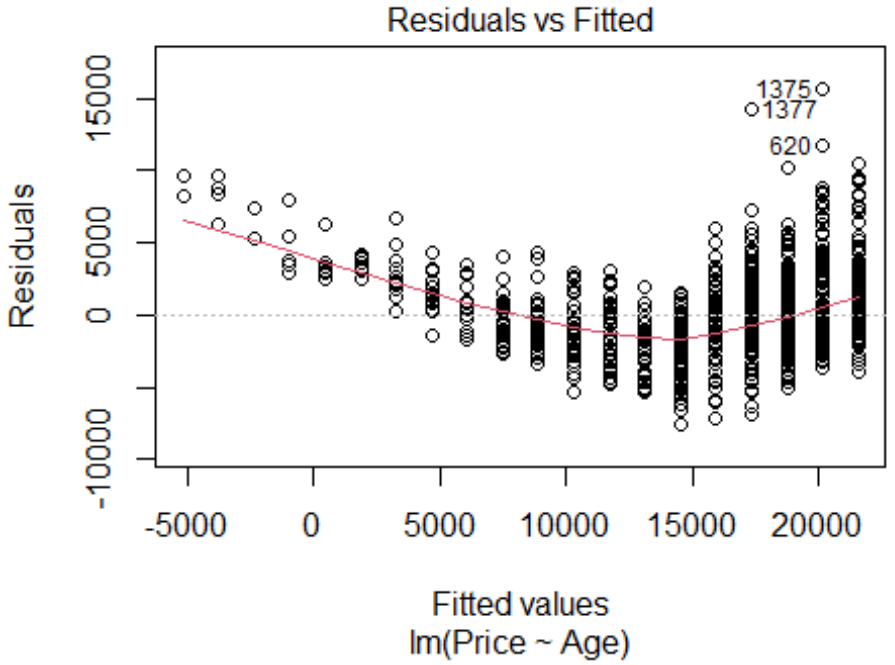
Inspecting the five conditions for a simple linear model: - Linearity: This data does not appear to fit a linear model very well. The data on the residuals vs fitted plot have an obvious left skew with a distinct curved pattern.
- Zero mean: According to the residual vs fitted plot, the residuals appear to be high on the low end and high on the extreme high end. Between about 7000 and 17000, the data appears to dip. I would say the zero mean condition is met since the residuals still hover around a 0 residual level.
- COnstant Variance: This model fails to satisfy the constant variance for the simple linear model condition. This is because there is an ovibous pattern to the residuals.

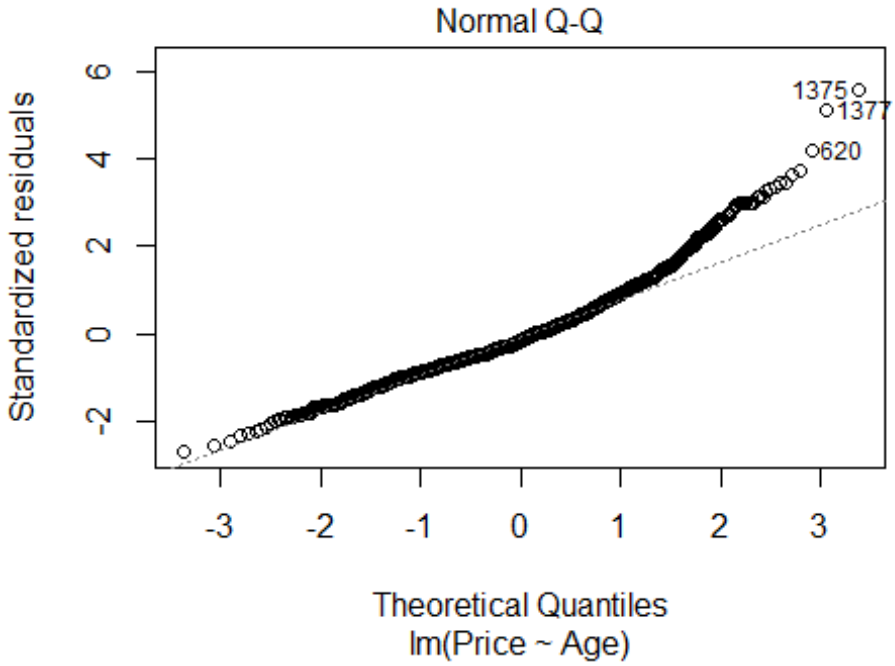
- Independence: I will assume that each car is independent of each other. - Normality: Looking at the histogram of the residuals, this method appears to have a right skew.

```
plot(Price~Age, MyCars)
abline(mod1)
```

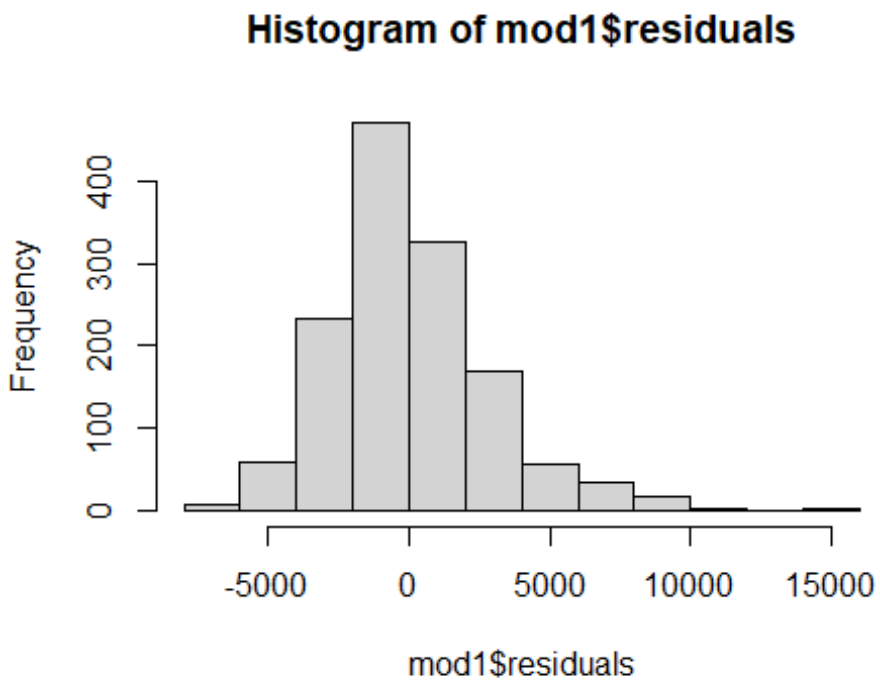



```
plot(mod1, 1:2)
```






```r
hist(mod1$residuals)
```



4. Find the car in your sample with the largest residual (in magnitude - positive or negative). For that car, find its standardized and studentized residual. Based on these residuals, could this value be considered influential?

The max residual of 15,619.29 is the largest residual in magnitude. Based on these residuals, this value is considered influential because it has a studentize and a standardized value of 5.6 or 5.5. The absolute value of these standardized and studentized values a re greater than 2, which should be the upper bound of the acceptable sutdentized and standardized values.

```
MyCars$residuals <- mod1$residuals
max(MyCars$residuals)

## [1] 15619.29

x <- which.max(MyCars$residuals)
MyCars[x,]

## # A tibble: 1 x 11
##       Id Price  Year Mileage City   State Vin        Make  Model   Age res
iduals
##    <dbl> <dbl> <dbl>   <dbl> <chr>  <chr> <chr>      <chr> <chr> <dbl>
<dbl>
## 1 795748 35782  2016    7000 Odessa TX    1HGCT2B03~ Honda Acco~     1
15619.

min(MyCars$residuals)
```

```
## [1] -7533.129
```

```
y <- which.min(MyCars$residuals)
MyCars[y,]
```

```
## # A tibble: 1 x 11
##       Id Price  Year Mileage City      State Vin     Make  Model   Age res
iduals
##    <dbl> <dbl> <dbl>   <dbl> <chr>     <chr> <chr>   <chr> <chr> <dbl>
<dbl>
## 1 382862  6995  2012  201002 Texarkana TX    1HGCP2~ Honda Acco~     5
-7533.
```

Finding the standardized and studentized residual:

```
rstudent(mod1)[x]
```

```
##     1375
## 5.611175
```

```
rstandard(mod1)[x]
```

```
##     1375
## 5.549988
```

5. Determine the leverages for the cars with the ten largest absolute residuals. What do these leverage values say about the potential for each of these ten cars to be influential on your model?

Looking at the residuals below, the 10 highest positive residuals are all absolutely greater than the 10 lowest negative residuals, so we will use the positive residuals.

```
test <- sort(MyCars$residuals, decreasing = TRUE)[1:10]
test # Gives you the row numbers for the highest postive residuals
```

```
##      1375      1377       620       594      1352      1290       683
1294
## 15619.293 14309.582 11736.293 10423.648 10240.937  9690.894  9684.249  942
7.648
##      1282      1374
##  9316.648  9228.648
```

```
test2 <- sort(MyCars$residuals, decreasing = FALSE)[1:10]
test2 # Gives you the row numbers for the lowest negative residuals
```

```
##       673       185       919       982        73      1006       496
804
## -7533.129 -7131.774 -6895.418 -6533.129 -6350.418 -6228.129 -6106.774 -594
1.774
##       463       508
## -5722.129 -5537.129
```

```
test
```

```
##      1375      1377       620       594      1352      1290       683
1294
## 15619.293 14309.582 11736.293 10423.648 10240.937  9690.894  9684.249  942
7.648
##      1282      1374
##  9316.648  9228.648

test2

##       673       185       919       982        73      1006       496
804
## -7533.129 -7131.774 -6895.418 -6533.129 -6350.418 -6228.129 -6106.774 -594
1.774
##       463       508
## -5722.129 -5537.129
```

The values below calculate the leverage for the high residual values we want to check to see if they are influential.

```
# mod1[[2]][c(1375,1377,620,594,1352, 1290, 683, 1294, 1282, 1374)]
# The above code just checks to make sure I'm calling the correct values
learning <- hatvalues(mod1)
chosen <- learning[c(1375,1377,620,594,1352, 1290, 683, 1294, 1282, 1374)]
```

The code below compares the chosen hatvalues from mod1 to leverage. The below inequality tells us that values from row 1290 and 683 have high leverage on the model. This tells us that of the 10 highest residuals in our model, only 2 have high leverage on the model.

```
lowlev <- 2*(2/nrow(MyCars)) # Low Leverage
highlev <- 3*(2/nrow(MyCars)) # High Leverage

# Low Leverage
chosen > lowlev

##  1375  1377   620   594  1352  1290   683  1294  1282  1374
## FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE

chosen > highlev

##  1375  1377   620   594  1352  1290   683  1294  1282  1374
## FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
```

6. Determine the Cook's distances for the cars with the ten largest absolute residuals. What do these Cook's distance values say about the influence of each of these ten cars on your model?

The low cook's distance values tell me that these values do not have high influence on the regression model. This means that the points have leverage, meaning they stand out a bit from the observed points, but they do not have high influence, meaning the removal of these values would result in a big change in teh regression model.

```
cookmod1 <- cooks.distance(mod1)
cookmod1[c(1375,1377,620,594,1352, 1290, 683, 1294, 1282, 1374)]
```

```
##        1375        1377         620         594        1352        1290
## 0.019060188 0.010025926 0.010761339 0.011328593 0.006261439 0.088674402
##         683        1294        1282        1374
## 0.077499746 0.009267086 0.009050152 0.008879994
```

```
cookmod1[c(1375,1377,620,594,1352, 1290, 683, 1294, 1282, 1374)] > 1
```

```
##  1375  1377   620   594  1352  1290   683  1294  1282  1374
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

7. Compute and interpret in context a 90% confidence interval for the slope of your regression line.

RESPONSE: On average, with a 1 year increase in age the price should decrease by 1444.727 to 1372.56 dollars. In other words, we are 90% confident with a 1 year increase in the age of a Texas Honda Accord, the price will decrease somewhere between 1444.72 to 1372.56 dollars.

```
# Personal Reminder: lm(Price~Age, MyCars)
confint(mod1, level = .90)
```

```
##                   5 %      95 %
## (Intercept) 21383.177 21759.527
## Age         -1444.727 -1372.562
```

8. Test the strength of the linear relationship between your variables using each of the three methods (test for correlation, test for slope, ANOVA for regression). Include hypotheses for each test and your conclusions in the context of the problem.

Test for correlation

RESPONSE Pt 1: This tells us that, of the MyCars data, there is a strong, negative correlation between age and price of a Texas Honda Accord.

```
cor(MyCars[c(2,10)])
```

```
##            Price        Age
## Price  1.0000000 -0.8661308
## Age   -0.8661308  1.0000000
```

Test for Slope

RESPONSE Pt 2: - Null Hypothesis: p = 0 - Alternative Hypothesis: p /=/ 0

Based on the p value of the correlation test, there is a really low chance that we would get this result by chance. As a result of the small p-value, we have evidence to suggest to reject the null hypothesis.

```
cor.test(MyCars$Price, MyCars$Age)
```

```
##
##  Pearson's product-moment correlation
##
## data:  MyCars$Price and MyCars$Age
## t = -64.257, df = 1375, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8787504 -0.8523011
## sample estimates:
##       cor
## -0.8661308
```

ANOVA for regression test

RESPONSE Pt 3:

- Null Hypothesis: p = 0
- Alternative Hypothesis: p /=/ 0

Key Question: Does the model on the price of Honda accords based on their age explain a "significant" amount of the total variability?

Because the f-test statistic is relatively small, not as much variability are being explained by the model. Based on the f-distribution, it's unlikely we would get this result by chance.

```
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    4129 < 2.2e-16 ***
## Residuals 1375 1.0904e+10 7.9300e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The below are just for my own notes and practice SSModel <- anova(mod1)[1,2] SSE <- anova(mod1)[2,2] MSModel <- SSModel/1 MSE <- SSE/(nrow(MyCars)-2) t.s <- MSModel/MSE t.s

MSModel MSE t.s

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

We are 90% confident that the true mean price for the population of 4 year old texas honda accords is somewhere between 15787.84 and 16085.70 dollars. We are 90% confident that the true mean price for a single 4 year old texas honda accord is somewhere between

10410.85 and 21462.97 dollars. 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 = 0.90)
```

```
##                   5 %      95 %
## (Intercept) 21383.177 21759.527
## Age         -1444.727 -1372.562
```

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

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

```
predict.lm(mod1, newx, interval="confidence") # All peope; for poualtion
```

```
##        fit      lwr     upr
## 1 15936.77 15787.84 16085.7
```

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

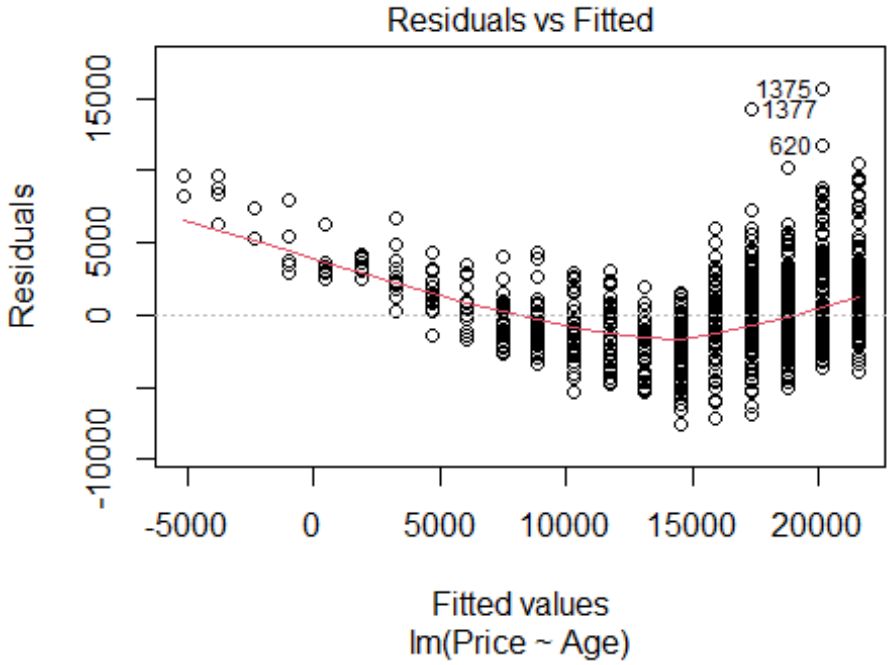
```
##        fit      lwr      upr
## 1 15936.77 10410.58 21462.97
```

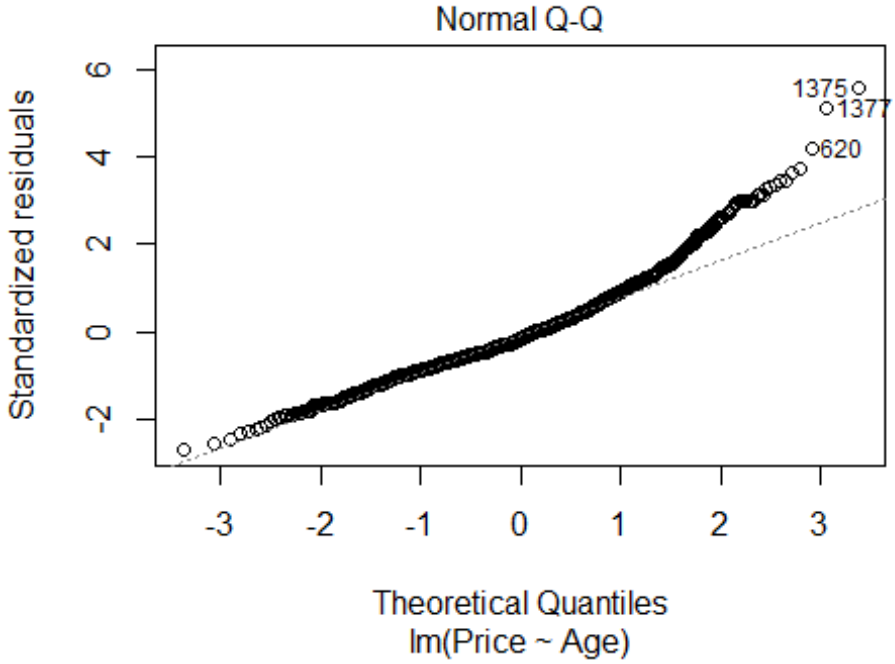
10. Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain why you think that this transformation does or does not improve satisfying the linear model conditions.

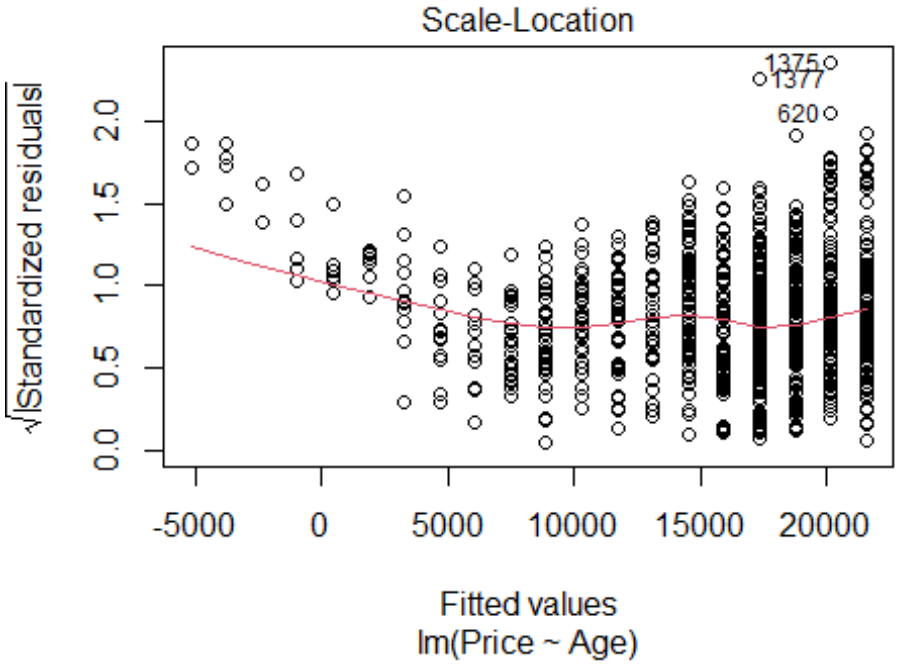
The square root age transformation appears to be the best transformation of the data. First, the scatter plot of the data with the regression line on the orginial data fits better because the square root transformation follows the data all the way the the right side of the graph (see the code after all these graphs). Second, the transformed data's fitted risidaul plot is much more centered around zero than the non transfomred data. Third, the transformed data follows the qqnorm plot more closely than the non transformed dat, with less deviation from the tails.

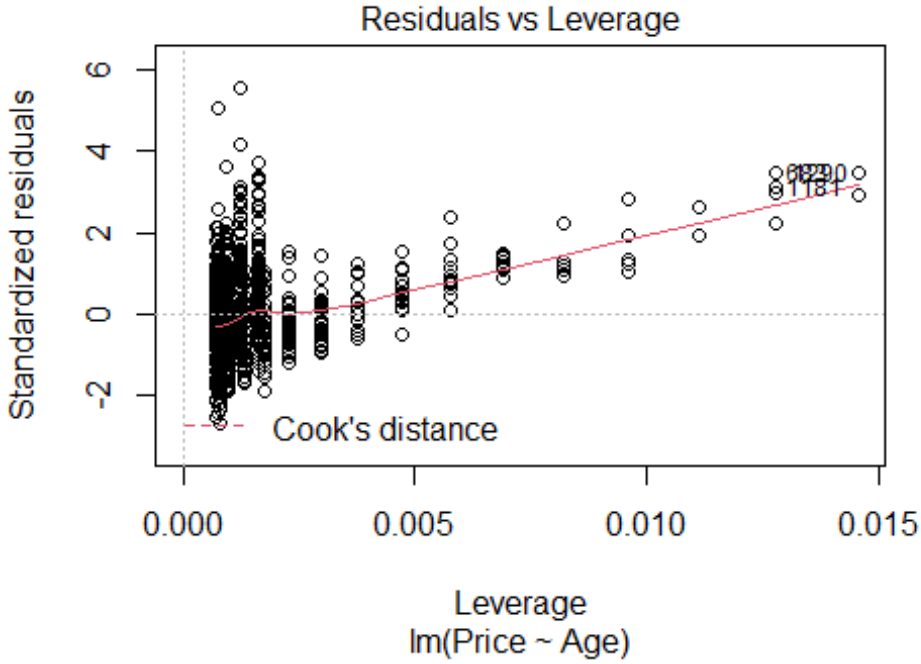
Lastly, the residual histogram plot of both of the transformed data and the non transformed data appear roughly the same. They are both right skew.

