STOR 455 Homework #2

40 points - Due Wednesday 2/9 at 5:00pm

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 and mileage.

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 will choose a vehicle *Model* from a US company for which there are at least 100 of that model listed for sale in North Carolina. Note that whether the companies are US companies or not is not contained within the data. It is up to you to determine which *Make* of vehicles are from US companies. 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 for your vehicle, with a range of at least six years.

Directions: The code below should walk you through the process of selecting data from a particular model vehicle 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}.

```
library(readr)

# This line will only run if the UsedCars.csv is stored in the same dir
ectory 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 da
ta.
## i Specify the column types or set `show_col_types = FALSE` to quiet
this message.

StateHW2 = "NC"

# Creates a dataframe with the number of each model for sale in North C
```

```r
arolina
Vehicles = as.data.frame(table(UsedCars$Model[UsedCars$State==StateHW2]
))

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

# Restricts the data to only models with at least 100 for sale
# Vehicles from non US companies are contained in this data
# Before submitting, comment this out so that it doesn't print while kn
itting
Enough_Vehicles = subset(Vehicles, Count>=100)

source("https://raw.githubusercontent.com/JA-McLean/STOR455/master/scri
pts/anova455.R")

# Delete the ** below and enter the model that you chose from the Enoug
h_Vehicles data.
ModelOfMyChoice = "EdgeSEL"

# Takes a subset of your model vehicle from North Carolina
MyVehicles = subset(UsedCars, Model==ModelOfMyChoice & State==StateHW2)

# Check to make sure that the vehicles span at least 6 years.
range(MyVehicles$Year)
```

```
## [1] 2007 2017
```

MODEL #1: Use Mileage as a predictor for Price

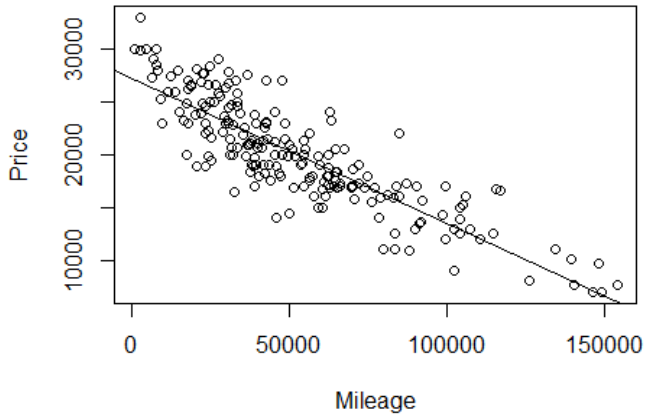
1. Calculate the least squares regression line that best fits your data using *Mileage* as the predictor and *Price* as the response. Interpret (in context) what the slope estimate tells you about prices and mileages of your used vehicle model. Explain why the sign (positive/negative) makes sense.

The price of a used car drops by $0.1375 for every additional mile of mileage on a used car. The older the car, the cheaper the price.

```
Myvehicles_Model=lm(Price~Mileage,data=MyVehicles)
summary(Myvehicles_Model)$coef[2,1]

## [1] -0.1374717
```

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

```
plot(Price~Mileage, MyVehicles)
abline(Myvehicles_Model)
```



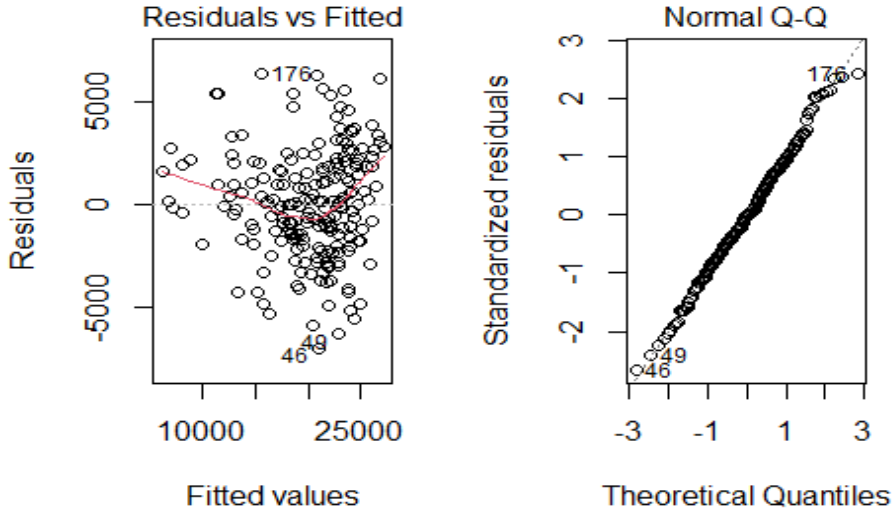
3. Produce appropriate residual plots and comment on how well your data appear to fit the conditions for a linear model. Don't worry about doing transformations at this point if there are problems with the conditions.

(1)Linearity: Not good nor bad. From the residuals vs Fitted plot, we can see that there are some skews.

(2)Constant variance(Homoscedasticity): Good. From the regression line we know that points fit similarly around the line.

(3)Normality: Good. From QQ plot we see that points are basically arranged on the dotted line.

```
par(mfrow=c(1,2))
plot(Myvehicles_Model,1:2)
```



4. Find the five vehicles in your sample with the largest residuals (in magnitude - positive or negative). For these vehicles, find their standardized and studentized residuals. Based on these specific residuals, would any of these vehicles be considered outliers? Based on these specific residuals, would any of these vehicles possibly be considered influential on your linear model?

They are neither outliers nor influential to the model.

The differences between their standardized and studentized residuals are small.

The difference between two plot is subtle.

All points are within 3 standard deviations.

From boxplot, all points are within Q1-1.5IQR to Q3+1.5IQR.

```
head(sort(Myvehicles_Model$residuals, decreasing=TRUE),n=5)

##      176      204        4      181      190
## 6335.459 6258.146 6109.223 5622.339 5536.492

MyVehicles[c(176,204,4,181,190),]

## # A tibble: 5 x 9
##       Id Price  Year Mileage City           State Vin
Make  Model
##    <dbl> <dbl> <dbl>   <dbl> <chr>          <chr> <chr>
<chr> <chr>
## 1 662044 21975  2015   84600 Raleigh        NC    2FMTK4J95FBB43217
Ford  EdgeS~
## 2 702924 26995  2015   47521 Southern Pines NC    2FMTK4J91FBB79437
Ford  EdgeS~
## 3  37734 32989  2017    2836 Wake Forest    NC    2FMPK3J98HBB95780
Ford  EdgeS~
## 4 665422 26995  2015   42896 Spruce Pine    NC    2FMTK4J8XFBB64023
Ford  EdgeS~
```

```
## 5 692710 28988  2016   27774 New Bern       NC    2FMPK4J85GBB79032
Ford  EdgeS~
```

