STOR 455 Homework 2

40 points - Due Thursday 2/9 at 12:00pm

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)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v dplyr   1.0.10
## v tibble  3.1.8      v stringr 1.5.0
## v tidyr   1.2.1      v forcats 0.5.2
## v purrr   1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# 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.
```

```
StateHW2 = "NC"

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

```r
# Vehicles from non US companies are contained in this data
# Before submitting, comment this out so that it doesn't print while knitting
Enough_Vehicles = subset(Vehicles, Count>=100)
Enough_Vehicles
```

```
##                Model Count
## 21        200Limited   191
## 34                 3   477
## 74                 5   174
## 130        AcadiaAWD   103
## 131        AcadiaFWD   259
## 139           Accord   776
## 141       AccordEX-L   132
## 149        Altima2.5   779
## 153        Altima4dr   131
## 245      CamaroCoupe   322
## 247         Camry4dr   106
## 251          CamrySE   133
## 284     ChallengerR/T   123
## 309  CherokeeLatitude   108
## 315            Civic   509
## 324          CivicLX   135
## 355     ColoradoCrew   112
## 384           Cooper   237
## 394       Corvette2dr   101
## 405           CR-VEX   127
## 406         CR-VEX-L   231
## 407           CR-VLX   115
## 423         Cruze1LT   120
## 434       CruzeSedan   185
## 438              CTS   132
## 464          DartSXT   124
## 500          EdgeSEL   205
## 504       Elantra4dr   178
## 508        ElantraSE   164
## 521    EnclaveLeather   144
## 545       EquinoxAWD   129
## 546       EquinoxFWD   454
## 550               ES   220
## 563        EscapeFWD   219
## 568         EscapeSE   230
## 570    EscapeTitanium   133
## 573             ESES   109
## 598   ExplorerLimited   138
## 603       ExplorerXLT   258
## 606         F-1502WD   225
## 607         F-1504WD   623
## 613      F-150Lariat   142
## 623         F-150XLT   332
## 685    FocusHatchback   161
## 689          FocusSE   181
## 690       FocusSedan   195
## 707          ForteLX   115
## 734         FusionSE   414
## 737    FusionTitanium   115
## 754              G37   124
```

```
## 801             Grand  1066
## 874                IS   158
## 876             Jetta   115
## 902       LaCrosseFWD   109
## 962         Malibu1LT   121
## 973          MalibuLS   121
## 974          MalibuLT   243
## 997           Mazda3i   128
## 1062       Mustang2dr   138
## 1070  MustangFastback   152
## 1071        MustangGT   151
## 1102      OdysseyEX-L   176
## 1109         OptimaEX   142
## 1111         OptimaLX   317
## 1161     PatriotSport   132
## 1166        PilotEX-L   122
## 1244              Ram   289
## 1305           RogueS   149
## 1307          RogueSV   148
## 1311            Rover   190
## 1316               RX   237
## 1318             RXRX   119
## 1352            Santa   386
## 1367         SedonaLX   111
## 1372          SentraS   149
## 1375         SentraSV   159
## 1389           Sierra   770
## 1390        Silverado  1807
## 1410       Sonata2.4L   224
## 1411        Sonata4dr   208
## 1428        SorentoLX   263
## 1431            Soul+   114
## 1433    SoulAutomatic   155
## 1463        SRXLuxury   109
## 1476      Suburban4WD   166
## 1479            Super   428
## 1483        Tacoma4WD   127
## 1488         Tahoe2WD   103
## 1490         Tahoe4WD   217
## 1506       TerrainFWD   212
## 1540             Town   250
## 1544          Transit   159
## 1548      TraverseFWD   162
## 1577           Tundra   109
## 1607            Versa   114
## 1625         Wrangler   604
## 1731            Yukon   176
## 1734         Yukon4WD   135
```

```r
# Delete the ** below and enter the model that you chose from the Enough_Vehicles data.
ModelOfMyChoice = "EquinoxFWD"

# 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] 2008 2018
```

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.

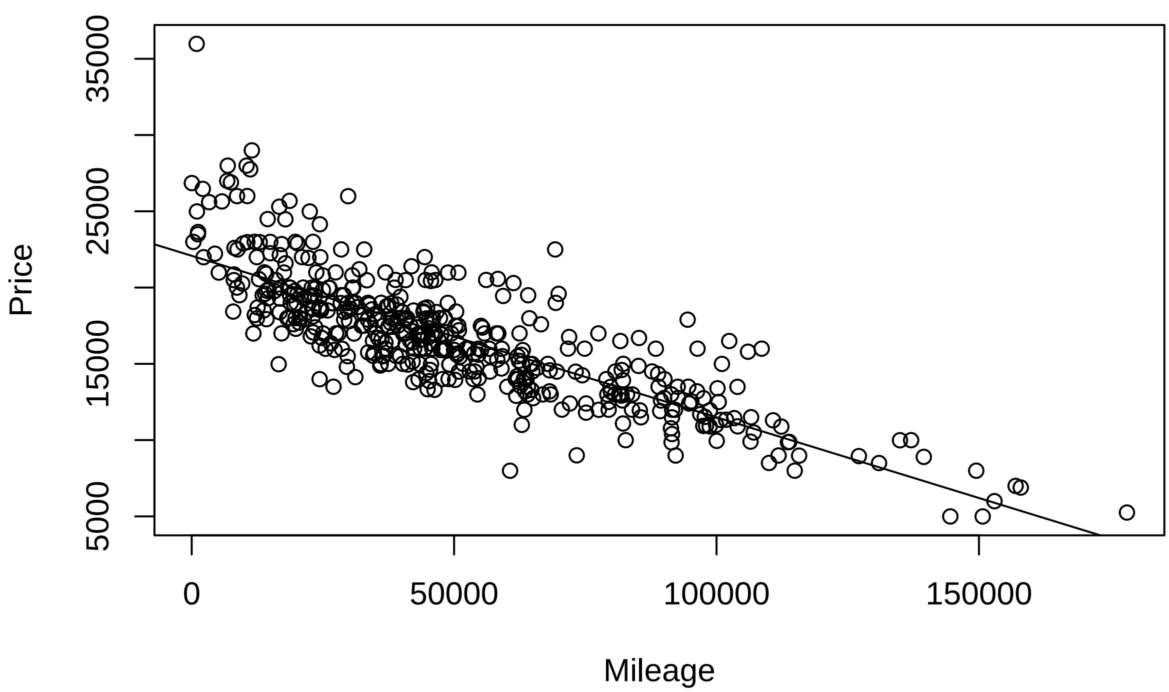
```
mileage_model = lm(Price~Mileage, data = MyVehicles)
summary(mileage_model)
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7669.9 -1532.7  -279.3  1090.9 13994.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.208e+04  2.091e+02  105.63   <2e-16 ***
## Mileage     -1.059e-01  3.526e-03  -30.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2388 on 452 degrees of freedom
## Multiple R-squared:  0.6661, Adjusted R-squared:  0.6653
## F-statistic: 901.6 on 1 and 452 DF,  p-value: < 2.2e-16
```

The slope of tell us that for an increase of one mile on the mileage, the predicted price of a used Equinox in North Carolina goes down by 0.105866 dollars. The sign is negative and this makes sense because we expect cars with higher mileages to be cheaper, or have been used more, so as the mileage goes up, the price goes down. This indicates an inverse relationship between Mileage and Price.

