Wilfred Lauier University

ST 362

How Much Is Your Car Worth

Authors:
William Ngana
195801950

Waterloo, ON

Date Started: July,2019

1 Analysis

This Project consists of a set of activities designed to build a multivariate regression model. The model will be used to describe the relationship between the retail price of 2005 used GM cars and various car characteristics, such as mileage, make, model, presence or absence of cruise control, and engine size.

1.1 Activity 1

The first part of this project we are looking at simple linear regression. That is we were asked to produce a scatter plot from the Cars data set to display the relationship between Mileage and Price [1a] is the resulting plot. From this we can see that there is not a strong relationship between the two variables.Although a quick glance at the graph show us that the two are negatively correlated. Meaning as the car mileage increases the price of the car tends downwards. For the next part we needed to calculate the least squares regression line.

$$Price = b_0 + b_1 * X_1 \tag{1}$$

Where the value of b_0 is the intercept and X_1 is the Mileage. Using the R function 'lm' with the two variable we are looking at mileage and price where price is the response variable we get that the equation becomes.

$$Price = 24764.559 - 0.1725205 * Mileage \tag{2}$$

With the other important statistics being

R^2	0.02046
R^2_{adj}	0.01924
Correlation Coefficient	-0.1430505
t-value Mileage	-4.093231
p-value Mileage	4.685e-05
t-value Intercept	27.383
p-value Intercept	$< 2 * e^{-16}$

Table 1: Table of the reported values gained form the simple Regression Model

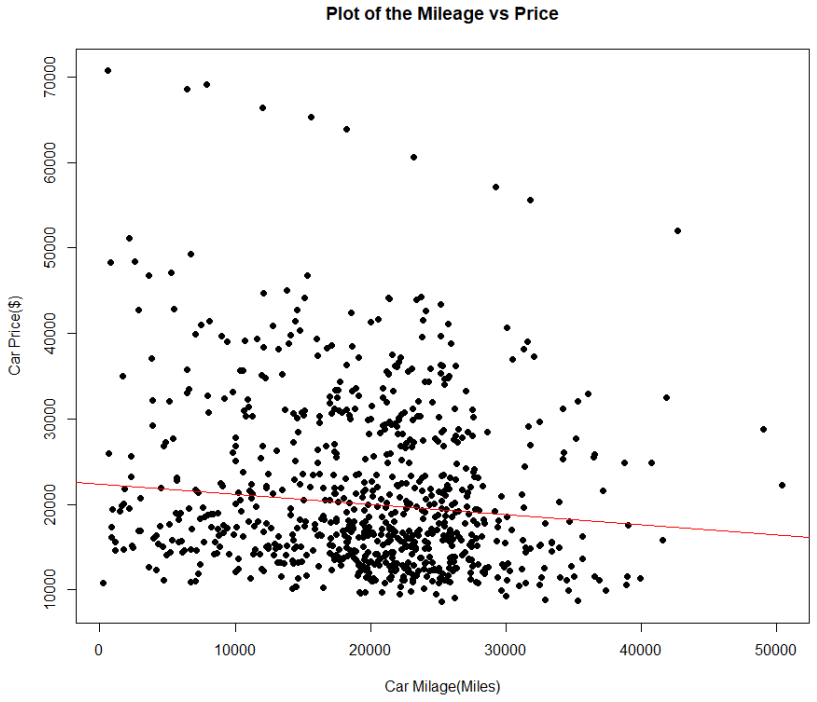
Looking at these values we see that the R^2 value and the $R^2_(adj)$ are both low but this does not necessarily mean that the model is bad so we look at the t-statistic and the p-value. The magnitude of the t-statistic is greater than 2 and the corresponding p-value is less than 0.5 so this means that mileage is a good indicator of price.

The first car in this data set is a Buick Century with 8221 miles at a price of 17314.10. Using the model equation we get that

$$Price = 24764.559 - 0.1725205 * 8221 = 23346.27 \tag{3}$$

The residual is $17314.10 - 23346.27 = -6032.17$. So the residual is -6032.17

(a) Here we have the scatter plot of the relationship between mileage and retail price



(b) This is the scatter plot of the relationship between mileage and retail price with the regression line created through R

1.2 Activity 2

The first part of this we are doing a stepwise regression analysis. To start off we are looking at each explanatory variable, Cy1, Liter, Doors, Cruise, Sound, Leather, and Mileage and identifying the one with the highest r^2 value. From this table we can see that the variable

Variable	R^2
Cyl	0.323859
Liter	0.3115267
Doors	0.01925147
Cruise	0.185633
Sound	0.01546239
Leather	0.02471085
Mileage	0.02046345

Table 2: Table of the R^2 values of each explanatory variable

with the highest R^2 value is Cyl, so we set that as X_1. Now we will calculate six more regression models that each have two explanatory variables each and look for the one with the highest R^2 value.

