Multiple Linear Regression

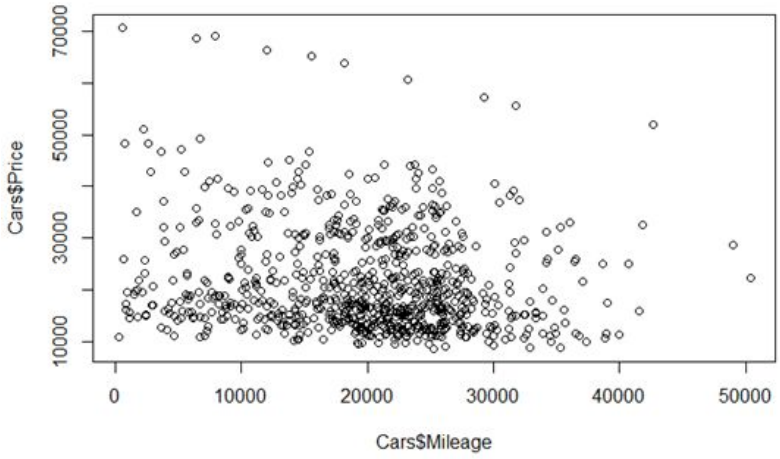
Introduction:

Within this project, we view different 2005 GM cars, their characteristics and the relationship between the different models and the effects it has on the retail value of the car based on the subset of the Kelly Blue Book dataset given. As we are given characteristics like mileage, make, model, cruise, engine size etc. we try to find the best fit model using different and all explanatory variables to predict the response variable.

By using R-Studio, we are able to plot and find relationships between all variables and how they affect the models that we use.

Activity 1: A Simple Linear Regression Model:

In this activity, a simple linear regression model is created with Price as the response variable and Mileage as the sole explanatory variable. This activity illustrates the limitations inherent to simple linear regression, especially with respect to vehicle prices, which are often influenced by a variety of factors, not just mileage.



The scatter plot above appears to indicate a weak linear relationship between Price and Mileage. It appears that as Mileage increases, Price generally decreases. However, it is worth noting that with 804 observations, it is difficult to conclude whether there is truly a strong relationship, especially given the large number of outliers.

Calculated manually and confirmed using R's built in regression function as shown in Appendix 1, the linear model is:

$$Y = 24664.56 - (0.1725205)X,$$

where Y is the Price of the vehicle in dollars and X is the Mileage of the vehicle.

This model has the following associated statistical values:

$R^2 = 0.02046345$
$R^2_{Adj.} = 0.01924208$
$Correlation\ Coefficient = -0.1430505$
$t - Stat_{Intercept} = 27.38342$
$t - Stat_{Slope} = -4.093231$
$\Pr(>|t|)_{Intercept} =< 2 * 10^{-16} \approx 0$
$\Pr(>|t|)_{Slope} = 4.68 * 10^{-5} \approx 0$

First and foremost, the R^2 statistic indicates that only approximately 2% of the variation in the data can be explained by the model, which does not indicate that the model is overly effective. The correlation coefficient indicates that there exists a relatively weak negative correlation, which is to be expected since in general, customers are less willing to pay as much as the mileage of a car increases. From the t values and the associated p values of the slope and the intercept, it can be concluded very confidently that the model is significant, that is, the value of the intercept and the slope are very likely to be non-zero. Although it is clear that there is some relationship between Price and Mileage, it should be noted that this relationship is not overly strong and thus more explanatory variables must be incorporated into the model.

The residual value of the first car in the data set is -$6032.17.

Simple linear regression, using Mileage to explain Price, is particularly weak since Mileage, although an important factor in purchasing a car, is just one of many factors. The presence of other explanatory variables is clear as a result of the preceding statistics. For example, consider two base model Honda Civics, both with 50,000 km, both built in 2005, however one Civic is heavily rusted and well damaged and the other is in pristine condition. If these cars are identical in features and other factors, then clearly no reasonable person is going to pay the same amount of money for the rusted Civic as they would the pristine Civic. However, the simple linear regression model would consider these two vehicles to be worth the same amount of money. The first vehicle in the dataset, which has a residual value of -$6032.17, indicating that this vehicle may have problems or be missing features, is a good example of the limitations of simple linear regression. The condition of the vehicle is just one of many missing explanatory variables. There are numerous other variables that influence the value of the car such as the model, engine size, leather or fabric seats, GPS or sound systems, cruise control, major collision history, equipment, etc. Therefore, it appears that in order to create a better model, a more holistic view of the vehicle is required, that is, more explanatory variables must be introduced.