```
plot(mod1)
```

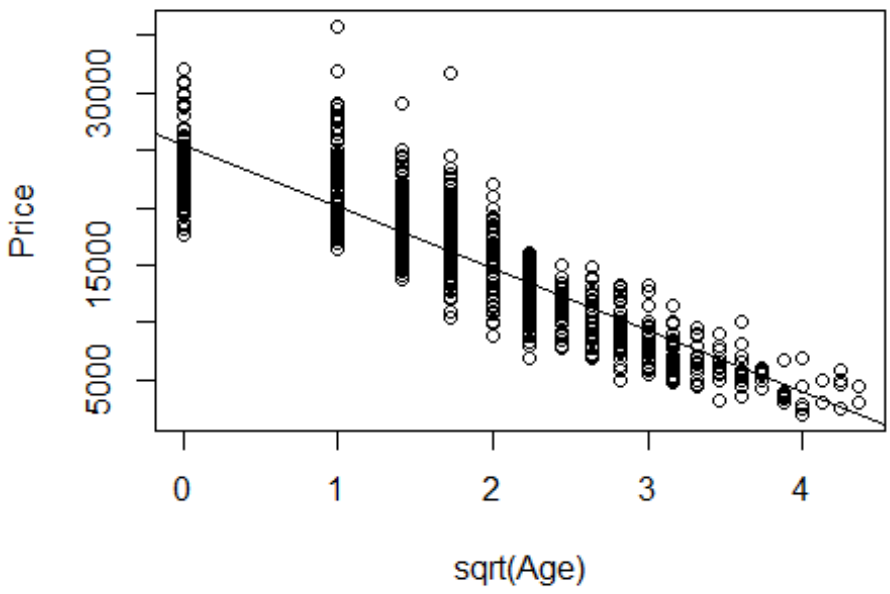

Residuals vs Fitted
Residuals
Fitted values
lm(Price ~ Age)
1375
1377
620
-10000
-5000
0
5000
10000
15000
20000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ Age)
1375
1377
620
-2
0
2
4
6
-3
-2
-1
0
1
2
3

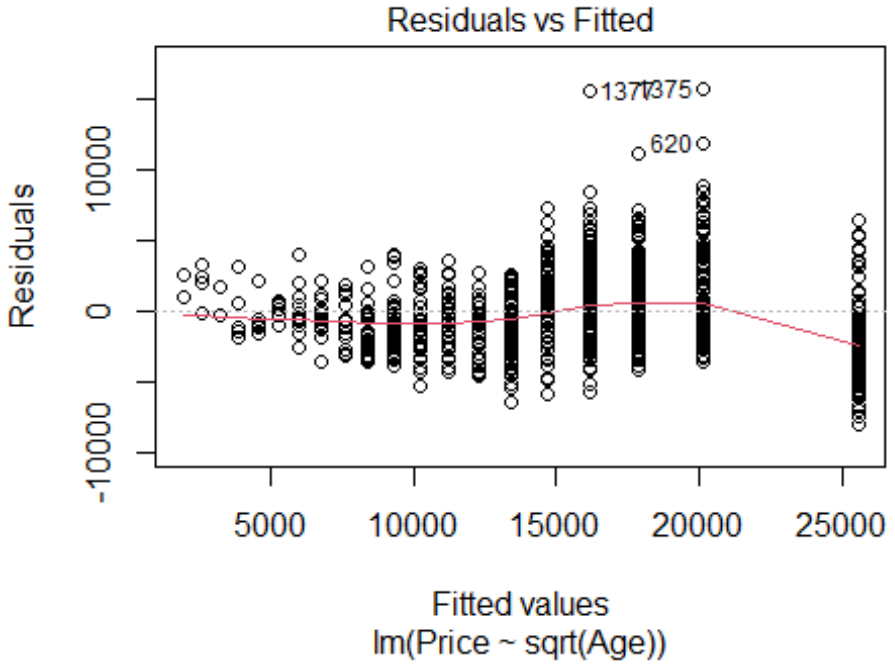


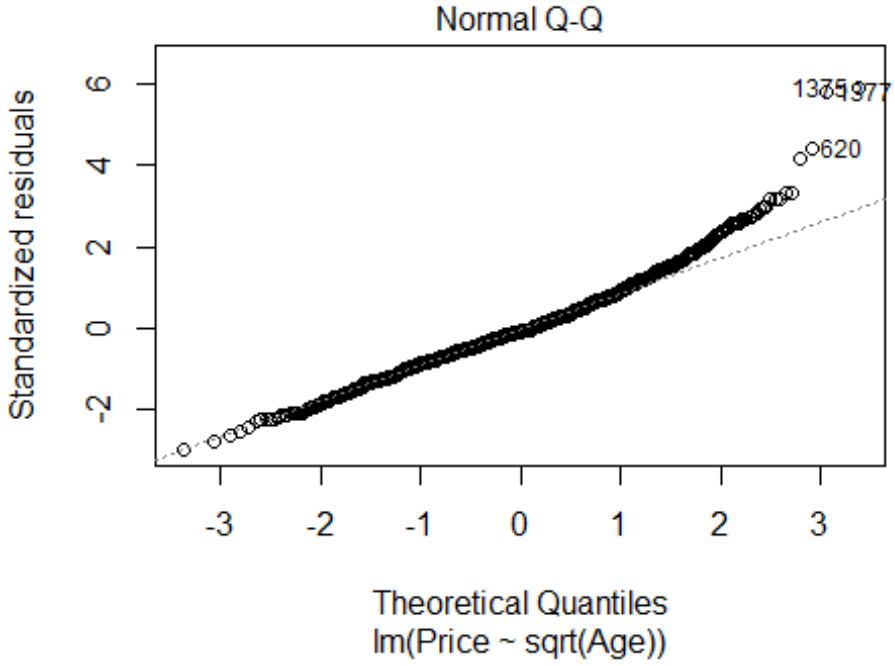


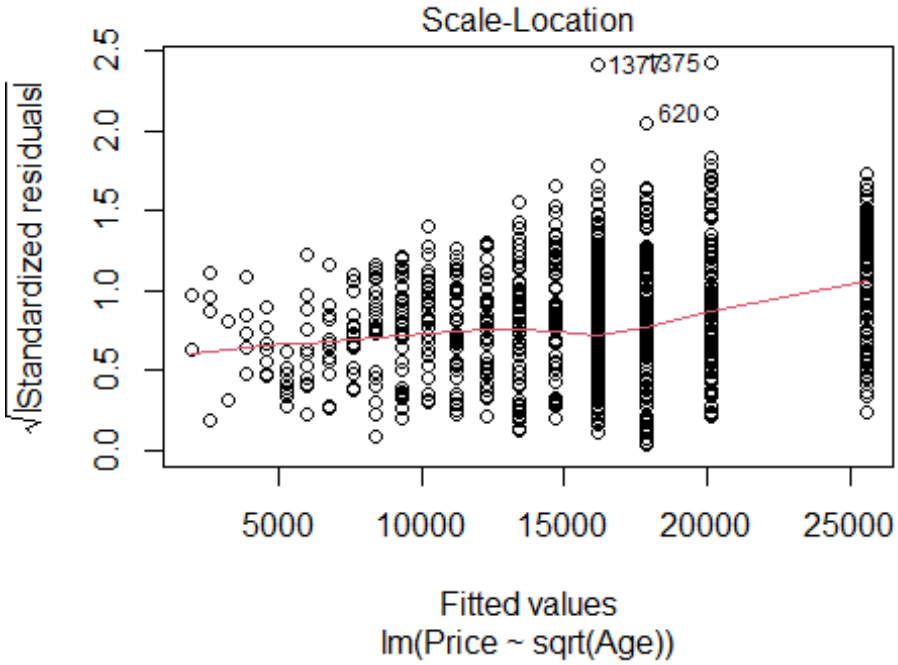

```r
# square root Age
plot(Price~sqrt(Age), data = MyCars)
abline(lm(Price~sqrt(Age), data = MyCars))
```

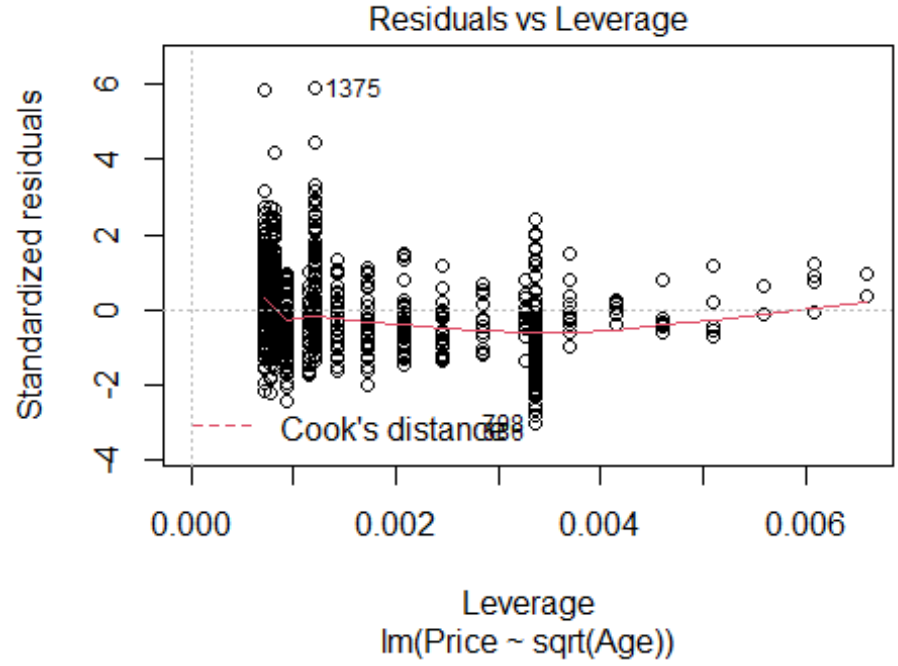



```
plot(lm(Price~sqrt(Age), data = MyCars))
```

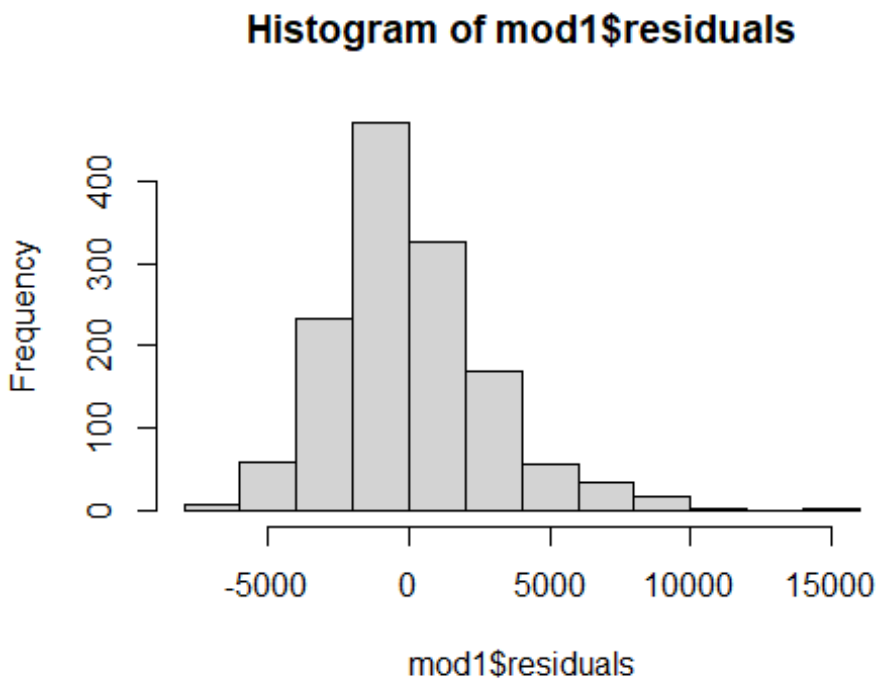

Residuals vs Fitted
Residuals
Fitted values
lm(Price ~ sqrt(Age))
620


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ sqrt(Age))
620

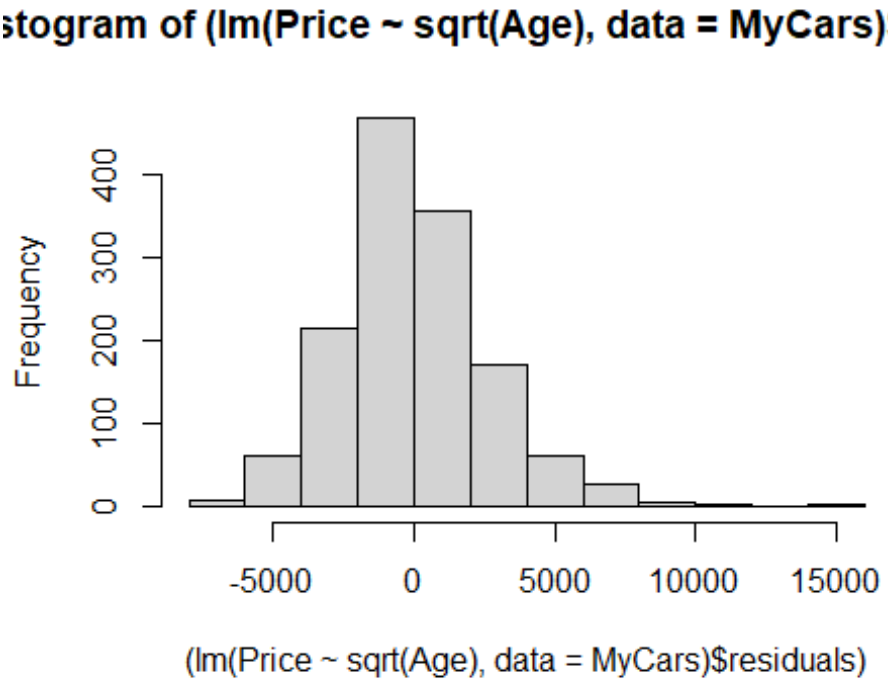





```
hist(mod1$residuals)
```

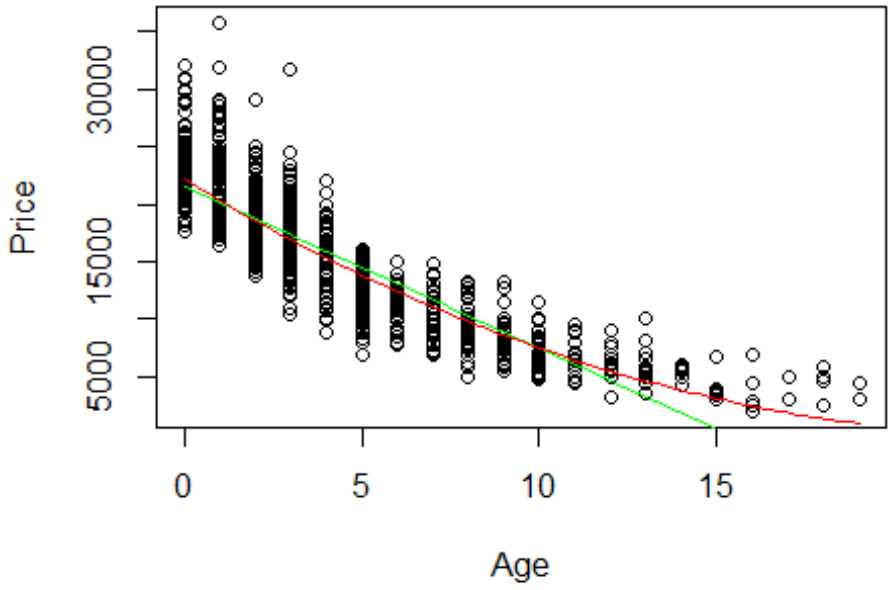



```r
hist((lm(Price~sqrt(Age), data = MyCars)$residuals))
```

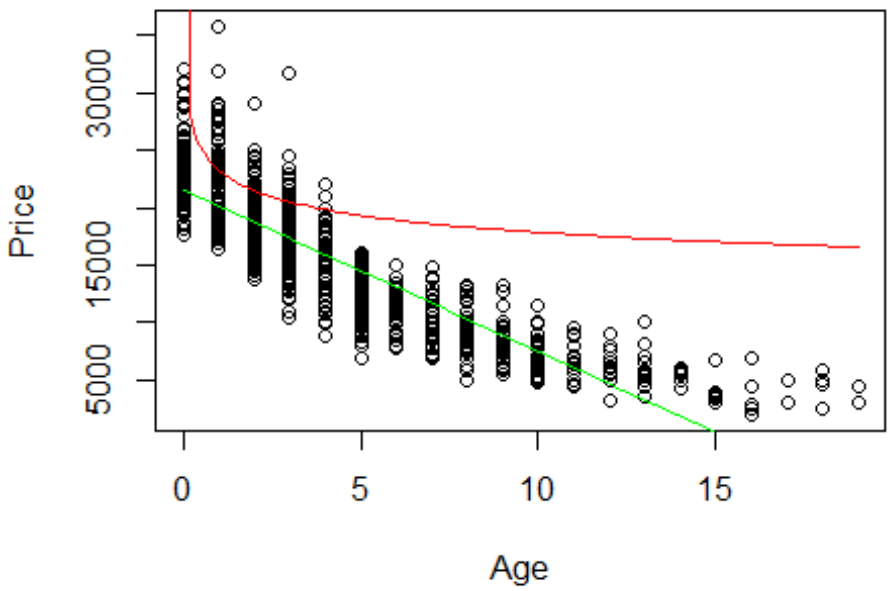



```r
transformmod <- lm(Price~sqrt(Age), data = MyCars)
```

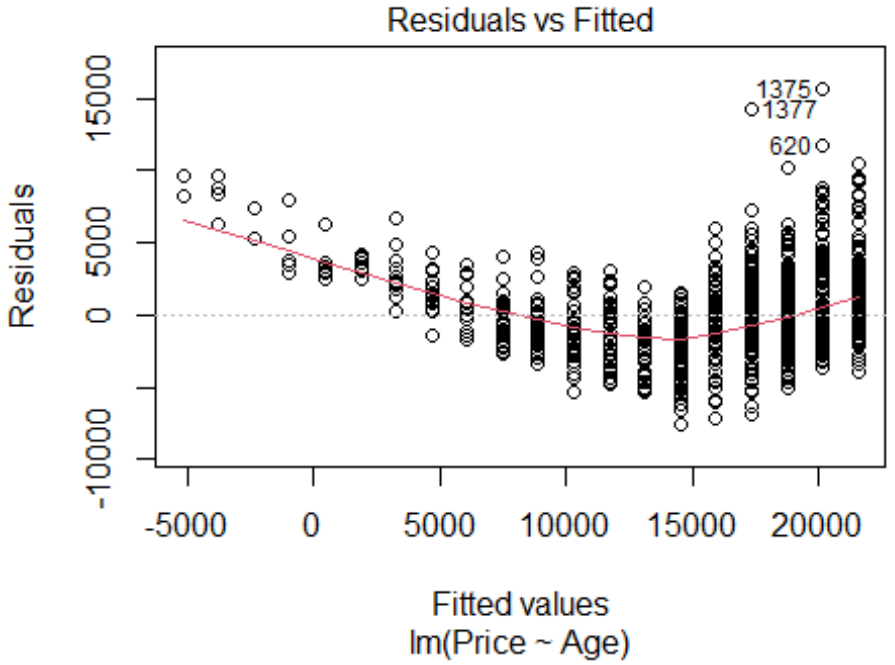
```r
# With square root age
B0.2 = summary(lm(Price~Age, MyCars))$coefficients[1,1] # Intercept
B1.2 = summary(lm(Price~Age, MyCars))$coefficients[2,1] # Slope

B0.2.trans = summary(lm(sqrt(Price)~Age, data = MyCars))$coefficients[1,1] #
Intercept
B1.2.trans = summary(lm(sqrt(Price)~Age, data = MyCars))$coefficients[2,1] #
Slope

plot(Price~Age, MyCars)
curve(B1.2*x+B0.2, col = "green", add=TRUE)
curve((B1.2.trans*x+B0.2.trans)^2, col = "red", add=TRUE)
```

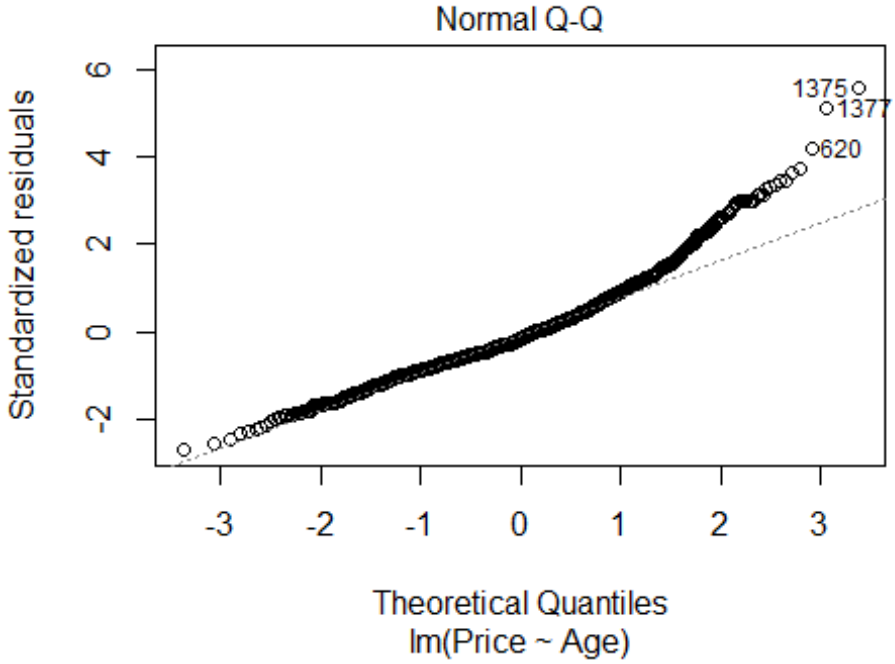


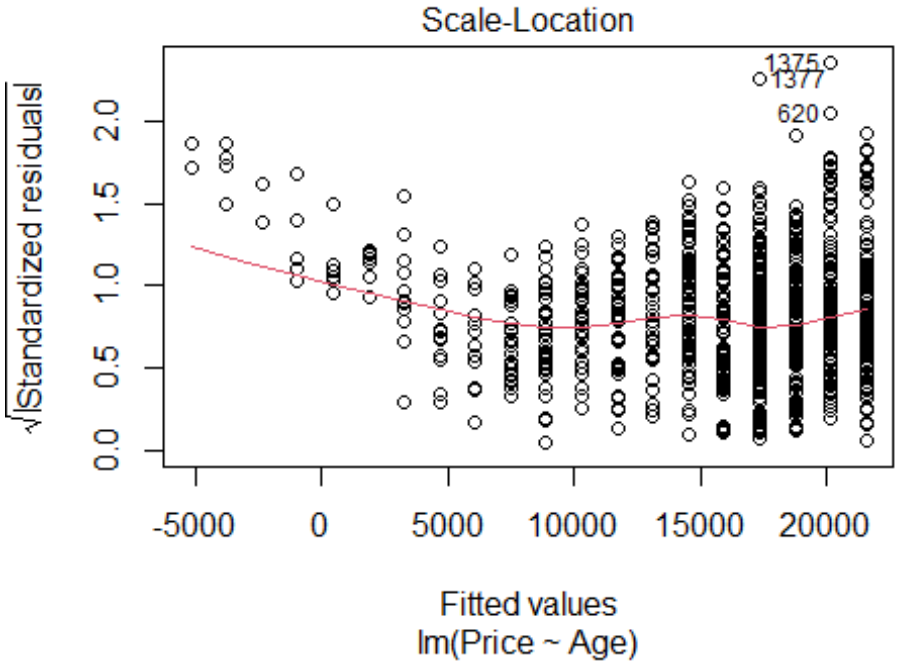

