University of Glasgow
School of Mathematics and Statistics

Investigating the KBB Automobile Dataset

By David Quinlan

Contents

1 Introduction

Kelley Blue Book or KBB is a vehicle and automotive research company based in Irvine, California, USA. KBB analyses and publishes data on new and used cars. Large automotive manufacturers, as well as individual car and truck buyers, use the Kelley Blue Book website to find in-depth pricing information when searching for a specific vehicle. KBB provides a Fair Purchase Price and Fair Market Range for both used and new vehicles. These values are based on actual transactions of what other drivers are paying for a vehicle and are adjusted regularly for changes in market conditions. Kelley Blue Book also provides information on the certified pre-owned value, retail value, trade-in value and private party value for used cars.

For this investigation, a sample dataset was collected from Kelly Blue Book. This data set contains a total of eight hundred and ten observations with each observation related to a 2005 used General Motors (GM) car. All vehicles contained within this dataset were less than one years old when priced and were considered to be in an excellent condition. For these observations, the following 12 variables characterising each car were collected:

- **Price:** Suggested retail price of the used 2005 GM car in excellent condition.
- **Mileage:** Number of miles the car has been driven.
- **Make:** Manufacturer of the car such as Saturn, Pontiac, and Chevrolet.
- **Model:** Specific models for each car manufacturer such as Ion, Vibe, Cavalier.
- **Trim(of car):** Specific type of car model such as SE Sedan 4D, Quad Coupé 2D etc.
- **Type:** Body type such as Sedan, Coupé, etc.
- **Cylinder:** Number of cylinders in the engine.
- **Liter:** A more specific measure of engine size.
- **Doors:** Number of doors.
- **Cruise:** Indicator variable representing whether the car has cruise control (1 = cruise).
- **Sound:** Indicator variable representing whether the car has upgraded speakers (1 = upgraded).
- **Leather:** Indicator variable representing whether the car has leather seats (1 = leather).

The main aims of this investigation are:

1. To fit a model between of the response variable price and the 11 other explanatory variables.
2. To find the explanatory variables which best model the price of the 810 GM cars.

2 Methods

2.1 Multiple Linear Regression

A multiple linear regression model is a modelling approach whereby a linear relationship is fitted between a continuous response variable and a number of explanatory variables.

The linear regression model is given by the following formula:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \quad \text{for } i = 1, ..., n \tag{1}$$

Where β_0 is the intercept term and $\beta_1, ..., \beta_p$ are the coefficients of the explanatory variables.

The linear regression assumptions are as follows:

- There is a linear relationship between the response variable and the explanatory variables.
- The values $x_{i1}, ..., x_{ip}$ are fixed values.
- The explanatory variables are independent (i.e. no presence of multicollinearity).
- Error terms are independent and $N(0, \sigma^2)$ distributed with zero mean and constant variance σ^2.

3 Results and Discussion

3.1 Initial Exploratory Analysis

Prior to conducting the exploratory analysis, the variables included in our data set were investigated. When examining the data set, tables were created of the different variables against each other, it was found that the type and liter variable contained the same characteristic information as the doors and cylinder variables, respectively. When these variables were also plotted against each other, it was clear to see that both type and doors and liter and cylinder were correlated. The plot of the car type against the number of doors as well as the box-plot of the liter variable split by the number of cylinders are shown below in Figure 1 and Figure 2, respectively.

