Multiple Linear Regression: How Much is Your Car Worth?

By: James Bui & Malcolm Beard

Abstract

In this paper, we answer a series of activities in which we create a multivariate regression model. From this model, we compare the price of 2005 General Motors vehicles to a number of varying vehicular characteristics. These characteristics include mileage, cylinders, engine size (liters), doors, cruise control, sound system, and leather seating. From the data, we are able to form a simple linear regression, compare variable selection techniques, check model assumptions and determine outliers and influential observations. As well, our conclusions from the data help to answer a number of questions surrounding regression analysis techniques.

For our dataset, it is important to note that the suggested price -*price* in our data set- will always be used for the Y variable.

Activity 1: A Simple Linear Regression Model

The purpose of this activity is to determine the strength of the relationship between suggested car prices and mileage through means of plotting and creating a simple linear regression model.

First, here is a scatterplot of the cars data set displaying the relation between the cars' suggested prices and their mileage.

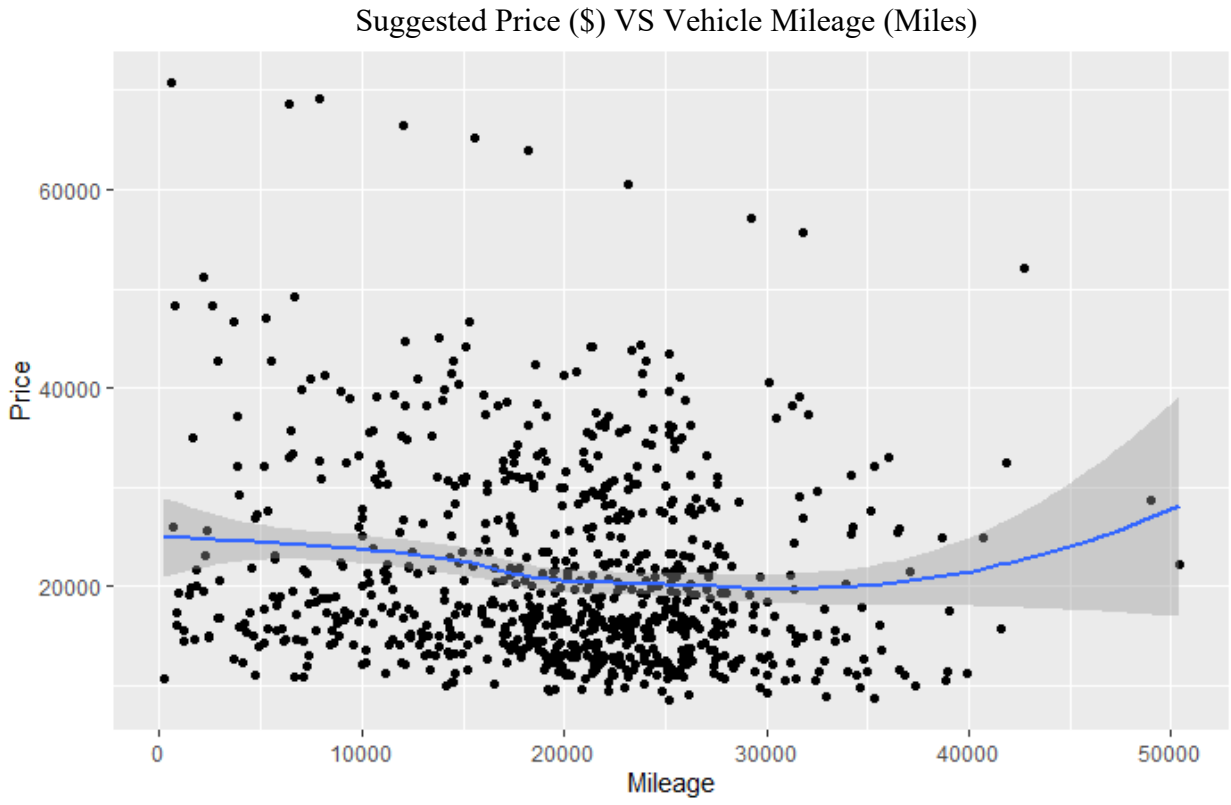


Figure 1: Activity 1.1 Scatterplot

From the scatterplot, there are no indications that the expected price of the cars is affected by the mileage. A lot of the data is centered in the middle of the plot, but there is no price

increase or decrease occurring at either end of the mileage variable. Logically, you would expect the suggested price to decrease when there is more mileage put on the vehicle, yet this plot and our data suggest no such results. Overall, the mileage and price relation does not seem very strong.