```r
# Trying with Log
B0.1 = summary(lm(Price~Age, MyCars))$coefficients[1,1] # Intercept
B1.1 = summary(lm(Price~Age, MyCars))$coefficients[2,1] # Slope

B0.1.trans = summary(lm(log(Price)~Age, data = MyCars))$coefficients[1,1] # I
ntercept
B1.1.trans = summary(lm(log(Price)~Age, data = MyCars))$coefficients[2,1] # S
lope

plot(Price~Age, MyCars)
curve(B1.1*x+B0.1, col = "green", add=TRUE)
curve(exp(B0.1.trans)*x^B1.1.trans, col = "red", add=TRUE)
```

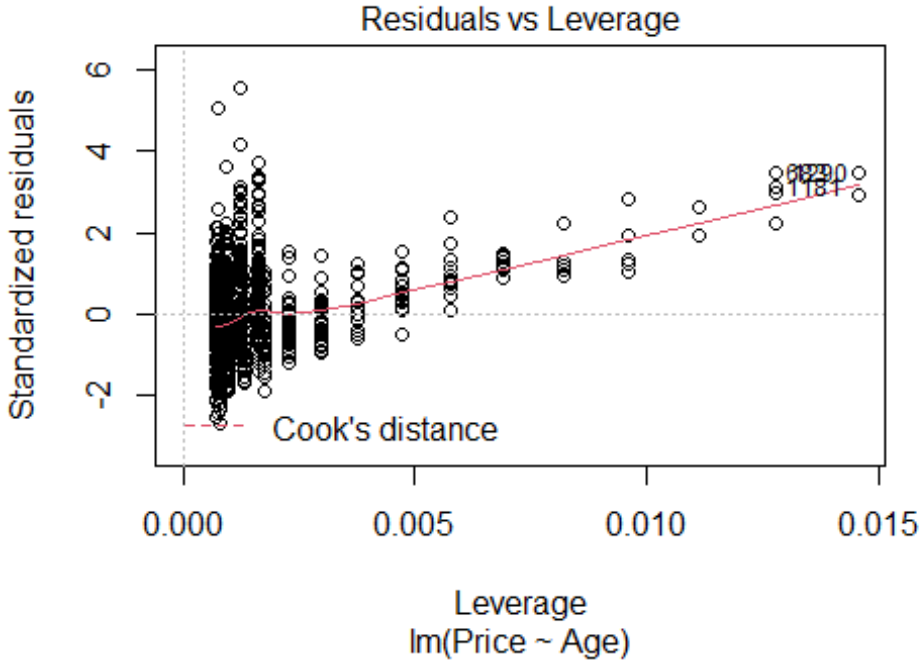



```
#Original
plot(mod1)
```

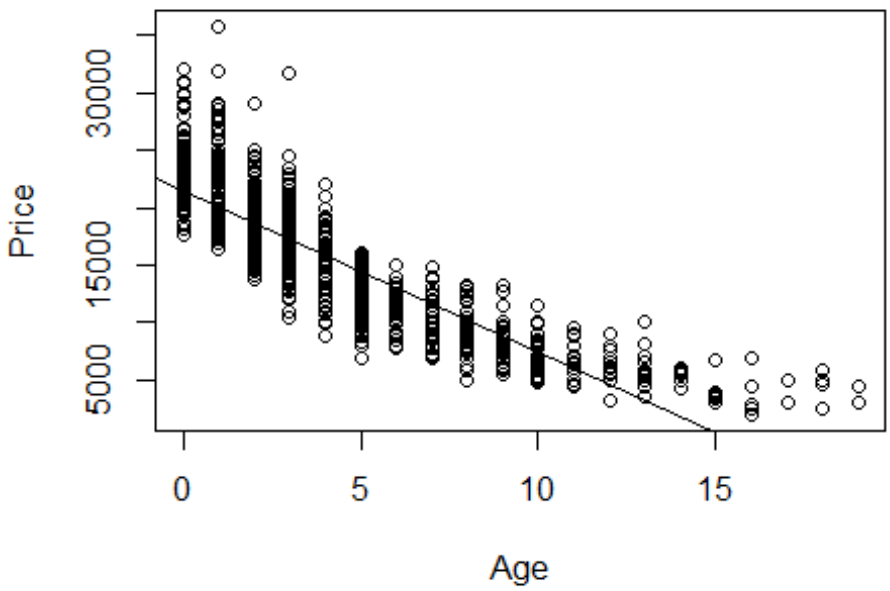

Residuals vs Fitted
Residuals
Fitted values
lm(Price ~ Age)
1375
1377
620
-10000
-5000
0
5000
10000
15000
20000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Price ~ Age)
1375
1377
620
-2
0
2
4
6
-3
-2
-1
0
1
2
3

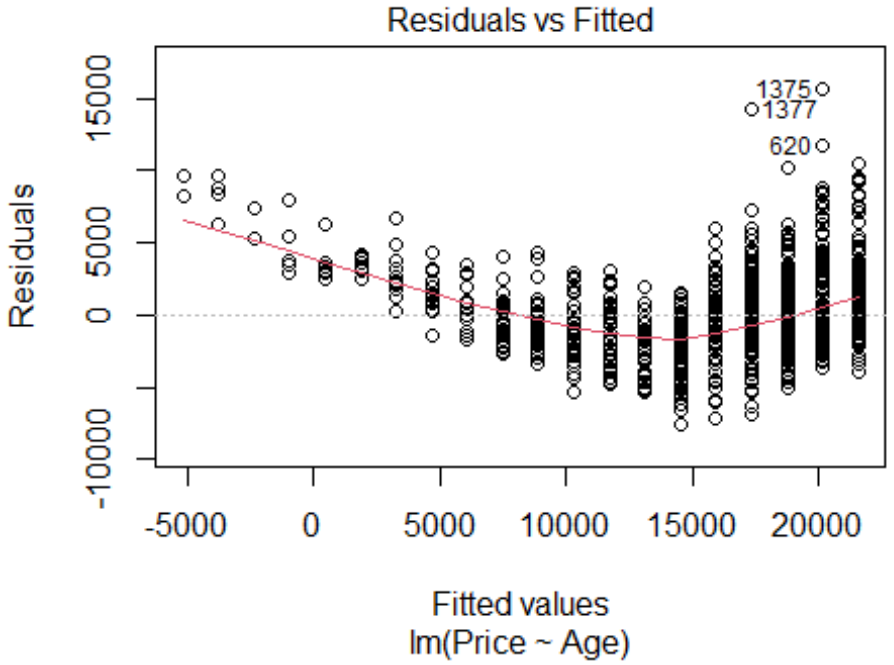


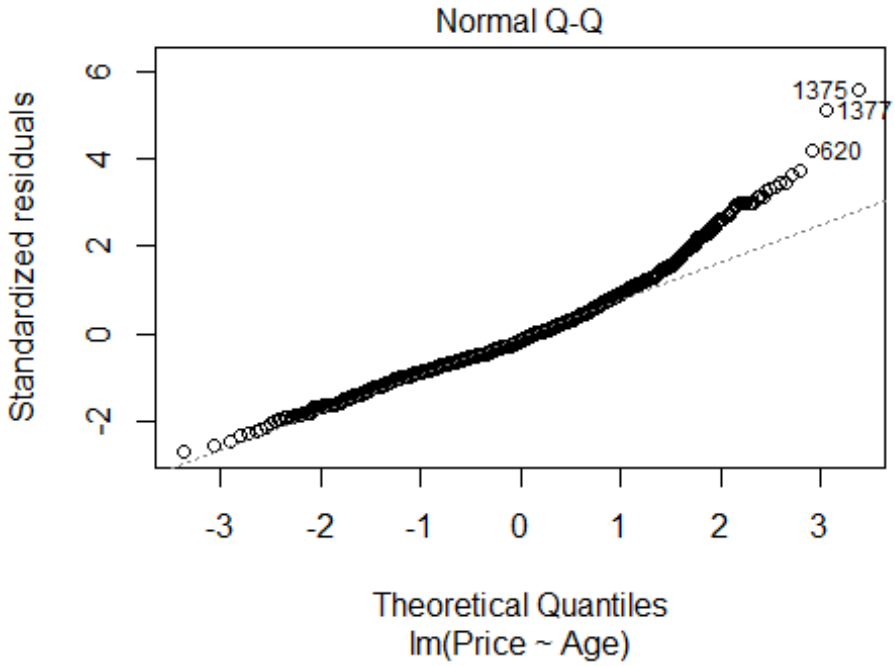



```r
plot(Price~Age, data = MyCars) # This is the original
abline(lm(Price~Age, data = MyCars))
```

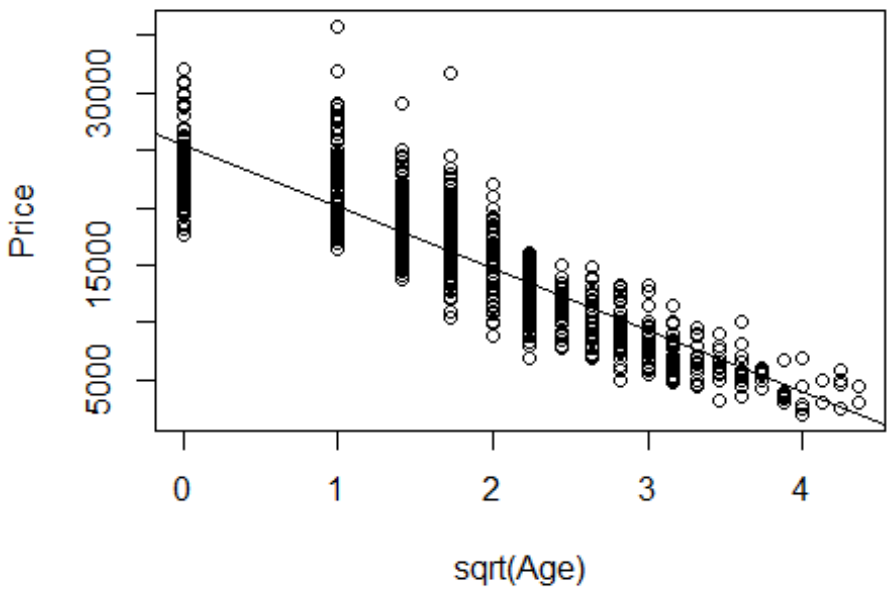



```
plot(lm(Price~Age, data = MyCars), 1:2)
```

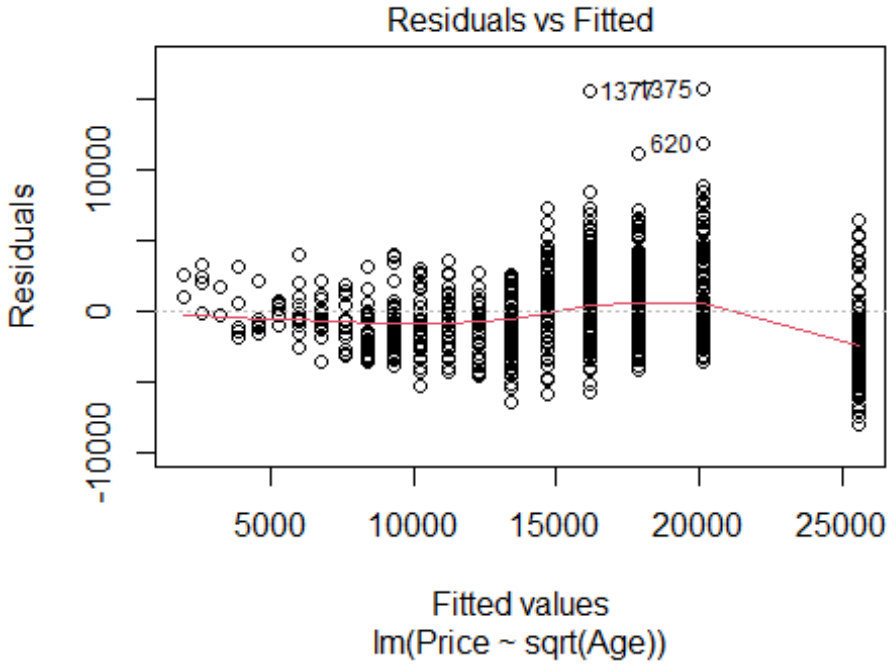


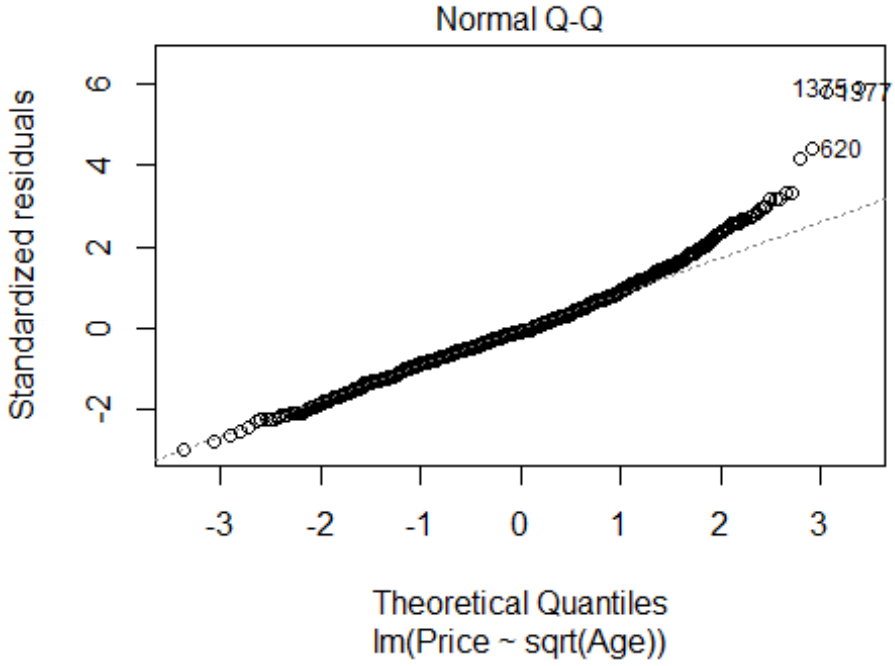



```r
# square root Age
plot(Price~sqrt(Age), data = MyCars)
abline(lm(Price~sqrt(Age), data = MyCars))
```

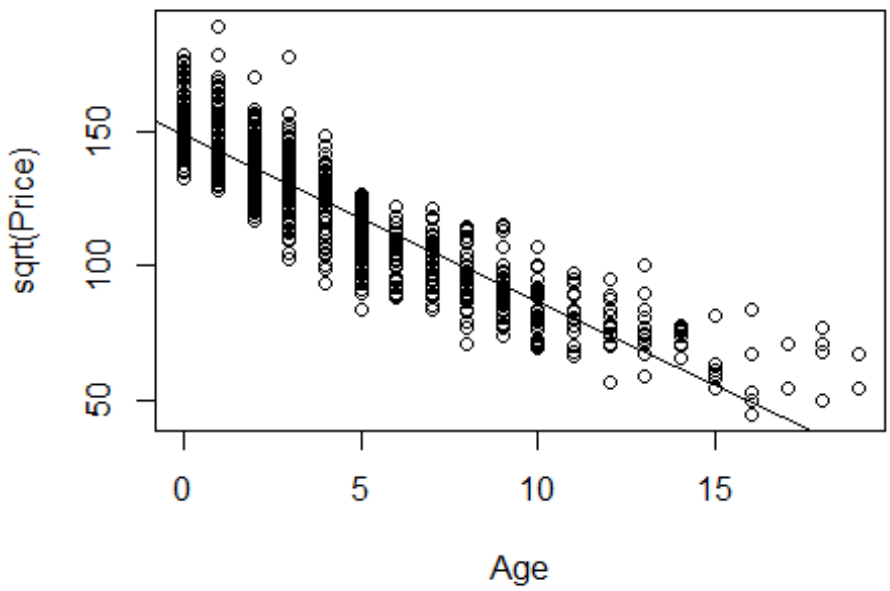



```
plot(lm(Price~sqrt(Age), data = MyCars), 1:2)
```

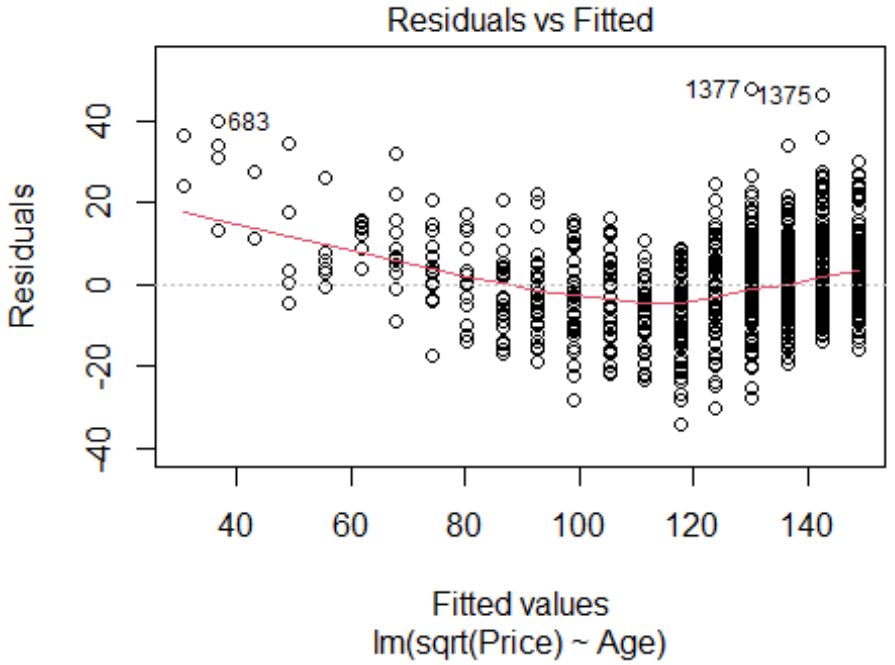


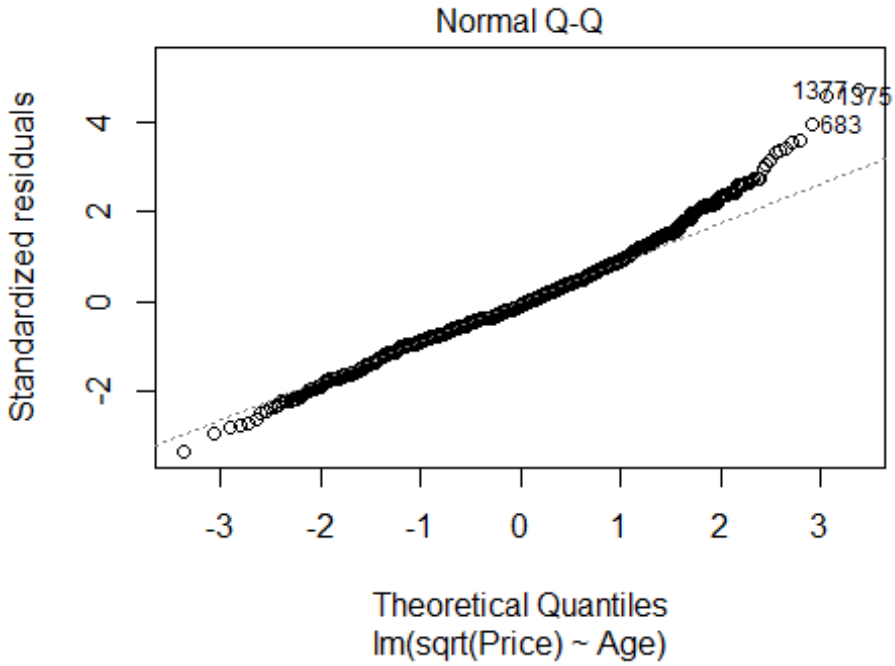