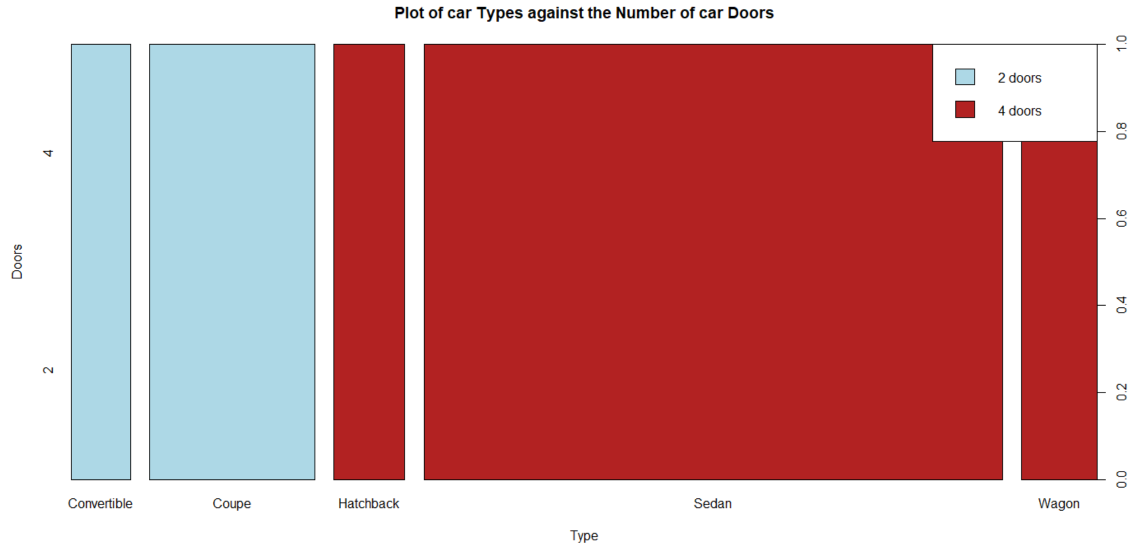

Figure 1: Plot the type against the number of doors

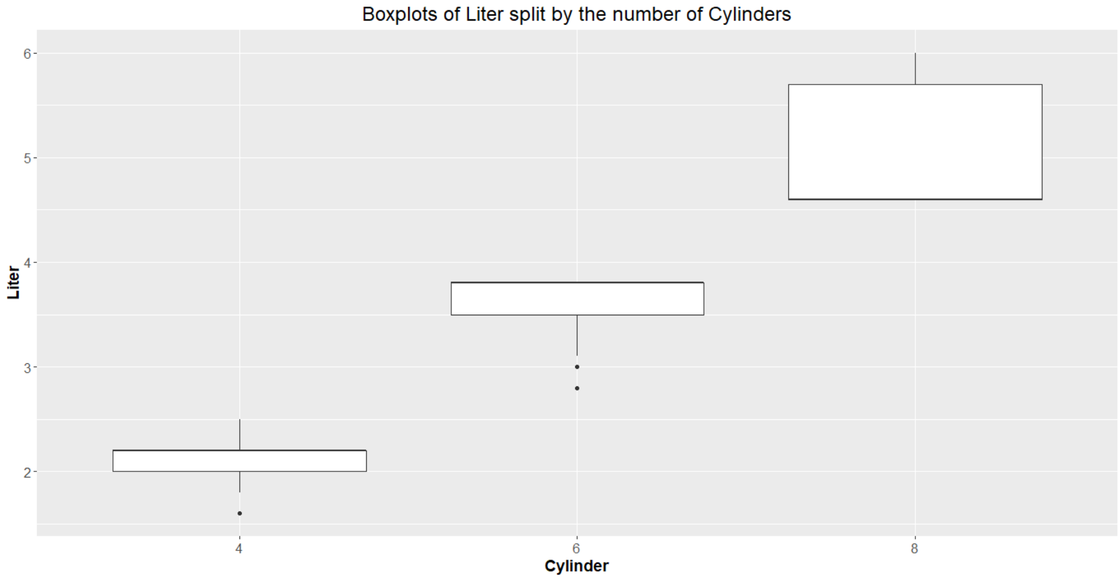

Figure 2: Box-plot of liter split by the number of cylinders

From the plot above it is clear that the number of doors is related to the type of car, i.e. a Coupé only has two doors. Therefore, the information contained within the number of doors variable was also contained within the type variable. Similarly, it was found that the cylinder variable was a subset of the liter variable, this can be seen visually above when examining the box-plots in Figure 2, i.e. we can see that there is no overlap between any of the box-plots. Thus, the variables doors and cylinder were excluded from the model fitting procedure as they would not have added any additional information to the model. The variables model and trim contained 32 and 47 different categories, respectively. These variables contained many levels with most having very few observations. It was decided that these two variables would also be removed as if they were included in a model they would cause the model to be over-parametrized.

A histogram were created to examine the distribution of the response variable, price. This histogram is shown below on the left in Figure 3.

