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}.

```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
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
# 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
```

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

# 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] 2012 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.

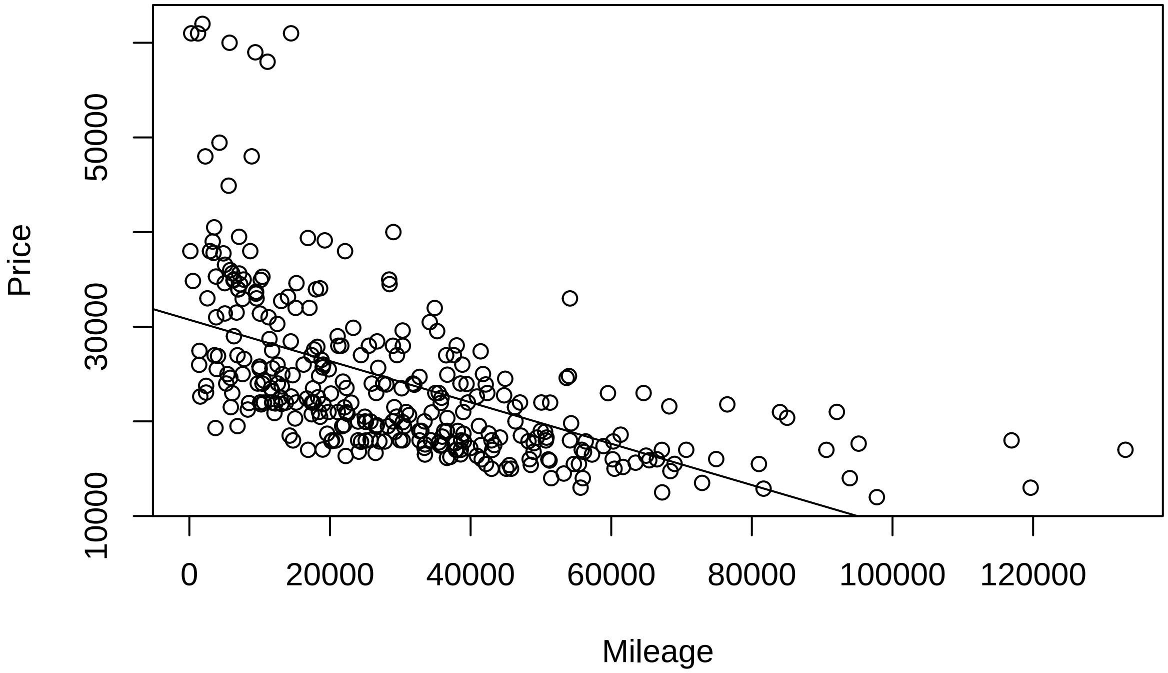
```
VehicleModel = lm(Price~Mileage, data = MyVehicles)
VehicleModel
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = MyVehicles)
##
## Coefficients:
## (Intercept)      Mileage
##  30736.6457      -0.2183
```

```
# The intercept is the price of the used car
# The mileage is the slope.
# For every mile driven the car is worth 0.2183 less
# The signs make sense because when a car has lots of miles it is worth less.
# Compare that to the same car with less miles
```