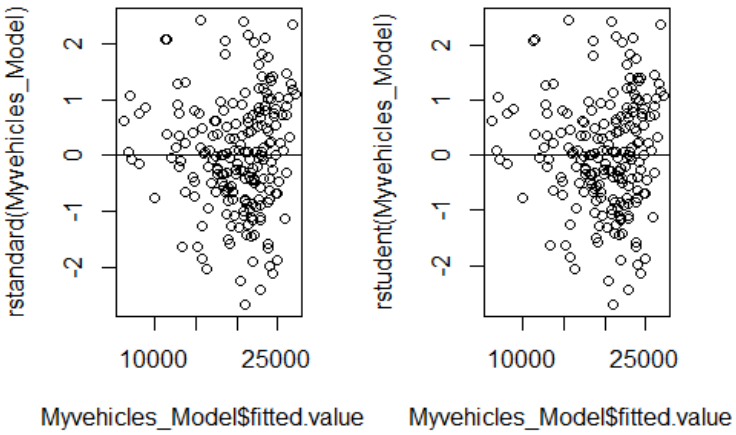
```
head(sort(rstandard(Myvehicles_Model), decreasing=TRUE),n=5)
```

```
##      176      204        4      181      190
## 2.429352 2.393954 2.350810 2.151088 2.120894
```

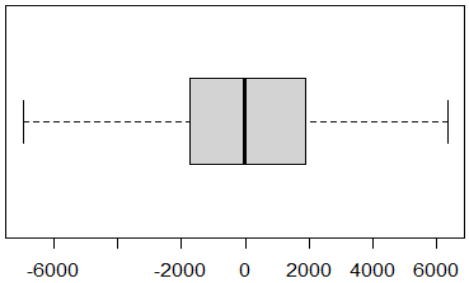
```
head(sort(rstudent(Myvehicles_Model), decreasing=TRUE),n=5)
```

```
##      176      204        4      181      190
## 2.459375 2.422491 2.377599 2.170665 2.139500
```

```
par(mfrow=c(1,2))
plot(rstandard(Myvehicles_Model)~Myvehicles_Model$fitted.values)
abline(0,0)
plot(rstudent(Myvehicles_Model)~Myvehicles_Model$fitted.values)
abline(0,0)
```




```
par(mfrow=c(1,1))
boxplot(Myvehicles_Model$residuals, horizontal=TRUE)
```



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

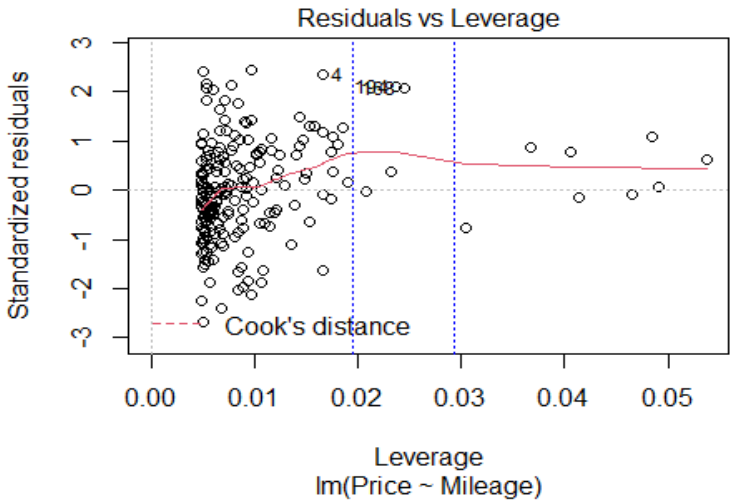
The first five items all have leverage values greater than the typical leverage values, so they MAY have an impact to my model. They only have an impact to the predictors, which influence horizontally.

```
2*(2/205)

## [1] 0.0195122

3*(2/205)

## [1] 0.02926829

head(sort(hatvalues(Myvehicles_Model), decreasing=TRUE),n=5)

##         200         141         187         122          85
## 0.05374010 0.04910199 0.04840700 0.04652949 0.04138682
```

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

They are not influential because they are within the lines of 0.5D1 and 1D2.

```
head(sort(cooks.distance(Myvehicles_Model), decreasing=TRUE), n=5)

##         168         194           4         176         187
## 0.05392434 0.05302410 0.04683496 0.02909679 0.02878701

plot(Myvehicles_Model,5)
abline(v = 4/205, col="blue", lty=3)
abline(v = 6/205, col="blue", lty=3)
```



7. Compute and interpret in context a 95% confidence interval for the slope of your regression line. Interpret (in context) what the confidence interval for the slope tells you about prices and mileages of your used vehicle model.

   There is a 95% chance that the intercept of the population is somewhere in the interval(26575,27963), which is when mile is 0, the price is somewhere in between(26575,27963).

   There is a 95% chance that the slope of the population is somewhere in the interval(-0.149,-0.126), which is the price of a used car drops by $0.149 to $0.126 for every additional mile of mileage on a used car.

```
confint(Myvehicles_Model, level=0.95)
```

```
##                      2.5 %        97.5 %
## (Intercept) 26575.4201992 27963.8737257
## Mileage        -0.1487511    -0.1261923
```

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.

   For singularity regression, they all have the same hypotheses.

   H0: β1=0, Ha:β1≠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.

   For singularity regression, they all have the same conclusions.

   All P values are small and the same.

   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.

```
summary(Myvehicles_Model)
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6969.9 -1719.8   -46.4  1870.9  6335.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.727e+04  3.521e+02   77.45   <2e-16 ***
## Mileage     -1.375e-01  5.721e-03  -24.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2621 on 203 degrees of freedom
```

```
## Multiple R-squared:  0.7399, Adjusted R-squared:  0.7386
## F-statistic: 577.5 on 1 and 203 DF,  p-value: < 2.2e-16

cor(MyVehicles[c(2:4)])

##              Price       Year    Mileage
## Price    1.0000000  0.8191843 -0.8601777
## Year     0.8191843  1.0000000 -0.7233227
## Mileage -0.8601777 -0.7233227  1.0000000

cor.test(MyVehicles$Price, MyVehicles$Mileage,use="complete.obs")

##
##  Pearson's product-moment correlation
##
## data:  MyVehicles$Price and MyVehicles$Mileage
## t = -24.031, df = 203, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8920614 -0.8197729
## sample estimates:
##        cor
## -0.8601777

anova455(Myvehicles_Model)

## ANOVA Table
## Model: Price ~ Mileage
##
##        Df     Sum Sq    Mean Sq F value     P(>F)
## Model   1 3966229713 3966229713  577.49 < 2.2e-16 ***
## Error 203 1394223824    6868098
## Total 204 5360453537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9. Suppose that you are interested in purchasing a vehicle of this model that has 50,000 miles on it (in 2017). Determine each of the following: 95% confidence interval for the mean price at this mileage and 95% prediction interval for the price of an individual vehicle at this mileage. Write sentences that carefully interpret each of the intervals (in terms of vehicles prices).

   For x of 500 miles in 2017, the average Y is in between $20033.99 to $20758.13

   For x of 500 miles in 2017, most Y are in between $15216.1 to $25576.03
   The first is the mean value of all data sets of 500 miles.

   The second one predicts the center of distribution and the variability around the center.