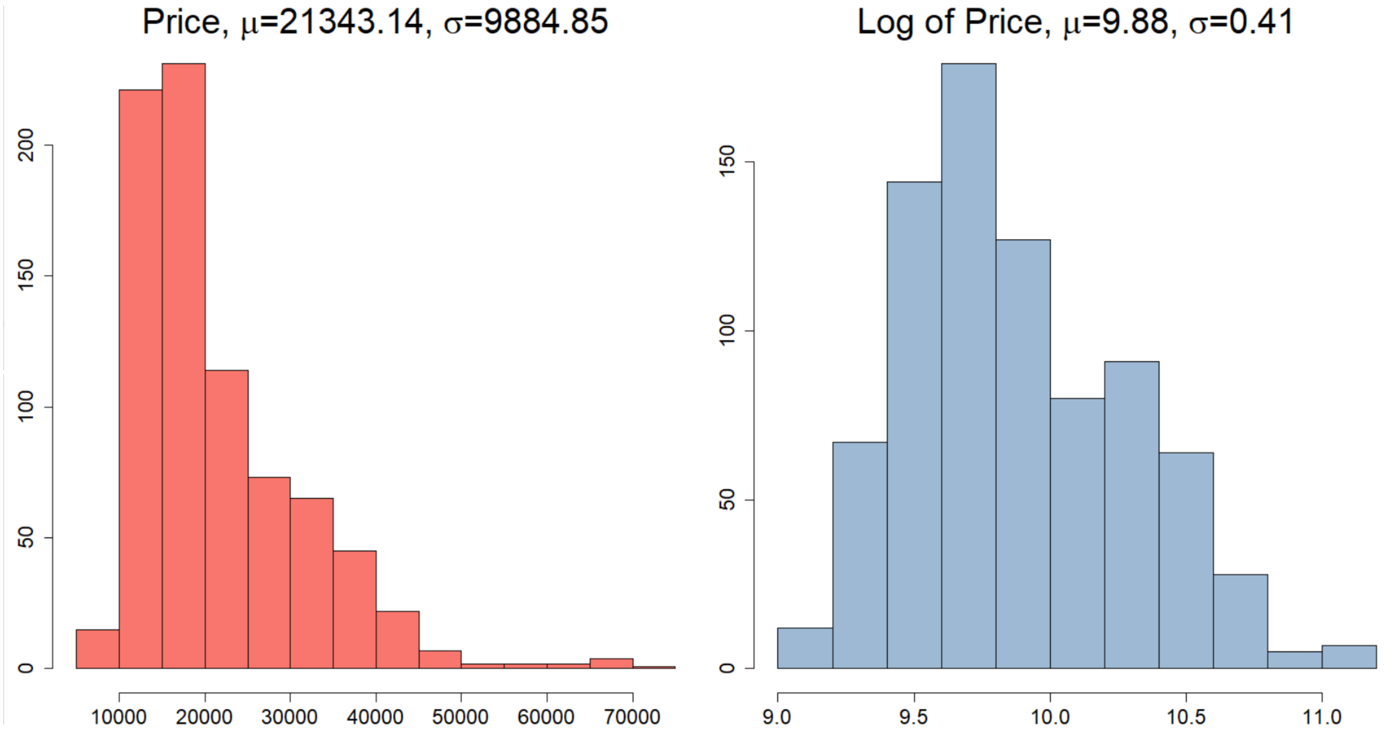

Figure 3: Histograms of the price and log of price variables

It is clear from the price histogram shown above that the price variable is right skewed with the vast majority of the car prices occurring within the interval of 10000 to 30000, with a median value of 18025 dollars.

Given that the distribution of price is skewed to the right it was decided to transform this variable using a log transformation. A log transformation would reduce the skewness and transform the data to an approximately normal distribution[1]. The histogram on the right in Figure 3 shows the log of the price variable, which, now appears to be approximately normally distributed.

It was also interesting to examine how the price variable varied across the different levels contained within the categorical variables. Figure 4 below is a plot of the price variable split by both car type and car make.