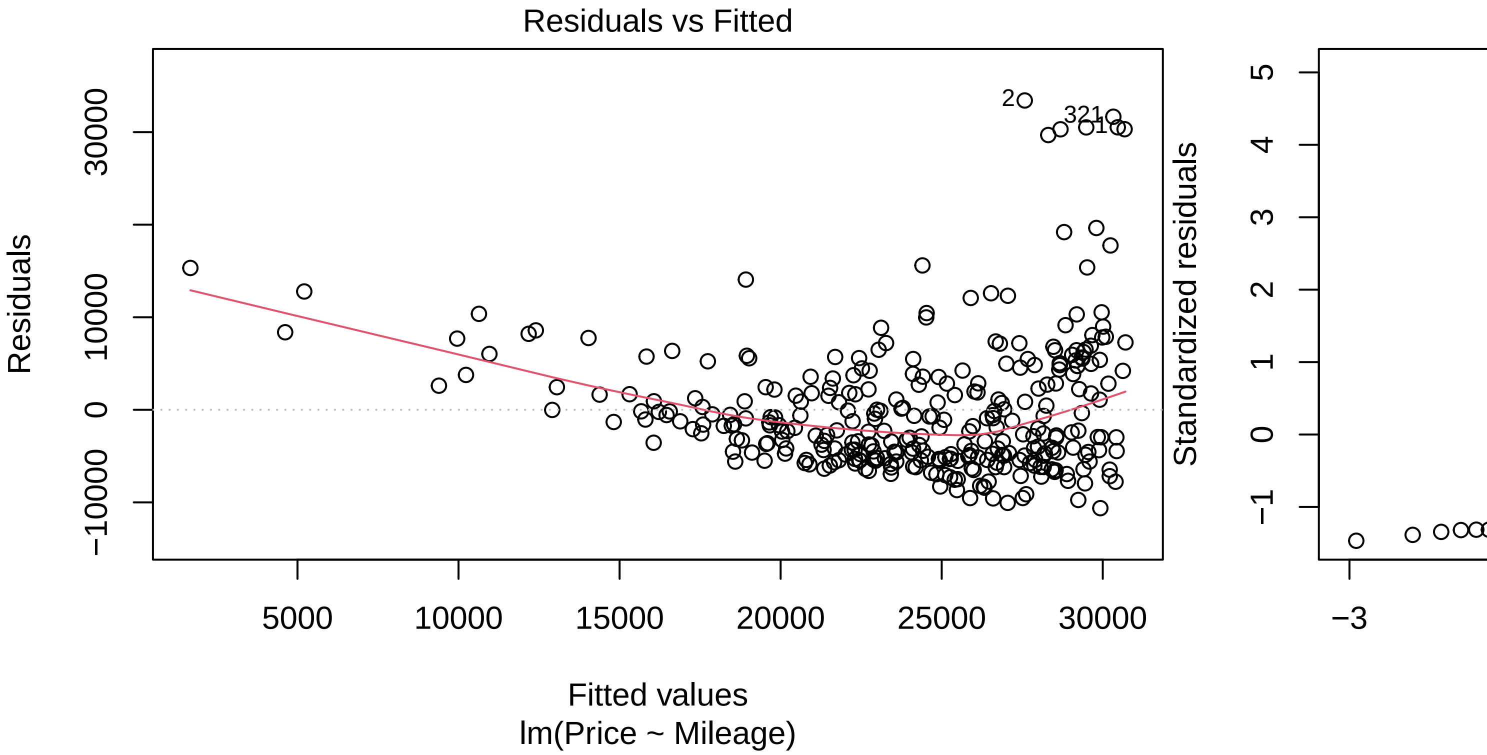
2. Produce a scatterplot of the relationship with the regression line on it.

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



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(VehicleModel, 1:2)
```



```
# The data appears to decently fit the conditions for a linear model
# That is shown with a large chunk of data points near the residual line
# But there is still a lot of data points far away from the residual line
```

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?

```
VehicleModel_resid = abs(VehicleModel$residuals)
largest_resid = which(VehicleModel_resid >= 30315)

rstandard(VehicleModel)[largest_resid]
```

```
##        1        2        3      320      321
## 4.215287 4.606849 4.187728 4.210600 4.372242
```

```
rstudent(VehicleModel)[largest_resid]
```

```
##        1        2        3      320      321
## 4.330646 4.760206 4.300686 4.325549 4.501947
```

```
# Since these values are larger than 3 they would all be outliers
# Yes they would significantly impact the slope making it higher or lower
```

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?

```
sort(hatvalues(VehicleModel), decreasing = TRUE)[1:5]
```

```
##        262        218        258        171        234
## 0.06848152 0.05247085 0.04951606 0.03127371 0.02915143
```

```
# The first 3 are heavy outliers and influence the model a lot
# The other 2 aren't much of an outlier when compared to the rest of the values
```

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?

```
sort(cooks.distance(VehicleModel), decreasing=TRUE)[1:5]
```

```
##        262        258        321          3          1
## 0.17527327 0.08471647 0.07785434 0.07656070 0.07437517
```

```
# Compared to the regular values all of these vehicles look like outliers
# Which makes it so they influence the model heavily
```

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.

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

```
##                     2.5 %        97.5 %
## (Intercept) 29398.0744179 32075.2170476
## Mileage        -0.2538483    -0.1827453
```

```
# 95% confident that the mean price is between 29,398.07 and 32,075.22
# 95% confident that the mean slope price per mile is between -0.25 and -0.18
# On average for every mile the car has, the price goes down -0.215
# The average base price of a used car is 30,736.65
```

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.