```
newx=data.frame(Mileage=50000,Year=2017)
head(newx)
```

```
##   Mileage Year
## 1   50000 2017
```

```
predict.lm(Myvehicles_Model, newx, interval="confidence")
```

```
##        fit      lwr      upr
## 1 20396.06 20033.99 20758.13
```

```
predict.lm(Myvehicles_Model, newx, interval="prediction")
```

```
##        fit     lwr      upr
## 1 20396.06 15216.1 25576.03
```

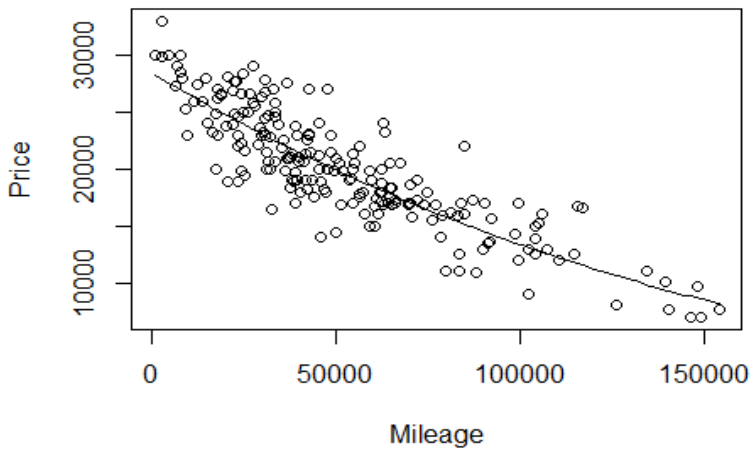
10. Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain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

    Power function transformation improves the linear model conditions.
    Multiple R-squared improves from 0.7399 to 0.7722.

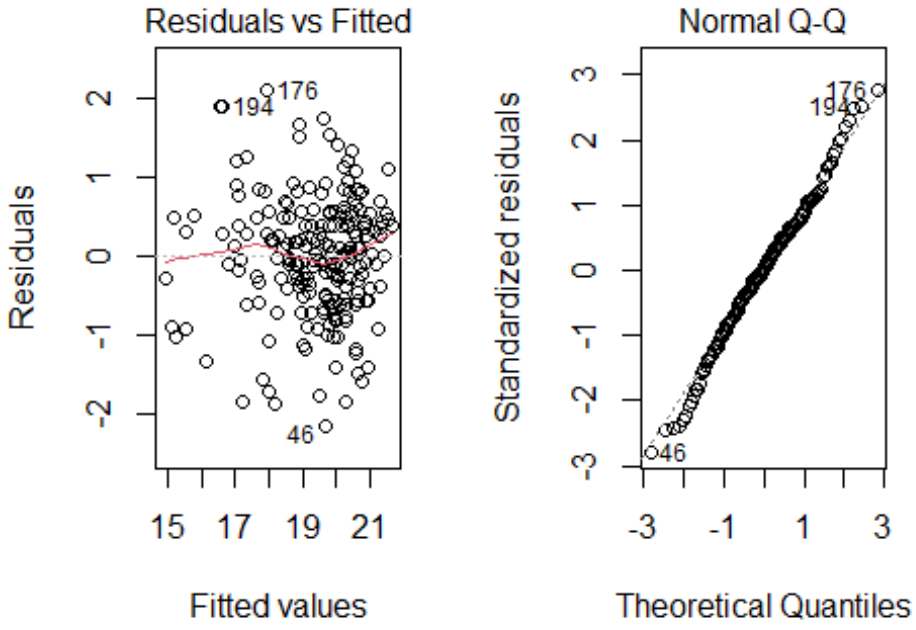
    The interpretation is better.

    The regression line fitted the scatterplot better. From the plot pictures, we know Linearity also improves.

```
Myvehicles_Model2=lm(Price^0.3~Mileage, data=MyVehicles)

B0 =summary(Myvehicles_Model2)$coef[1,1]
B1 = summary(Myvehicles_Model2)$coef[2,1]

plot(Price~Mileage, data=MyVehicles)
curve((B0+B1*x)^(10/3), add=TRUE)
```



```
summary(Myvehicles_Model2)

##
## Call:
## lm(formula = Price^0.3 ~ Mileage, data = MyVehicles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14055 -0.49156  0.01095  0.49032  2.10429
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.169e+01  1.032e-01  210.18   <2e-16 ***
## Mileage     -4.398e-05  1.677e-06  -26.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7681 on 203 degrees of freedom
## Multiple R-squared:  0.7722, Adjusted R-squared:  0.7711
## F-statistic: 688.2 on 1 and 203 DF,  p-value: < 2.2e-16

par(mfrow=c(1,2))
plot(Myvehicles_Model2,1:2)
```



11. According to your transformed model, is there a mileage at which the vehicle should be free? If so, find this mileage and comment on what the "free vehicle" phenomenon says about the appropriateness of your model.
    Myvehicles_Model2: y^0.3=B0+B1*x,
    when y=0, x=(-B0)/B1
    When mileage number is 493,106.3, the price is free.
    Typically standard cars can run up to 200,000 miles in a lifespan.
    The fitted value is much higher so the model is not that good.

```
(-B0)/B1

## [1] 493106.3
```

12. Again suppose that you are interested in purchasing a vehicle of this model that has 50,000 miles on it (in 2017). Determine each of the following using your transformed model: 95% confidence interval for the mean price at this mileage and 95% prediction interval for the price of an individual vehicle at this mileage. Write sentences that carefully interpret each of the intervals (in terms of vehicle prices).
    For x of 500 miles in 2017, the average Y is in between $19561.01 to $20284.09
    For x of 500 miles in 2017, most Y are in between $15201.81 to $25579
    The first is the mean value of all data sets of 500 miles.
    The second one predicts the center of distribution and the variability around the center.