Variable	R^2
Cyl + Liter	0.3259155
Cyl + Doors	0.3434605
Cyl + Cruise	0.3839491
Cyl + Sound	0.3292753
Cyl + Leather	0.33698
Cyl + Mileage	0.3398207

Table 3: Table of the R^2 values of each explanatory using Cyl as X_1 variable

Looking at the table we see that the model that now has the highest R^2 value is with the two explanatory variable Cyl and Cruise. The R^2 value increased by 0.0600901 when we added the second variable Cruise. Instead of continuing to add more and more variables like this the next part R was used to perform a stepwise regression analysis. From this the explanatory variables that were suggested are mileage + cyl + doors + sound + leather + cruise with Multiple R-squared: 0.4457 and Adjusted R-squared: 0.4415.

Next we will use R to develop a model using best subsets techniques for the entire data set. Which produces the following table summary.

		mileage	cyl	liter	doors	sound	leather	cruise
1	(1)	" "	"*"	" "	" "	" "	" "	" "
2	(1)	" "	"*"	" "	" "	" "	" "	"*"
3	(1)	" "	"*"	" "	" "	" "	"*"	"*"
4	(1)	"*"	"*"	" "	" "	" "	"*"	"*"
5	(1)	"*"	"*"	" "	"*"	" "	"*"	"*"
6	(1)	"*"	"*"	" "	"*"	"*"	"*"	"*"
7	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"

Table 4: Table of subsets produced when running the best subsets techniques in R

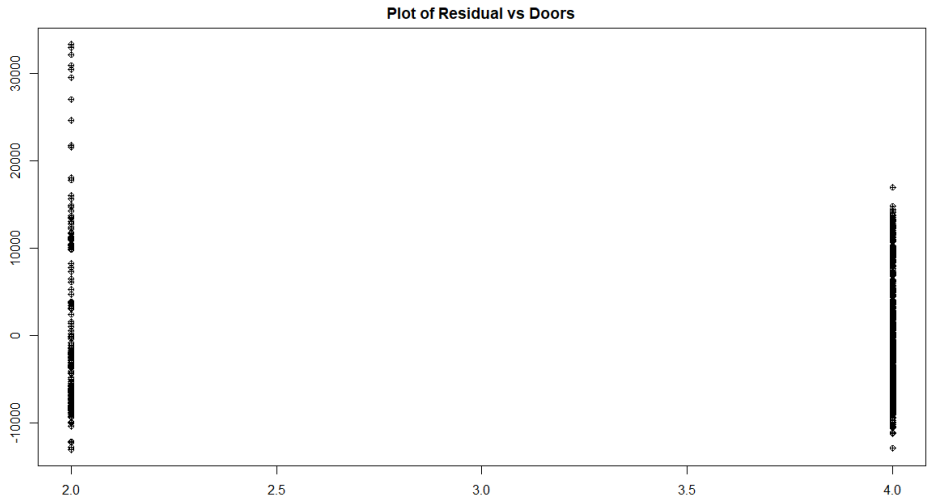
Now from this summary we have to choose which explanatory variables to choose. The variable that is included in each model, 1-7, is the one that has an * in it. So for model 1 the variable is Cyl, model 2 the variables are Cyl and Cruise, so on. The best model is one with a high R^2 value for a low number of variables. To do this I used the following R code

```
res.sum <- summary(models)
data.frame(
  Adj.R2 = which.max(res.sum$adjr2))
```