By creating a least squares regression line, we are able to more clearly determine whether or not mileage is a good predictor of suggested car price. We know the model must follow the format: $rice = \beta 0 + \beta 1 * Mileage$.

We also know that $\beta 0 = meanPrice - \beta 1 * meanMileage$ and$\beta 1 = SXY/SXX$, where X represents Mileage while Y represents Price. From these equations, we determined the regression model is as follows:

$$Price = 24764.56 - 0.1725205 * Mileage$$

With the assistance of an ANOVA table, we summarized the R^2 value, correlation coefficient, t-statistics, and p-values in the table below.

Observations	Values
R^2 value	0.02
Corr. Coeff.	0.14
t-statistics	27.38-4.09*Mileage
p-values	(2e-16)+(4.68e-05)*Mileage

Assuming a standard confidence interval of 0.95, the p-values of both models are under 0.05; therefore, we can reject the null hypothesis. Additionally, looking at the t-statistics, we can compare the absolute t-statistic to its critical t value. Therefore, mileage is not a good indicator of price.

To exemplify the weak relation between price and mileage, we took the residual value of the first car in the cars dataset, a Buick Century with 8221 miles on it. The following formula is what gives us the residual value:

Residual Value=Observed retail price-Expected price (calculated by regression line)

Residual Value=17314.103-23346.269

Residual value=-6032.139

In activity 1, the last thing we must discuss is the limitation of using simple linear regression to measure car price. The limitation of using this technique is that it assumes there is a linear (a straight line) relationship between both the independent and dependent variable. Other assumptions include homoscedasticity, which is the assumption that the variance of residuals is the same for any value of X; independence, in which observations are independent of each other; and normality, such that for any fixed value of X, Y is normally distributed. Another common limitation to simple linear regressions is 'under fitting" and 'overfitting'. If any of these assumptions are violated we cannot do a simple linear regression.

Activity 2: Comparing Variable Selection Techniques

In activity 2, we must create and compare two regression models using stepwise regression analysis as well as a best subsets technique.

In the first technique, we calculate seven regression models using the seven characteristics described in the abstract. From these, we determine the explanatory variable (X1) to be the variable used in the model with the largest R^2 value. *Table A* in the *Appendix* has the R^2 values for each characteristic; however, cylinders is the variable with the highest R^2 value. The R^2 value for cyl is 0.3239. Therefore, we will denote cyl as X1.

Next, we must calculate 6 more regression models with two explanatory variables (X1 and one of the other six characteristics). Again, we search for the model with the largest R^2 value. *Table B* has the information for all the two-variable models. The model with the largest R^2 value is the one including the cruise characteristic. Its R^2 value is 0.3839.

The difference of the R^2 values between the models with and without the inclusion of the cruise variable is 0.06. The regression model improved by 0.06 with cruise included, thus, it is necessary for a more accurate regression model.