Activity 2: Comparing Variable Selection Techniques:

Within this activity we use the stepwise regression analysis to find the model that would have the highest R-Squared value to predict the response variable.

During the first step of the activity we calculate every single explanatory variable by using the test.EXPLANATORY command found in Appendix 2 i). This command gives information about the R-Squared values for each of the variables and it is found that Cyl has the largest R-Squared value (seen in Table 2.1) and thus making X1 = Cyl for the next step of the activity.

Table 2.1: Regression Models of One Explanatory Variable and Its R-Squared Values

Explanatory Variable	Multiple R-Squared	Adjusted R-Squared
Cyl	0.3239	0.323
Liter	0.3115	0.3107
Doors	0.01925	0.01803
Cruise	0.1856	0.1846
Sound	0.01546	0.01423
Leather	0.02471	0.02349
Mileage	0.02046	0.01924

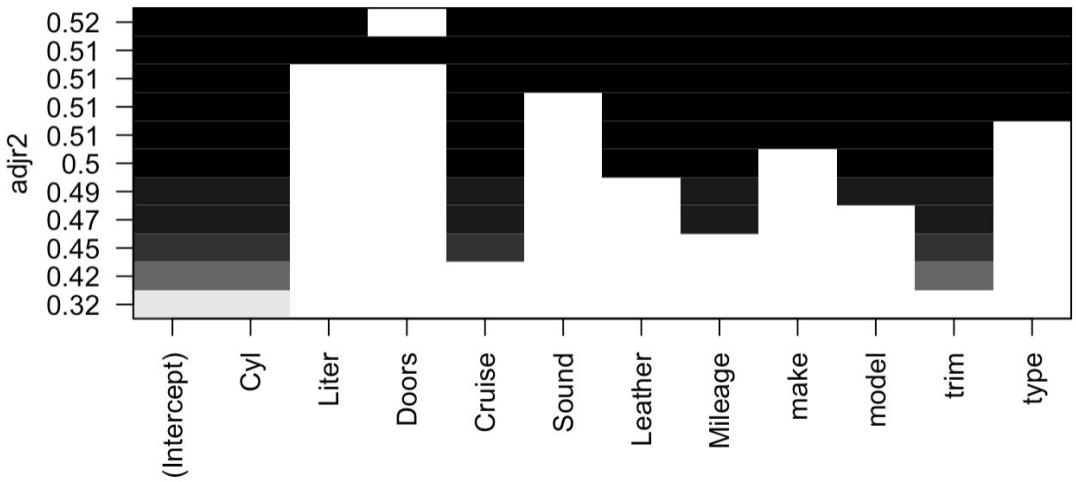
The second part of activity 2 is much like the first part in that we must calculate the highest R-Squared value for all explanatory variables. However, since we have found that Cyl has the highest R-Squared value in the first part, we must see if adding another extra explanatory variable in addition to Cyl would make the model better. After testing all other 6 explanatory variables using the command found in Appendix 2 ii), we found that the best model 2 is Cyl + Cruise, which has the largest R-Squared value out of all the others (found in Table 2.2).

Table 2.2: Regression Models of Cyl (X1) + Another One Explanatory Variable and Its R-Squared Values

Explanatory Variables (2)	Multiple R-Squared	Adjusted R-Squared
Cyl + Liter	0.3259	0.3242
Cyl + Doors	0.3435	0.3418
Cyl + Cruise	0.3839	0.3824
Cyl + Sound	0.3293	0.3276
Cyl + Leather	0.337	0.3353
Cyl + Mileage	0.3398	0.3382

Finally, it is obvious that by continuing this further with the third, forth etc. variables, it would take a long time to find the best model as we would need to code each and every additional variable separately and find its R-Squared value. By using the code found in Appendix 2 iii), the RStudio software shows that the best model suggested by the stepwise regression procedure is Price ~ Cyl + Cruise + Leather + Mileage + Doors + Sound.

Another technique to find the best subsets for the whole dataset is by converting all qualitative explanatory variables into numerical variables in order to build a subset model using the leaps package from RStudio, this can be done by using the factor() and as.numeric() commands found in Appendix 2 iv).