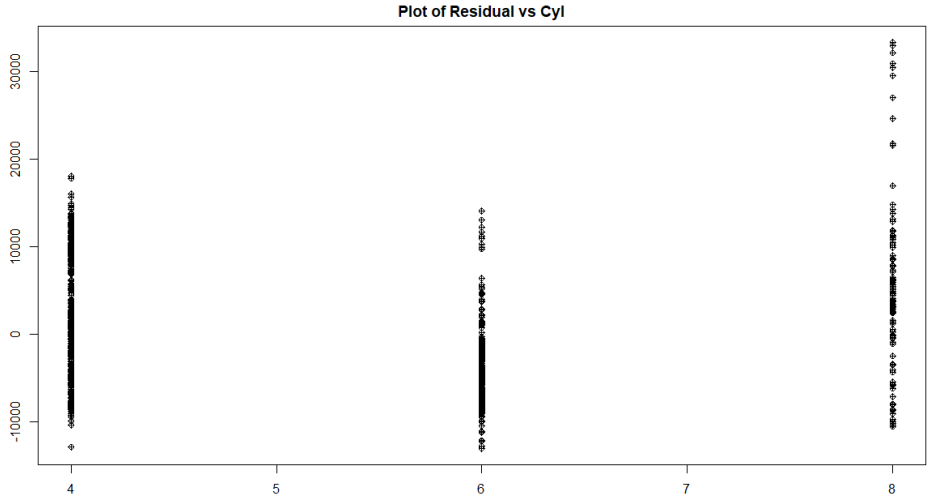
This will show me the best model based on the adjusted R^2 value and that model is 6 which include the variables, mileage, cyl, doors, sound, leather, and cruise. Comparing this to the model formed in question 4 with the stepwise analysis we see that they include the same variables. From the stepwise part of this activity we see that liter is not a useful in predicting price at all. Same with the subset method. If you look at 4 it is shown that the only time liter is added is in model 7 with all the other ones. Meaning that it is the least useful variable in predicting price.
Stepwise analysis follows very structured algorithms that do not allow for nuance of the study being done and produces one model. The real world is complicated and although the algorithms can give you an idea of what to look at you will most often need more to go on. This is where best subset comes in. The best subset method compares all possible models and returns the best models that contain one variable, two variables and so on. So from this you can gather a lot more information about each variable. Which are more useful in predicting your response variable and which aren't that useful.

1.3 Activity 3

In this Activity we will start by creating plots of the residuals versus each explanatory variable in the models, as well as the plot residual versus the predicted retail price.