```
cor(MyVehicles$Mileage, MyVehicles$Price)
```

```
## [1] -0.5596537
```

```
# HO = 0, HA != 0
# The correlation is -0.56 meaning that there is a decently strong negative correlation
# This makes it so we reject the null
# Proving a relationship between mileage and price
# As the mileage rises, the price drops
summary(VehicleModel)
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = MyVehicles)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10625  -4952  -1751   3503  33420
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.074e+04  6.804e+02   45.18   <2e-16 ***
## Mileage     -2.183e-01  1.807e-02  -12.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7271 on 320 degrees of freedom
## Multiple R-squared:  0.3132, Adjusted R-squared:  0.3111
## F-statistic: 145.9 on 1 and 320 DF,  p-value: < 2.2e-16
```

```
# HO = 0, HA > 0
# We reject the null hypothesis since the p-value is less than 0.05
# Showing that there is a relationship between mileage and price
anova(VehicleModel)
```

```
## Analysis of Variance Table
##
## Response: Price
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## Mileage     1 7.7162e+09 7716212161  145.94 < 2.2e-16 ***
## Residuals 320 1.6920e+10   52873485
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# HO = 0, HA != 0
# We reject the null hypothesis since the p-value is less than 0.05
```

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).

```
newx=data.frame(Mileage=50000)

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

```
##        fit      lwr      upr
## 1 19821.81 18759.35 20884.26
```

```
# 95% confident that the mean price is between 18,759.35 and 20,884.26

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

```
##        fit      lwr      upr
## 1 19821.81 5476.587 34167.02
```

```
# 95% confident that the next vehicle price will fall within 5,476.59 and 34,167.02
```

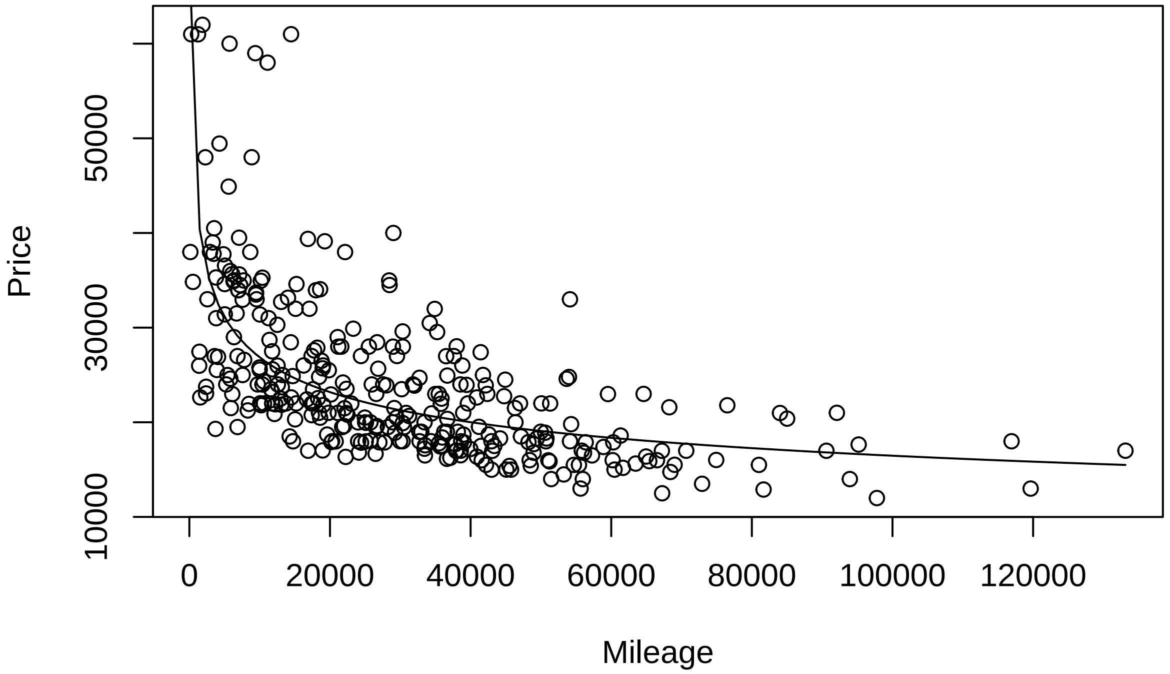
10. Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain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

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