We can continue following this technique until we find the model with the largest R^2 value overall. Stepwise regression recommends a model involving the following explanatory variables in this respective order: Cyl, Cruise, Leather, Mileage, Doors and Sound. The following is what the model looks like in the R code:

$$lm(formula = Price \sim Cyl + Cruise + Leather + Mileage + Doors + Sound, data = CARS)$$

Another method we can use to find the best model is the in-class method. In this method, the model with the lowest AIC value is the model that we should use according to this Stepwise Regression. This is because the AIC, Akaike information criterion, is an estimator of out-of-sample prediction error, which makes it a good measure to judge the relative quality of statistical models from a given dataset. The suggested model with an AIC value of 14330.22 is as follows:

$$Price \sim Cyl + Doors + Cruise + Sound + Leather + Mileage$$

With the coefficients inserted in the model, we see the stepwise regression suggests the following as our model:

$$Price = 7323 + 3200 * Cyl - 1463 * Doors + 6206 * Cruise - 2024 * Sound + 3327 * Leather - 0.17 * Mileage$$

The next technique we use to get another regression model is the *best subsets technique* for the entire cars dataset. This means that we must include both the quantitative variables used in the last technique and the categorical variables provided in the data. Namely, these categorical variables are make, model, trim and type. To include categorical variables into a regression model we must first begin by converting each categorical variable from a string factor to a numeric variable. Once this is done, these variables will then be fit into the model for a best subsets regression because R has the intelligence to process these variables individually. This is the process we did to transform variables Trim, Make, Type and Model. Table C in the appendix gives the table of the variables in the model with best predictive power. Of the 11 variables the best subset was 10 variables and this was based on the subset having the highest Adjusted R^2 value and lowest Cp Value.

Now the question is which explanatory variables should be included in our regression model to get the most accurate predictor of price. It is important to note that we can look more in-depth to see which order and number of variables is best. The adjusted R^2 value and the CP

value is subset 10, further indicating that 10 variables are best for increasing the accuracy of our regression model. The data table shows that the regression model should be as follows:

$$Price = 4248.35 + 5029.18 * Cyl - 165.43 * Trim + 4479.90 * Cruise$$
$$-0.19 * Mileage - 144.94 * Model + 2551.46 * Leather + 749.68 * Make$$
$$-683.97 * Type - 1270.26 * Sound - 1295.97 * Liter$$

Previously the quantitative variables: y = Cyl, Doors, Cruise, Sound, Leather and Mileage were suggested by the stepwise regression; however, when we add the categorical variables for a best subsets regression, we see that most of these the number of doors has no predictive power. Additionally and surprisingly, we see that Liter has more predictive power than doors, although not by much.

A reason why the best subsets technique is more informative than the stepwise technique is because it considers all the data in the cars dataset. Not only does it consider the quantitative points important, but it also considers the characteristic variables. These variables are very important to analyze because, logically, the brand of an item is very important to consumers. The best subsets technique will give the most useful information for the given dataset.

Activity 3: Checking the Model Assumption

In the third activity, we check our model assumptions of our regression model obtained through the stepwise technique. As a reminder, that regression model is as follows:

$$Price = 7323 + 3200 * Cyl - 1463 * Doors + 6206 * Cruise$$
$$-2024 * Sound + 3327 * Leather - 0.17 * Mileage$$

After plotting the residuals against the explanatory variables and price, we came to a number of conclusions. First, it generally appears that as mileage increases, residuals stay relatively constant. Linearity exists though, which can be seen in the outliers of Figure 2. Next, the predicted price has signs of heteroscedasticity. In figure 3, the residuals increase in a funnel shape as the model increases. While it is increasing, the points are still scattered without form, suggesting heteroscedasticity exists.