```
newx2=data.frame(Mileage=50000,Year=2017)
head(newx2)

##   Mileage Year
## 1   50000 2017

predict.lm(Myvehicles_Model2, newx2, interval="confidence")^(10/3)

##        fit      lwr      upr
## 1 19920.25 19561.01 20284.09

predict.lm(Myvehicles_Model2, newx2, interval="prediction")^(10/3)

##        fit      lwr   upr
## 1 19920.25 15201.81 25579
```

MODEL #2: Again use Mileage as a predictor for Price, but now for new data

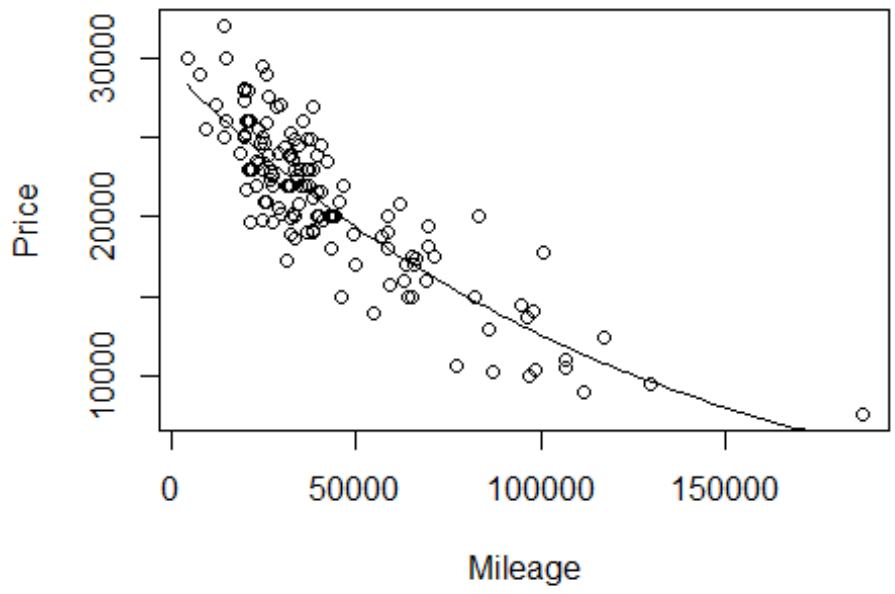
13. Select a new sample from the UsedCar dataset using the same *Model* vehicle that was used in the previous sections, but now from vehicles for sale in a different US state. You can mimic the code used above to select this new sample. You should select a state such that there are at least 100 of that model listed for sale in the new state.

    PENNSYLVANIA: 138 observations

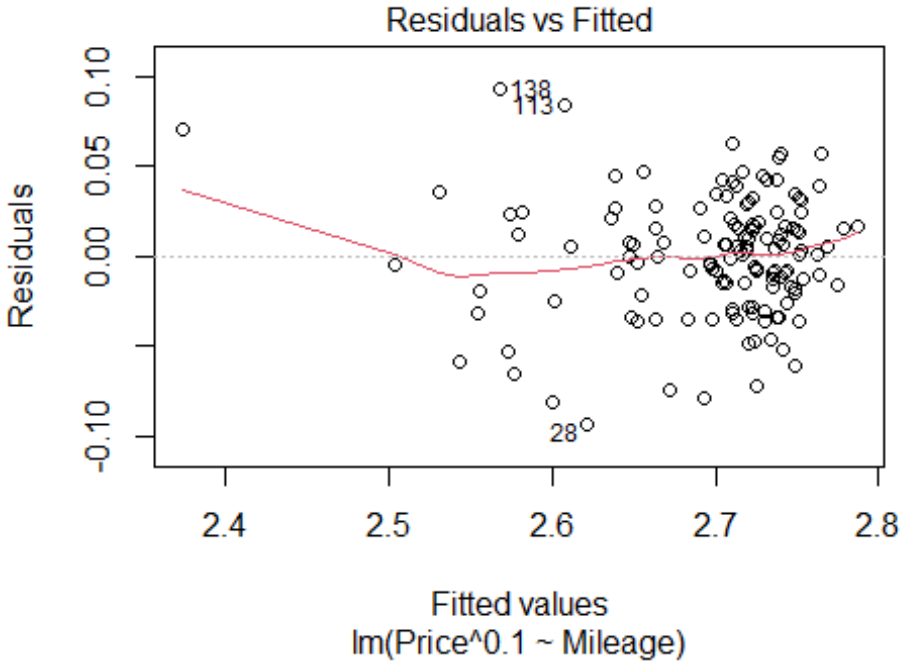
```
StateHW2.1 = "PA"
PAVehicles = subset(UsedCars, Model==ModelOfMyChoice & State==StateHW2.
1)
range(PAVehicles$Year)

## [1] 2007 2017
```

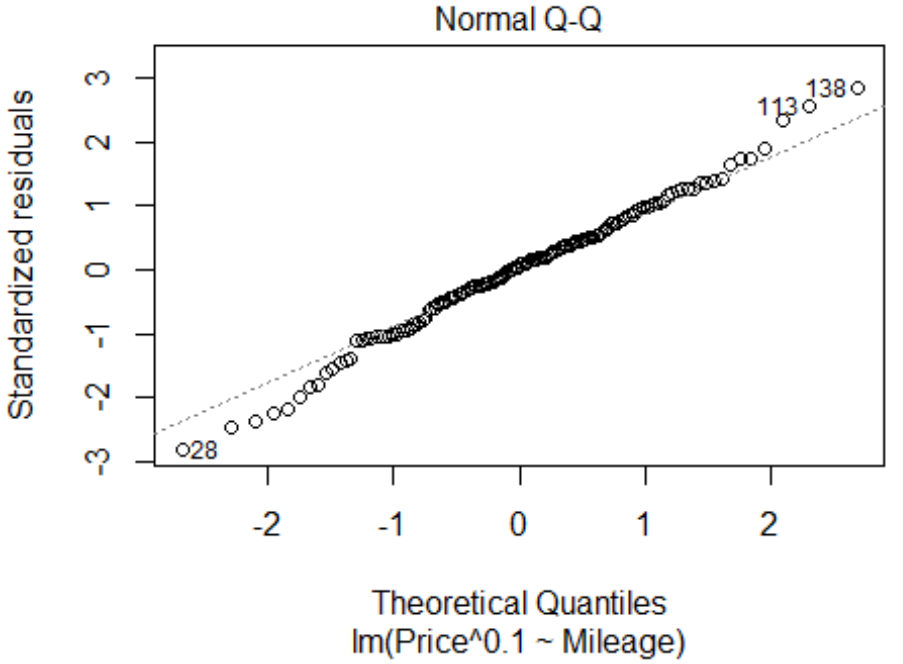
14. Calculate the least squares regression line that best fits your new data and produce a scatterplot of the relationship with the regression line on it.

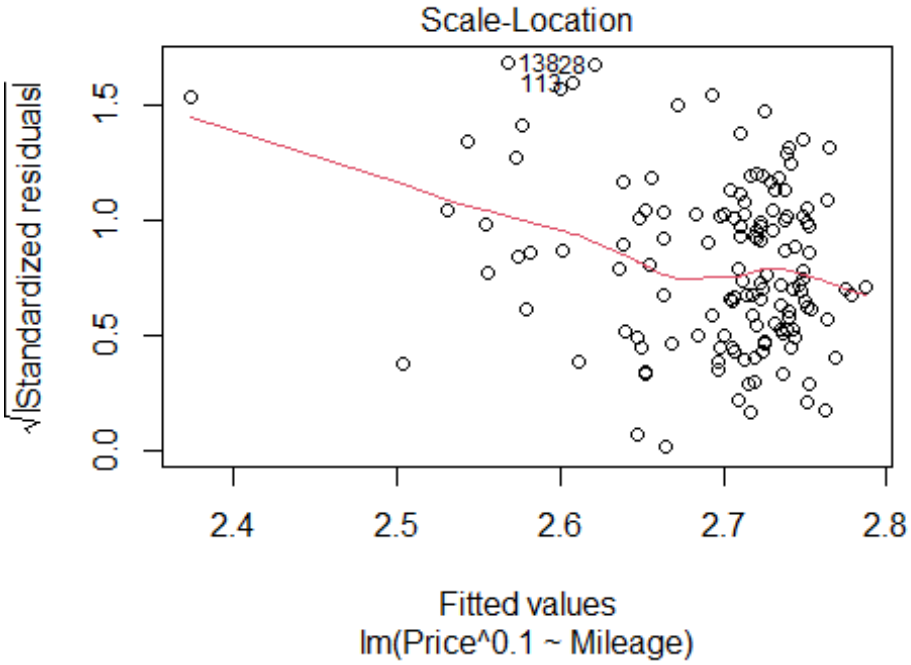
```
PAvehicles_Model=lm(Price^0.1~Mileage,data=PAVehicles)