```r
# Square root Price
plot(sqrt(Price)~Age, data = MyCars)
abline(lm(sqrt(Price)~Age, data = MyCars))
```

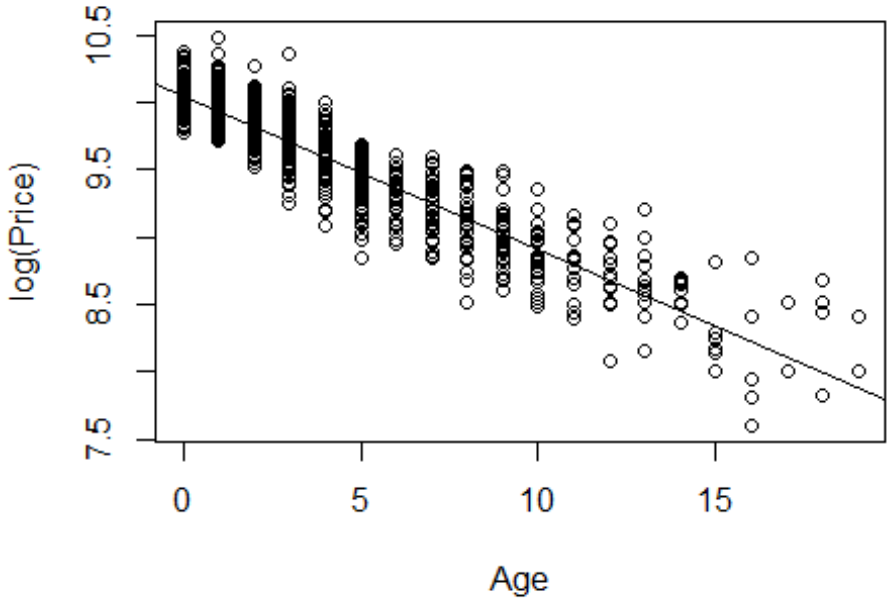



```
plot(lm(sqrt(Price)~Age, data = MyCars), 1:2)
```

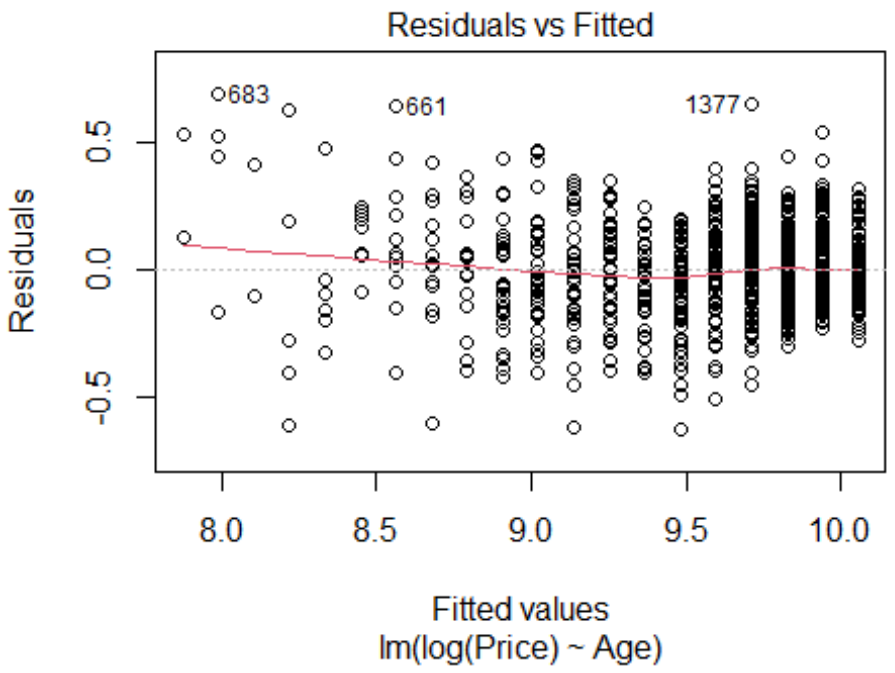


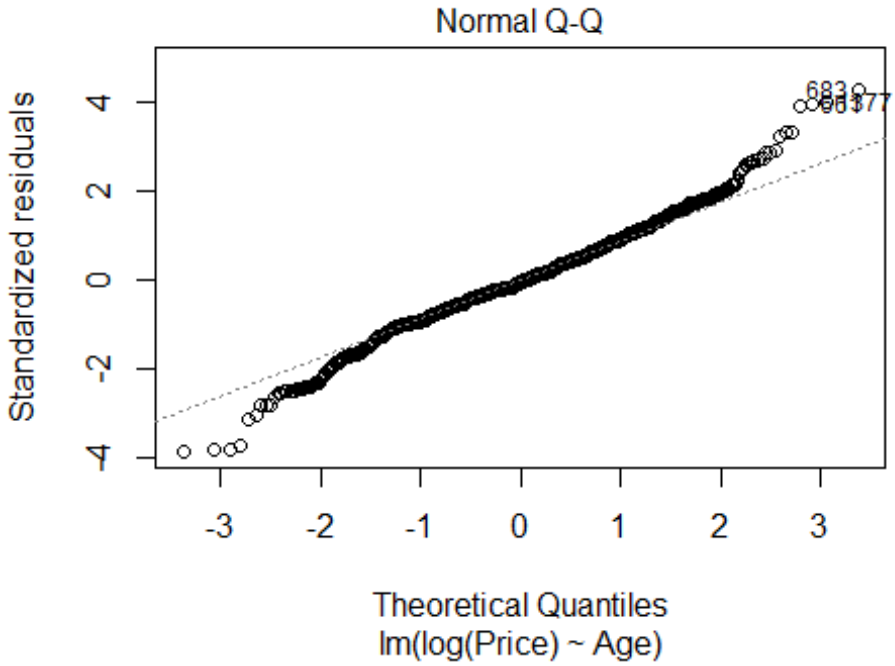



```r
# Log Price
plot(log(Price)~Age, data = MyCars)
abline(lm(log(Price)~Age, data = MyCars))
```

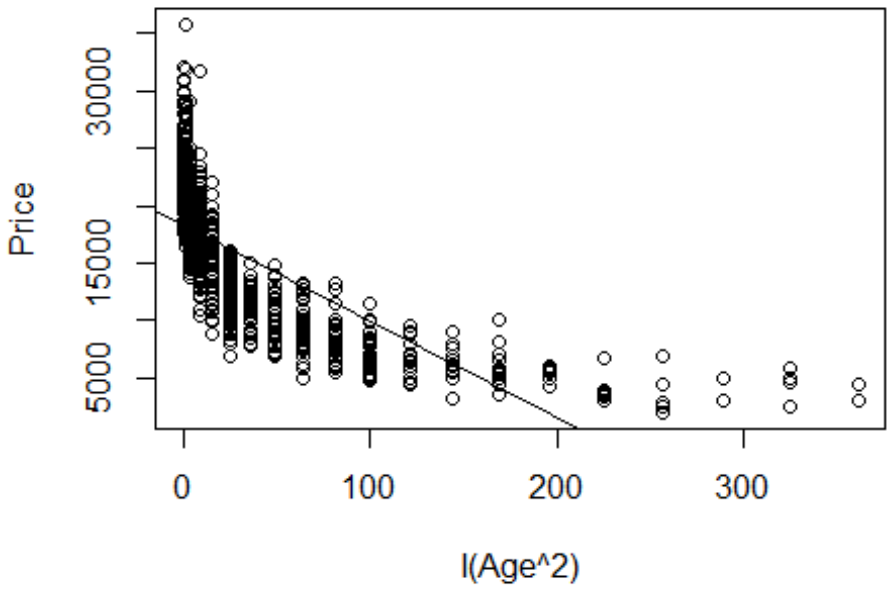



```
plot(lm(log(Price)~Age, data = MyCars), 1:2)
```

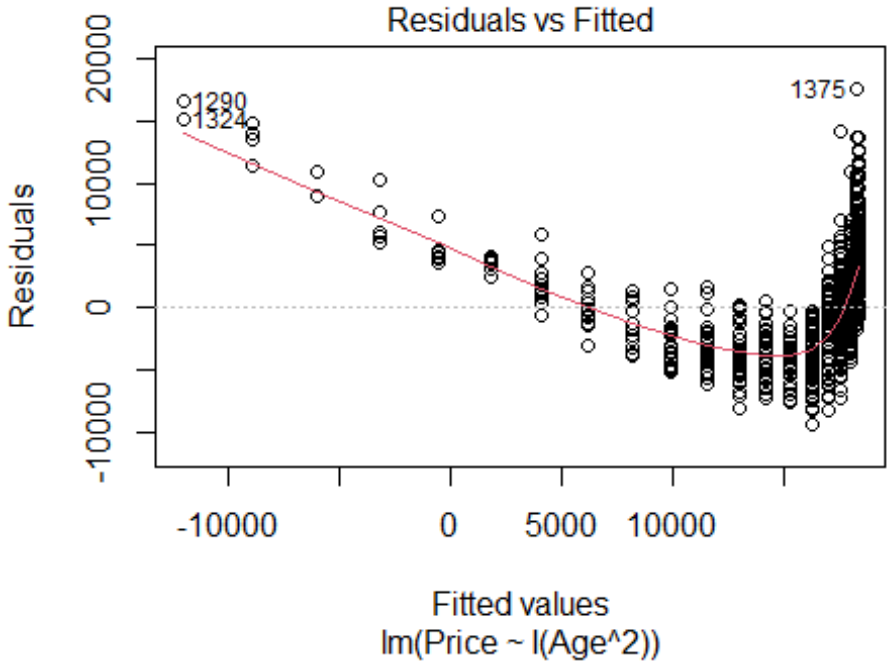


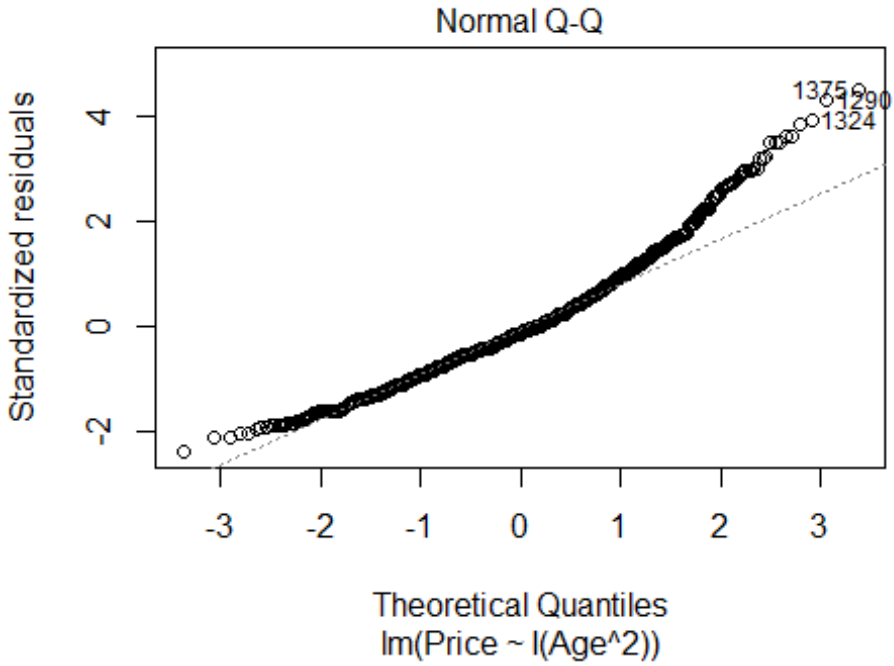



```r
# Add Exponent to Age
plot(Price~I(Age^2), data = MyCars)
abline(lm(Price~I(Age^2), data = MyCars))
```

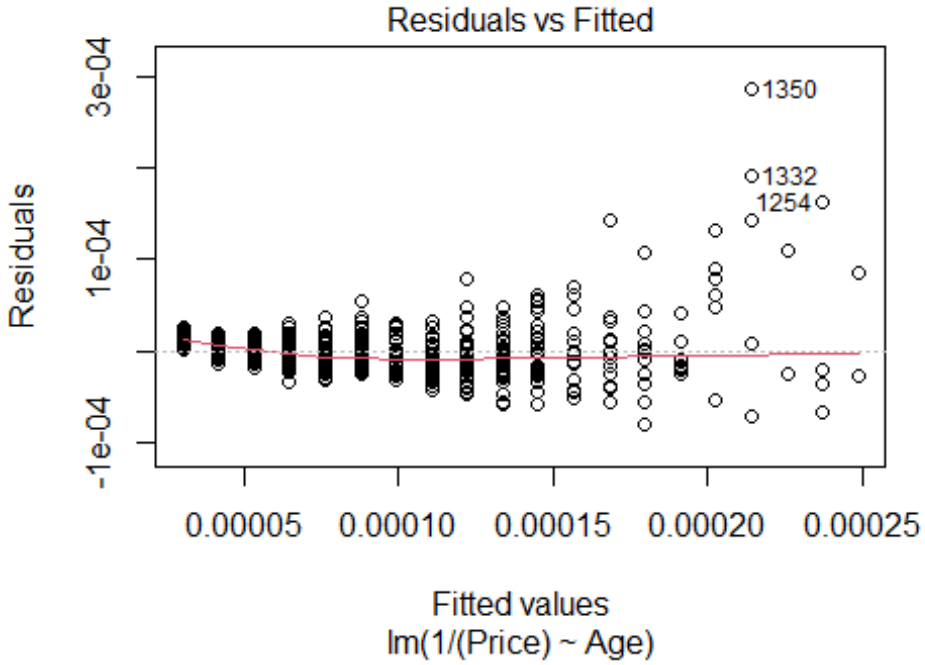


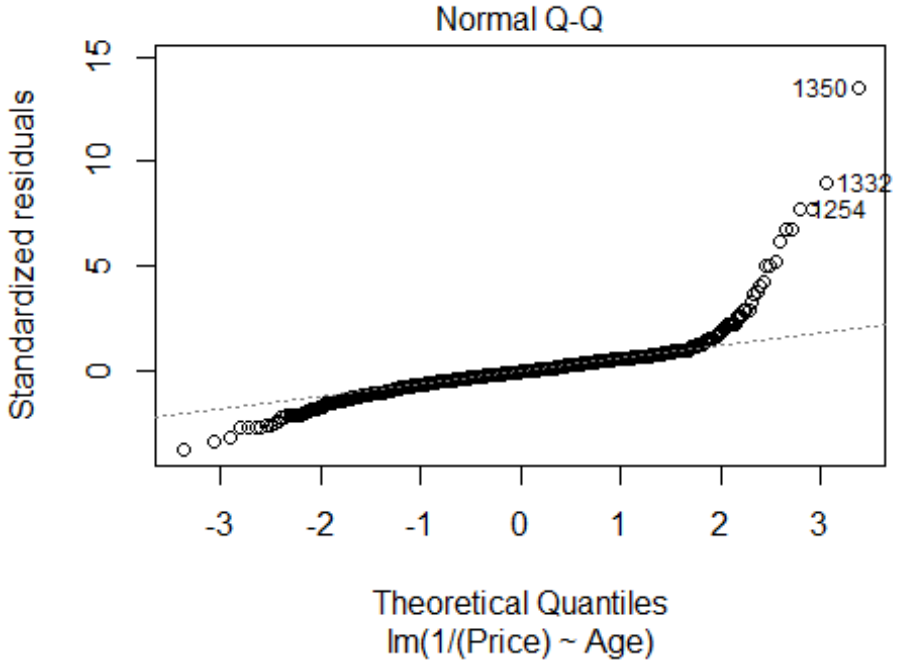

```
plot(lm(Price~I(Age^2), data = MyCars), 1:2)
```

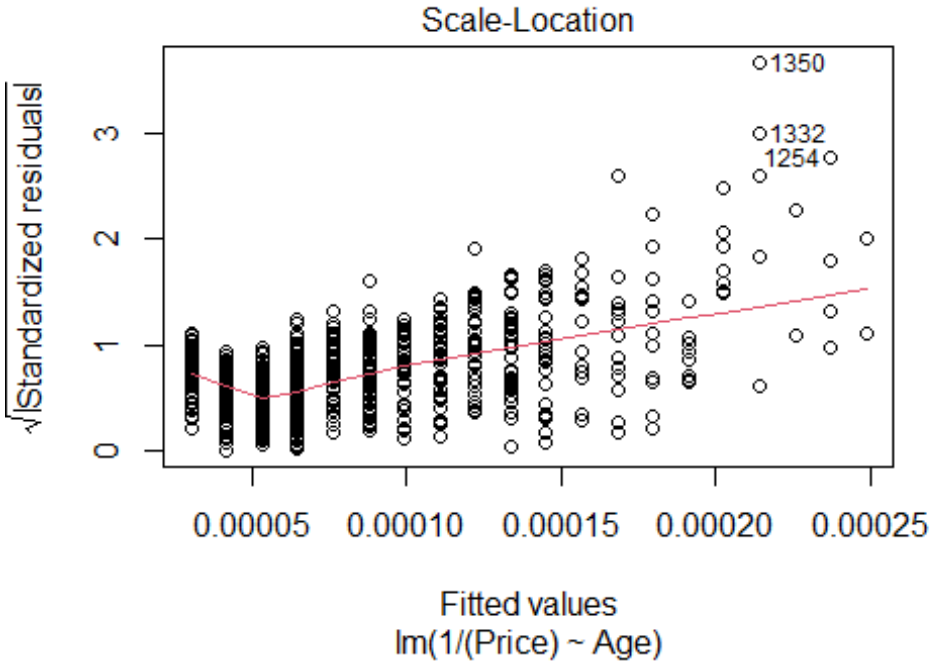




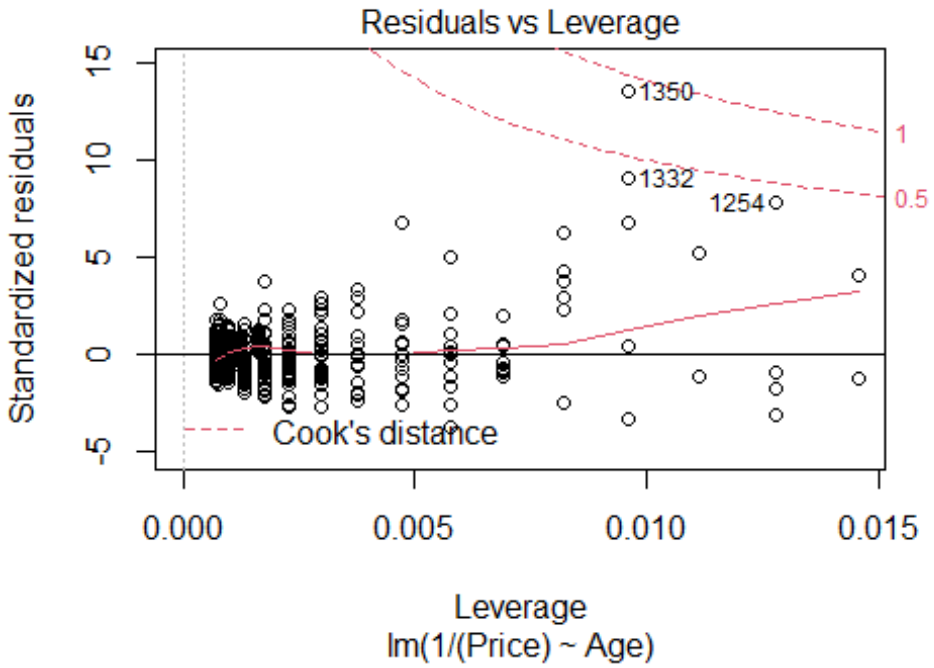

```r
# Division Price
plot(lm(1/(Price)~Age, data=MyCars))
```


Residuals vs Fitted
Residuals
3e-04
1e-04
-1e-04
1350
1332
1254
0.00005
0.00010
0.00015
0.00020
0.00025
Fitted values
lm(1/(Price) ~ Age)

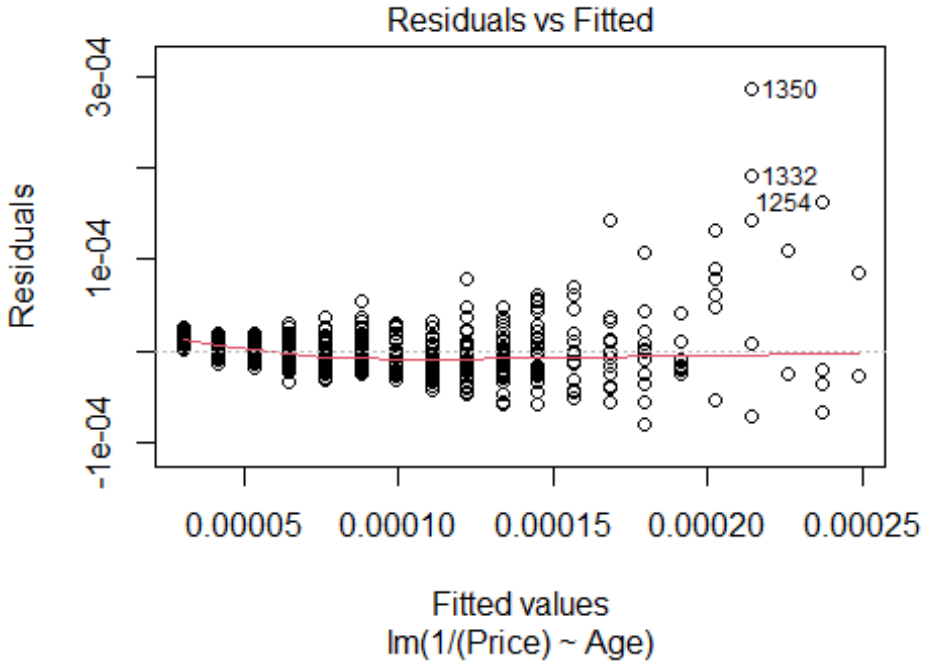

Normal Q-Q
Standardized residuals
15
10
5
0
1350
1332
1254
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(1/(Price) ~ Age)

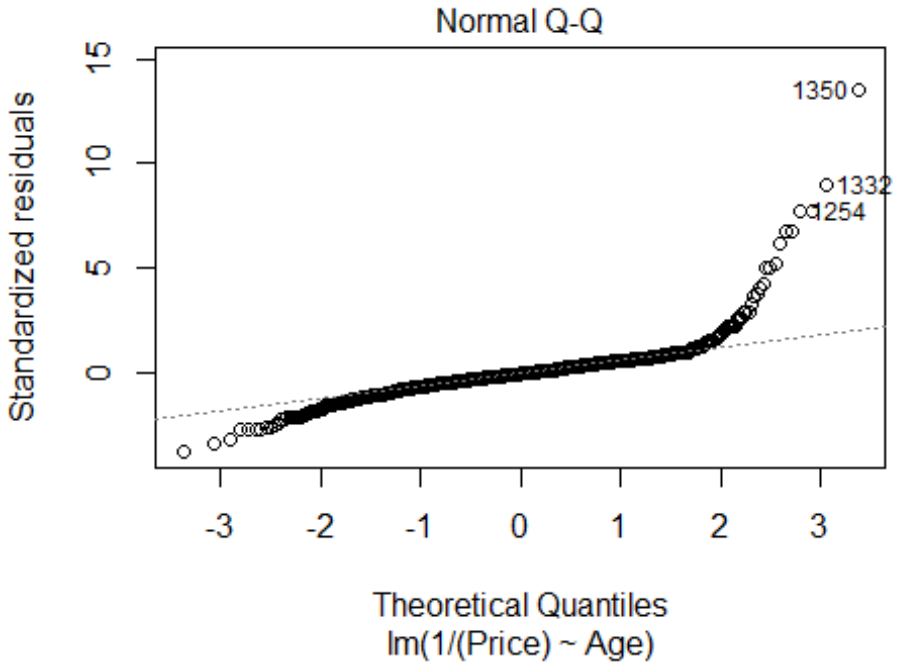