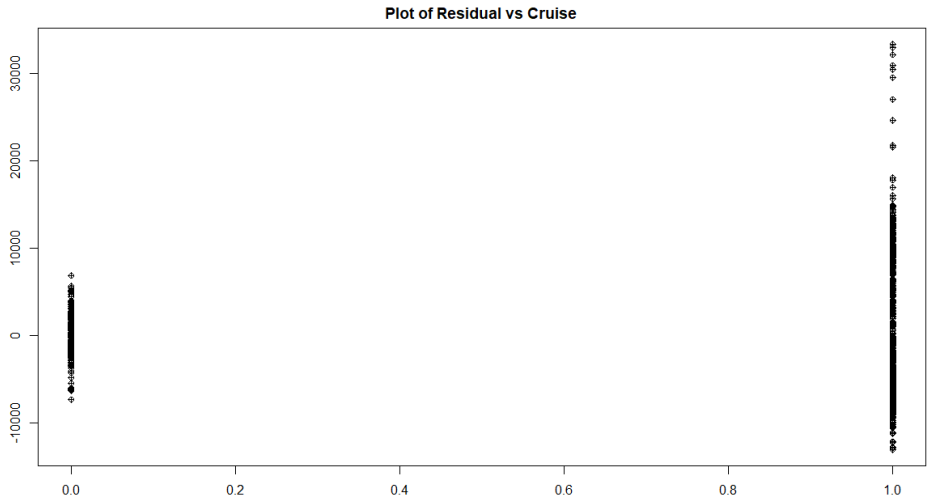


(a) Residual plot vs the doors

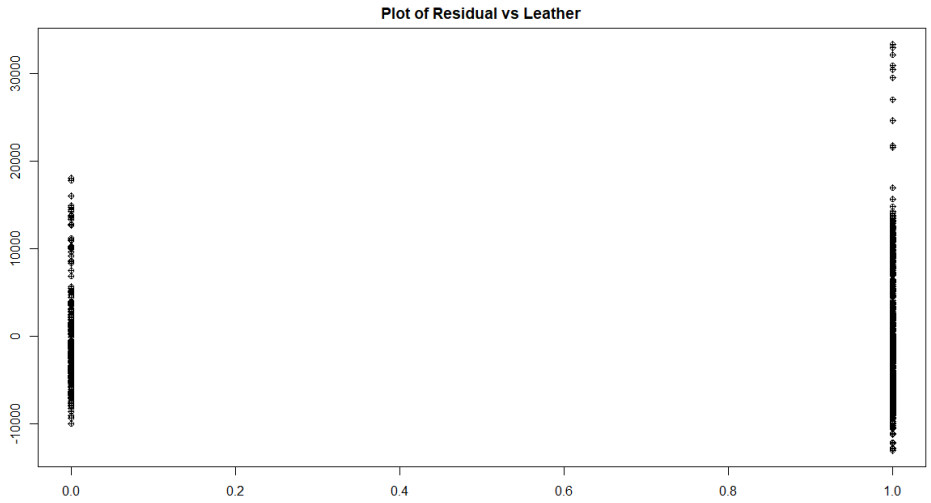


(b) Residual plot vs Cyl

Figure 2: Residual vs Explanatory Variables

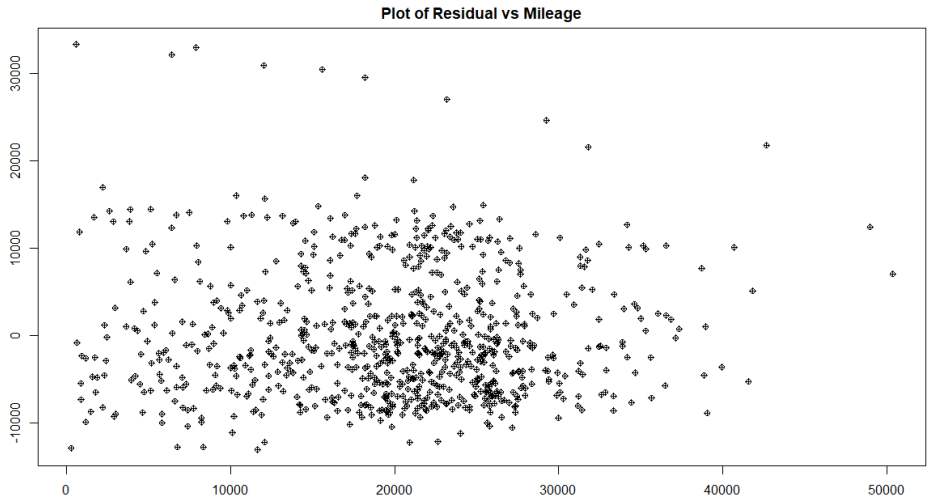


(a) Residual plot vs Cruise

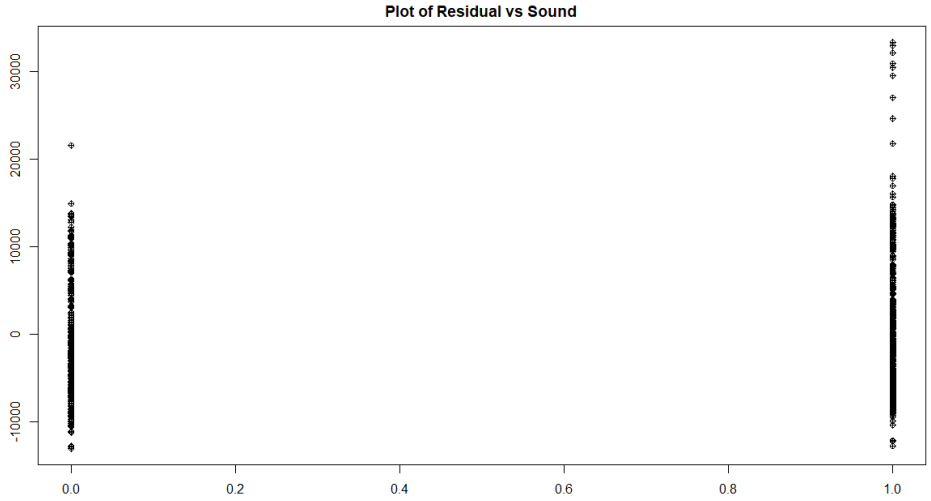


(b) Residual plot vs Leather

Figure 3: Residual vs Explanatory Variables



(a) Residual plot vs Mileage



(b) Residual plot vs Sound

Figure 4: Residual vs Explanatory Variables

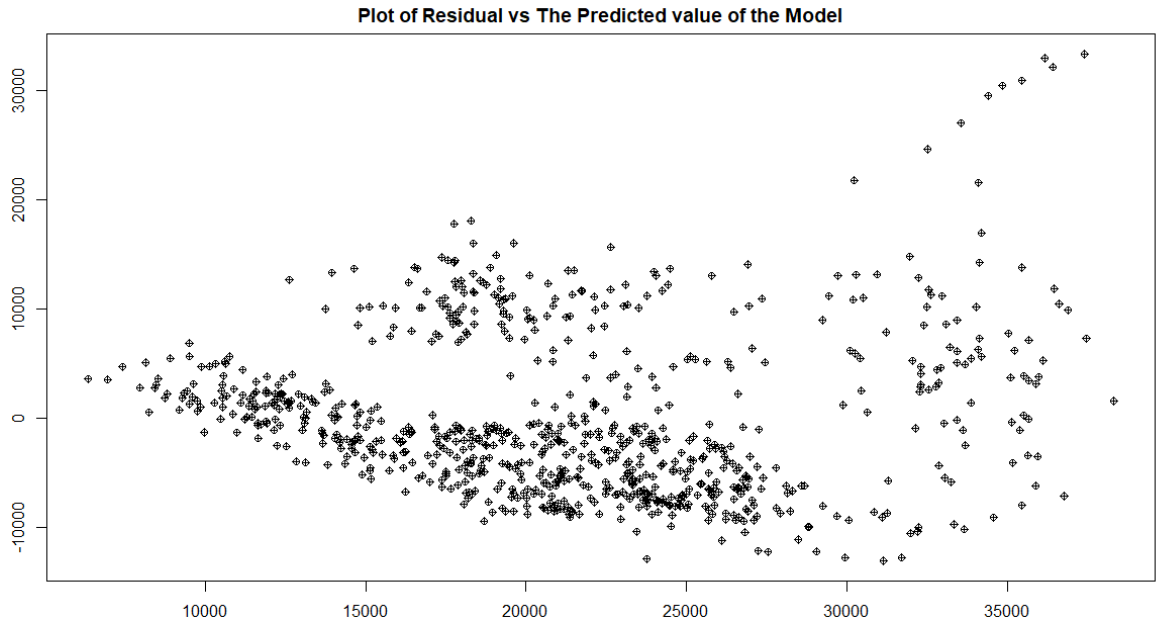


Figure 5: Plot of residual vs the predicted value of the model