Residuals vs Mileage

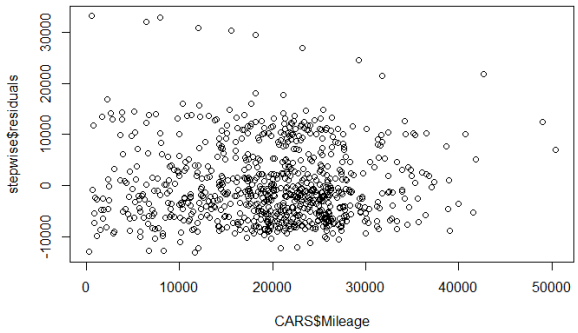


Figure 2: Activity 3.8 Residuals vs Mileage

Residuals vs Price

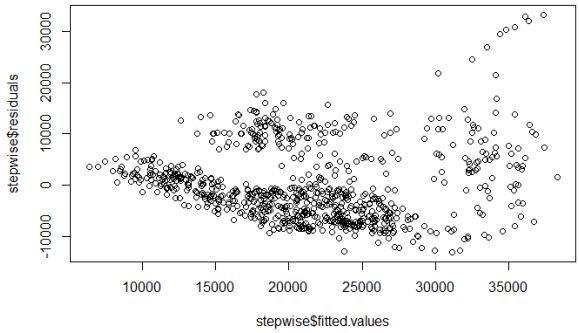


Figure 3: Activity 3.8 Residuals vs Price

In our instructions, we learn that the residual vs mileage plot is right skewed. Below is a Q-Q plot of residuals vs mileage, a visual indication of the right skewness. Also, taking a Q-Q

plot of residuals vs price, we can see that right skewness also exists in the Y variable. This can also be seen in the residual vs price.

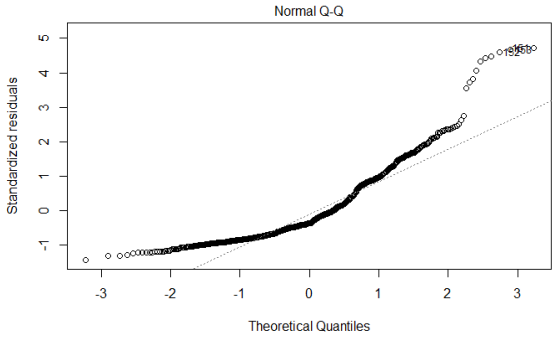


Figure 4: Activity 3.8.c normal QQ/mileage

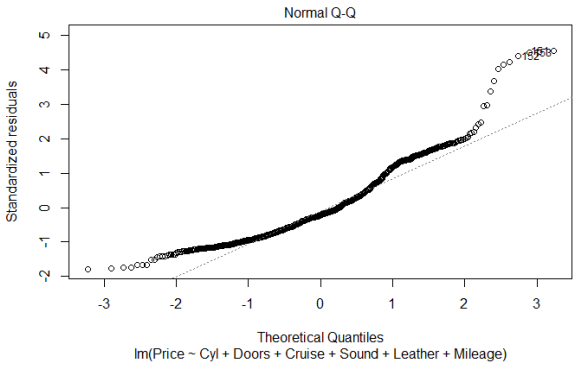


Figure 5: Activity 3.8.c normal QQ/price

When we look at the residual vs mileage plot at the fitted value 8000, the residuals are clearly not centered at Y=0, implying that the skewness will be to the right rather than centered.

Residuals vs Mileage at fitted value 8000

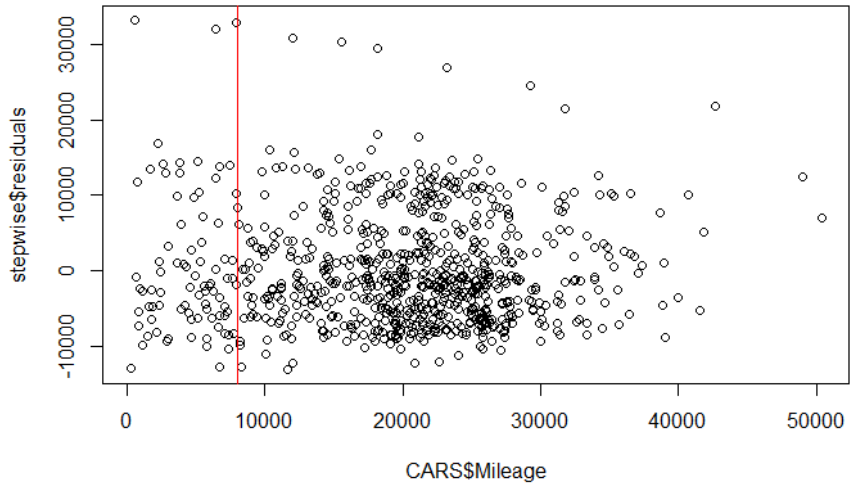


Figure 6: Activity 3.8.c Fitted Value Residual Plot

Before we move on to the next portion, we would like to note that the other residual plots provide little information for the regression model as a whole. In fact, all values in each plot only exist on a few numbers, along long vertical lines. While the tables may not be necessary to the current topic, we have included the plots in the appendix.