```
abline(lm(1/(Price)~Age, data=MyCars))
```




```
plot(lm(1/(Price)~Age, data=MyCars), 1:2)
```

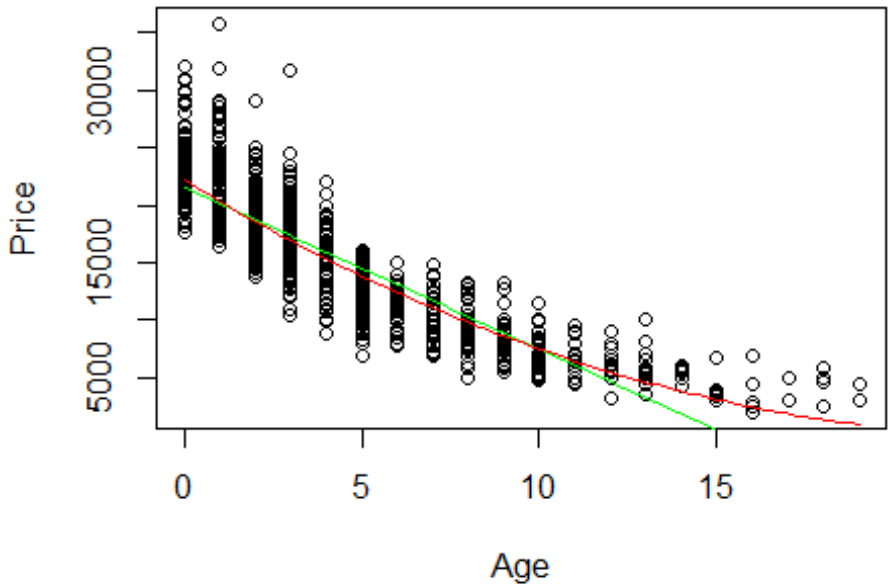




11. According to your transformed model, is there an age at which the car should be free? If so, find this age and comment on what the "free car" phenomenon says about the appropriateness of your model.

According to my model, when a car reaches 23.9 years old the car would be predicted as "free". This free car phenomenon says that the model I chose may not be the best. This also means that I may need to add certain constraints to my model to ensure that it doesn't reach zero. These constraints may be through the form of a different type of transformation or it may be through phyiscally coding into R a constraint that the model should not be able to achieve. I could do this by looking into the average baseline price for TX honda accords and write the constraint based on what the sample population looks like.

```
#Free Car Age
# When y is zero, what is x
freecarage <- (sqrt(0)-B0.2.trans)/B1.2.trans
freecarage

## [1] 23.93987

B0.2 = summary(lm(Price~Age, MyCars))$coefficients[1,1] # Intercept
B1.2 = summary(lm(Price~Age, MyCars))$coefficients[2,1] # Slope

B0.2.trans = summary(lm(sqrt(Price)~Age, data = MyCars))$coefficients[1,1] #
Intercept
B1.2.trans = summary(lm(sqrt(Price)~Age, data = MyCars))$coefficients[2,1] #
Slope

plot(Price~Age, MyCars)
curve(B1.2*x+B0.2, col = "green", add=TRUE)
curve((B1.2.trans*x+B0.2.trans)^2, col = "red", add=TRUE)
```



12. Again suppose that you are interested in purchasing a car of this model that is four years old (in 2017). Determine each of the following using your model constructed in question 9: 90% confidence interval for the mean price at this age and 90% prediction interval for the price of an individual car at this age. Write sentences that carefully interpret each of the intervals (in terms of car prices).

We are 90% confident that the true mean price of a single 4 year old honda accord in texas is between 10801.38 and 20711.64 dollars. We are 90% confident that the ture mean price of the sample population for 4 year old honda accords in texas is between 15223.54 and 15490.62 dollars.

```
confint(lm(sqrt(Price)~Age, data = MyCars), level = 0.90)
```

```
##                     5 %       95 %
## (Intercept) 148.101008 149.462581
## Age          -6.345354  -6.084272
```

```
148.10^2
```

```
## [1] 21933.61
```

```
149.45^2
```

```
## [1] 22335.3
```

```
newx.2=data.frame(Age = 4)
head(newx.2)
```

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

```
predict.lm(lm(sqrt(Price)~Age, data = MyCars), newx.2, interval="confidence")
# All peope; for poualtion
```

```
##        fit      lwr      upr
## 1 123.9225 123.3837 124.4613
```

```
123.3837^2
```

```
## [1] 15223.54
```

```
124.4613^2
```

```
## [1] 15490.62
```

```
predict.lm(lm(sqrt(Price)~Age, data = MyCars), newx.2, interval="prediction")
#For one perosn
```

```
##        fit      lwr      upr
## 1 123.9225 103.9297 143.9154
```

```
103.9297^2
```

```
## [1] 10801.38
```

```
143.9154^2
```

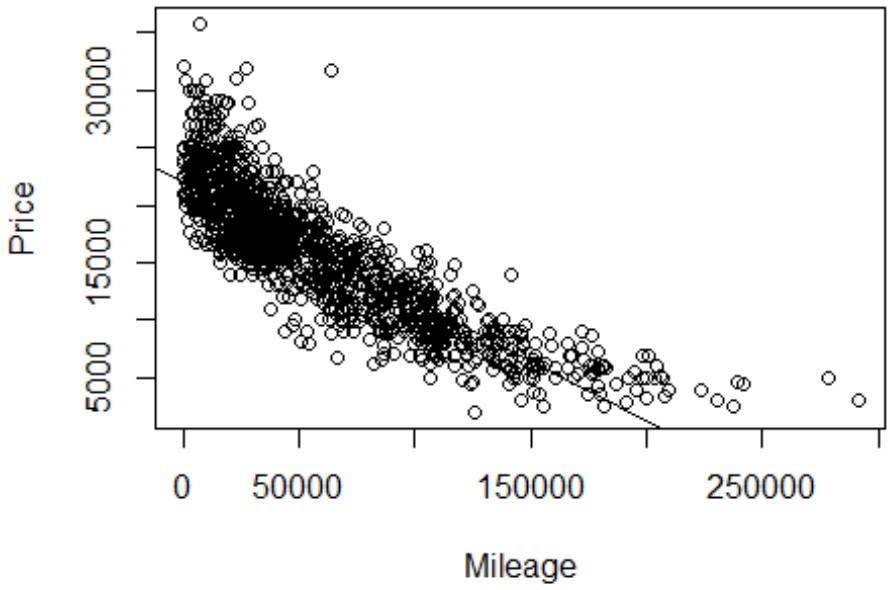
```
## [1] 20711.64
```

MODEL #2: Use Mileage as a predictor for Price

13. Calculate the least squares regression line that best fits your data (with *Mileage* now as the predictor) and produce a scatterplot of the relationship with the regression line on it.

The LSRL of the linear model for car mileage by age is yhat = (-1.039x10$^{-1)x+(2.200ex10}$04). This means for each increase in age, the car's price decreases by -1.039x10^-1 dollars.

```
mod3 <- lm(Price~Mileage, MyCars)
plot(Price~Mileage, MyCars)
abline(mod3)
```




```
summary(mod3)
```

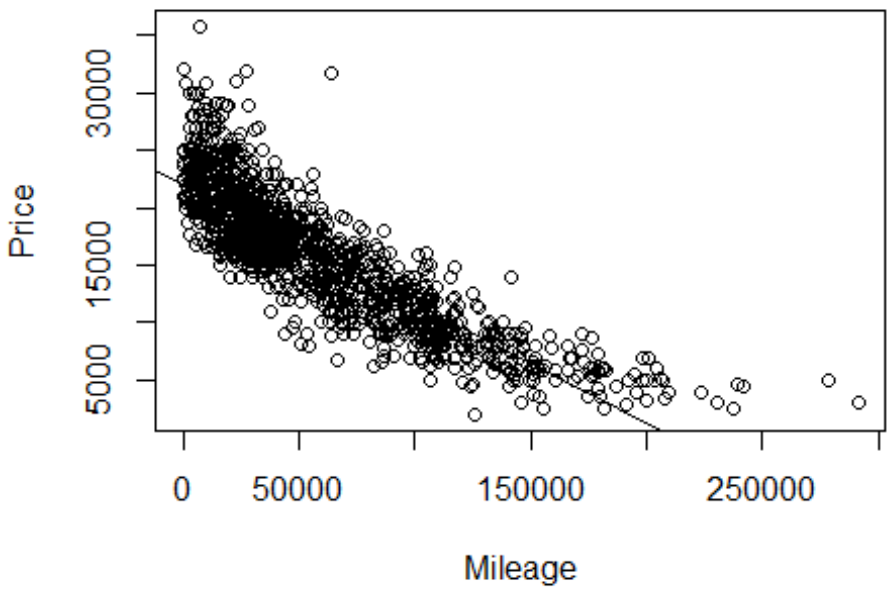
```
##
## Call:
## lm(formula = Price ~ Mileage, data = MyCars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8572.4 -1878.2  -365.7  1543.6 16294.9
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.200e+04  1.245e+02   176.7   <2e-16 ***
## Mileage     -1.039e-01  1.694e-03   -61.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2916 on 1375 degrees of freedom
## Multiple R-squared:  0.7321, Adjusted R-squared:  0.7319
## F-statistic:  3758 on 1 and 1375 DF,  p-value: < 2.2e-16
```

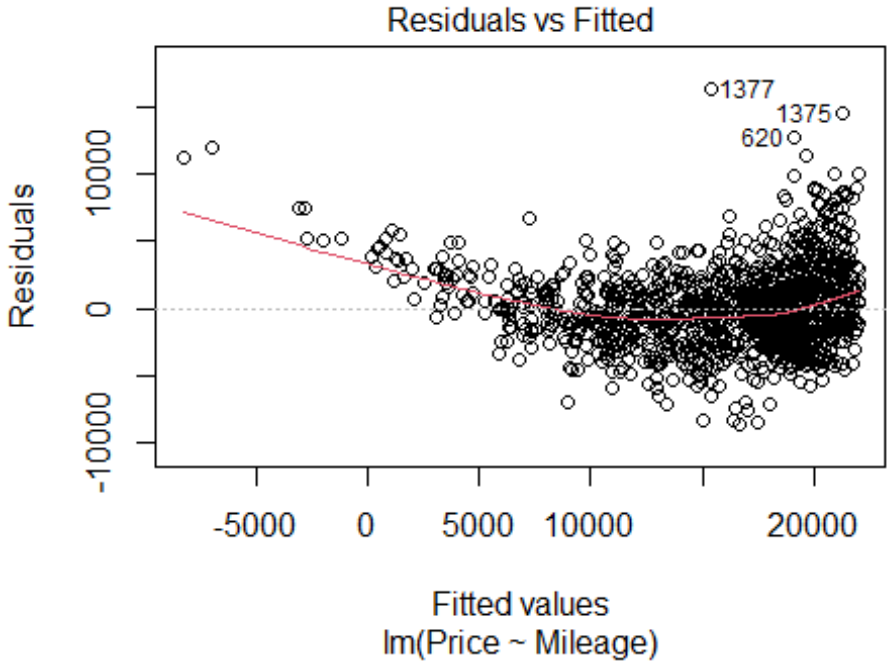
14. Produce appropriate residual plots and comment on how well your data appear to fit the conditions for a simple linear model. Don't worry about doing transformations at this point if there are problems with the conditions.

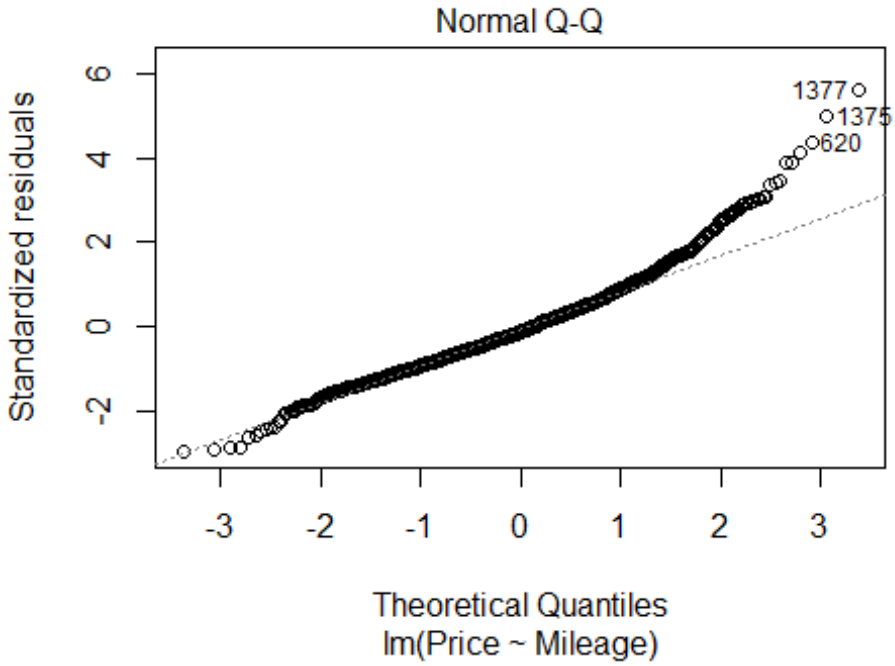
Inspecting the five conditions for a simple linear model: - Linearity: This data does not fit the linear model very well. The simple linear model shows the data largely favoring the left side of the graph, while the fitted residuals plot shows the data heavily hugging the right side of the graph.
- Zero mean: The histogram and fitted residuals appears to be centered around zero, but there is a definite skew towards the right.
- COnstant Variance: The data is heavily clustered towards the left side of the graph. It spreads out slightly as you move more to the right, but not very much.
- Independence: I will assume that each car is independent of each other. - Normality: Looking at the histogram of the residuals, this method appears to have a right skew. The qq-norm plot also suggests slight deviation from normality at the tails of the regression line.

```
plot(Price~Mileage, MyCars)
abline(mod3)
```

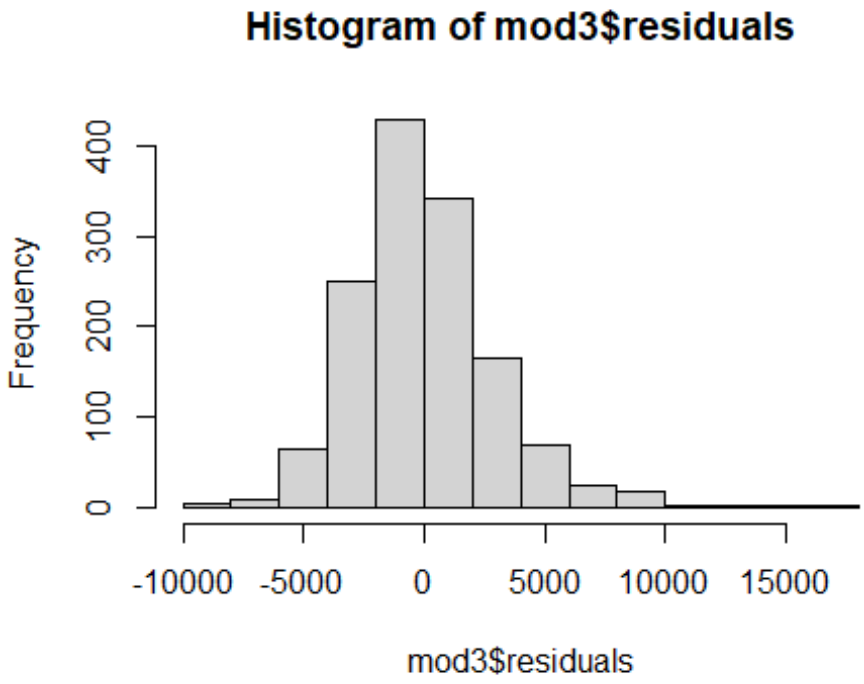



```
plot(mod3, 1:2)
```






```
hist(mod3$residuals)
```



15. Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain why you think that this transformation does or does not improve satisfying the linear model conditions.

I think the linear model is the best to represent the regression for Price by car Mileage. If I was to choose a different approach, I think that the squareroot Price approach would be the second best. I think this due to the data conforming to many of the five conditions for a simple linear model:

- Linearity: The SLRL does not appear to fit the data as closely as the non-transformed simple linear model. Nonetheless, the data appears to follow it in a roughly linear progression, with a slight downwards curve as the graph progresses from left to right.

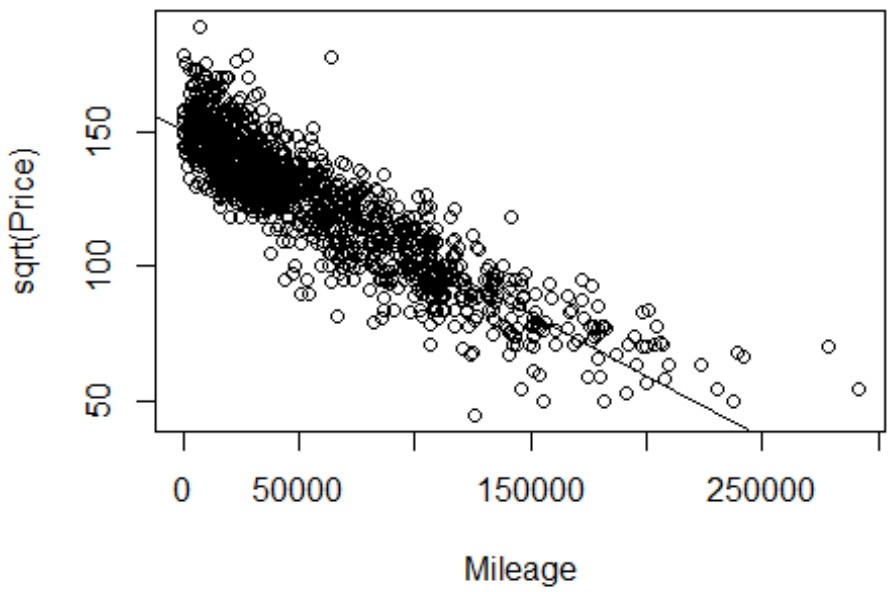
- Zero mean: The residuals vs fitted plot and the histogram are both centered around a zero mean.

- COnstant Variance: There is roughly constant variance. I think the transformed data has more data varation than the untransformed simple linear model because the transformed model conforms closer to a normal distribution than the untransformed simple linear model.

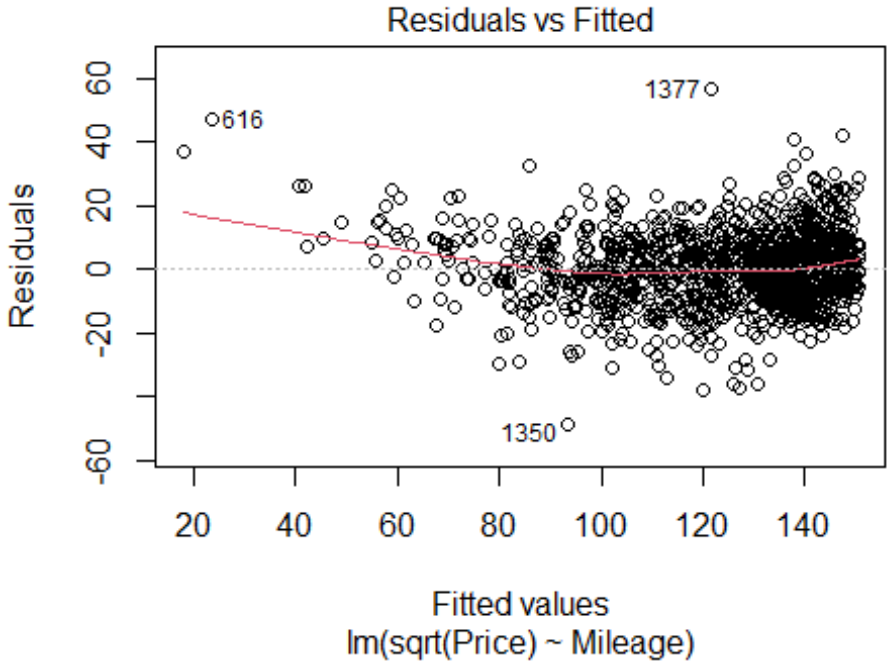
- Independence: I will assume that each car is independent of each other.

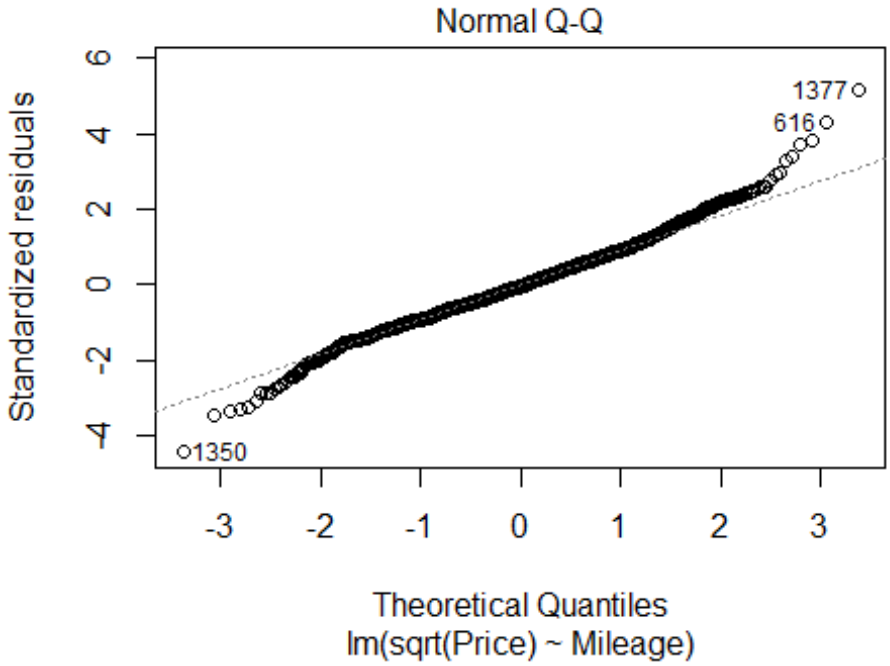
- Normality: Looking at the histogram of the residuals, this method appears to have a slight right skew. This is a better histogram than the un-transformed simple linear regression that mod3 represents. The qq-norm plot also hugs the line much closer than the simple linear model.