B2 =summary(PAvehicles_Model)$coef[1,1]
B3 = summary(PAvehicles_Model)$coef[2,1]

plot(Price~Mileage, data=PAVehicles)
curve((B2+B3*x)^10, add=TRUE)
```

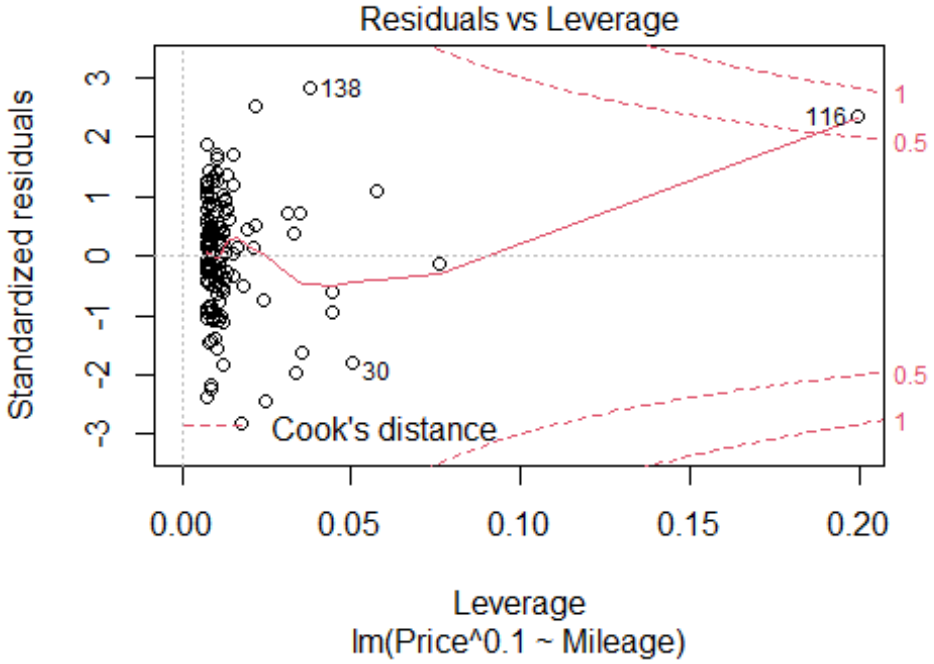



```
plot(PAvehicles_Model)
```


Residuals vs Fitted
Residuals
0.10
0.05
0.00
-0.10
138
113
28
2.4
2.5
2.6
2.7
2.8
Fitted values
lm(Price^0.1 ~ Mileage)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
138
113
28
-2
-1
0
1
2
Theoretical Quantiles
lm(Price^0.1 ~ Mileage)





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

They are still negatively correlated.
The relationship is similar.
The values of multiple R^2 from two states are close.
So mileage can explain 77%-78% variability of price.

```
summary(PAvehicles_Model)

##
## Call:
## lm(formula = Price^0.1 ~ Mileage, data = PAVehicles)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.093054 -0.019152  0.002085  0.020196  0.092830
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.796e+00  5.296e-03  527.95   <2e-16 ***
## Mileage     -2.259e-06  1.023e-07  -22.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03348 on 136 degrees of freedom
## Multiple R-squared:  0.7819, Adjusted R-squared:  0.7803
## F-statistic: 487.7 on 1 and 136 DF,  p-value: < 2.2e-16
```

16. Again suppose that you are interested in purchasing a vehicle of this model that has 50,000 miles on it (in 2017) from your new state. How useful do you think that your model will be? What are some possible cons of using this model?

    My model can provide a rough estimated range of prices, but it's far from accurate.

    Its flaws are that mileage can not fully explain price, and there are other predictors exist; and the point with high leverage (index 116) significantly affects the model.

```
PAnewx=data.frame(Mileage=50000,Year=2017)
head(PAnewx)

##   Mileage Year
## 1   50000 2017

predict.lm(PAvehicles_Model, PAnewx, interval="confidence")^10

##        fit     lwr      upr
## 1 19343.42 18930.7 19764.23

predict.lm(PAvehicles_Model, PAnewx, interval="prediction")^10

##        fit     lwr      upr
## 1 19343.42 15052.7 24705.14
```

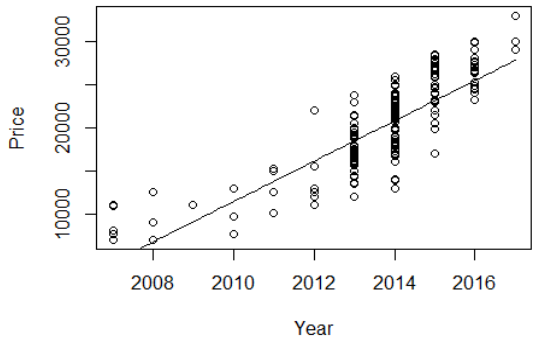
MODEL #3: Use Year as a predictor for Price

17. What proportion of the variability in the *Mileage* of your North Carolina vehicles' sale prices is explained by the *Year* of the vehicles? 52.32%

```
mod3=lm(Mileage~Year, data=MyVehicles)
summary(mod3)