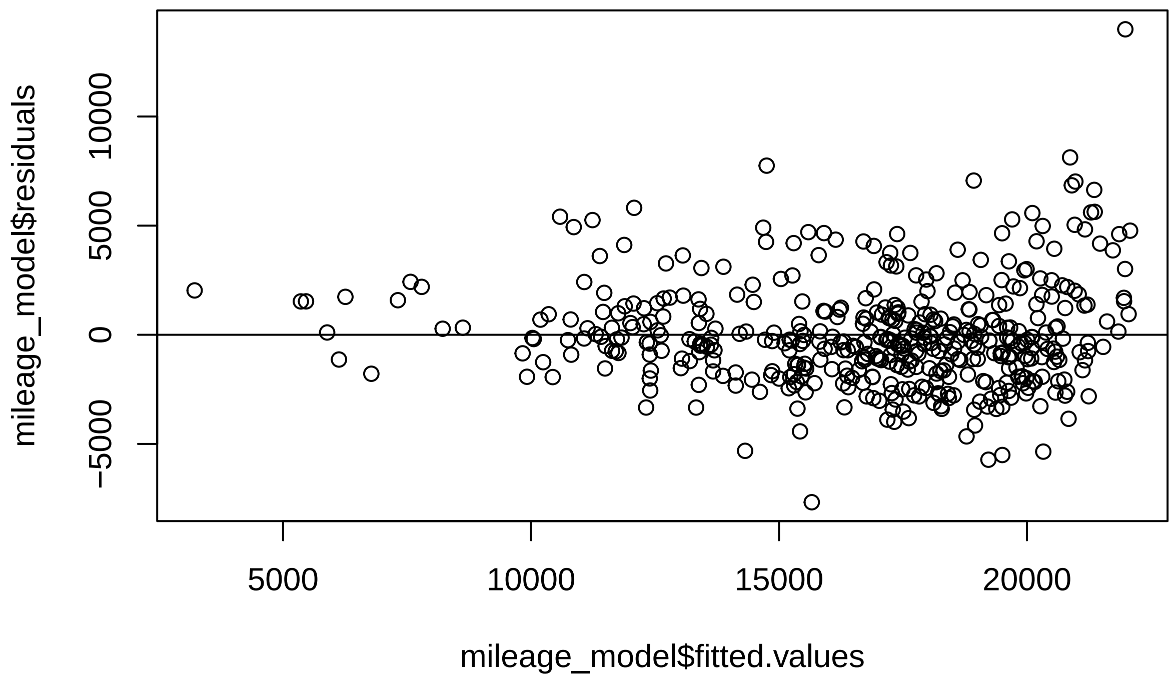
2. Produce a scatterplot of the relationship with the regression line on it.

```
plot(Price~Mileage, data = MyVehicles)
abline(mileage_model)
```

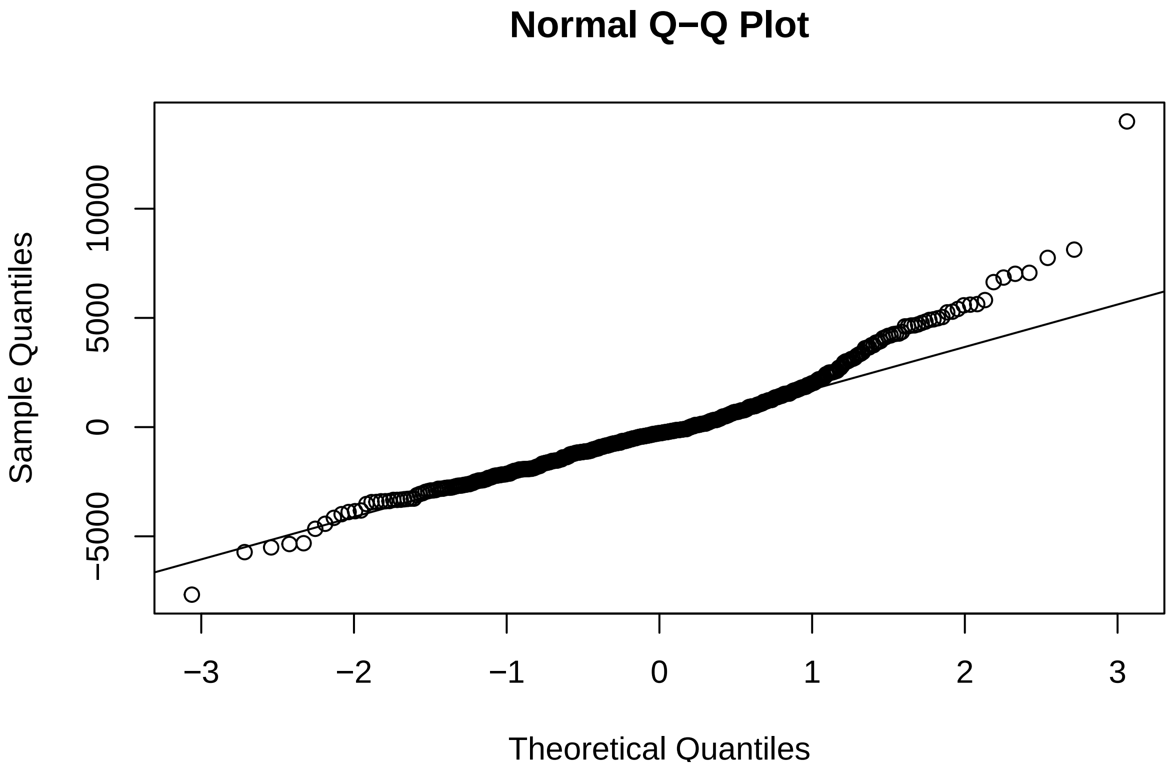


3. Produce appropriate residual plots and comment on how well your data appear to fit the conditions for a linear model. Don't worry about doing transformations at this point if there are problems with the conditions.

```
plot(mileage_model$residuals~mileage_model$fitted.values)
abline(0, 0)
```




```
qqnorm(mileage_model$residuals)
qqline(mileage_model$residuals)
```



In the residual plot there seems to be no noticeable pattern, except that all the fitted values are positive, which makes sense as our center is the average used Equinox price. In the residuals vs fitted values graph, there is not uniform spread in the error and seems to be a megaphone shape, as the predictor changes so does the spread. Also through our qqnorm plot we can see that the data is approximately normal so we can do inference on the data.

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?

```
head(sort(abs(mileage_model$residuals), decreasing = TRUE), 5)
```

```
##       438       425         2        13        15
## 13994.419  8126.127  7748.004  7669.912  7064.553
```

```
rstandard(mileage_model)[c(438, 425, 2, 13, 15)]
```

```
##       438       425         2        13        15
##  5.883506  3.412913  3.250114 -3.216455  2.963554
```

```
rstudent(mileage_model)[c(438, 425, 2, 13, 15)]
```

```
##       438       425         2        13        15
##  6.115844  3.453930  3.285130 -3.250307  2.989460
```

Based on these residuals I would consider all of these to be potential outliers with the Equinox at 438 to be a definite outlier. These vehicles all have the potential to be influential on my linear model because of their large distance away from the model.

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?

```
2 / 454
```

```
## [1] 0.004405286
```

```
2 * (2 / 454)
```

```
## [1] 0.008810573
```

```
3 * (2 / 454)
```

```
## [1] 0.01321586
```

```
hatvalues(mileage_model)[c(438, 425, 2, 13, 15)]
```

```
##         438         425           2          13          15
## 0.007462227 0.005450223 0.003006511 0.002447875 0.003096179
```

These leverage values show that the used Equinoxes with the highest residuals do not seem to have unusual leverages and all are contained within the normal boundaries. There is low potential for each of these five vehicles to be influential on my model.

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?

```
cooks.distance(mileage_model)[c(438, 425, 2, 13, 15)]
```

```
##        438        425          2         13         15
## 0.13012594 0.03191598 0.01592713 0.01269341 0.01363856
```

Cook's distance takes both the leverage (deviation on x) and residuals (deviation on y) into account and gives us a value that shows how a point influences the regression fit. Any Cook's distance under 0.5 is normal so these values tells us that the vehicles have normal influence over the model and there is nothing unusual about them.