```
mod4 <- lm(sqrt(Price)~Mileage, data = MyCars)
# Square root Price
plot(sqrt(Price)~Mileage, data = MyCars)
abline(mod4)
```




```
plot(mod4, 1:2)
```

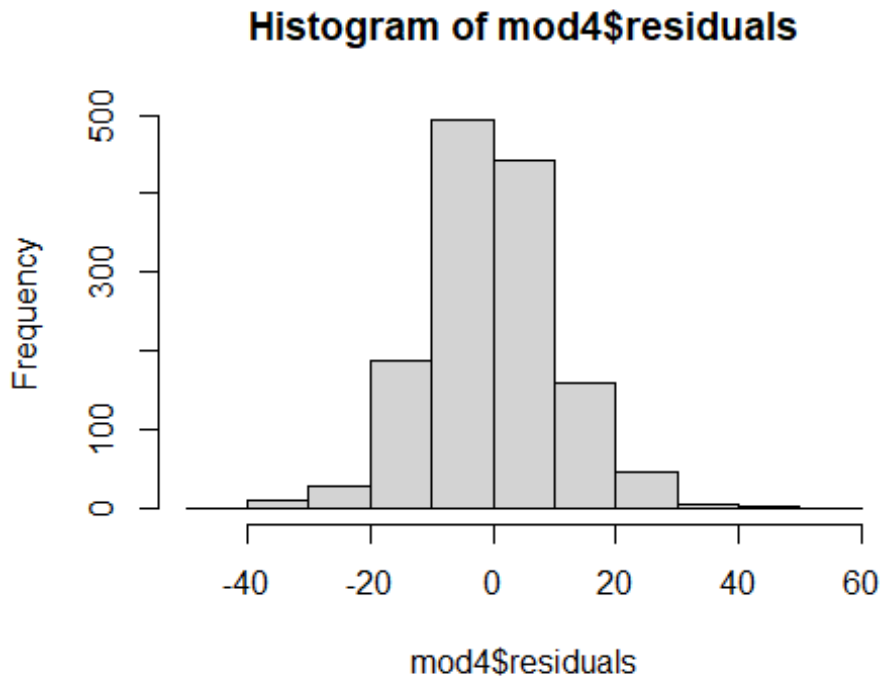





```
summary(mod4)

##
## Call:
```

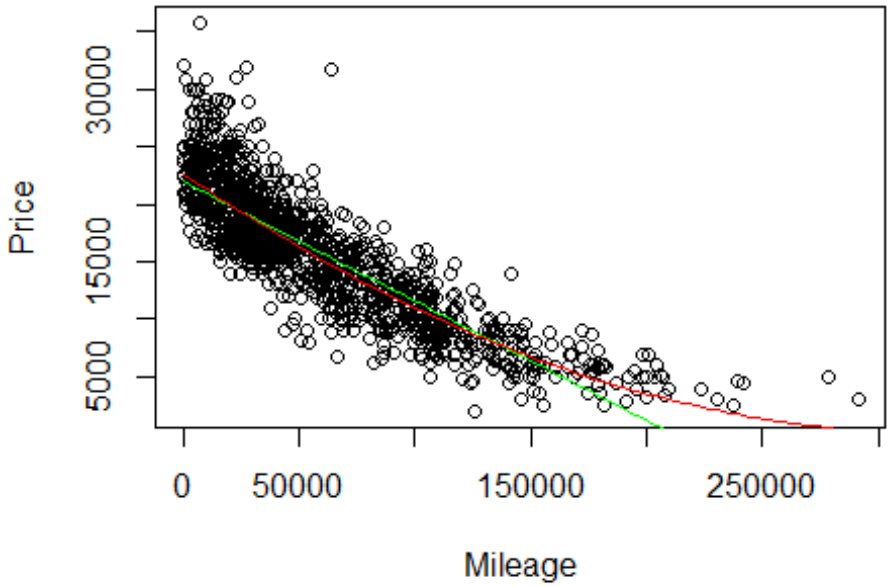
```
## lm(formula = sqrt(Price) ~ Mileage, data = MyCars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.627  -6.885  -0.540   6.618  56.520
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.505e+02  4.693e-01  320.67   <2e-16 ***
## Mileage     -4.552e-04  6.386e-06  -71.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.99 on 1375 degrees of freedom
## Multiple R-squared:  0.787,  Adjusted R-squared:  0.7869
## F-statistic:  5081 on 1 and 1375 DF,  p-value: < 2.2e-16
```

```
hist(mod4$residuals)
```

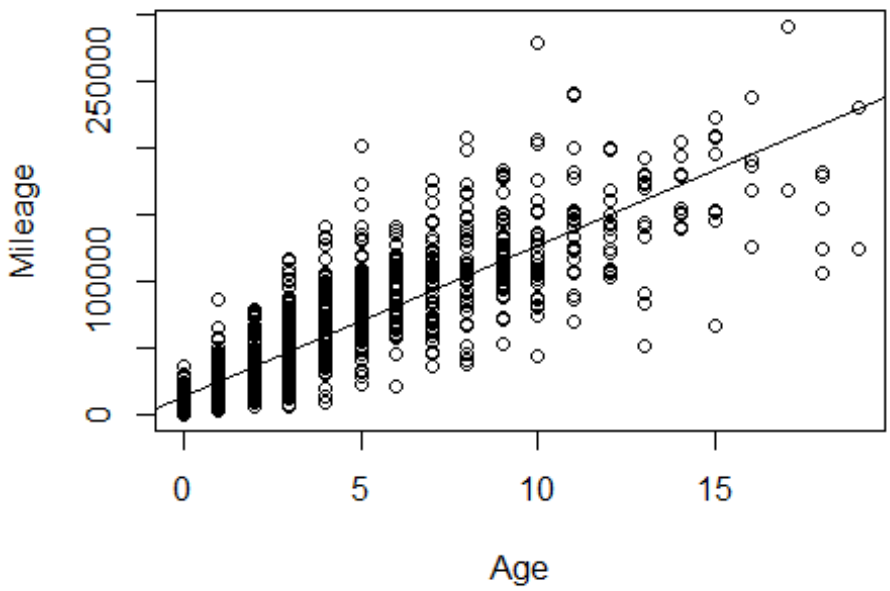



```
B0 = summary(lm(Price~Mileage, MyCars))$coefficients[1,1] # Intercept
B1 = summary(lm(Price~Mileage, MyCars))$coefficients[2,1] # Slope

B0.trans = summary(lm(sqrt(Price)~Mileage, data = MyCars))$coefficients[1,1]
# Intercept
B1.trans = summary(lm(sqrt(Price)~Mileage, data = MyCars))$coefficients[2,1]
# Slope
```

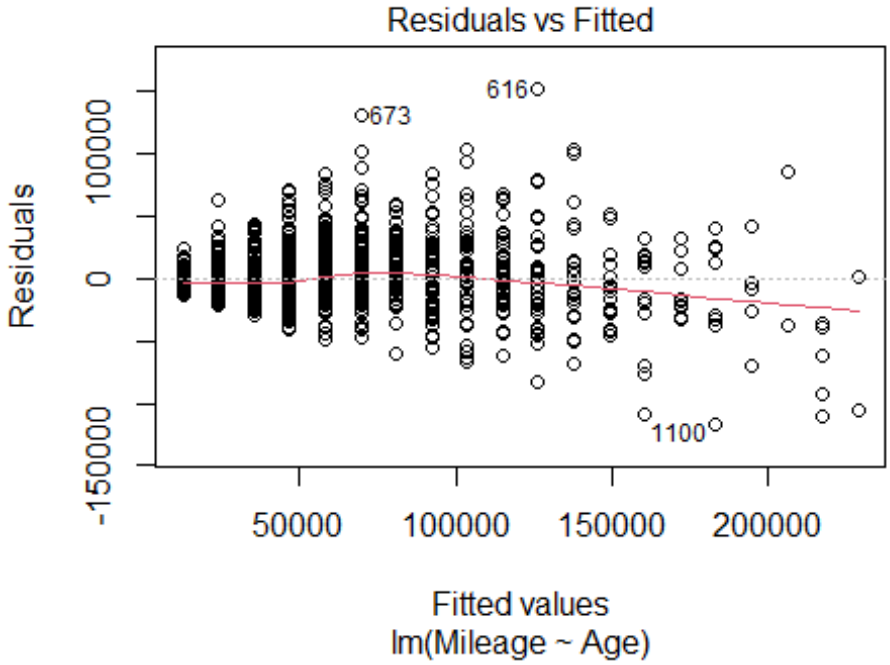
```r
plot(Price~Mileage, MyCars)
curve(B1*x+B0, col = "green", add=TRUE)
curve((B1.trans*x+B0.trans)^2, col = "red", add=TRUE)
```

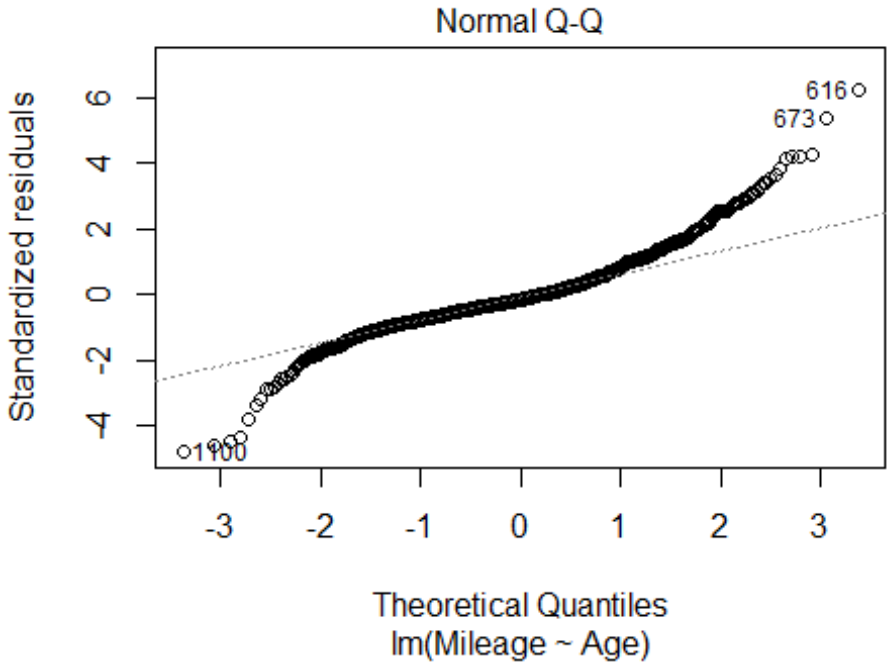



```r
plot(Mileage~Age, data = MyCars) # This is the original
abline(lm(Mileage~Age, data = MyCars))
```

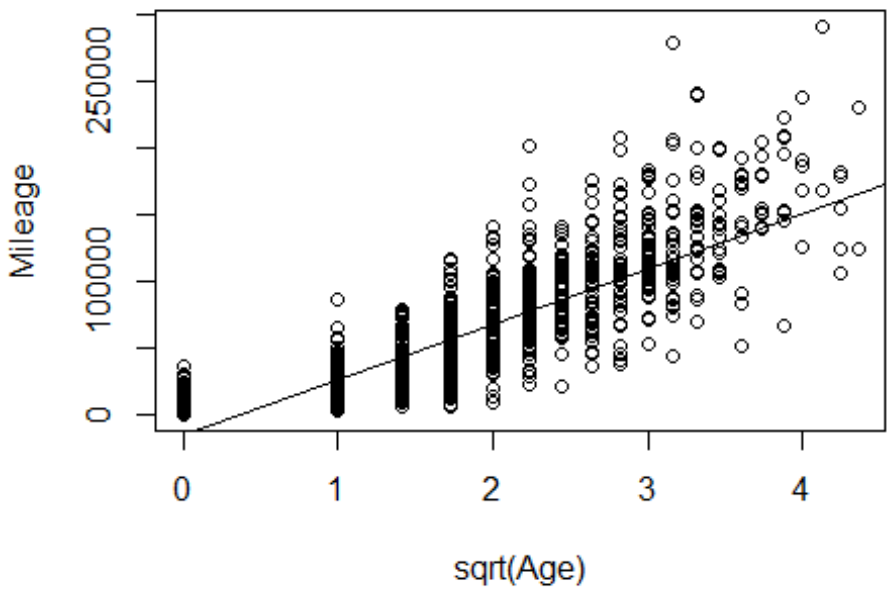



```
plot(lm(Mileage~Age, data = MyCars), 1:2)
```

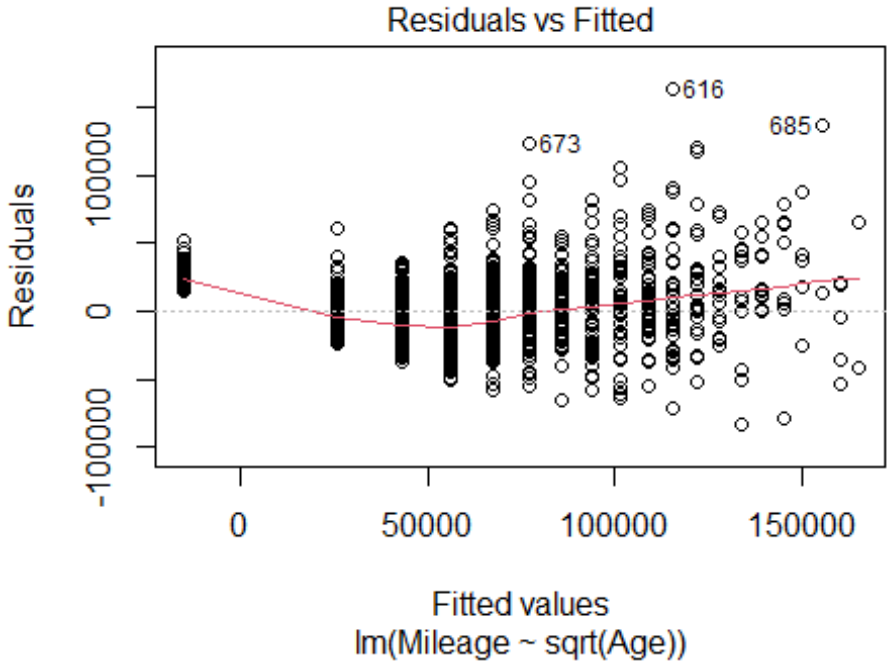


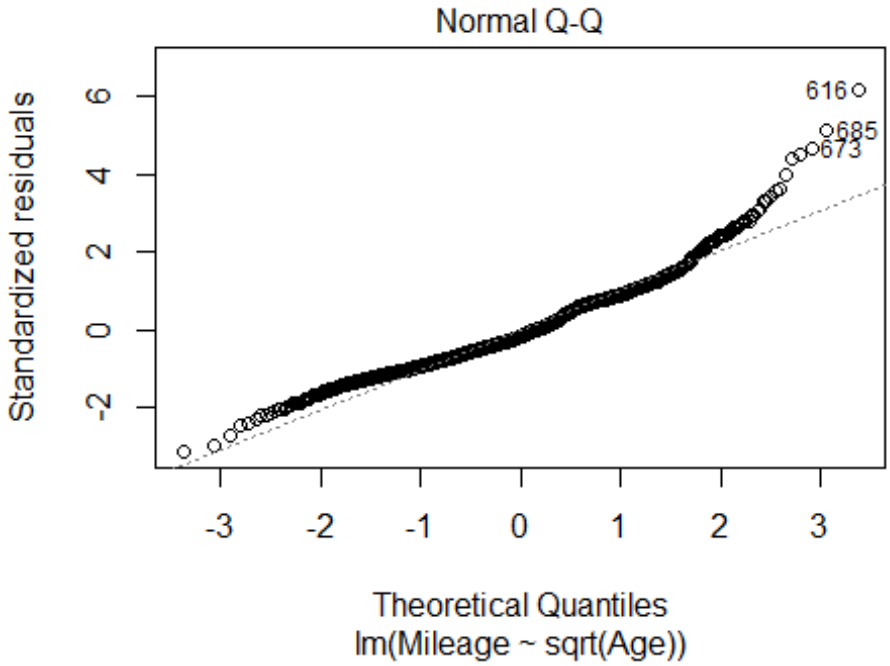



```r
# square root Age
plot(Mileage~sqrt(Age), data = MyCars)
abline(lm(Mileage~sqrt(Age), data = MyCars))
```

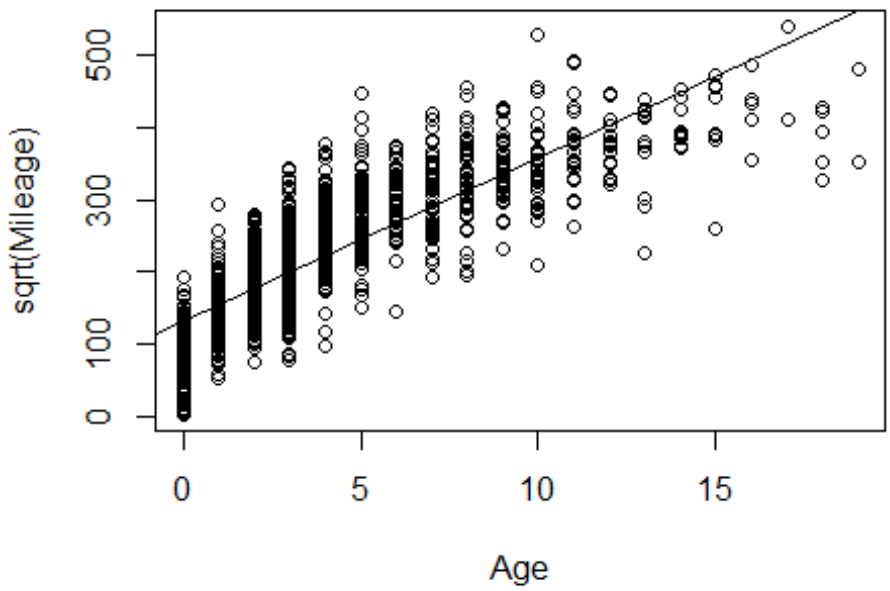



```
plot(lm(Mileage~sqrt(Age), data = MyCars), 1:2)
```

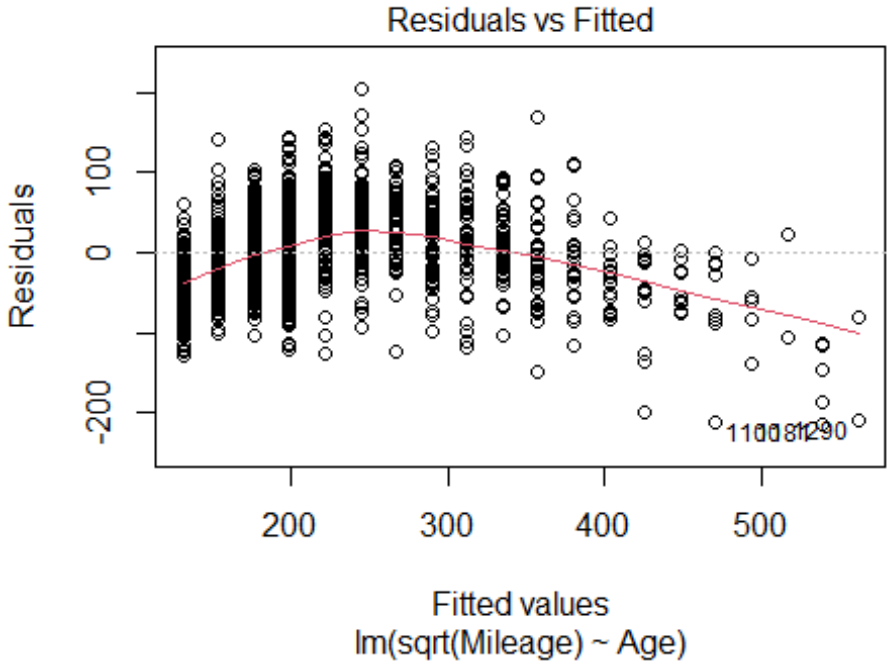


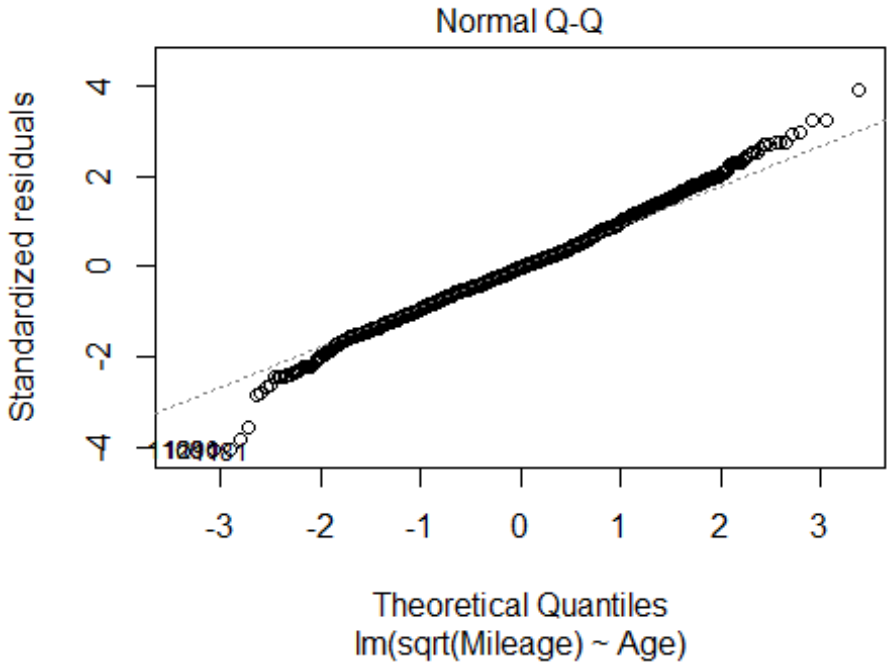