If we take a look at [4a] which is the plot of the residual versus the mileage of the car, looking at this plot at first glance there really isn't a lot of patterns. As the residual increases the size of the residual tends to stay the same. That is until you draw a line at mileage = 8000. From this you can see that most of the data is skewed to the right. As

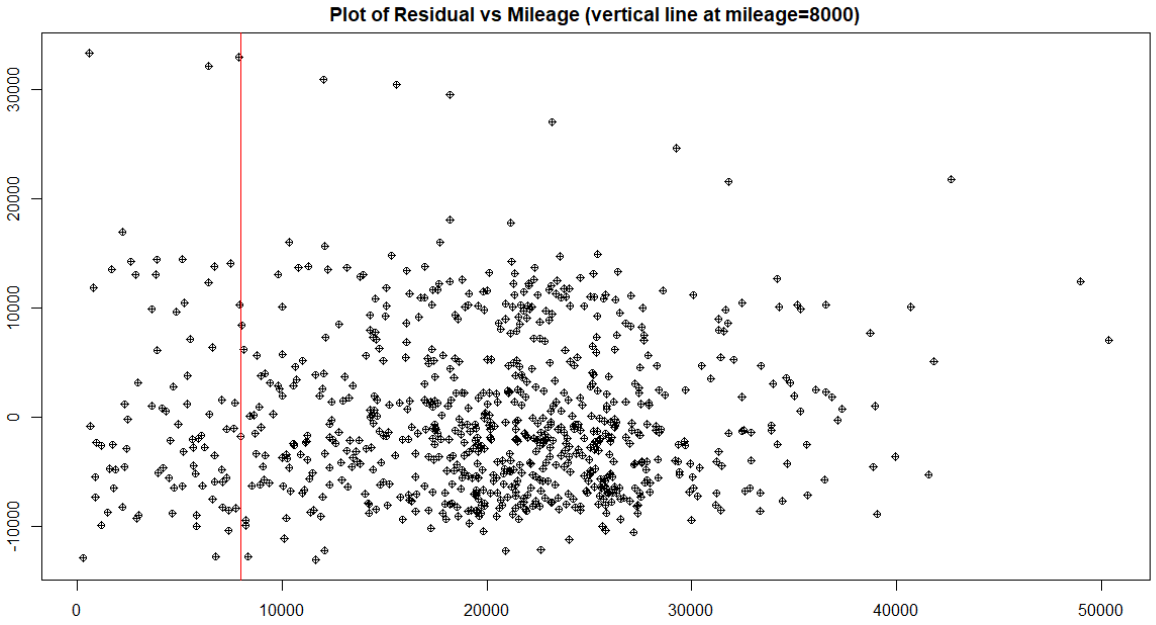


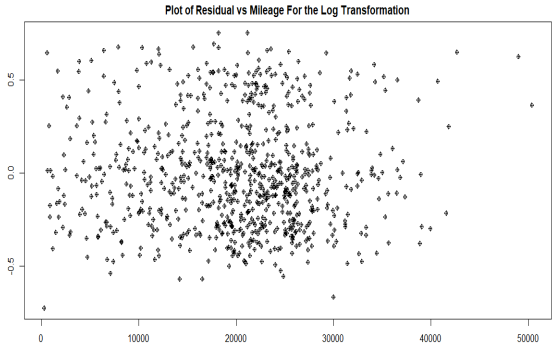
Figure 6: Plot of residual vs mileage with a vertical line at mileage=8000

well you can see that the majority of the data is centered around Y=0 except for the outliers.