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.

```r
summary(mileage_model)
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7669.9 -1532.7  -279.3  1090.9 13994.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.208e+04  2.091e+02  105.63   <2e-16 ***
## Mileage     -1.059e-01  3.526e-03  -30.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2388 on 452 degrees of freedom
## Multiple R-squared:  0.6661, Adjusted R-squared:  0.6653
## F-statistic: 901.6 on 1 and 452 DF,  p-value: < 2.2e-16
```

```r
t_score = qt(0.025, mileage_model$df.residual)
upper_bound <- summary(mileage_model)$coef[2, 1] + abs(t_score) * summary(mileage_model)$coef[2, 2]
lower_bound <- summary(mileage_model)$coef[2, 1] - abs(t_score) * summary(mileage_model)$coef[2, 2]
sprintf("[%f, %f]", lower_bound, upper_bound)
```

```
## [1] "[-0.112795, -0.098937]"
```

```r
confint(mileage_model, level = 0.95)
```

```
##                     2.5 %        97.5 %
## (Intercept) 21671.9968264  2.249367e+04
## Mileage        -0.1127947 -9.893685e-02
```

This confidence interval tells us that we are 95% confident that the true slope of mileage vs price is between -0.112795 and -0.098937 Because the entire confidence interval is negative and does not contain zero, we can know there is a significant association between mileage and price in our used car model.

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.

```r
cor.test(MyVehicles$Mileage,
         MyVehicles$Price)
```

```
##
##  Pearson's product-moment correlation
##
## data:  MyVehicles$Mileage and MyVehicles$Price
## t = -30.026, df = 452, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8447162 -0.7829046
## sample estimates:
##        cor
## -0.8161317
```

The first test is the test for correlation and our null hypothesis is that there is r is equal to zero or there is no correlation between Mileage and Price, while our alternative is that r does not equal zero. After doing our test, we got a test statistic of -30.026 and a p-value that is approximately zero. That mean we can reject that r is equal to zero and have convincing evidence that there is some correlation between Mileage and Price.

```
summary(mileage_model)$coef[2,]
```

```
##       Estimate     Std. Error        t value       Pr(>|t|)
##  -1.058658e-01   3.525764e-03  -3.002634e+01  1.017974e-109
```

In the test for slope we assume that our null hypothesis is that the slope is equal to zero and our alternative hypothesis is that the slope does not equal zero. Through the test, we get a test statistics of -30.026, which gives us a p-value that is approximately zero. So we can reject that the slope is equal to zero and have convincing evidence that the slope does not equal zero. Although for the correlation and slope we eliminate the possibility of them being zero, the test gives us no idea of the direction or strength of the relation.

```
anova(mileage_model)
```

```
## Analysis of Variance Table
##
## Response: Price
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## Mileage     1 5139194396 5139194396  901.58 < 2.2e-16 ***
## Residuals 452 2576491547    5700203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis and alternative hypothesis for the anova test are the same as the ones for the regression test: null hypothesis is the slope is zero and the alternative hypothesis is the slope is not equal to zero. Through the test we get a F value of 901.58 with a p-value of approximately zero, so we have the same results as the regression test. We can reject the fact that the slope is equal to zero and have convincing evidence that the slope does not equal zero.

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).

```
sample_mileage <- data.frame(Mileage = 50000)

predict.lm(mileage_model, sample_mileage, level = 0.95, interval="confidence")
```

```
##        fit      lwr      upr
## 1 16789.55 16569.34 17009.75
```

We are 95% that the true mean price of used Equinoxes with 50000 miles on them is between 16569.34 dollars and 17009.75 dollar. This means if we drew multiple samples from the population of used Equinoxes and did a CI on the true mean price of Equinoxes with 50000 mileage for each sample, then we expect that true mean price to be within 95% of these intervals.

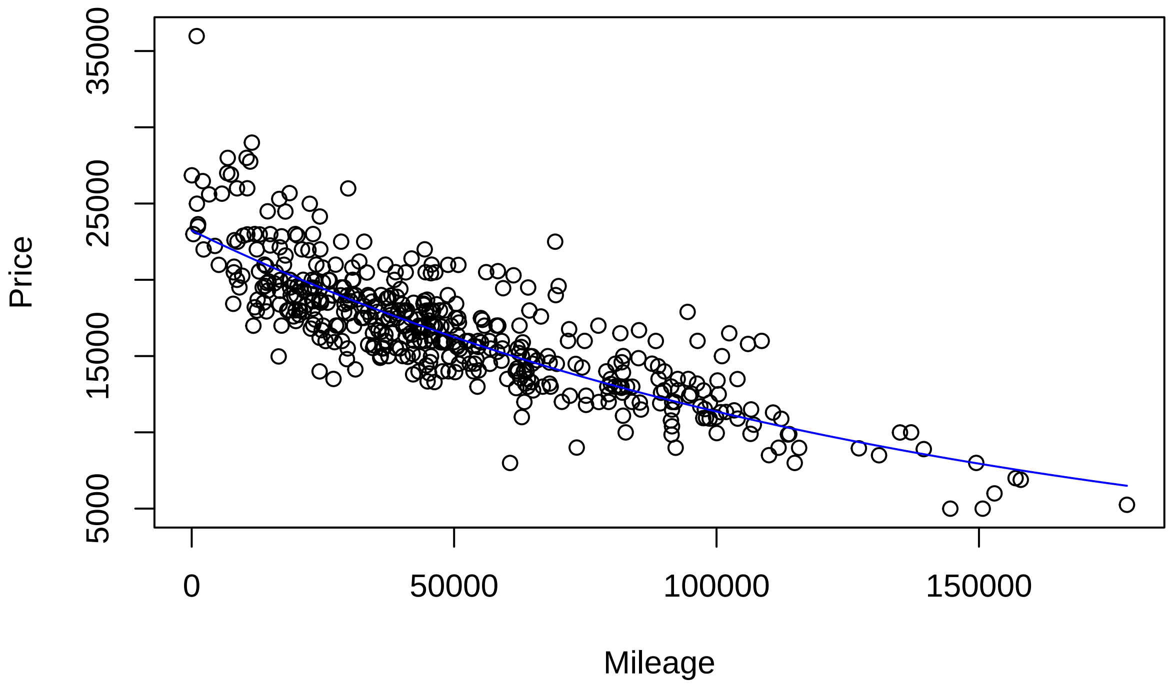
```
predict.lm(mileage_model, sample_mileage, level = 0.95, interval="prediction")
```