```
##
## Call:
## lm(formula = log(Price) ~ log(Mileage), data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5679 -0.1601 -0.0534  0.1574  0.8985
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.15684    0.12959   93.81   <2e-16 ***
## log(Mileage) -0.21261    0.01294  -16.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2302 on 320 degrees of freedom
## Multiple R-squared:  0.4575, Adjusted R-squared:  0.4558
## F-statistic: 269.8 on 1 and 320 DF,  p-value: < 2.2e-16
```

```
plot(Price~Mileage, data = MyVehicles)
curve(exp(12.15684)*(x^-0.21261), add=TRUE)
```



```
# The log model improves on the linear conditions because the r^2 is 0.4575
# This is much better compared to the untransformed linear r^2 of 0.3132
```

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.

```
# With the transformed model of e^(12.15684)*(x^-0.21261) the car can't be free at any mileage
```

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(NewModel, newx, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 9.856457 9.823092 9.889822
```

```
# 95% confident that the mean log price is between 9.82 and 9.89

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

```
##        fit      lwr      upr
## 1 9.856457 9.402263 10.31065
```

```
# 95% confident that the next vehicle log price will fall within 9.40 and 10.31
```

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.

```
StateHW3 = "GA"

# Creates a dataframe with the number of each model for sale in Georgia
GAvehicles = as.data.frame(table(UsedCars$Model[UsedCars$State==StateHW3]))

# Renames the variables
names(GAvehicles)[1] = "Model"
names(GAvehicles)[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 knitting
#Enough_Vehicles = subset(GAvehicles, Count>=100)
#Enough_Vehicles

ModelOfMyChoice = "CamaroCoupe"

# Takes a subset of your model vehicle from Georgia
MyGAvehicles = subset(UsedCars, Model==ModelOfMyChoice & State==StateHW3)