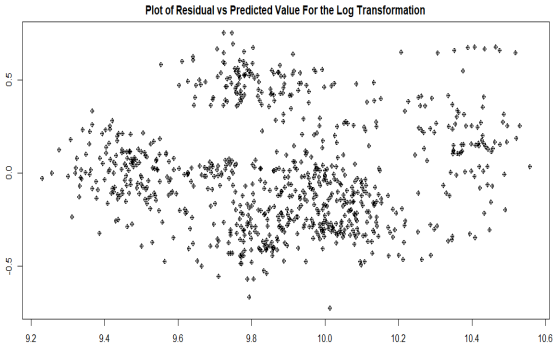
Though when looking at [5] you can see that as the predicted value increases the residual does change. The pattern that we see is heteroskedasticity. You can see this because in the plot you see that the scatter of the residual is in sort of the cone shape, as it widens when the predicted value increases and narrows when the predicted value decreases.

In the other residual plots there aren't many patters that you can see. Due to the fact that the values of those variables are discrete, you can't tell how the residual changes and the value of the of the variable increases or decreases. The residuals in these plots aren't centered around Y=0.

In the next part we transformed the response variable into log(price) and $\sqrt{price}$ and redrew the residual plots which are displayed below. The rest of the plots can be found in the appendix.

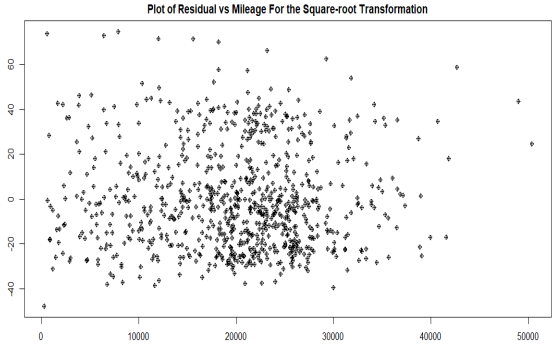


(a) Residual plot of the variable Mileage with the log transformed price

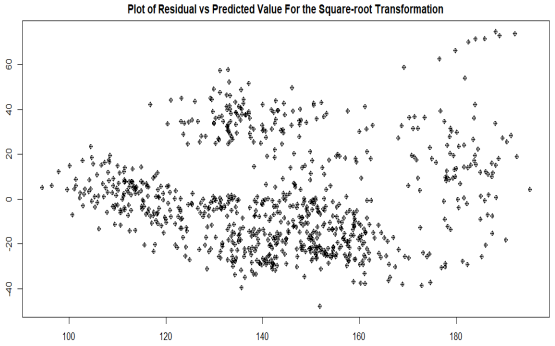


(b) Residual plot of the Predictor with the log transformed price

Figure 7: Residual Pots



(a) Residual plot of the variable Mileage with the Square-root transformed price



(b) Residual plot of the Predictor with the Square-root transformed price

Figure 8: Residual Pots

As we can see from the plot the transformation that did the best job of reducing the skewness and the heteroskedasticity is the the log transformation. The R^2 values for each were, for the log transformation it was 0.4836282, and for the square root transformation it was 0.4689459. From this we see that the best r-squared values correspond with the best residual plots, this is because the r-squared value represents the proportion of the variance in the dependent variable that is predictable from the independent variable. Which means that the better the the residual plot (plot of real value - predicted value) the higher the r-squared value will be. Some other transformations that could be done is you could transform the data with a cube root, or you could the log(price+1) this would make it so that if there were any zero values then the transformation would handle it.

1.4 Activity 4

In this section we will look at the outliers of the data. We will use the regression equation that was developed in question 5. To find the outliers I will use the built in function in R for Cook's Distance. Cook's distance measures the effect of deleting a given data point.

```
CarsData$cooksd <- cooks.distance(final.model)
CarsData$outlier <- ifelse(CarsData$cooksd < 4/nrow(CarsData),
'keep','delete')
```

This will go and identify all outliers according to Cook distance. For this particular case I will be looking at all values higher than 4/n. This analysis deemed that 34 of the variable were outliers. Looking at the specific data that was deemed as outliers they are almost all convertibles. When you look at the outliers you notice that the prices of all of them are high regardless of the mileage. This makes sense because looking at the residual plot against the predicted value you see that the model over predicts the price the higher the price goes. That is for a cluster which would be these cars because according to the model these cars should all cost less then they actually do. Due to that you get the cone shape that as the price increases the plot spreads.

To Determine just how much the outliers effect the data I will rerun the analysis with out them and compare it to the previous model.