As seen from above, we can see the best models provided by the software and each of it's adjusted R-Square value. It is important to note that in general staticians prefer a model with less variables and a high R-Squared value. In this graph we can clearly see that the best model includes all but the "Doors" explanatory variable, however, the next few best models all have the same R-Squared value of 0.51 with only a difference of 0.01 from the best model suggested by the software. This suggests to us that it might be best to pick the lowest amount of variables from the models with a R-Squared value of 0.51 since the difference is so small. From this conclusion, it may be best to suggest the model Price ~ Cyl + Cruise + Leather + Mileage + Model + Trim. This model suggested contains only 6 out of the 11 explanatory variables needed and has the second highest R-Squared value obtainable.

Comparing Techniques:

After comparing the two models given, one from the manual stepwise regression suggested by question 5 and the other with the leaps package that gives out all best subset models by using all of the dataset suggest by question 6, we see that both models give out out a similar conclusion in that the best model to use would most likely need to include the following explanatory variables: Cyl, Cruise, Leather and Mileage.

In these models, it is believed that different explanatory variables are considered important as different variables give different R-Squared values which decide how reliable the value is when determining the response variable. It is common that as we see more different explanatory variables, it would lead to a more accurate result of the response variable, however, it is hard sometimes to find a large amount of explanatory variables.

The stepwise regression in question 5 provided by the RStudio software concluded that the explanatory variable "Liter" was not useful in predicting the price. It had resulted in a suggestion of a model that uses 6 out of the 7 explanatory variables (all but Liter) to be the best model. Furthermore, the models developed by the software that included all of the dataset in question 6 suggested that if the objective was to have a model with the highest R-Squared value, Liter would be needed, however, almost all the other models suggested that also has a high R-Squared value (but not the highest value) does not include "Liter" as a variable within its model. To continue, the model that we recommended from the software's suggestions does not include that of "Liter" either. As a conclusion, this explanatory variable is important if you would like to sacrifice an additional variable for an increase in R-Squared value but it is not needed/important if the amount of variables is important when creating a model.

In these two different ways of finding the best models, the subsets techniques gives the user more options by assessing all the best models and determining the best ones based on the amount of variables included. Even though both techniques had a similar conclusion on which explanatory variable is a must, the subset technique gives the user more information on

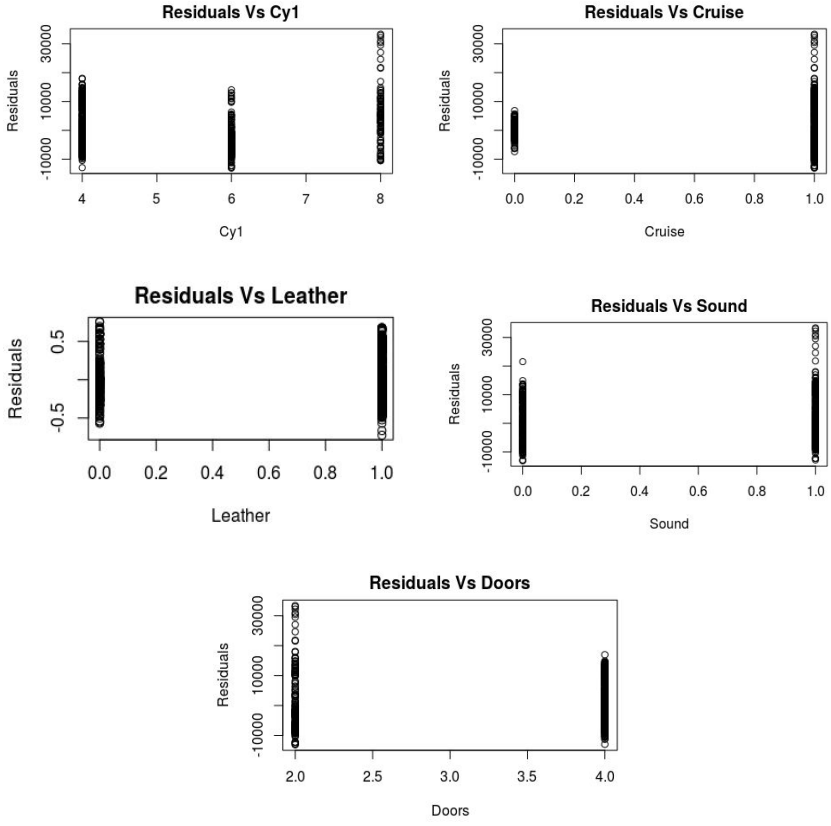
determining which model to use based on R-Squared value and amount of explanatory variables. The sequential technique on the other hand, only builds one model at a time which can be time consuming and will allow the user to only view the details of a singular model thus NOT allowing it to compare with other models that may be more/less/as effective.