```
##        fit      lwr      upr
## 1 16789.55 12092.39 21486.71
```

We expect that 95% of Equinoxes with 50000 miles on them will have a price between 12092.39 and 21486.71 This interval is always larger than the confidence interval because you are trying to predict individual values, which means you need to account for values further from the fitted value. This also means if we get future used Equinoxes that have 50000 miles, then there is 95% chance that it will be contained in this interval.

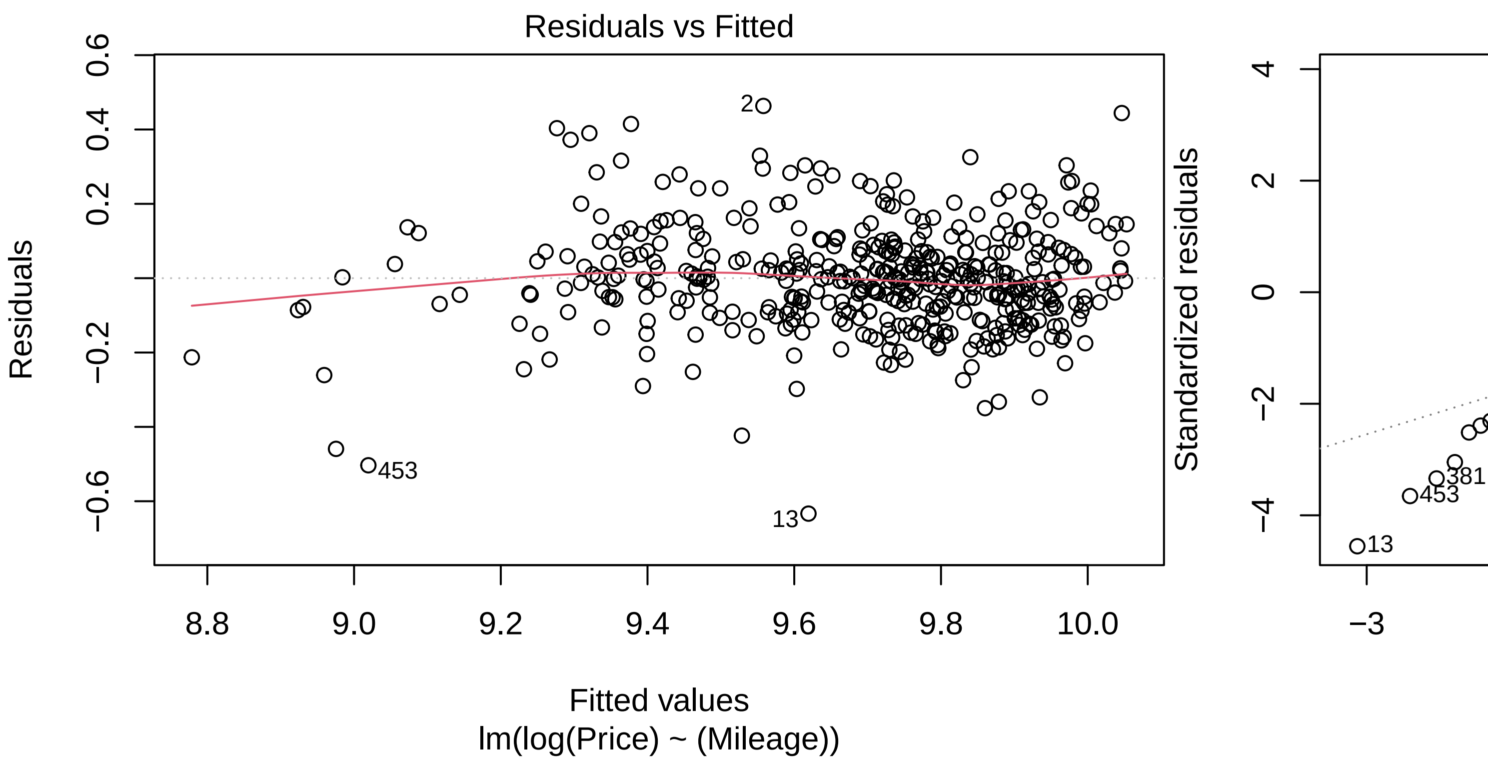
10. Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain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

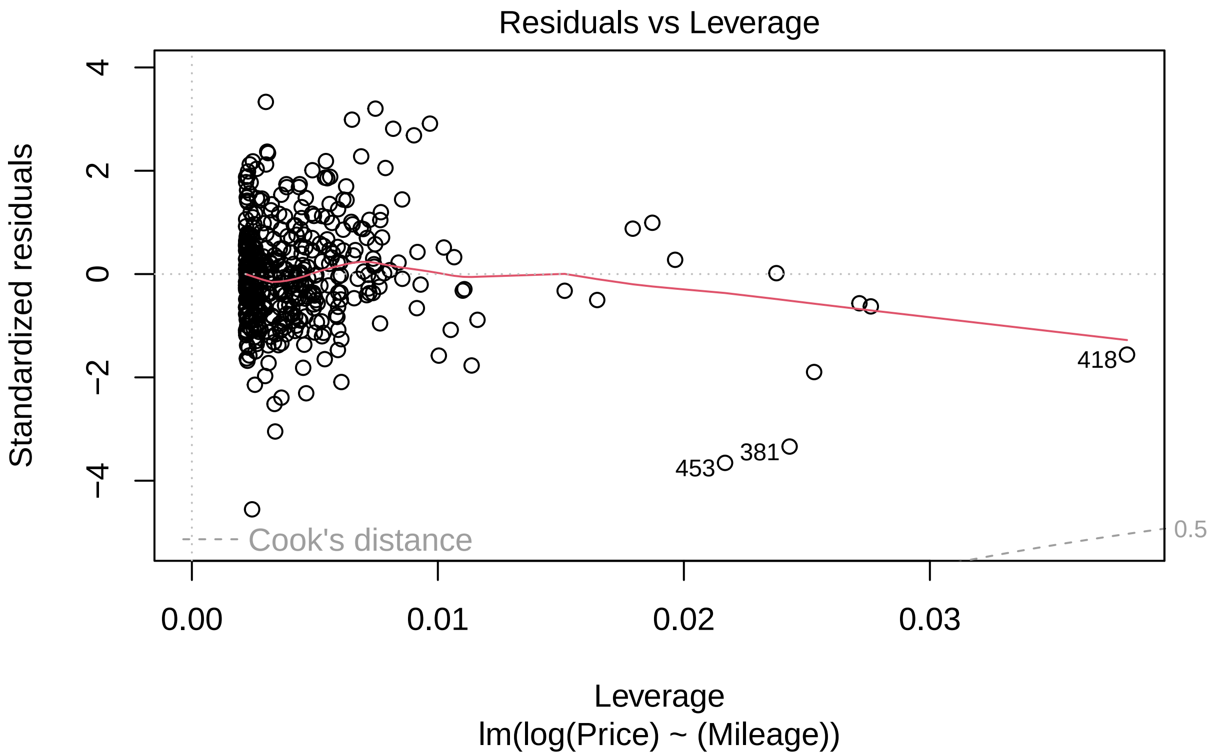
```
mileage_model2 <- lm(log(Price)~(Mileage), MyVehicles)
plot(Price~Mileage, MyVehicles)
curve(exp(mileage_model2$coefficients[1])/exp(abs(mileage_model2$coefficients[2]) * x), add = TRUE, col =
```



log(Predicted Price) = 10.053 - 0.00000715(Mileage) Predicted Price = e^(10.053 - 0.00000715(Mileage) Predicted Price = ($e^{10.053)/(e}$(0.00000715(Mileage))) Predicted Price = exp(intercept)/exp(slope(Mileage)))

```
plot(mileage_model2, c(1, 2, 5))
```





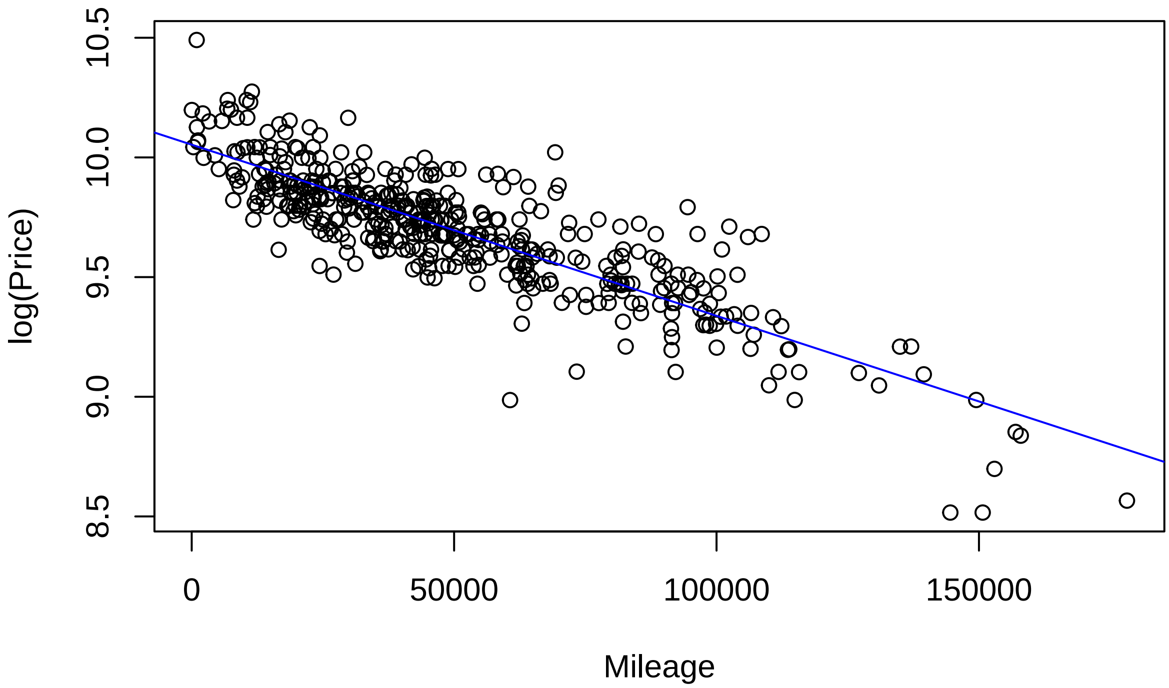