When running the step wise procedure for the data without the outliers the coeffieents that are chosen are Mileage, Liter, Doors, Sound, Leather, Cruise. This differs to the variables that were chosen when the outliers were present. This model has a Multiple R-squared value of 0.4827 and an adjusted R-squared value of 0.4786. This is an increase from the previous model which had Multiple R-squared: 0.4457 and Adjusted R-squared: 0.4415.

Next we run the best subset again, as well as the code that allows us to choose the best model based on the R^2 value. From this we get that the best model is price liter + cruise + leather + sound + mileage + doors. Which again is a change from the previous model determined by best subset with the outliers present. This model has Multiple R-squared: 0.4248 and Adjusted R-squared: 0.4205. While the best subset with outliers had Multiple R-squared: 0.4457 and Adjusted R-squared: 0.4415. Here we see that the R^2 value actually went down.

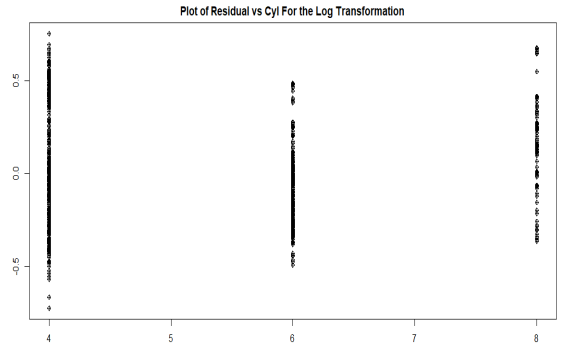
		mileage	cyl	liter	doors	sound	leather	cruise
1	(1)	" "	" "	"*"	" "	" "	" "	" "
2	(1)	" "	" "	"*"	" "	" "	" "	"*"
3	(1)	" "	" "	"*"	" "	" "	"*"	"*"
4	(1)	" "	" "	"*"	" "	"*"	"*"	"*"
5	(1)	"*"	" "	"*"	" "	"*"	"*"	"*"
6	(1)	"*"	" "	"*"	"*"	"*"	"*"	"*"
7	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"

Table 5: Table of subsets produced when running the best subsets techniques in R on the data set without outliers

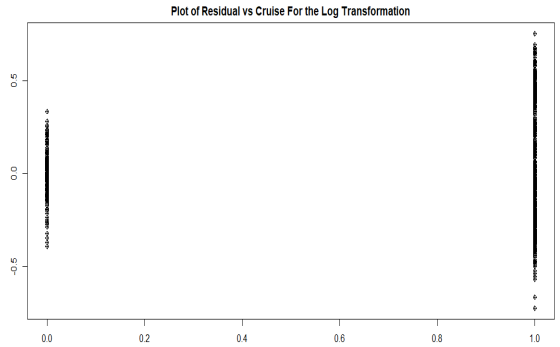
The main difference in the new models and the previous ones where he outliers were includes is the inclusion of Liter as a variable. For these models the variable that was deemed to be of no importance in determining the price was Cyl, which in the previous model was the variable with the highest R^2 value.

Appendices

A Log Transformation

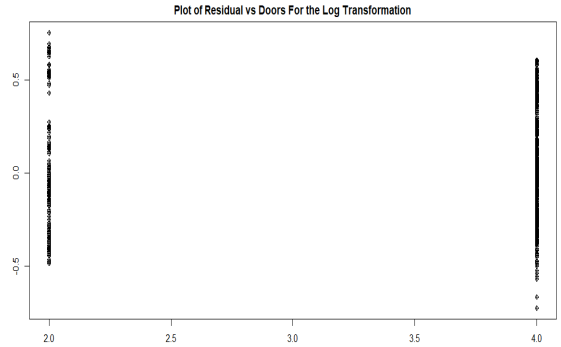


(a) Residual plot of the variable Cyl with the log transformed price

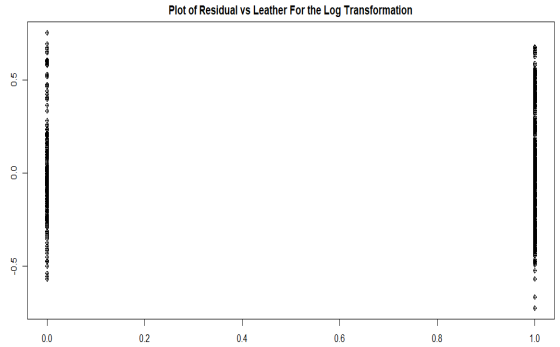


(b) Residual plot of the variable Cruise with the log transformed price

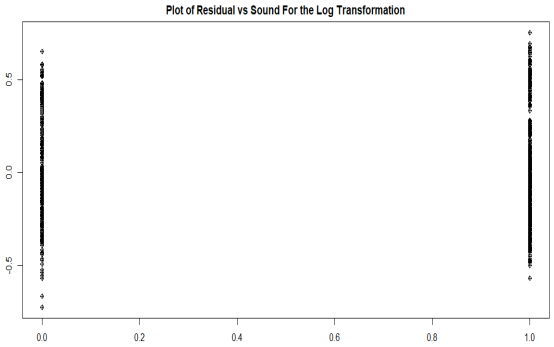
Figure 9: Explanatory Variable Log Transformation