##
## Call:
## lm(formula = Mileage ~ Year, data = MyVehicles)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59269 -14491  -3051  12870  63834
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26097179.1  1745050.1   14.96   <2e-16 ***
## Year          -12933.8      866.6  -14.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22200 on 203 degrees of freedom
## Multiple R-squared:  0.5232, Adjusted R-squared:  0.5208
## F-statistic: 222.8 on 1 and 203 DF,  p-value: < 2.2e-16
```

18. Calculate the least squares regression line that best fits your data using *Year* as the predictor and *Price* as the response. Produce a scatterplot of the relationship with the regression line on it.

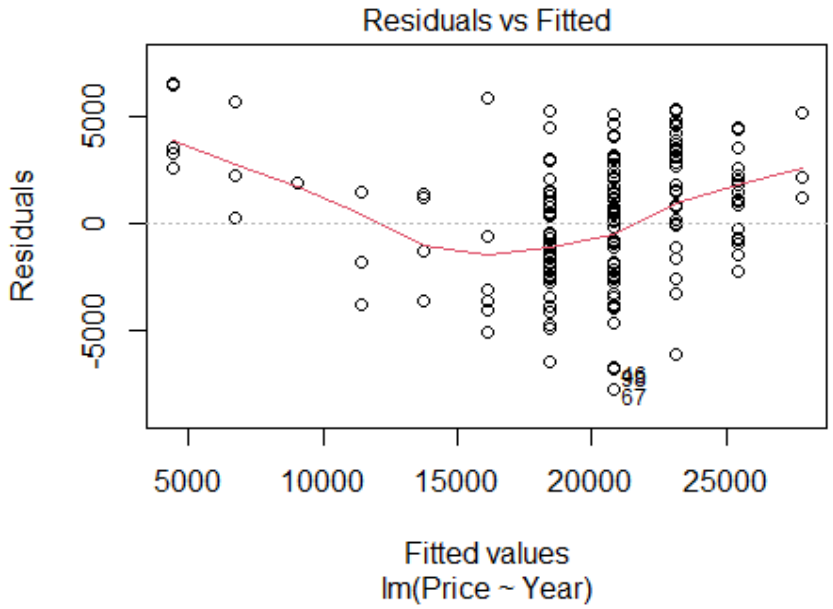
```
Model3=lm(Price~Year, data=MyVehicles)

B4 =summary(Model3)$coef[1,1]
B5 = summary(Model3)$coef[2,1]

plot(Price~Year, data=MyVehicles)
curve(B4+B5*x, add=TRUE)
```

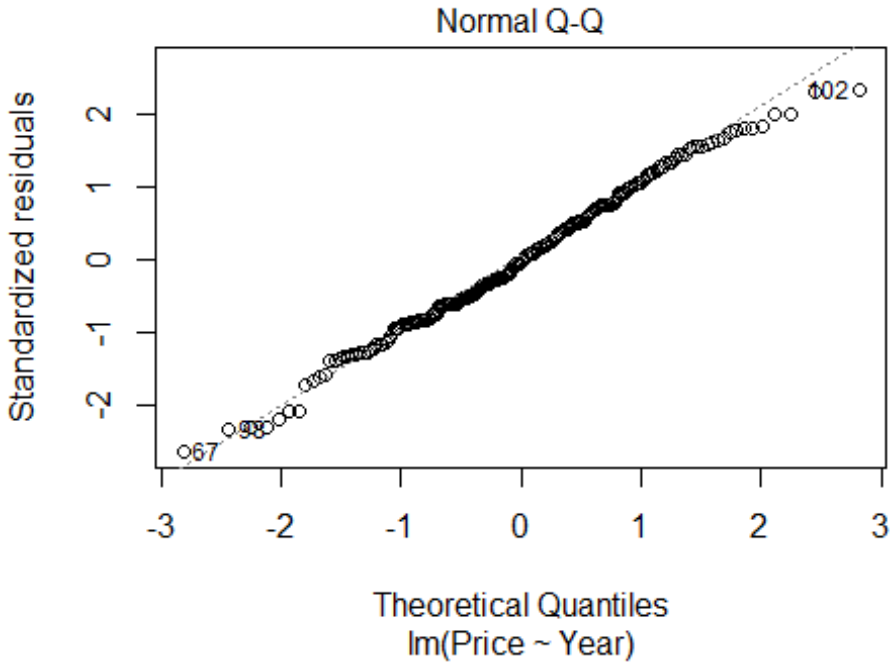


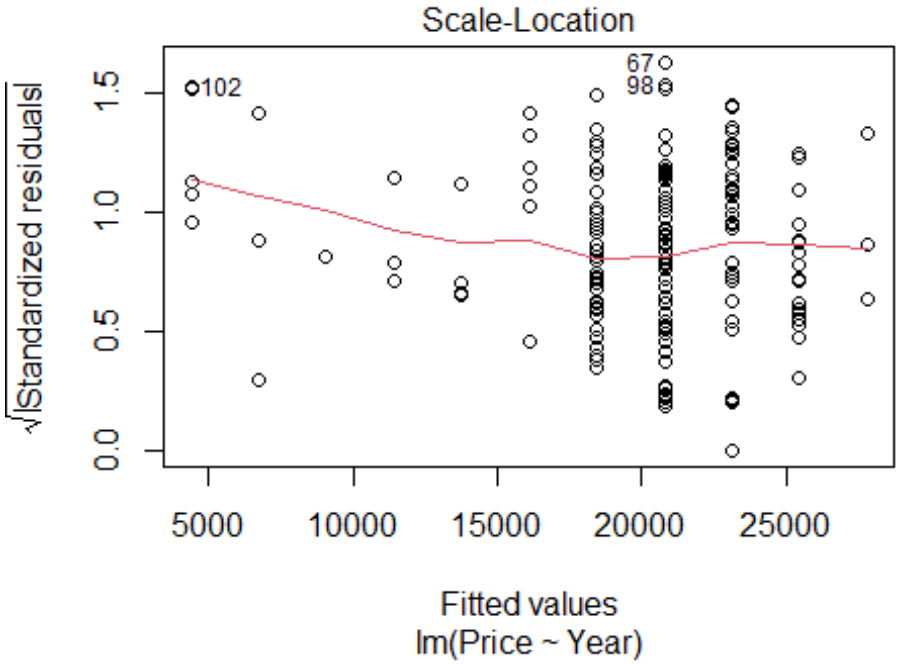
```
summary(Model3)