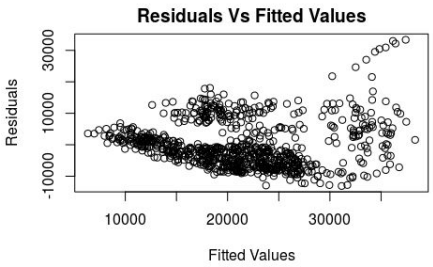
Activity 3: Checking the Model Assumption:

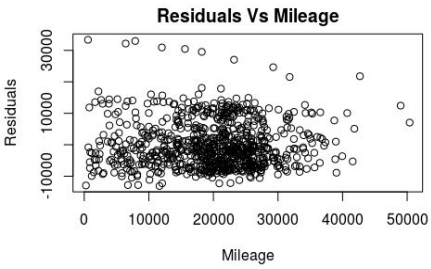
The regression model obtained in Question 5 is:

$$Y = 7323.16 + 3200.12X_1 + 6205.51X_2 + 3327.14X_3 - 2024.40X_4 - 1463.39X_5 - 0.1705X_6$$

Where $X_1 \rightarrow$ Cy1, $X_2 \rightarrow$ Cruise, $X_3 \rightarrow$ Leather , $X_4 \rightarrow$ Sound , $X_5 \rightarrow$ Doors, $X_6 \rightarrow$ Mileage







By looking at the graph: Residuals Vs Mileage, the size of the residuals do not seem to be changing as the mileage changes. There does not seem to be a very clear trend. At a mileage value of about 20,000, the graph seems to have a dense spot. The size of the residuals seems to decrease as the price increases. Graph: Residuals Vs Fitted Values also shows signs of heteroskedasticity, as the fitted values increase, the residual values 'fan' out more and more. The most frequently successful method to fix heteroskedasticity is to transform a variable, which is done later in this section. This is a bad residual plot as the plot has high density far away from the origin and a low density near the origin.