```r
# Square root Mileage
plot(sqrt(Mileage)~Age, data = MyCars)
abline(lm(sqrt(Mileage)~Age, data = MyCars))
```

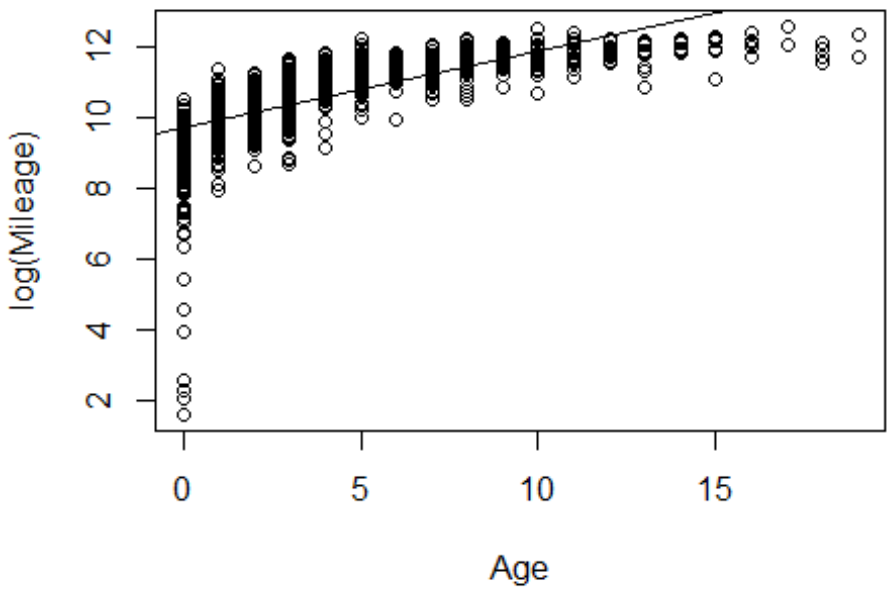



```r
plot(lm(sqrt(Mileage)~Age, data = MyCars), 1:2)
```

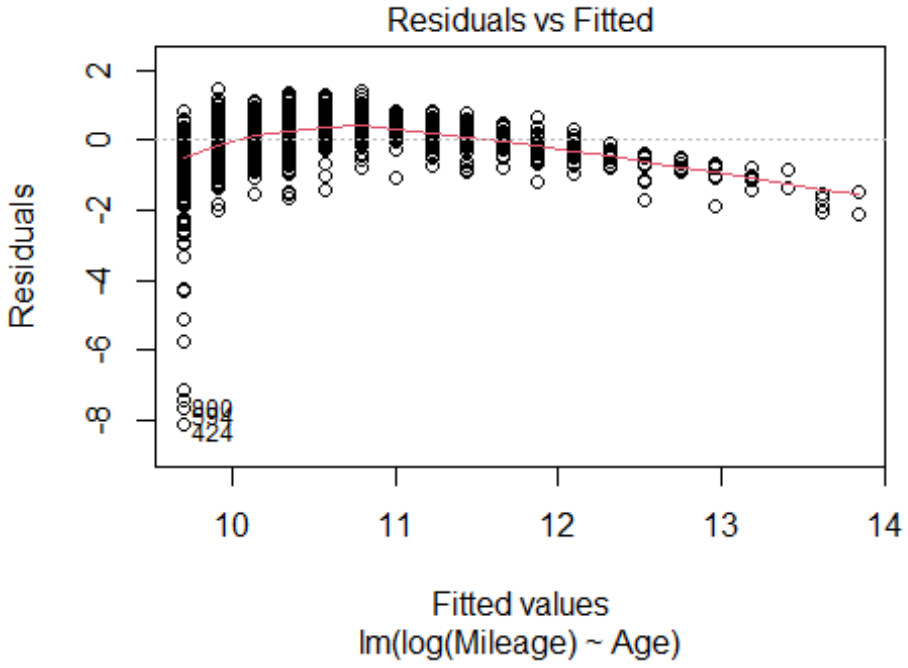


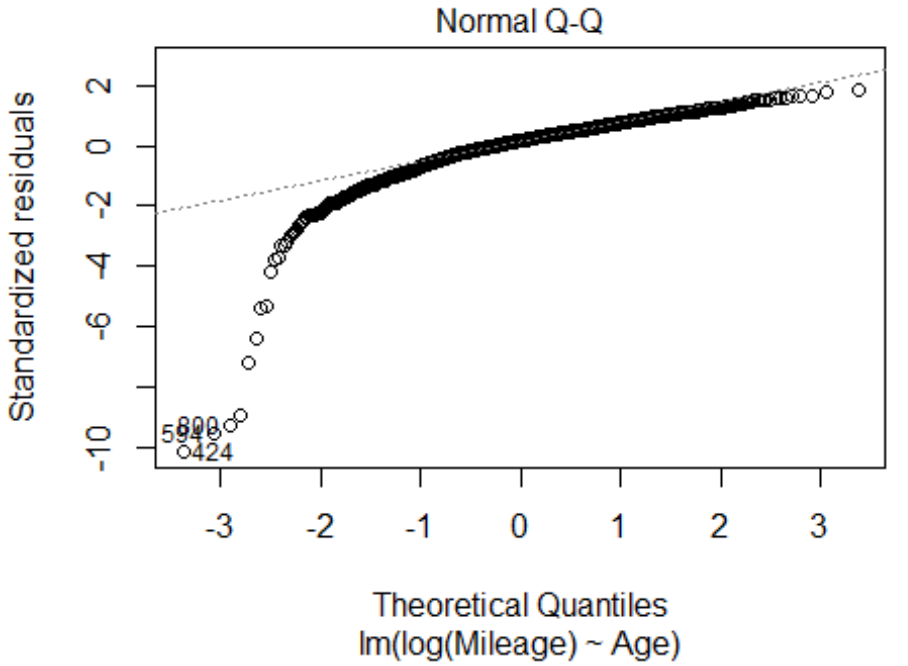



```r
# Log Mileage
plot(log(Mileage)~Age, data = MyCars)
abline(lm(log(Mileage)~Age, data = MyCars))
```

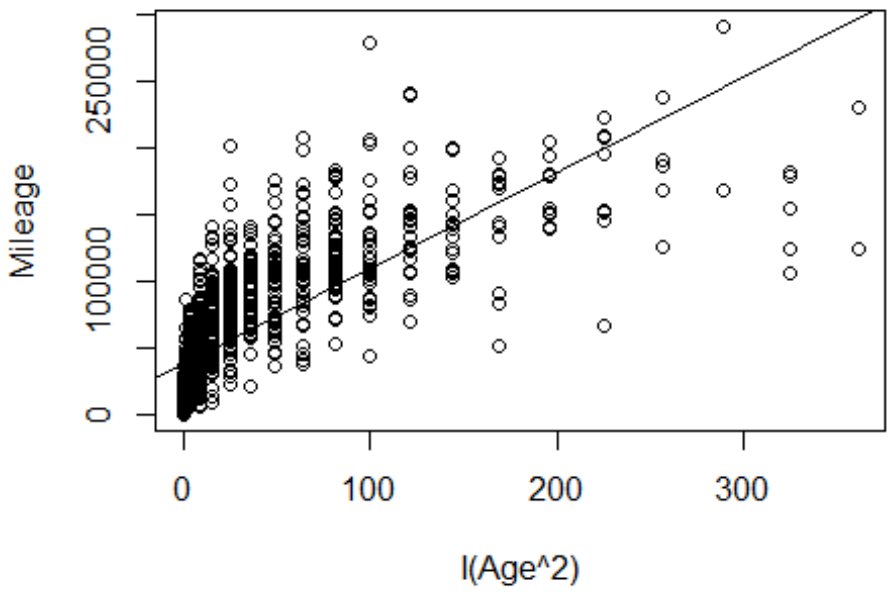



```r
plot(lm(log(Mileage)~Age, data = MyCars), 1:2)
```

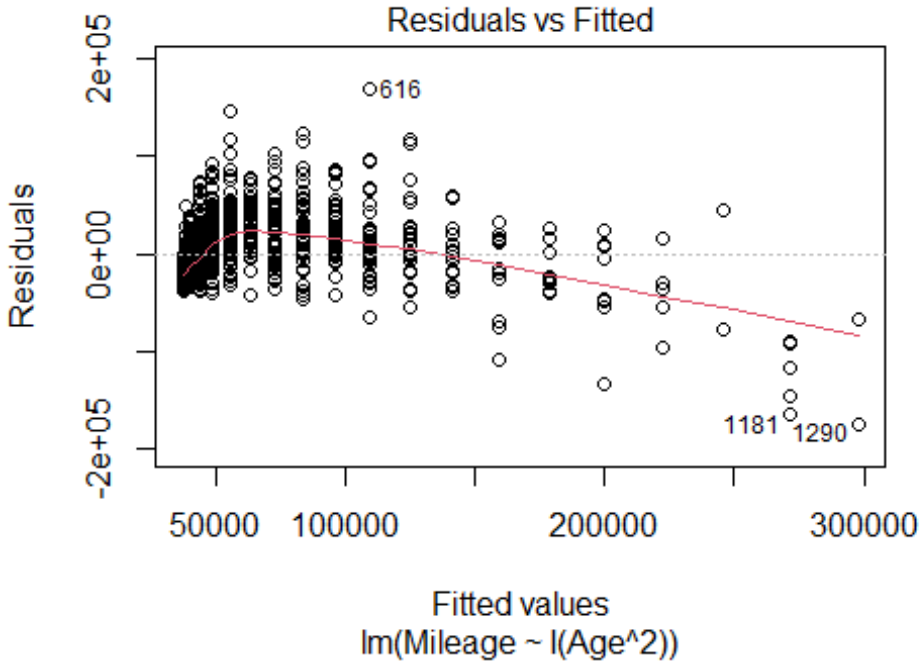


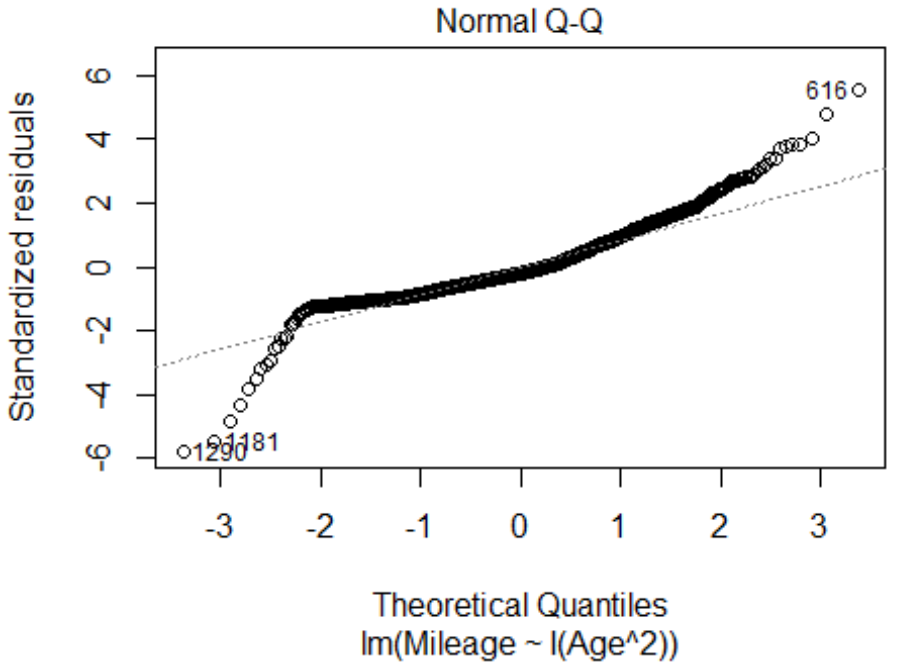



```r
# Add Exponent to Age
plot(Mileage~I(Age^2), data = MyCars)
abline(lm(Mileage~I(Age^2), data = MyCars))
```

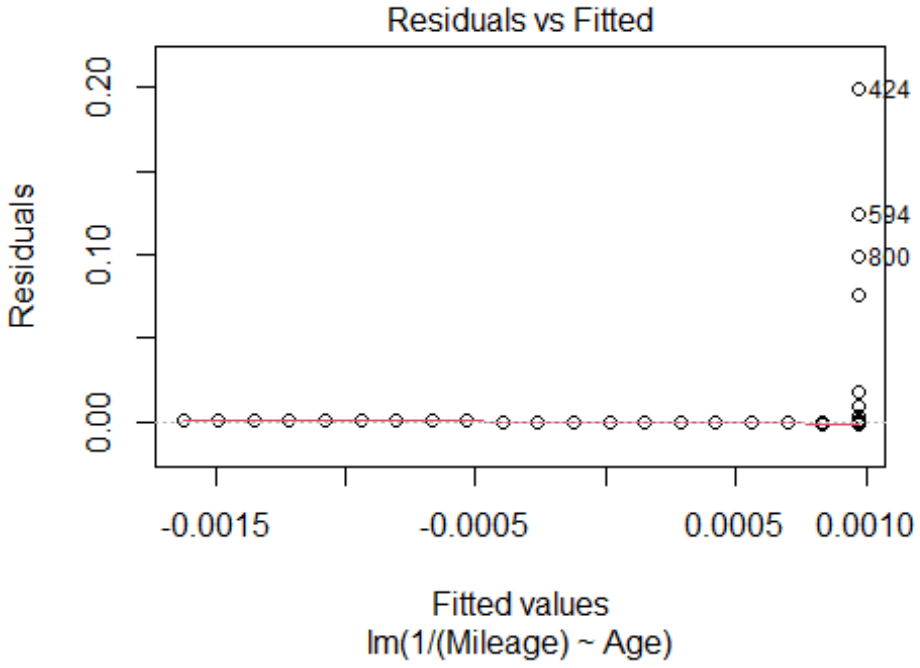


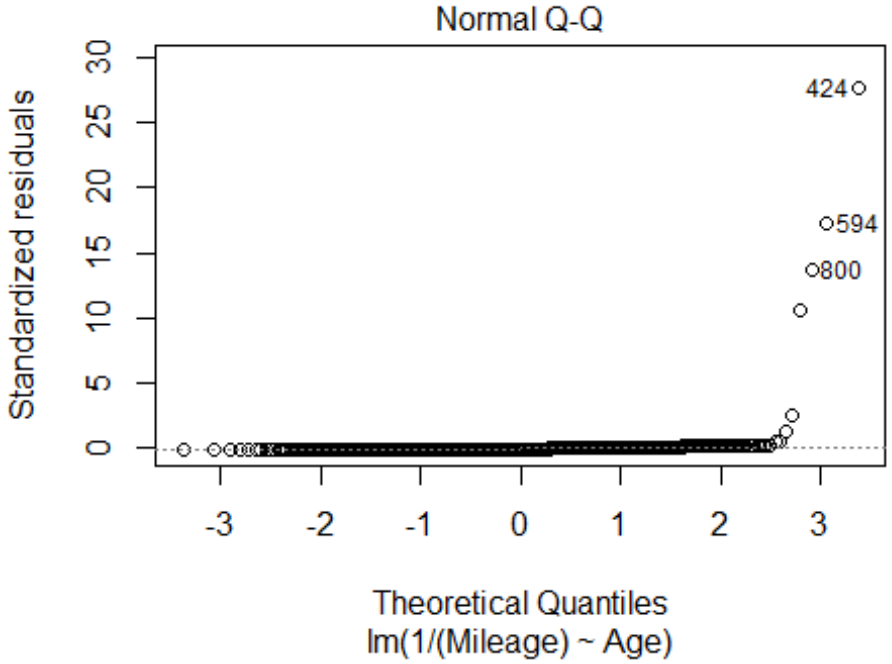

```
plot(lm(Mileage~I(Age^2), data = MyCars), 1:2)
```

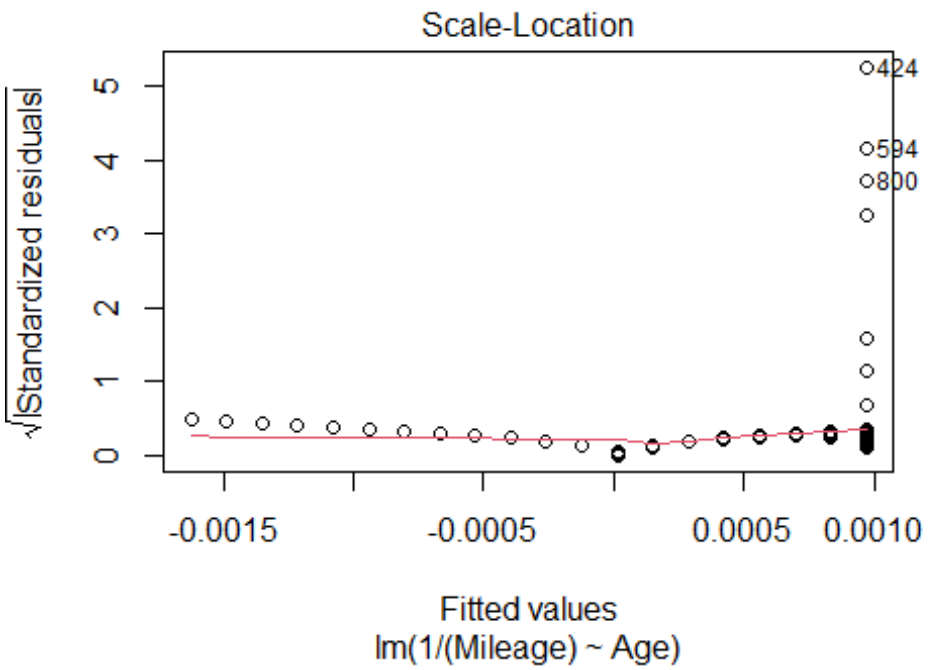




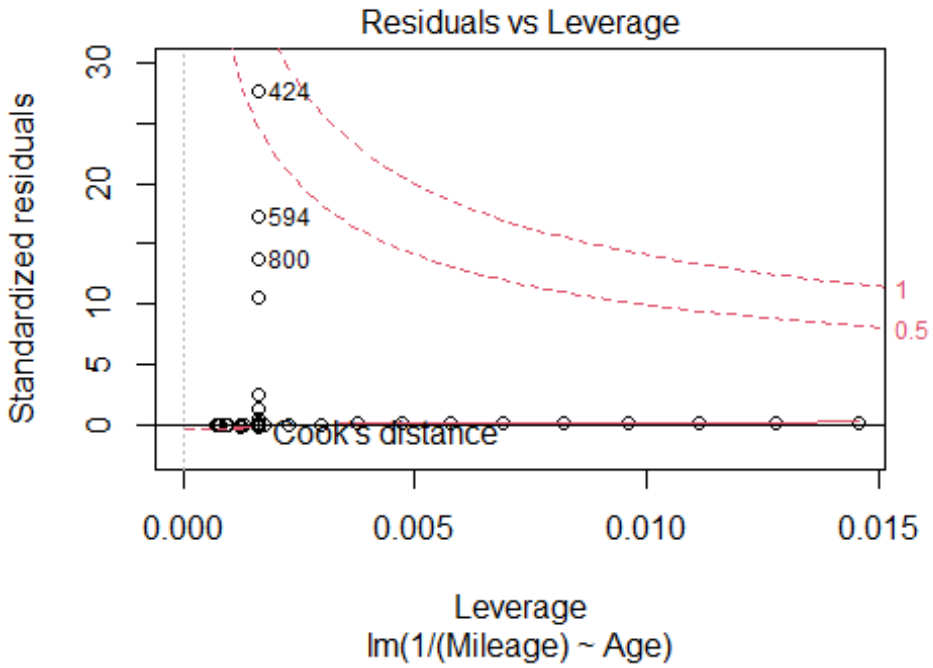

```r
# Division Mileage
plot(lm(1/(Mileage)~Age, data=MyCars))
```


Residuals vs Fitted
Residuals
0.00
0.10
0.20
424
594
800
-0.0015
-0.0005
0.0005
0.0010
Fitted values
lm(1/(Mileage) ~ Age)

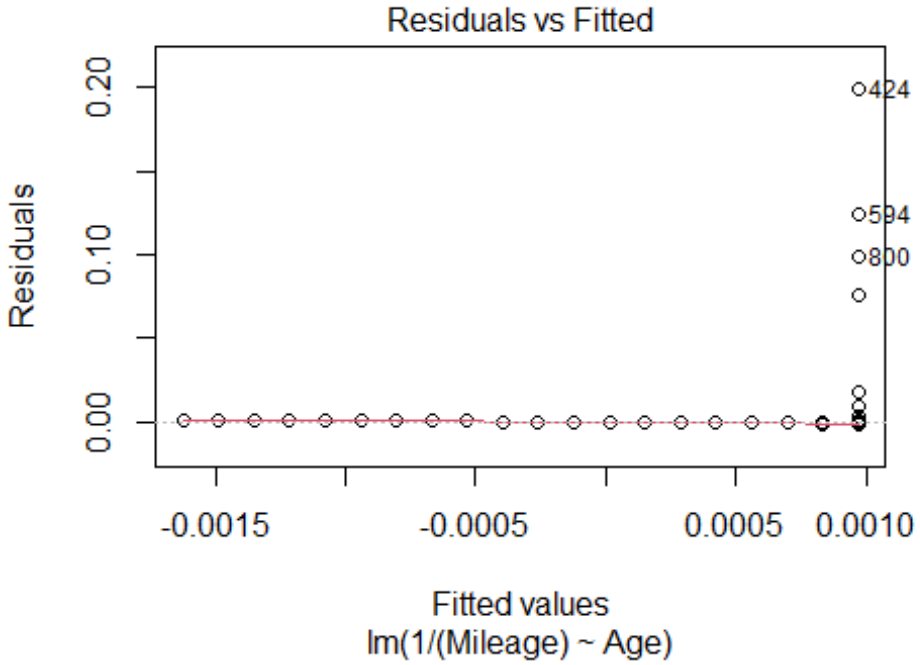

Normal Q-Q
Standardized residuals
0
5
10
15
20
25
30
424
594
800
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(1/(Mileage) ~ Age)

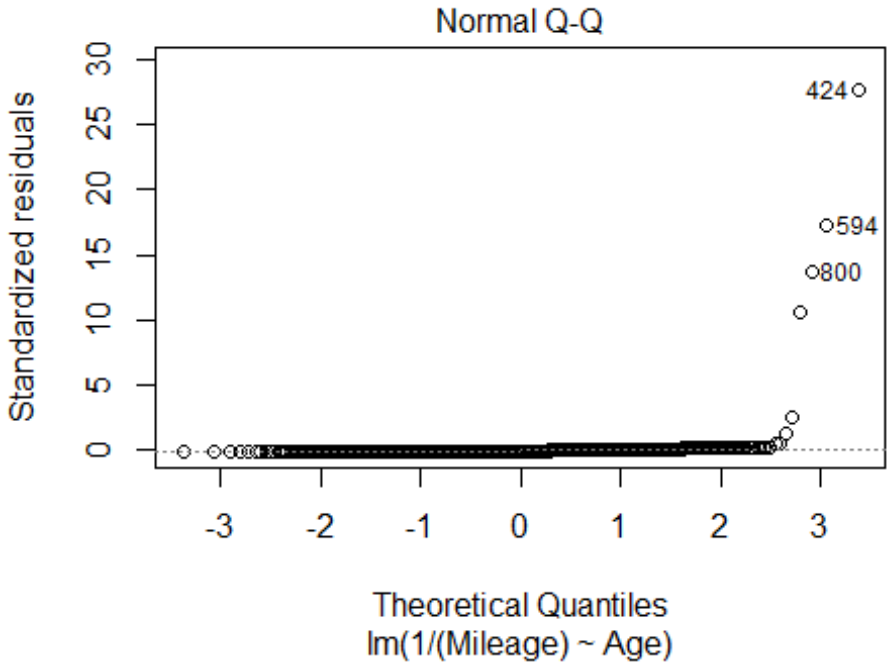



```
abline(lm(1/(Mileage)~Age, data=MyCars))
```




```
plot(lm(1/(Mileage)~Age, data=MyCars), 1:2)
```





16. How do the models, using either *Age* or *Mileage* as the predictor compare? Does one of the models seem "better" or do they seem similar in their ability to predict *Price*? Explain.

Age and Mileage both have about the same correlation to price. When running a slope test, both pvalues are very low - meaning that we have evidence to reject the null hypothesis that the true correlation is equal to 0. Also looking at the anova test shows us that the SSModel and SSE are very similar, meaning that the model has about the same amount of error regardless if you cast with age or mileage. Lastly, looking at the residual plot, the residuals of using age are more skewed right htan the residuals of using mileage. Based on these observations, I would say that mileage is the better predictor of price, but only by very little.