```
summary(mileage_model2)
```

```
##
## Call:
## lm(formula = log(Price) ~ (Mileage), data = MyVehicles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63333 -0.08415 -0.00366  0.07246  0.46330
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.005e+01  1.219e-02  824.67   <2e-16 ***
## Mileage     -7.152e-06  2.056e-07  -34.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1392 on 452 degrees of freedom
## Multiple R-squared:  0.728,  Adjusted R-squared:  0.7274
## F-statistic:  1210 on 1 and 452 DF,  p-value: < 2.2e-16
```

This transformation has improved the linear model conditions because the spread of the residuals vs the fitted values is roughly more uniform. The conditions for normality seem to have worsened because there are more departures from the linear pattern of the qqnorm plot.

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.

```
plot(log(Price)~(Mileage), MyVehicles)
abline(mileage_model2, col = "blue")
```



log(Predicted Price) = 10.053 - 0.00000715(Mileage) 0 = 10.053 - 0.00000715x 0.00000715x = 10.053 x = 1406013.98601

There is a mileage that the vehicle will become essentially free because it is a linear model and that would be 1.4 million miles on the car. This shows that our model cannot be extrapolated far past what the data predicts because it does not make sense for a car to be sold for free so we need to be careful when using this model for high mileages.

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).

```
predict.lm(mileage_model2, sample_mileage, level = 0.95, interval = "confidence")
```

```
##       fit      lwr     upr
## 1 9.69564 9.682799 9.70848
```

We are 95% confidence that the true mean log(price) of used Equinoxes with 50000 miles is between 9.682799 and 9.70848.

```
predict.lm(mileage_model2, sample_mileage, level = 0.95, interval = "prediction")
```

```
##       fit      lwr      upr
## 1 9.69564 9.421735 9.969544
```

We expect that 95% of Equinoxes with 50000 miles on them will have a log(price) between 9.421735 and 9.969544

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

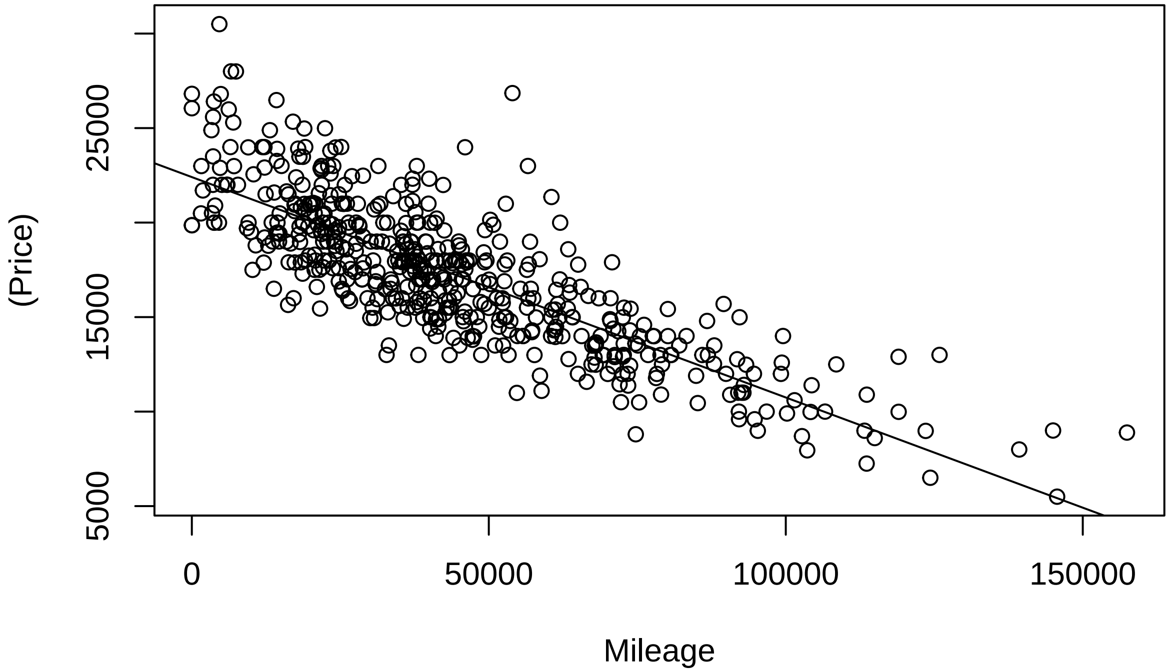
13. Select a new sample from the UsedCar dataset using the same *Model* vehicle that was used in the previous sections, but now from vehicles for sale in a different US state. You can mimic the code used above to select this new sample. You should select a state such that there are at least 100 of that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