The next things to explore are the regression models and residual plots of the price variable transformed to $log(Price)$and$\sqrt{Price}$. The regression model for the log function is$Tprice = log(Price) \sim Cyl + Doors + Cruise + Sound + Leather + Mileage$; whereas, the sqrt function is$sqrt(Price) \sim Cyl + Doors + Cruise + Sound + Leather + Mileage$. The residual plots are as follows:

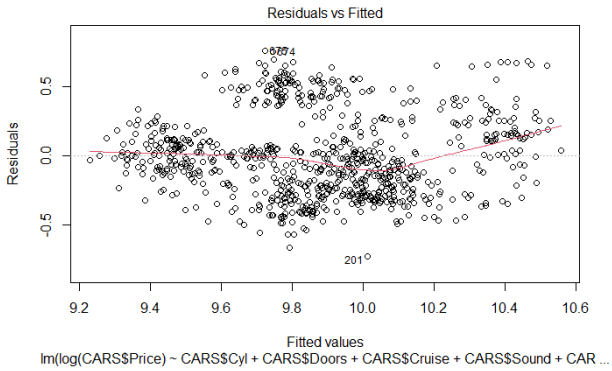


Figure 7: Activity 3.9 residual vs log

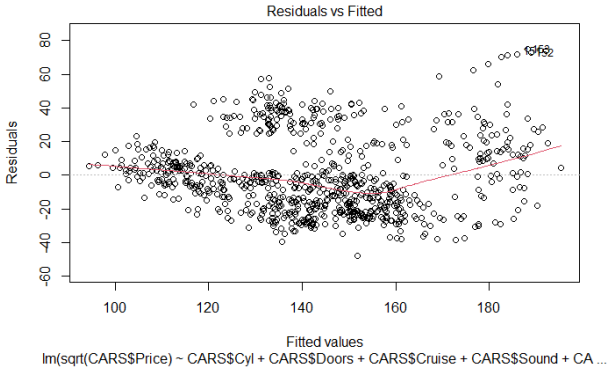


Figure 8: Activity 3.9 residual vs sqrt

For the Tprice function, Multiple R-squared: 0.4836, Adjusted R-squared: 0.4797.
For the sqrt function, Multiple R-squared: 0.4689, Adjusted R-squared: 0.4649.

The Tprice function seems to have more constant residual values as price increases while the sqrt function has an increasing residual value as price increases. From the Q-Q plots, we can tell that Tprice has less of a right skewness and thus, we can reduce more heteroscedasticity from the Tprice function. Based on this heteroscedasticity, we can also see that the Tprice function has the better R^2 value, which further corresponds to the better residual plot.