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

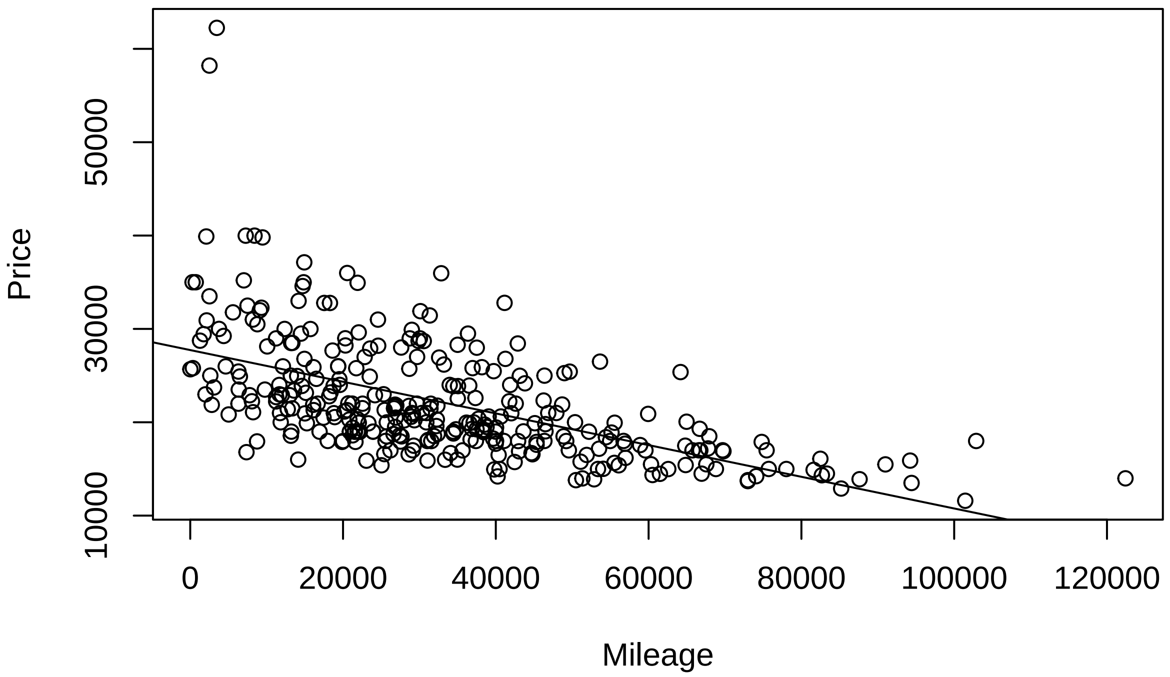
```
## [1] 2012 2018
```

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.

```
GAvehicleModel = lm(Price~Mileage, data = MyGAvehicles)
GAvehicleModel
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = MyGAvehicles)
##
## Coefficients:
## (Intercept)      Mileage
##  27734.5444      -0.1697
```

```
plot(Price~Mileage, data = MyGAvehicles)
abline(GAvehicleModel)
```



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.

```
GAvehicleModel$coefficients
```

```
##   (Intercept)       Mileage
## 27734.5444087    -0.1696889
```

```
VehicleModel$coefficients
```

```
##   (Intercept)       Mileage
## 30736.6457327    -0.2182968
```

```
# The relationship is similar as both are negatively correlated
# The relationships are slightly different as in Georgia the price and miles are cheaper
```

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?

```
# The model will be super helpful as it will allow me to save lots of money
# A con is that there isn't the greatest correlation between mileage and price
# That lack of correlation could make the model inaccurate
```

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?

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

```
##
## Call:
## lm(formula = Mileage ~ Year, data = MyVehicles)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49854 -10127  -2752   9464  77003
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19221443.4  1281584.4   15.00   <2e-16 ***
## Year           -9525.5      636.1  -14.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17250 on 320 degrees of freedom
## Multiple R-squared:  0.412,  Adjusted R-squared:  0.4102
## F-statistic: 224.2 on 1 and 320 DF,  p-value: < 2.2e-16
```

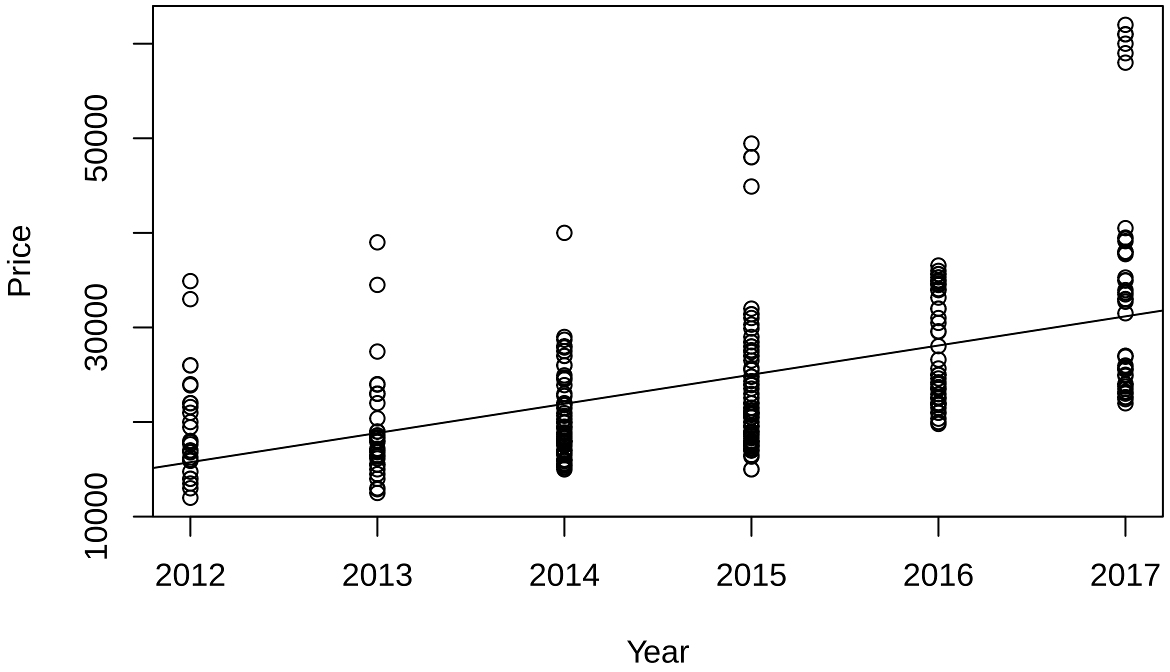
```
# The r squared or 0.412
```

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.