```
MyVehiclesCA <- subset(UsedCars, State == "CA" & Model == ModelOfMyChoice)
nrow(MyVehiclesCA)
```

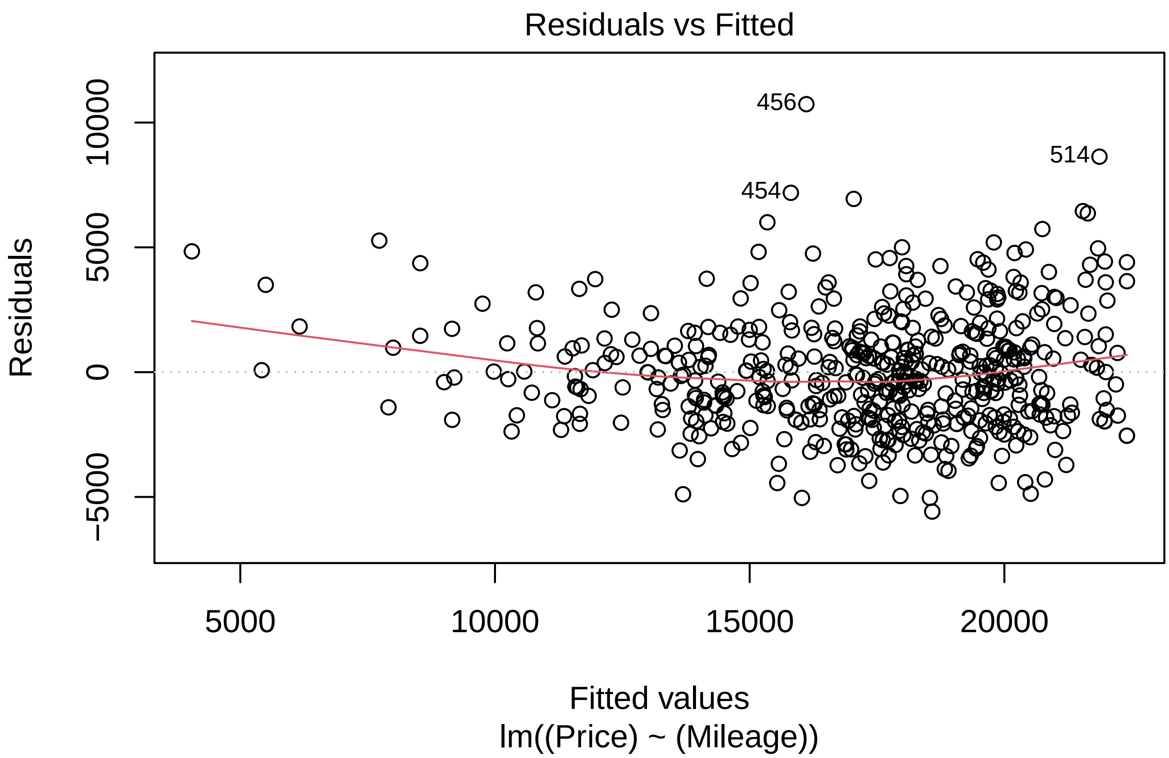
```
## [1] 516
```

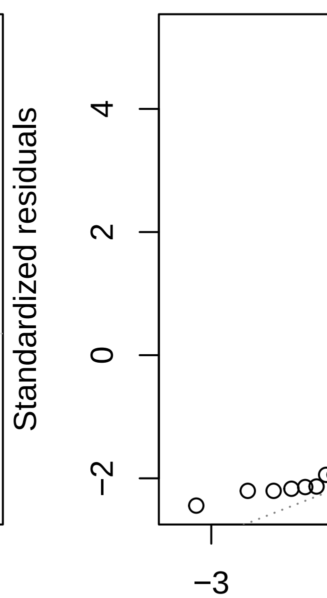
14. Calculate the least squares regression line that best fits your new data and produce a scatterplot of the relationship with the regression line on it.

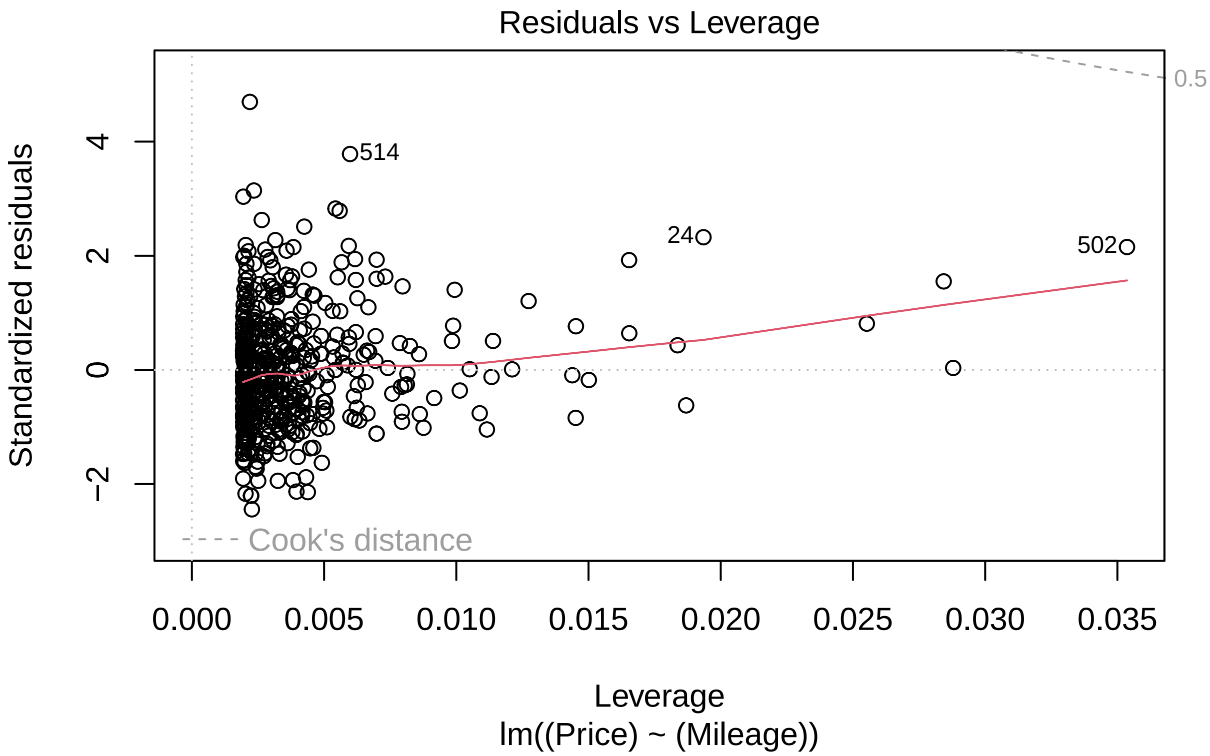
```r
plot((Price)~(Mileage), data = MyVehiclesCA)
mileage_model3 <- lm((Price)~(Mileage), data = MyVehiclesCA)
abline(mileage_model3)
```




```r
plot(mileage_model3, c(1, 2, 5))
```







Predicted Price = 22407.491791 - 0.116605(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.

```
summary(mileage_model)
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7669.9 -1532.7  -279.3  1090.9 13994.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.208e+04  2.091e+02  105.63   <2e-16 ***
## Mileage     -1.059e-01  3.526e-03  -30.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2388 on 452 degrees of freedom
## Multiple R-squared:  0.6661, Adjusted R-squared:  0.6653
## F-statistic: 901.6 on 1 and 452 DF,  p-value: < 2.2e-16
```

```
summary(mileage_model3)
```

```
##
## Call:
## lm(formula = (Price) ~ (Mileage), data = MyVehiclesCA)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
##  -5586  -1694   -195   1246  10737
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.241e+04  1.914e+02  117.10   <2e-16 ***
## Mileage     -1.166e-01  3.691e-03  -31.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2289 on 514 degrees of freedom
## Multiple R-squared:   0.66,  Adjusted R-squared:  0.6593
## F-statistic: 997.8 on 1 and 514 DF,  p-value: < 2.2e-16
```

The relationship between price and mileage on this new data is very similar to the relationship seen in the original data. Both have a strong negative linear relationship and both models are very similar. I can spot small discrepancies in the calculated slope and intercept but those are negligible.

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?

```
predict.lm(mileage_model3, sample_mileage, level = 0.95, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 16577.26 16374.67 16779.86
```

```
predict.lm(mileage_model3, sample_mileage, level = 0.95, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 16577.26 12075.64 21078.88
```

I think the model will be very useful because we are not extrapolating our model as there lots of data points representing used Equinoxes with 50000 mileage. Also through the confidence and prediction intervals, we can say that the model will return a positive value with 95% confidence and that makes sense for our situation. I think the model will be useful for predicting prices for mileages that are within its bounds. Some cons of using this model is that you can eventually get a negative price once the mileage gets to a certain point, and that makes no sense for our situation.

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?

```
mileage_year_model <- lm(Mileage~Year, data = MyVehicles)
anova(mileage_year_model)
```

```
## Analysis of Variance Table
##
## Response: Mileage
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## Year        1 2.7385e+11 2.7385e+11  670.16 < 2.2e-16 ***
## Residuals 452 1.8470e+11 4.0863e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mileage_year_model)
```

```
##
## Call:
## lm(formula = Mileage ~ Year, data = MyVehicles)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -66286 -13203  -1777  10671  74107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25232718     972774   25.94   <2e-16 ***
## Year          -12503        483  -25.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20210 on 452 degrees of freedom
## Multiple R-squared:  0.5972, Adjusted R-squared:  0.5963
## F-statistic: 670.2 on 1 and 452 DF,  p-value: < 2.2e-16
```

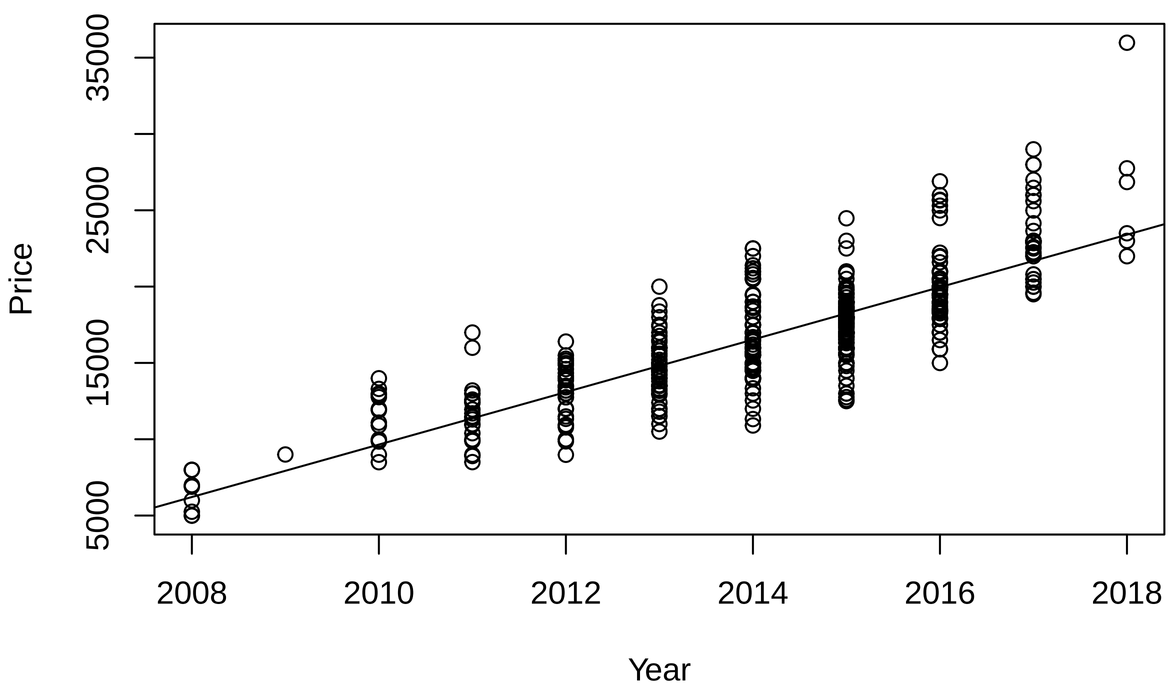
```r
273846989804 / (273846989804 + 184700089694)
```