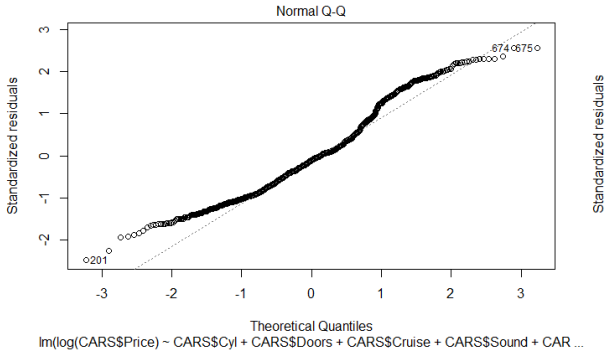


Figure 9: Activity 3.9 normal QQ/log

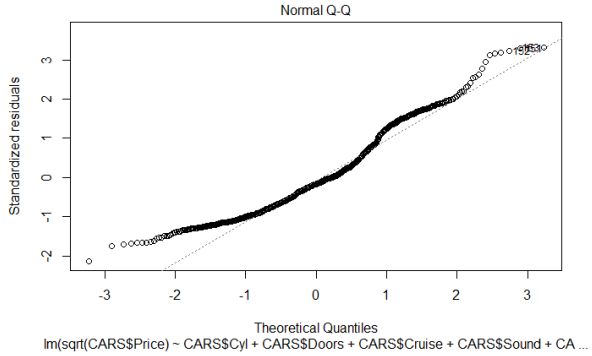


Figure 10: Activity 3.9 normal QQ/sqrt

Now, let us compare Tprice's regression model to the original regression model, which has the multiple R^2 value at 0.4457 and adjusted R^2 value at 0.4415.

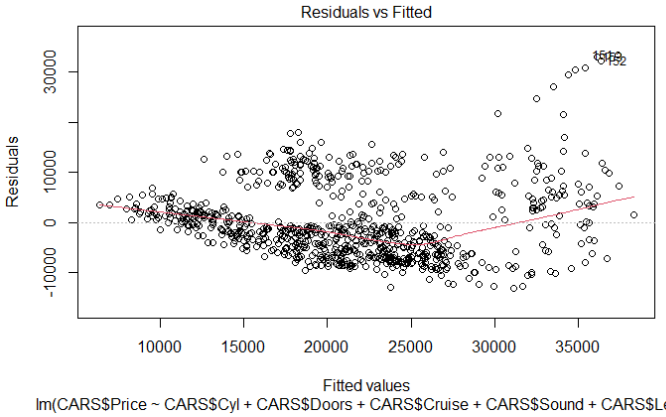


Figure 9: Activity 3.9 residual vs price

The Tprice function has the highest R^2 value. It is also evident by the right-skewness that the original residual plot has increasing residuals, implying more heteroscedasticity. The Tprice function model seems to improve on both of these aspects.

Activity 4: Outliers and Influential Observations

In this last activity, we must determine the outliers and their influence in the regression equation suggested by the stepwise regression model used in activity 2. The equation is as follows, with y representing the suggested price variable:

$$y = b0 + b1(Cyl) + b2(Doors) + b3(Cruise) + b4(Sound) + b5(Leather) + b6(Mileage)$$

A suitable starting point from which we can begin identifying any outlier points is defining the criteria of what is considered an 'outlier'. Outliers are data points that lie far away from the majority of the data and its predicted value (y-value), possessing larger residual values. More specifically, the traditional rule of thumb has always been that any point with a studentized residual value beyond 3 deviations of the dataset (in absolute value) is considered an outlier. Knowing this, we can begin the removal of data values. In figure 3, there visually appears to be a cluster of outliers at the top right of the residual plot. Rather than guess if those are the outliers of the cars dataset, we calculate the studentized residual value. By doing so, we conclude that the following data points from the cars data can be considered outliers:

Make and Model	Price ($)
Cadillac XLR-V8	70755.47
Cadillac XLR-V8	68566.19
Cadillac XLR-V8	69133.73
Cadillac XLR-V8	66374.31
Cadillac XLR-V8	65281.48
Cadillac XLR-V8	63913.12
Cadillac XLR-V8	60567.55
Cadillac XLR-V8	57154.44

These outliers occur in rows 151-158 in cars data. Looking at the entirety of that data, we see that these eight outliers are the eight most expensive cars in total. Between one another, every characteristic is the same except for the price and mileage, in which the vehicles with the least mileage cost the absolute most. This occurrence in the data tells us that this specific make and model, the Cadillac XLR-V8, is the highest end GM vehicle in the dataset. Because of this, we chose to remove observation 159 and 160 from the dataset since they are also categorized as Cadillac XLR-V8.

Now we must analyze the influential points to determine which points we need to remove. We will examine the 'leverage' of the dataset, which measures any unusual influence of a dataset over its X-axis. We will begin first by calculating the leverage in this dataset. Once this is done we can use 'Cook's Distance', one of the most commonly used tools in the field of statistical regression, to identify the net effect of removing an outlier or points that are worth further investigations. These points are called "influential points".