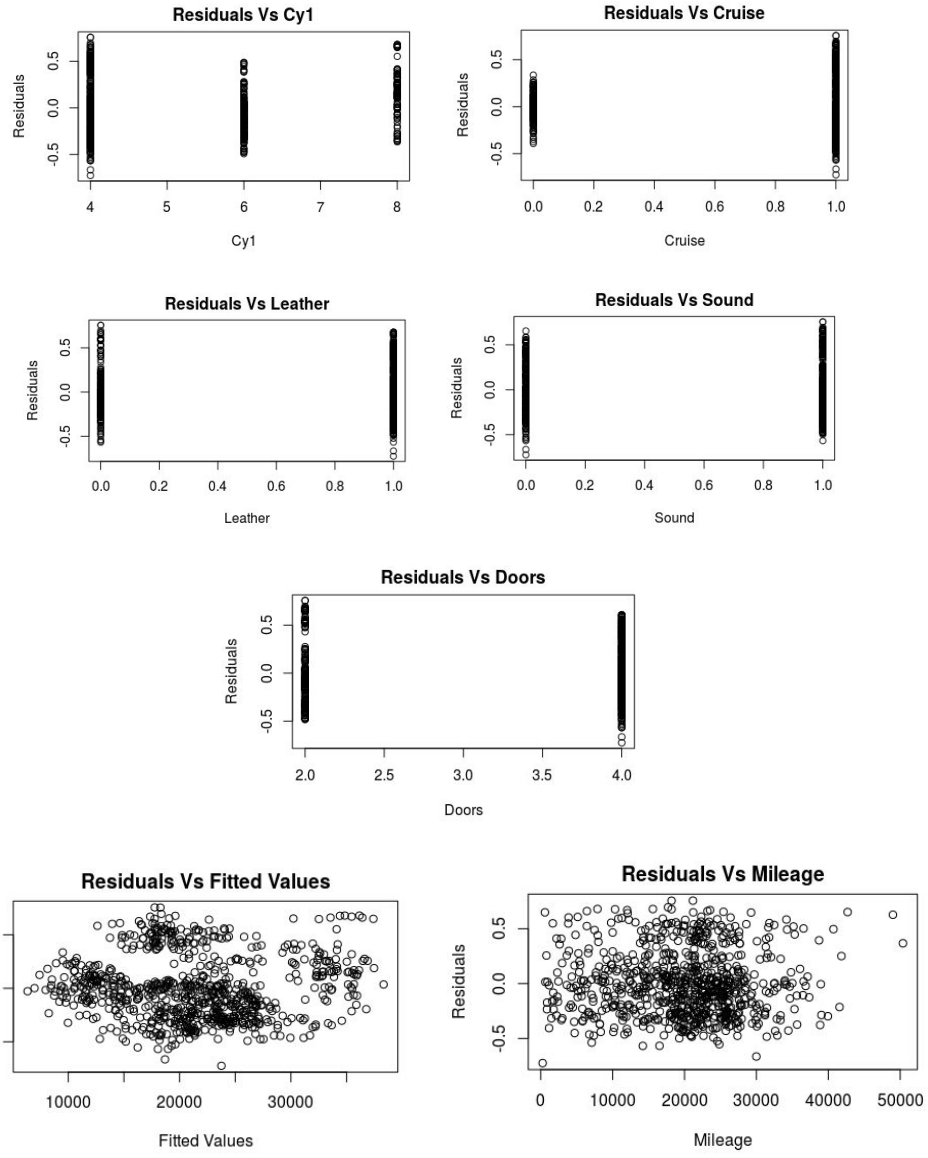
The Residual V.S. Mileage graph shows a pattern of right skewness as the data points are balanced around the line Y= 0. The residual plots for Cy1, Cruise, Leather, Doors and Sounds all seem to have no obvious trends which implies that the error terms are considered independent and the regression assumption satisfied. The residual plots for Leather and Doors seem to be positively skewed, while the residual plots for Sound, Cy1 and Cruise seem to be negatively skewed.

Log Transformation:

The regression model after taking the log transformation of the price values is:

$$Y = 9.20 + 0.1302X_1 + 0.3208X_2 + 0.1214X_3 - 0.087X_4 - 0.0371X_5 - 7.38e-6X_6$$

Where $X_1 \rightarrow$ Cy1, $X_2 \rightarrow$ Cruise, $X_3 \rightarrow$ Leather , $X_4 \rightarrow$ Sound , $X_5 \rightarrow$ Doors, $X_6 \rightarrow$ Mileage



Square Root Transformation:

The regression model after taking the square root transformation of the price values is:

$$Y = 94.75 + 9.958X_1 + 22.15X_2 + 9.934X_3 - 6.659X_4 - 3.679X_5 - 0.00054X_6$$

Where $X_1 \rightarrow$ Cy1, $X_2 \rightarrow$ Cruise, $X_3 \rightarrow$ Leather , $X_4 \rightarrow$ Sound , $X_5 \rightarrow$ Doors, $X_6 \rightarrow$ Mileage