```
## [1] 0.5972058
```

59.72% of the variability the mileage of North Carolina used Equinoxes can be explained by the Year of the vehicle.

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.

```r
price_year_model <- lm(Price~Year, data = MyVehicles)
plot(Price~Year, data = MyVehicles)
abline(price_year_model)
```




```r
summary(price_year_model)
```

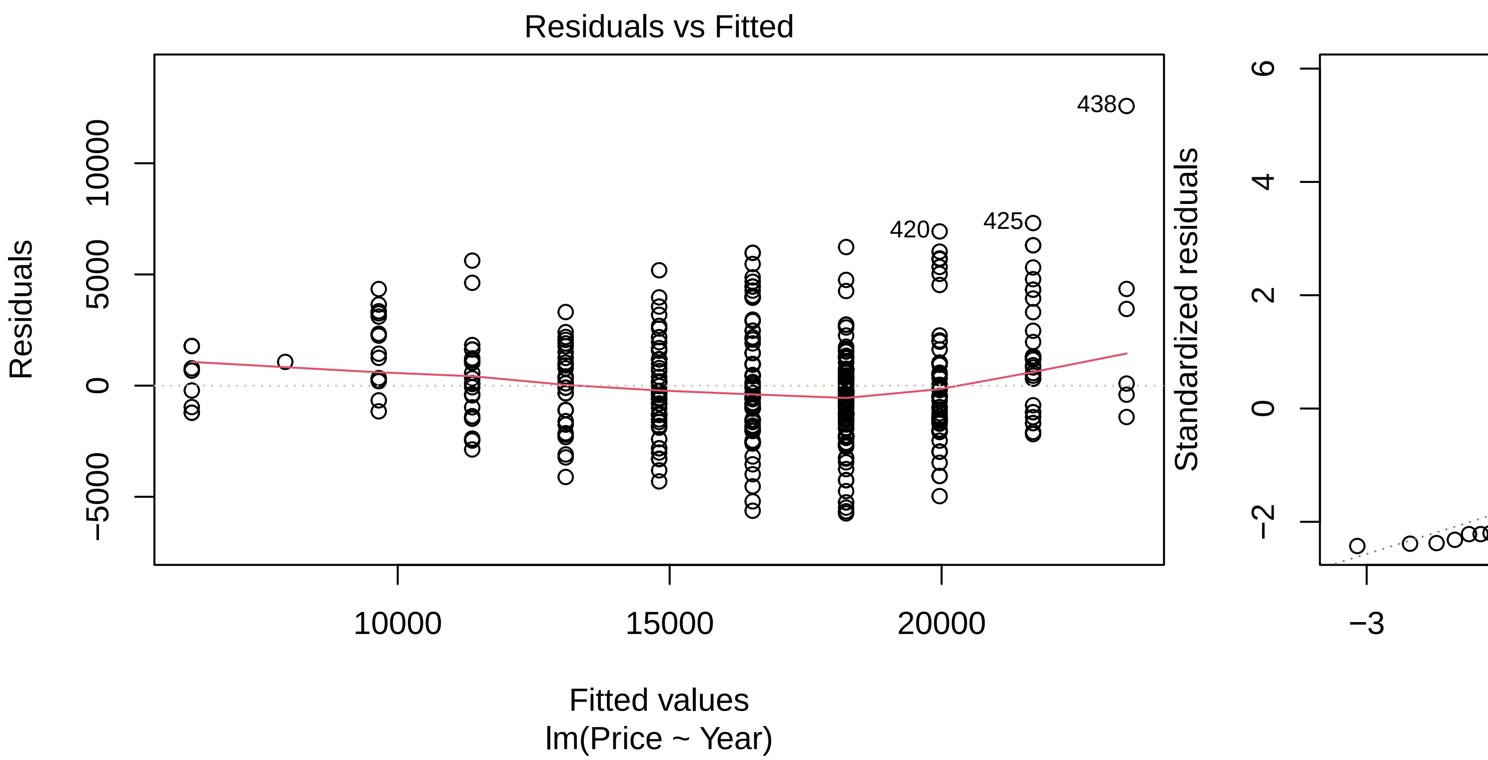
```
##
## Call:
## lm(formula = Price ~ Year, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5745.9 -1483.2  -248.4  1175.2 12575.7
```

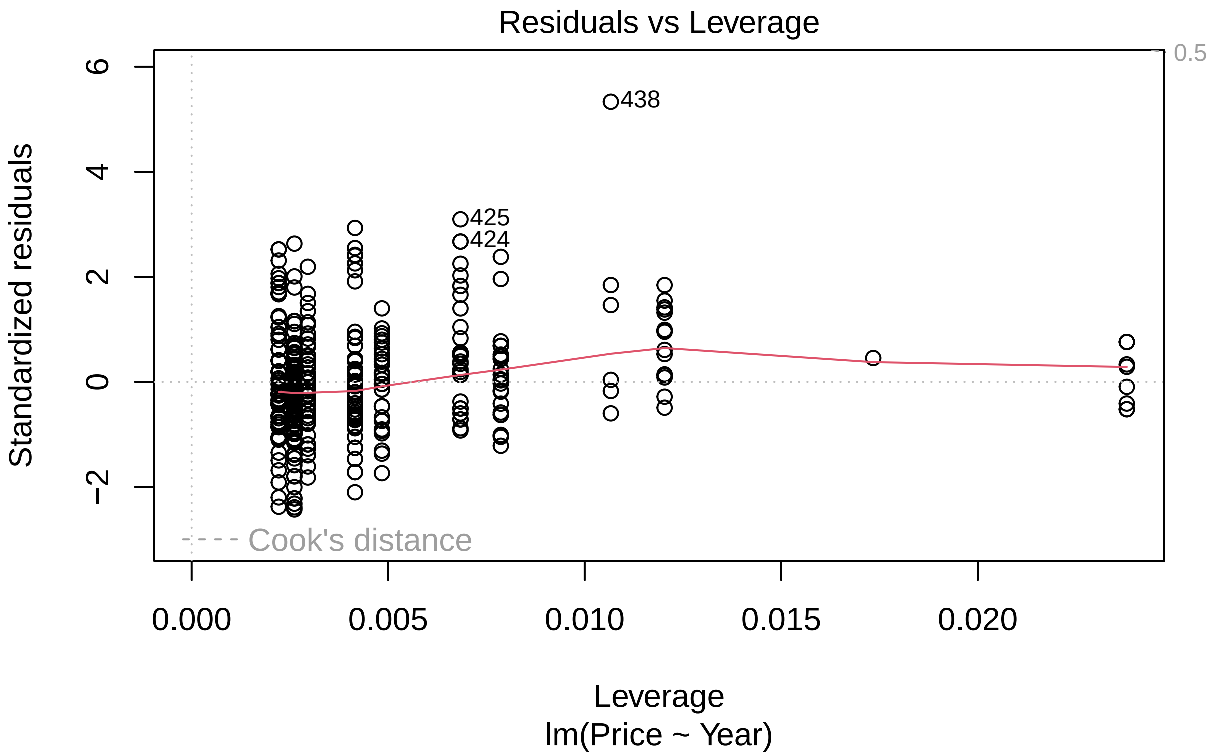
```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.445e+06  1.141e+05  -30.20   <2e-16 ***
## Year         1.719e+03  5.664e+01   30.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2371 on 452 degrees of freedom
## Multiple R-squared:  0.6708, Adjusted R-squared:   0.67
## F-statistic: 920.9 on 1 and 452 DF,  p-value: < 2.2e-16
```

Predicted Price = -3445154.03 + 1718.81 (Year)

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.