We are able to determine that there are no additional outliers over the X-axis to additionally exclude. Now we can use Cook's distance analysis to see the influence of the outliers on the regression line.

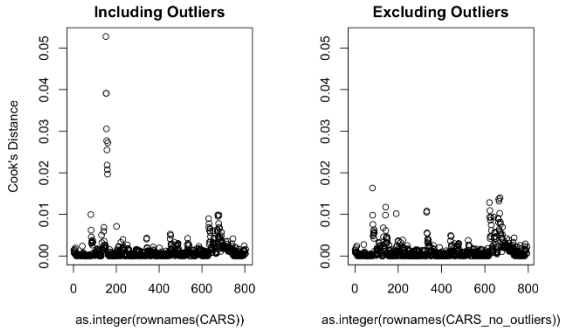


Figure 10: Activity 4.11 Outlier's influence

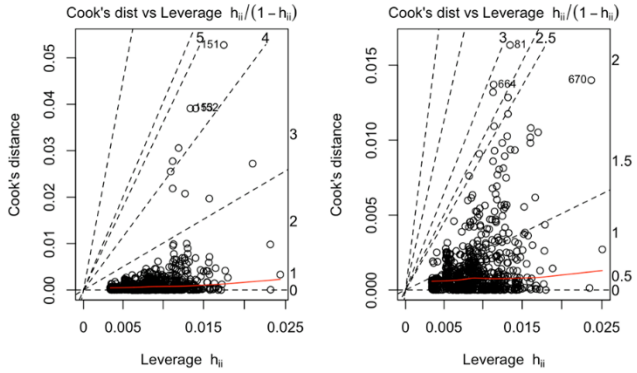


Figure 11: Activity 4.11 Cook's vs Leverage with & without Outliers

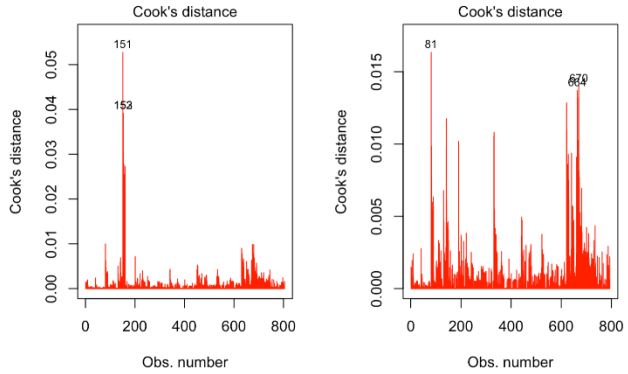


Figure 12: Activity 4.11 Outliers and no Outliers

Looking at the plots that compare cook's distance to the dataset with and without the outliers, we see a few points pulling the data upwards, but nothing as influential as without the outliers. We can see in the Cook's distance vs Leverage plot that the red line is pulled upwards at

the right end due to the outliers. There is less of a pull in the second plot, implying that outliers are influential to the dataset. From these plots, we can see that there is a drastic difference at observations 81, 151 and 152. Influential observations and outliers are observed.

Lastly, we must look at the coefficients of the regression line in order to see if the outliers are truly influential or not.

Coefficients	With Outliers	Without Outliers
Intercept	7323.16	6925.85
Cyl	3200.12	2713.29
Doors	-1463.40	-788.08
Cruise	6205.51	6310.23
Sound	-2024.40	-2512.72
Leather	3327.14	3156.76
Mileage	-0.17	-0.14

From this table, it is clear that the outliers are influential observations in the cars dataset.

Conclusion:

All in all, a lot of information can be observed from the cars dataset in terms of regression analysis. This information helps us to further understand how regression is utilized and why it is so important.