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

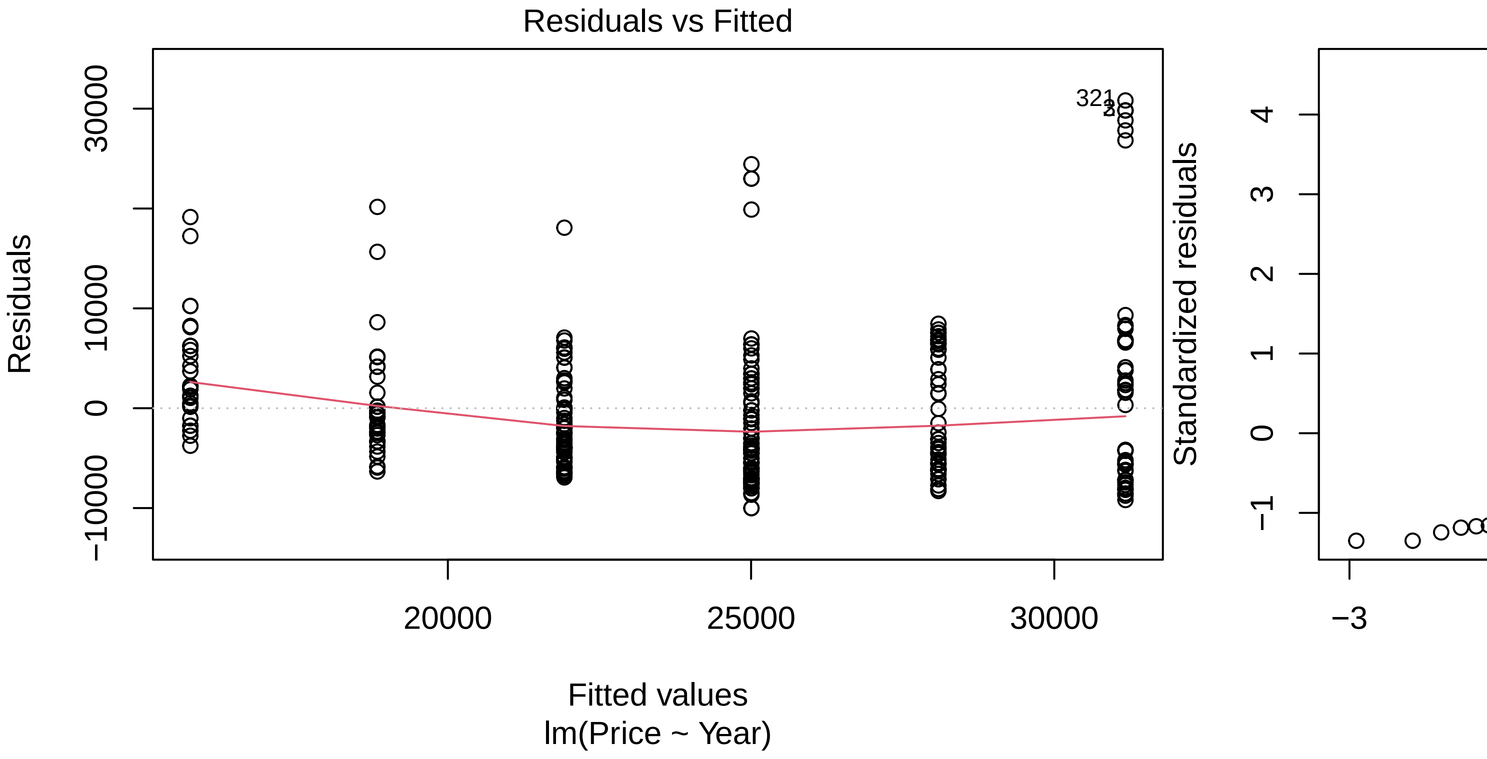
```
##
## Call:
## lm(formula = Price ~ Year, data = MyVehicles)
##
## Coefficients:
## (Intercept)         Year
##    -6189224         3084
```

```
plot(Price~Year, data = MyVehicles)
abline(NCVehicleModel)
```