```
plot(price_year_model, c(1, 2, 5))
```





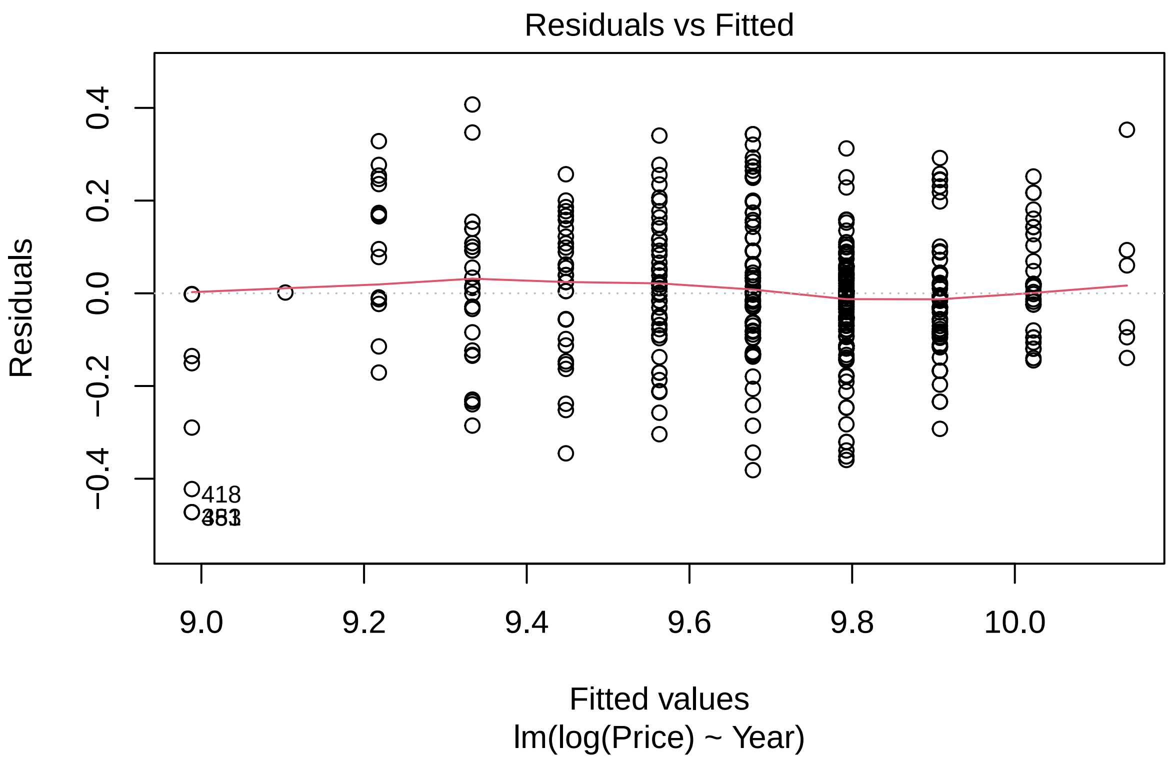
The data seems to be approximately normal as seen by the roughly linear pattern in the qqnorm plot and the residuals vs fitted values graph has no noticeable pattern and uniform spread. The model seems to fit the data effectively.

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.

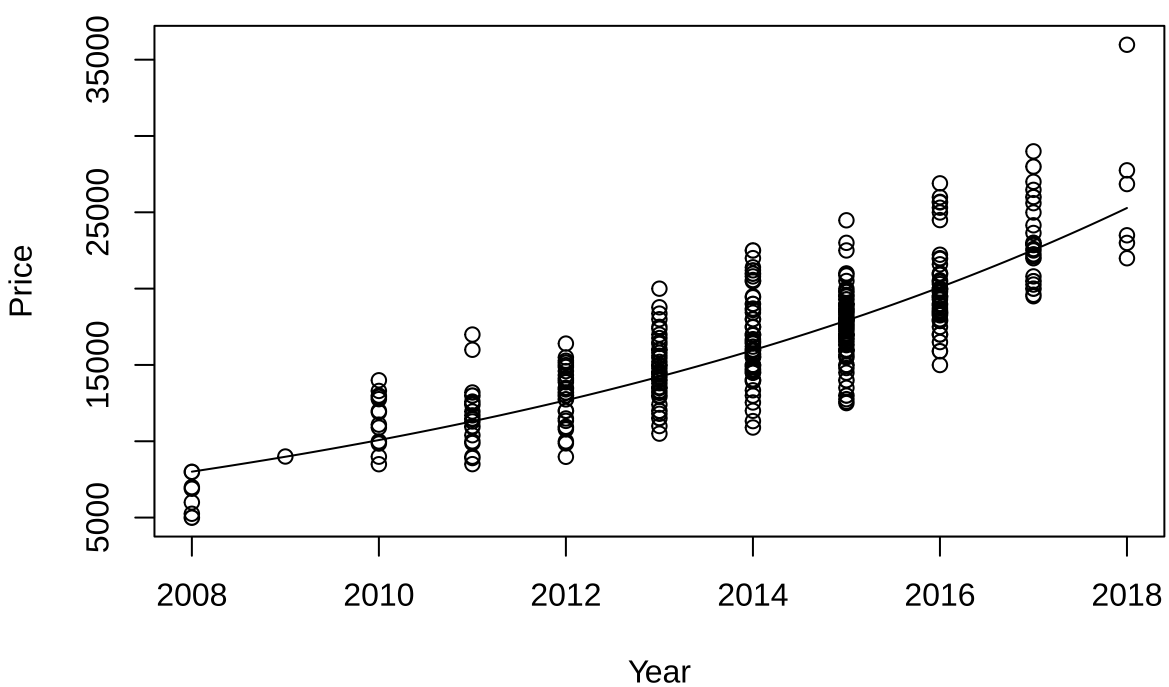
```r
transformed_model <- lm(log(Price)~Year, data = MyVehicles)
summary(transformed_model)
```

```
##
## Call:
## lm(formula = log(Price) ~ Year, data = MyVehicles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47213 -0.08540 -0.00047  0.07741  0.40733
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.218e+02  6.815e+00  -32.55   <2e-16 ***
## Year         1.149e-01  3.384e-03   33.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1416 on 452 degrees of freedom
## Multiple R-squared:  0.7186, Adjusted R-squared:  0.718
## F-statistic:  1154 on 1 and 452 DF,  p-value: < 2.2e-16
```

```
plot(transformed_model, 1)
```




```
plot(Price~Year, data = MyVehicles)
curve(exp(transformed_model$coefficients[2] * x)/exp(abs(transformed_model$coefficients[1])), add = TRUE)
```



This transformation does improve satisfying the linear model conditions because the spread of the residuals vs fitted values is more uniform when compared to the original model. All of these analysis are compared to the original model, so this transformed model does a slightly better job at satisfying the condition for a simple linear model.