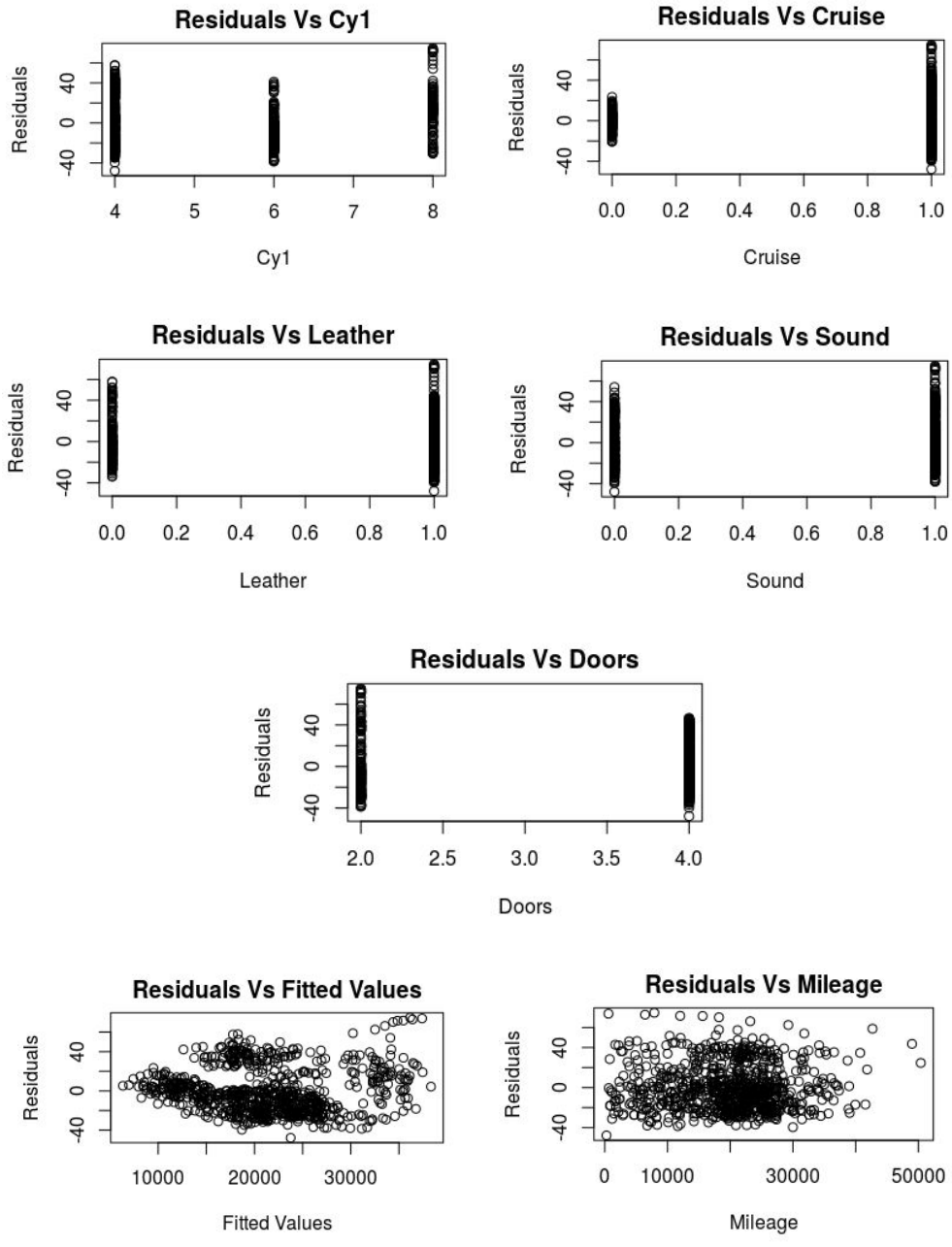


Table 3.1: Comparing Models based on R Value:

	R-Squared	Adjusted R-value
Original Model	0.4457	0.4415
Log Price Transformation	0.4836	0.4797
Square-root Transformation	0.4689	0.4649

The log price transformation did the best job of reducing the heteroskedasticity and skewness in the residual plots. The log transformation model has an R-squared value of 0.4836 while the square root transformation model has an R-squared value of 0.4689. The residual vs fitted values plot for the log transformation is more symmetric along the y-axis compared to the original model and the square root transformation model.

In this case, the best residual plots do correspond to the highest R-squared value. The log transformation has the best residual plots as well as the highest R-squared value. This is not always the case as one can have a large R-squared value but bad residual plots.

Table 3.2: Comparison of Results Obtained By Transformations:

	Log transformation compared to original model	Square root transformation compared to original model
Residuals Vs Mileage plot	The plot for the log transformation is less dense than that of the original model	The plot for the square root transformation is less dense than that of the original model
Residuals Vs Fitted Values plot	The plot for the log transformation shows less heteroskedasticity than the original model plot	The plot for the square root transformation shows less heteroskedasticity than the original model plot but more heteroskedasticity than the log transformation plot

Conclusion:

In conclusion, simple linear regression proved highly ineffective in modeling the dataset, necessitating a more sophisticated model. Furthermore, two different techniques were used to find the best models to predict the price of a car. This resulted in two different but similar regression models that had similar explanatory variables. Using the model found by the stepwise regression technique, two different transformations were applied to the response variable in order to achieve better residual plots. The log transformation produced a better regression model than the square root transformation based on the R-squared values and the residual plots.