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(NCVehicleModel, 1:2)
```

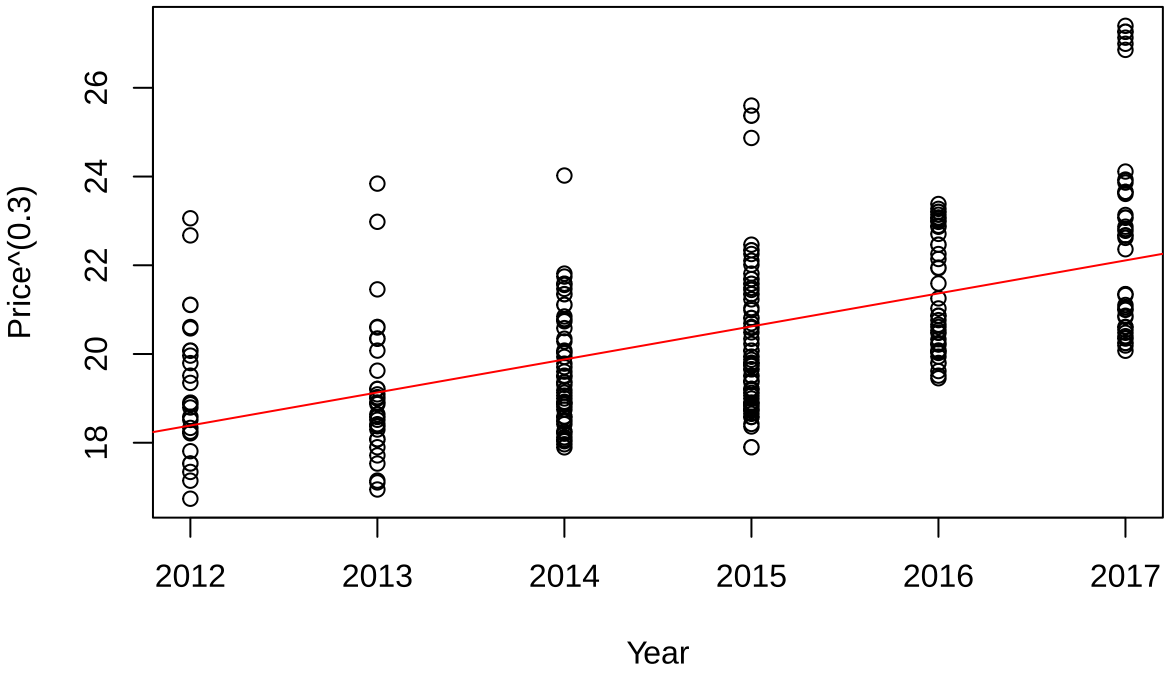


```
# In the residual and in the QQ plot the data looks to fit mediocrely well
# There looks to be some correlation in the residual graph and good correlation in the QQ plot
```

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.

```
NCModel = lm(Price^(0.3)~Year, data=MyVehicles)

plot(Price^(0.3)~Year, data=MyVehicles)
abline(NCModel, col="red")
```



```
summary(NCVehicleModel)
```

```
##
## Call:
## lm(formula = Price ~ Year, data = MyVehicles)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10008  -5354  -1929   3810  30819
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6189223.6   551686.0  -11.22   <2e-16 ***
## Year            3084.0      273.8   11.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7425 on 320 degrees of freedom
## Multiple R-squared:  0.2839, Adjusted R-squared:  0.2816
## F-statistic: 126.8 on 1 and 320 DF,  p-value: < 2.2e-16
```

```
summary(NCModel)
```

```
##
## Call:
## lm(formula = Price^(0.3) ~ Year, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7229 -1.2466 -0.4429  1.0562  5.2874
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1478.9407   121.8968  -12.13   <2e-16 ***
## Year            0.7442     0.0605   12.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.641 on 320 degrees of freedom
## Multiple R-squared:  0.321,  Adjusted R-squared:  0.3189
## F-statistic: 151.3 on 1 and 320 DF,  p-value: < 2.2e-16
```

```
# There is slight improvement with satisfying the linear model conditions
# The r^2 of the untransformed is 0.2839 which is less than the transformed r^2 of 0.321
```