(a) Residual plot of the variable Door with the log transformed price



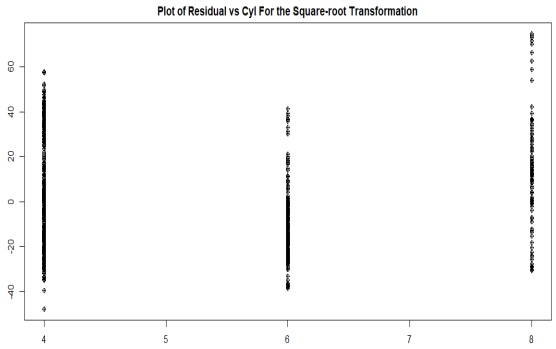
(b) Residual plot of the variable Leather with the log transformed price



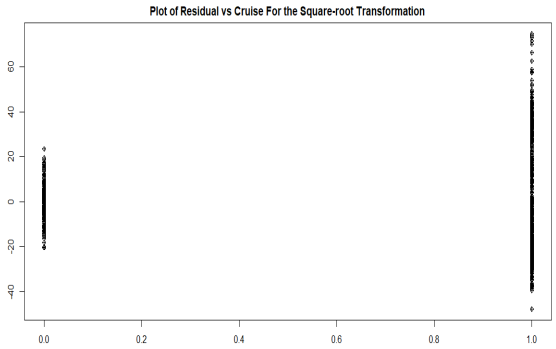
(a) Residual plot of the variable Cyl with the log transformed price

Figure 10: Explanatory Variable Log Transformation

B Square-root Transformation

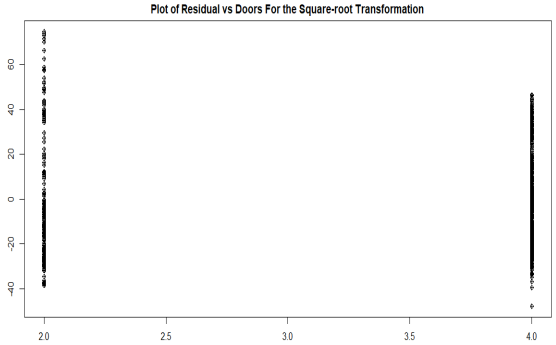


(a) Residual plot of the variable Cyl with the Square-root transformed price

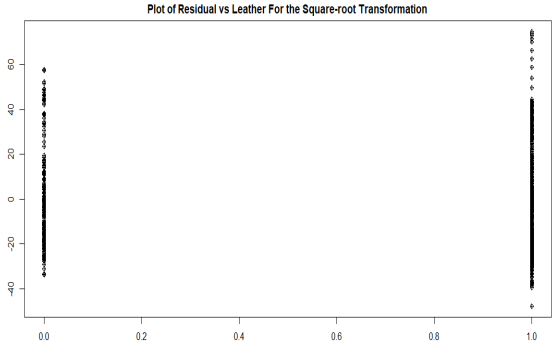


(b) Residual plot of the variable Cruise with the Square-root transformed price

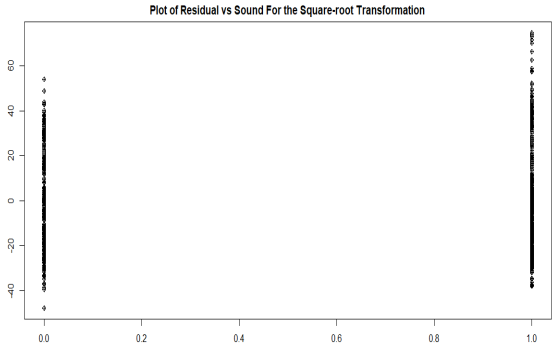
Figure 12: Explanatory Variable Square-root Transformation



(a) Residual plot of the variable Door with the Square-root transformed price



(b) Residual plot of the variable Leather with the Square-root transformed price



(c) Residual plot of the variable Sound with the Square-root transformed price

Figure 13: Explanatory Variable sqrt Transformation