```
head(MyCars)
```

```
## # A tibble: 6 x 11
##      Id Price  Year Mileage City           State Vin   Make  Model   Age r
esiduals
##   <dbl> <dbl> <dbl>   <dbl> <chr>          <chr> <chr> <chr> <chr> <dbl>
<dbl>
## 1  1476 21408  2017    4084 Corpus Christi TX    1HGC~ Honda Acco~     0
-163.
## 2  1477 15995  2014   30914 Houston        TX    1HGC~ Honda Acco~     3
-1350.
## 3  1498 16997  2015   21028 Fort Worth     TX    1HGC~ Honda Acco~     2
-1757.
## 4  1521 15000  2014   47632 McKinney       TX    1HGC~ Honda Acco~     3
-2345.
## 5  1601 16000  2014   35785 Frisco         TX    1HGC~ Honda Acco~     3
-1345.
## 6  1629 17500  2014   27131 San Antonio    TX    1HGC~ Honda Acco~     3
155.
```

```
cor(MyCars[c(2,4,10)])
```

```
##              Price    Mileage        Age
## Price    1.0000000 -0.8556332 -0.8661308
## Mileage -0.8556332  1.0000000  0.8494867
## Age     -0.8661308  0.8494867  1.0000000
```

```
cor.test(MyCars$Price, MyCars$Age)
```

```
##
##  Pearson's product-moment correlation
##
## data:  MyCars$Price and MyCars$Age
## t = -64.257, df = 1375, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8787504 -0.8523011
## sample estimates:
##        cor
## -0.8661308
```

```
cor.test(MyCars$Price, MyCars$Mileage)
```

```
##
##  Pearson's product-moment correlation
##
## data:  MyCars$Price and MyCars$Mileage
## t = -61.3, df = 1375, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8691730 -0.8408116
## sample estimates:
##       cor
## -0.8556332
```

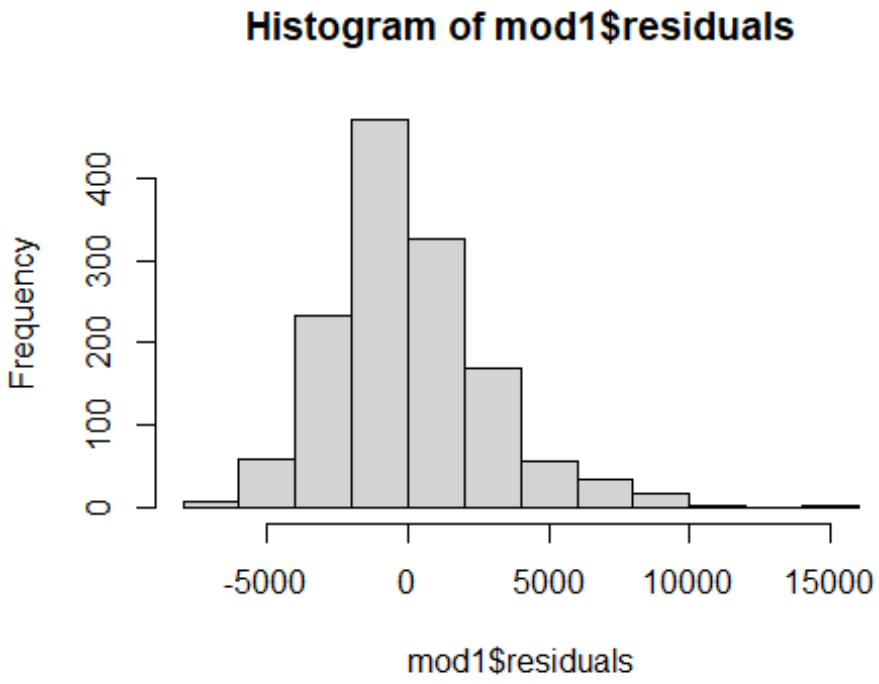
```
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: Price
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## Age          1 3.2743e+10 3.2743e+10    4129 < 2.2e-16 ***
## Residuals 1375 1.0904e+10 7.9300e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

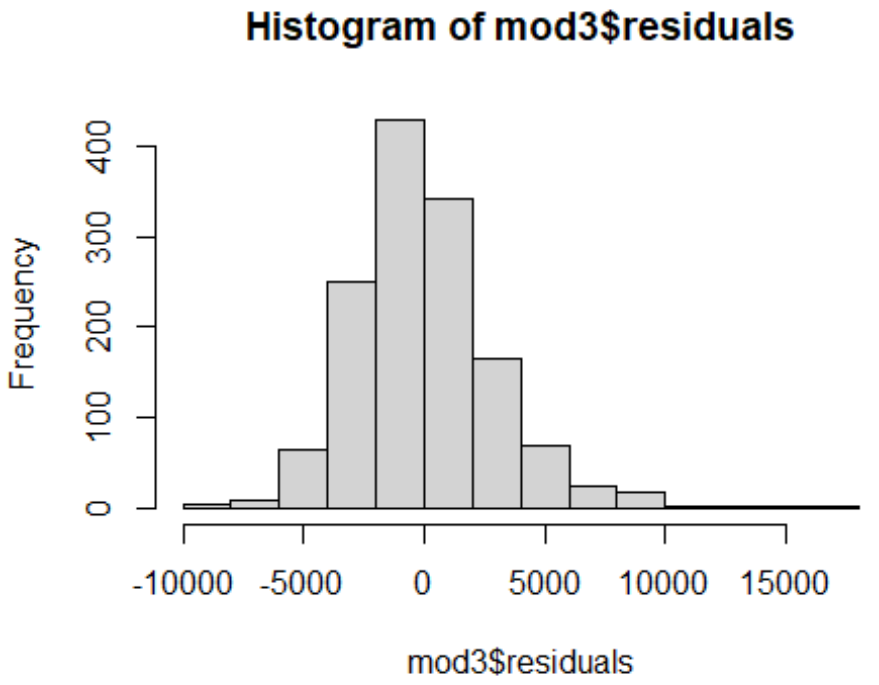
```
anova(mod3)
```

```
## Analysis of Variance Table
##
## Response: Price
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## Mileage      1 3.1954e+10 3.1954e+10  3757.7 < 2.2e-16 ***
## Residuals 1375 1.1693e+10 8.5038e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
hist(mod1$residuals)
```




```r
hist(mod3$residuals)
```



MODEL #3: Again use Age as a predictor for Price, but now for new data

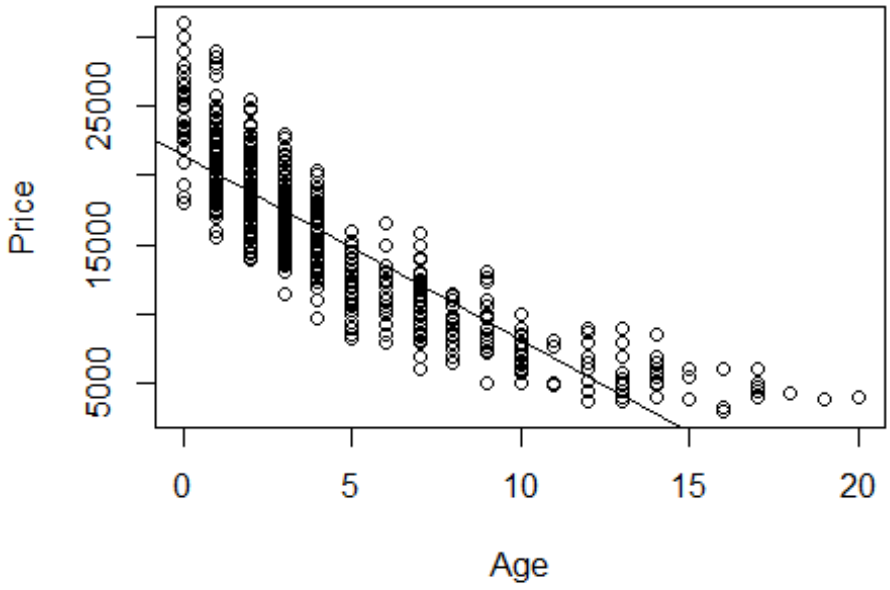
17. Select a new sample from the UsedCar dataset using the same *Model* car that was used in the previous sections, but now from cars for sale in North Carolina. You can mimic the code used above to select this new sample.

```
library(readr)

# This line will only run if the UsedCars.csv is stored in the same directory
as this notebook!
UsedCars2 <- 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 data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this m
essage.

# Delete the ** below and enter the two letter abbreviation for the state of
your choice.
StateOfMyChoice2 = "NC"

# Creates a dataframe with the number of each model for sale in your state
Cars2 = as.data.frame(table(UsedCars2$Model[UsedCars2$State==StateOfMyChoice2
]))

# Renames the variables
names(Cars2)[1] = "Model"
names(Cars2)[2] = "Count"

# Restricts the data to only models with at least 100 for sale
# Before submitting, comment this out so that it doesn't print while knitting
# Enough_Cars = subset(Cars, Count>=100)
# Enough_Cars

# Delete the ** below and enter the model that you chose from the Enough_Cars
data.
ModelOfMyChoice2 = "Accord"

# Takes a subset of your model car from your state
MyCars2 = subset(UsedCars2, Model==ModelOfMyChoice2 & State==StateOfMyChoice2
)

# Check to make sure that the cars span at least 5 years.
range(MyCars2$Year)
```

```
## [1] 1997 2017

# Add a new variable for the age of the cars.
MyCars2$Age = 2017 - MyCars2$Year
```

18. Calculate the least squares regression line that best fits your new data and produce a scatterplot of the relationship with the regression line on it.

```
modnc <- lm(Price~Age, MyCars2)
plot(Price~Age, MyCars2)
abline(modnc)
```




```
summary(lm(Price~Age, MyCars2))

##
## Call:
## lm(formula = Price ~ Age, data = MyCars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6494.8 -1849.0  -297.4  1552.1  9575.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21419.54     148.52  144.22   <2e-16 ***
## Age         -1324.95      27.17  -48.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2720 on 774 degrees of freedom
## Multiple R-squared:  0.7544, Adjusted R-squared:  0.7541
## F-statistic:  2378 on 1 and 774 DF,  p-value: < 2.2e-16
```

19. How does the relationship between *Price* and *Age* for this new data compare to the regression model constructed in the first section? Does it appear that the relationship between *Age* and *Price* for your *Model* of car is similar or different for the data from your two states? Explain.

The correlation between price, mileage and age are still fairly similar between states. The only notable difference would be that NC has slightly lower correlation rates between mileage and age. Texas's correlation rates are 0.84 between mileage and age, while NC's correlation rates are 0.82 between age and mileage.

When running a cor.test for the slope of the two variables still have very similar correlations, and both have very low pvalues - which suggest that we have evidence that suggests we can reject the null hypothesis that the true correlation is equal to zero.

When looking at the anova testing for the two states, NC has a significantly higher SSE than texas and texas has a slightly higher SSModel than NC. This is significant because this tells us that the Rsquared measurement is going to be effected differently for NC and TX.

Lastly, looking at teh residual plot of the linear regression models for TX and NC, this shows us that TX has a stronger right skew than NC. NC's model, while also right skewed, is signifigantly less skewed than TX's model.

Overall, I would say that the relationship between age and price for the Honda Accord in both TX and NC are similar based on the reasons above.

```
#TX
cor(MyCars[c(2,4,10)])
```

```
##              Price    Mileage        Age
## Price    1.0000000 -0.8556332 -0.8661308
## Mileage -0.8556332  1.0000000  0.8494867
## Age     -0.8661308  0.8494867  1.0000000
```

```
#NC
cor(MyCars2[c(2,4,10)])
```

```
##              Price    Mileage        Age
## Price    1.0000000 -0.8458657 -0.8685637
## Mileage -0.8458657  1.0000000  0.8221806
## Age     -0.8685637  0.8221806  1.0000000
```

```
#TX
cor.test(MyCars$Price, MyCars$Age)
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  MyCars$Price and MyCars$Age
## t = -64.257, df = 1375, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8787504 -0.8523011
## sample estimates:
##        cor
## -0.8661308
```

```
#NC
cor.test(MyCars2$Price, MyCars2$Age)
```

```
##
##  Pearson's product-moment correlation
##
## data:  MyCars2$Price and MyCars2$Age
## t = -48.76, df = 774, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8848527 -0.8501536
## sample estimates:
##        cor
## -0.8685637
```

```
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    4129 < 2.2e-16 ***
## Residuals 1375 1.0904e+10 7.9300e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modnc)
```

```
## Analysis of Variance Table
##
## Response: Price
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## Age         1 1.7588e+10 1.7588e+10  2377.5 < 2.2e-16 ***
## Residuals 774 5.7259e+09 7.3978e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod1)
```

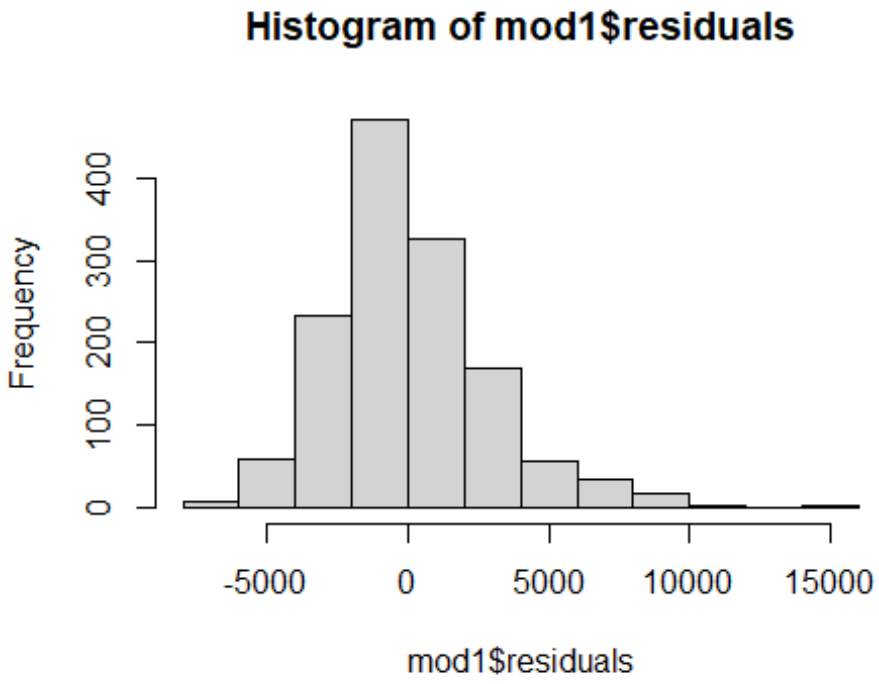
```
##
## Call:
## lm(formula = Price ~ Age, data = MyCars)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7533.1 -1820.8  -376.3  1423.6 15619.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21571.35     114.33  188.68   <2e-16 ***
## Age         -1408.64      21.92  -64.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2816 on 1375 degrees of freedom
## Multiple R-squared:  0.7502, Adjusted R-squared:   0.75
## F-statistic:  4129 on 1 and 1375 DF,  p-value: < 2.2e-16
```

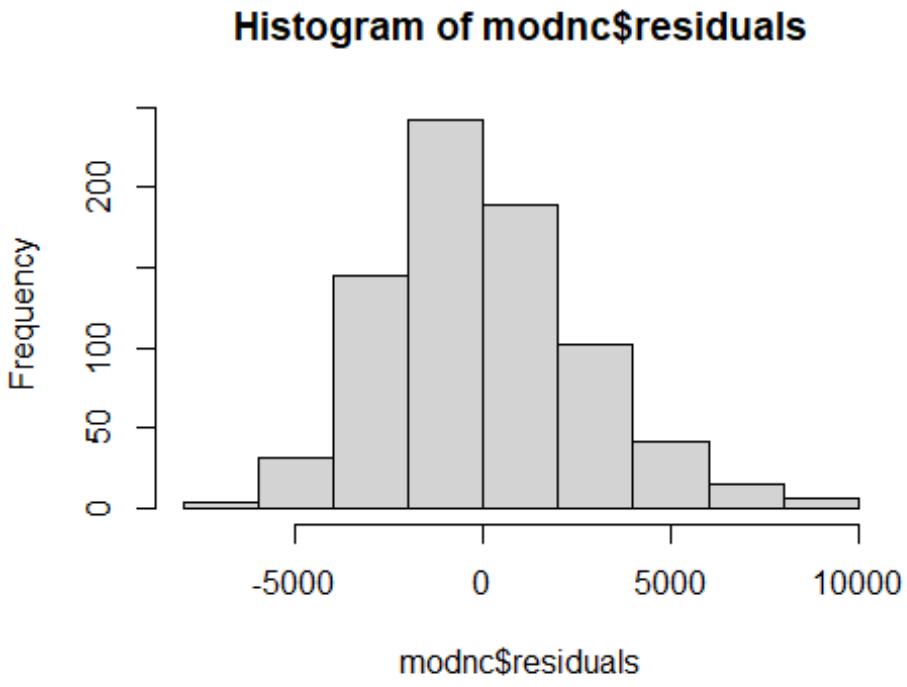
```
summary(modnc)
```

```
##
## Call:
## lm(formula = Price ~ Age, data = MyCars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6494.8 -1849.0  -297.4  1552.1  9575.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21419.54     148.52  144.22   <2e-16 ***
## Age         -1324.95      27.17  -48.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2720 on 774 degrees of freedom
## Multiple R-squared:  0.7544, Adjusted R-squared:  0.7541
## F-statistic:  2378 on 1 and 774 DF,  p-value: < 2.2e-16
```

```
hist(mod1$residuals)
```




```r
hist(modnc$residuals)
```



20. Again suppose that you are interested in purchasing a car of this model that is four years old (in 2017) from North Carolina. How useful do you think that your model will be? What are some possible cons of using this model?

For a 4 year old car I would feel okay with this model. The linear regression line runs through about 4 years old at about the middle of the distribution of 4 year old cars in the plot below. I would prefer a more transformed model, perhaps the squre root model I used previously for texas. Both the transformed model and the lienar model run into eventually predicting a "free car", which is a pretty big con of the data.

```
summary(modnc)
```

```
##
## Call:
## lm(formula = Price ~ Age, data = MyCars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6494.8 -1849.0  -297.4  1552.1  9575.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21419.54     148.52  144.22   <2e-16 ***
## Age         -1324.95      27.17  -48.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2720 on 774 degrees of freedom
## Multiple R-squared:  0.7544, Adjusted R-squared:  0.7541
## F-statistic:  2378 on 1 and 774 DF,  p-value: < 2.2e-16
```

```
confint(modnc, level = 0.90)
```

```
##                    5 %      95 %
## (Intercept) 21174.953 21664.124
## Age         -1369.696 -1280.198
```

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

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

```
predict.lm(modnc, newx.nc, interval="confidence") # All peope; for poualtion
```

```
##        fit      lwr      upr
## 1 16119.75 15927.98 16311.52
```

```
predict.lm(modnc, newx.nc, interval="prediction") #For one perosn
```

```
##        fit      lwr      upr
## 1 16119.75 10777.06 21462.44
```

```r
B0.nc = summary(modnc)$coefficients[1,1] # Intercept
B1.nc = summary(modnc)$coefficients[2,1] # Slope

plot(Price~Age, MyCars2)
curve(B1.nc*x+B0.nc, col = "green", add=TRUE)
```