##
## Call:
## lm(formula = Price ~ Year, data = MyVehicles)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7789  -1916   -127   2183   6593
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4693989     231643  -20.26   <2e-16 ***
## Year            2341        115   20.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2947 on 203 degrees of freedom
## Multiple R-squared:  0.6711, Adjusted R-squared:  0.6694
## F-statistic: 414.1 on 1 and 203 DF,  p-value: < 2.2e-16
```

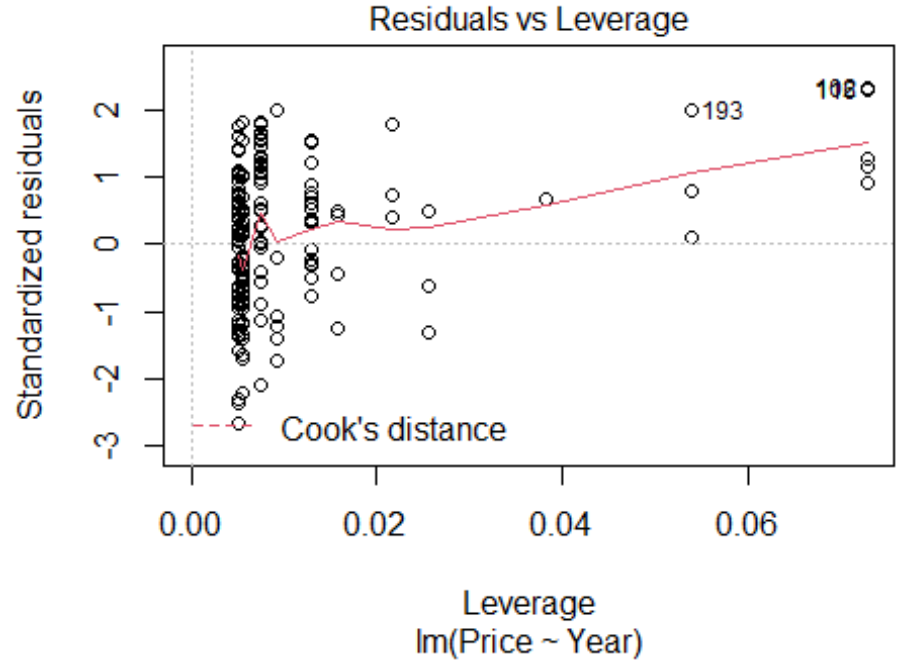
19. Produce appropriate residual plots and comment on how well your data appear to fit the conditions for a simple linear model. Don't worry about doing transformations at this point if there are problems with the conditions.
    (1)Linearity: Not good. There are some skews and points are not clustered on the horizontal line.
    (2)Constant variance(Homoscedasticity): Not Good.Points does not fit similarly around the regression line.
    (3)Normality:Not good.Points fall out from the line.

```
plot(Model3)
```


Residuals vs Fitted
Residuals
5000
0
-5000
5000
10000
15000
20000
25000
Fitted values
lm(Price ~ Year)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Price ~ Year)



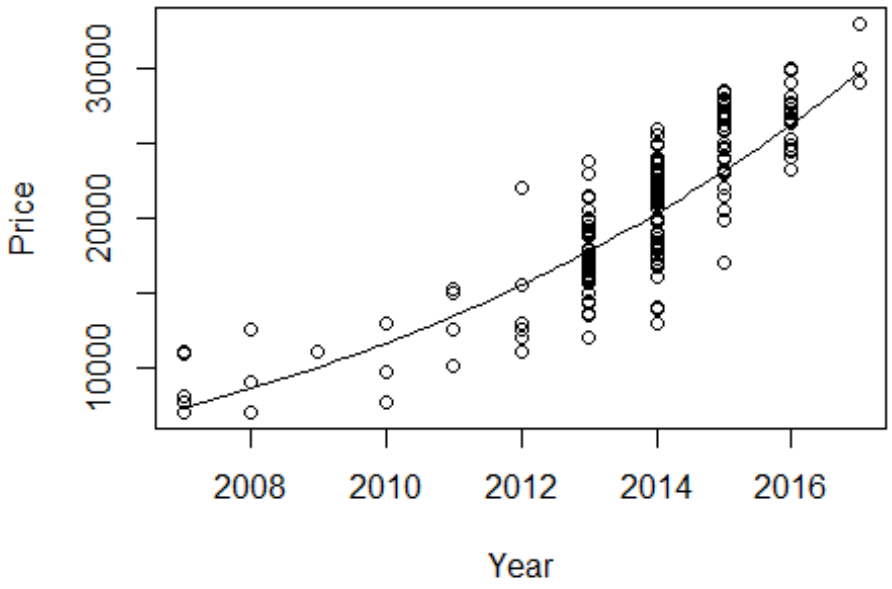


20. Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain why you

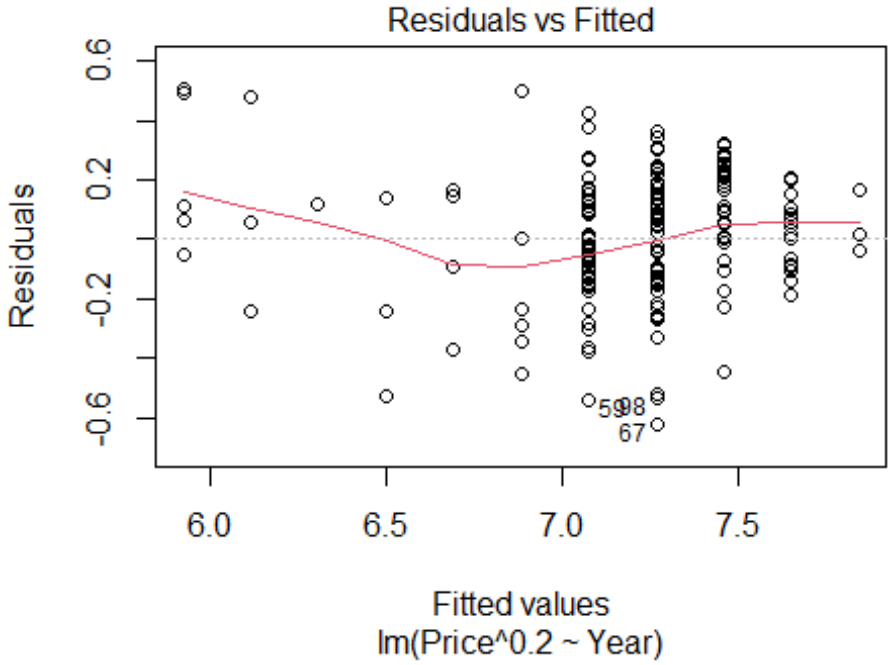
think that this transformation does or does not improve satisfying the linear model conditions.

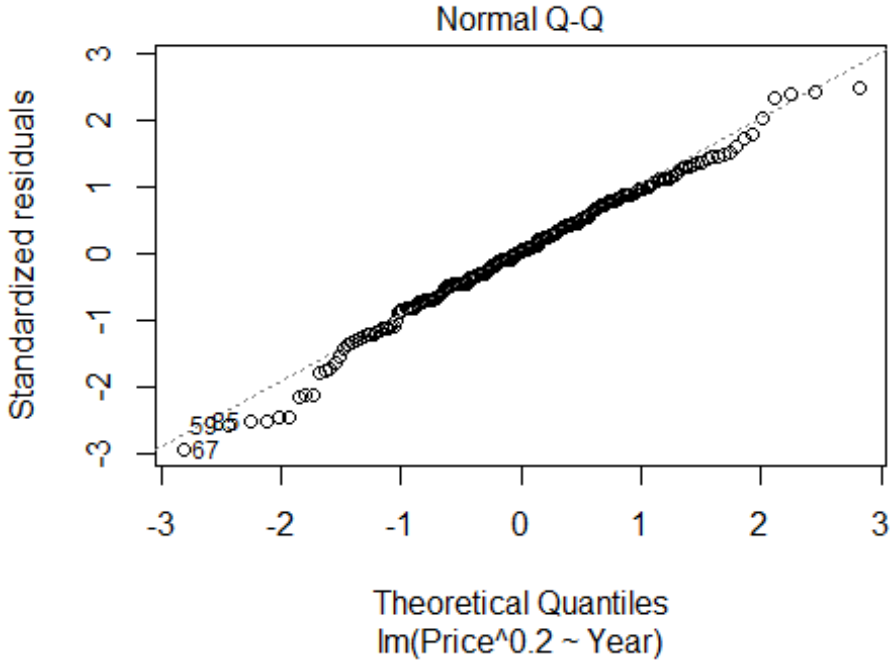
It does not improve. Although the linearity and multiple R^2 improve a little bit, other conditions like constant variance and normality do not improve, and residual vs leverage become worse.

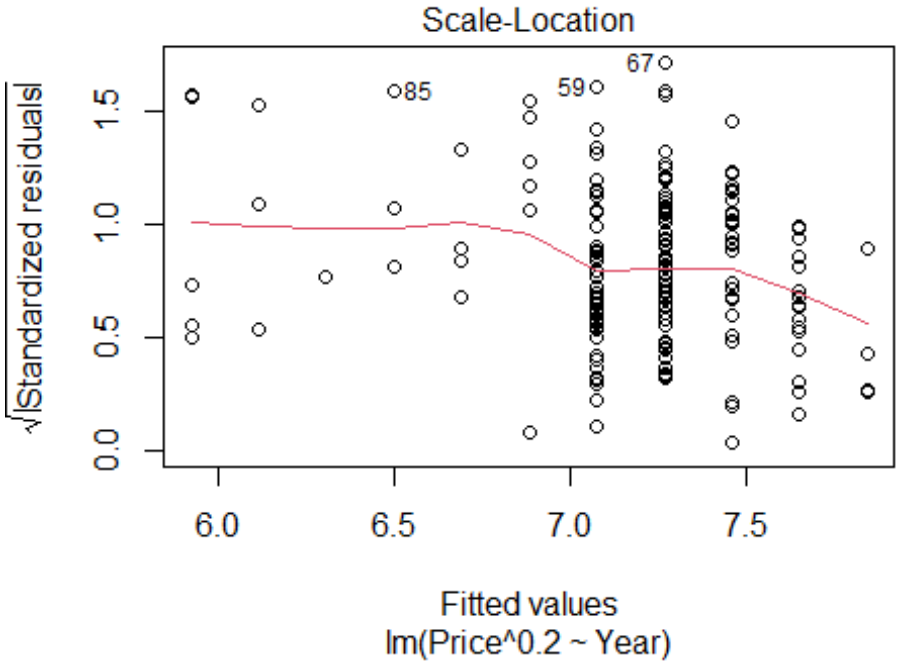
```
NewModel3=lm(Price^0.2~Year, data=MyVehicles)