Each activity has led us to the conclusion that the best regression model for the cars data is Tprice (the log function). The specific $b_0 ... b_6$ values are found through summarizing the Tprice function.

$$Log(Price) = 9.2 + 0.130 * Cyl - 0.037 * Doors + 0.321 * Cruise$$
$$-0.087 * Sound + 0.121 * Leather - (7.382e - 06) * Mileage$$

Appendix:

Table A: Activity 2.5.a

Characteristics	Intercept	X variable	R^2 value
Cylinders	-17.06	4054.20	0.3239
Liter	6186	4990	0.3115
Doors	27034	-1613	0.01925
Cruise	13922	9862	0.1856
Sound	23130	-2631	0.01546
Leather	18829	3473	0.02471
Mileage	24764.56	-0.1725	0.02046

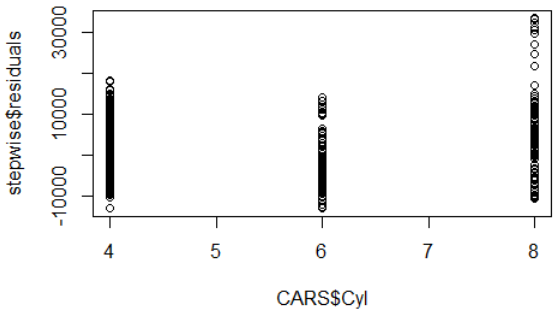
Table B: Activity 2.5.b

Characteristic	R^2 value
Liter	0.3256
Doors	0.3435
Cruise	0.3839
Sound	0.3293
Leather	0.3370
Mileage	0.3398

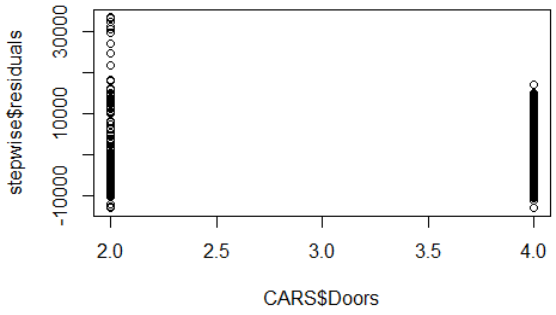
Table C: Activity 2.6

	Cyl	Liters	Doors	Cruise	Sound	Leather	Mileage	Make	Model	Trim	Type
1	*										
2	*									*	
3	*			*						*	
4	*			*			*			*	
5	*			*			*		*	*	
6	*			*		*	*		*	*	
7	*			*		*	*	*	*	*	
8	*			*		*	*	*	*	*	*
9	*			*	*	*	*	*	*	*	*
10	*	*		*	*	*	*	*	*	*	*

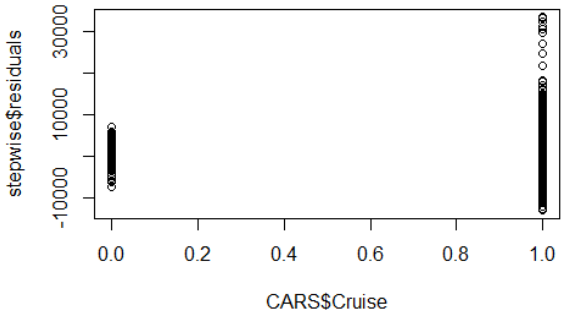
Plot A: Activity 3.8 residuals vs cylinders



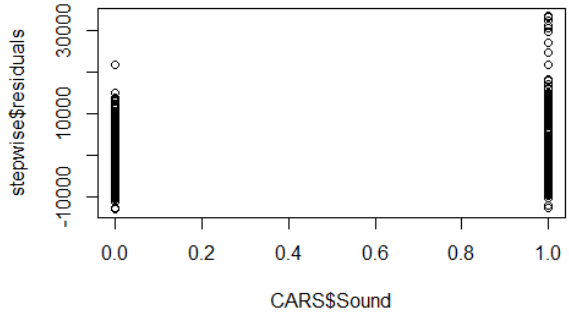
Plot B: Activity 3.8 residuals vs doors



Plot C: Activity 3.8 residuals vs cruise



Plot D: Activity 3.8 residuals vs sound



Plot E: Activity 3.8 residuals vs leather