B6 = summary(NewModel3)$coefficients[1,1]
B7 = summary(NewModel3)$coefficients[2,1]

plot(Price~Year, data=MyVehicles)
curve((B6+B7*x)^5, add=TRUE)
```

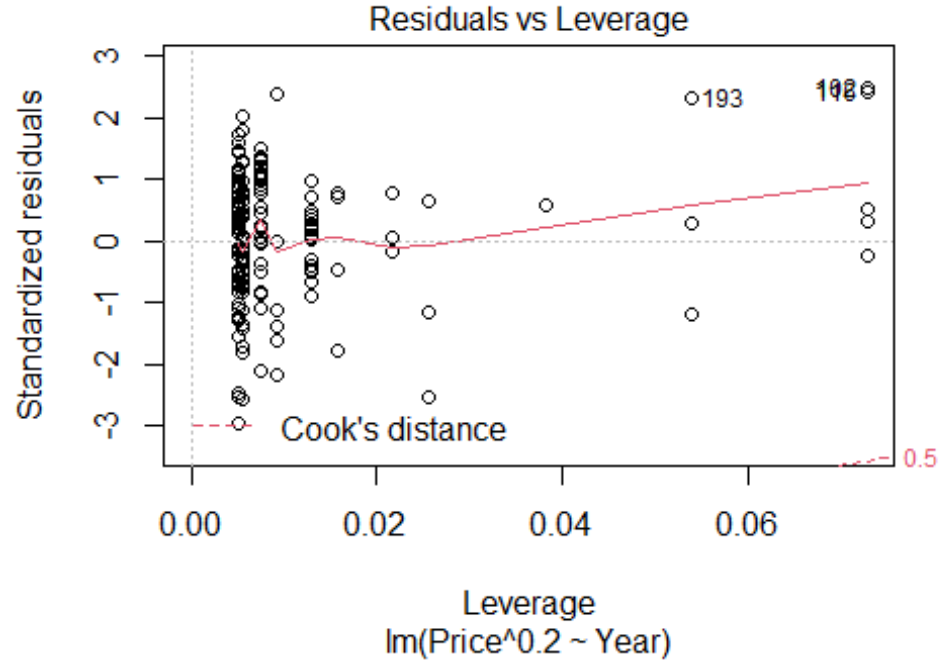



```
plot(NewModel3)
```


Residuals vs Fitted
Residuals
0.6
0.2
-0.2
-0.6
598
67
6.0
6.5
7.0
7.5
Fitted values
lm(Price^0.2 ~ Year)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
598
67
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Price^0.2 ~ Year)






```
summary(NewModel3)

##
## Call:
```

```
## lm(formula = Price^0.2 ~ Year, data = MyVehicles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62011 -0.12765  0.00929  0.15326  0.50446
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.793e+02  1.658e+01  -22.87   <2e-16 ***
## Year         1.920e-01  8.236e-03   23.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.211 on 203 degrees of freedom
## Multiple R-squared:  0.728,  Adjusted R-squared:  0.7266
## F-statistic: 543.3 on 1 and 203 DF,  p-value: < 2.2e-16
```

21. How do the transformed models, using either *Year* or *Mileage* as the predictor for your model of vehicle for sale in North Carolina compare? Does one of the models seem "better" or do they seem similar in their ability to predict *Price*? Explain.

    The model using mileage is better.

    Because the linear condition is better and the multiple r^2 value is larger. And Mileage can explain 52.32% of the variability of year.

    And compared with year, mileage is a better indicator of how worn a car is.