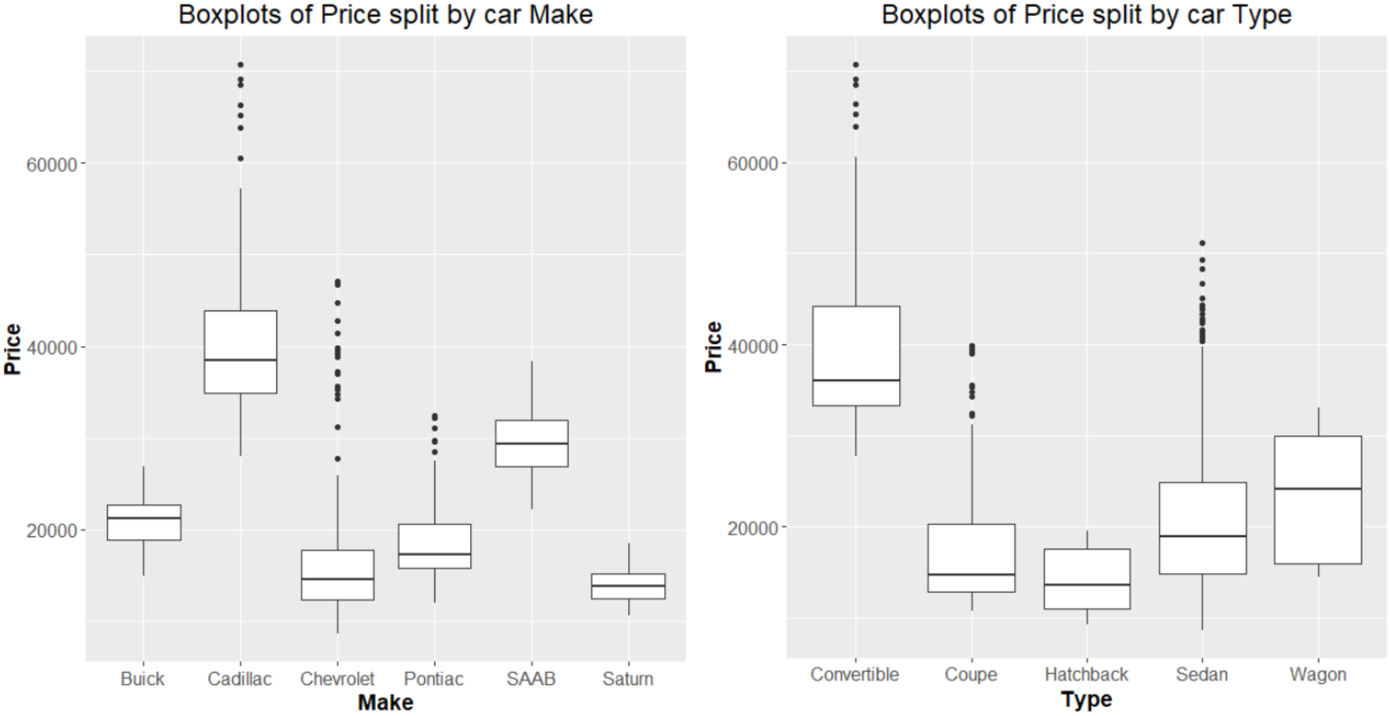

Figure 4: Box-plots of the price variable split by car make and type

It is clear from the plot of the price against car type that the price of a Cadillac is much greater than the price of the other types of car present in this data set. Similarly, in the plot of the price split by the type of car, it is clear that convertible cars are more expensive than say a Coupé or a Sedan. Thus, one can see from the plots above that the type and make of car one owns would have an affect on the price of that car, if this was not the case all of the box-plots shown above would have stayed around the same position, possessing an approximately similar median value. The results found above would suggest that the variables make and type may prove to be strong indicators of the price of a car. Similar plots box-plots were also created of the price of a car individually split by the binary variables sound, leather and cruise. Of these three plots, the plot of the price split by the cruise variable showed the most difference in price between a car that contained cruise control and one that did not. It was seen that cars that possessed cruise control appeared to be more expensive then those that did not. However, the presence or lack of leather seats or a sound system did not appear to show a significant different from each other. This would suggest that neither the presence of a sound system or leather seats in a car would significantly increase the price of such a car. This indicates that these variables may not be great indicators when modelling the price of a car.

[1] Later on in this project, when fitting a linear model to the data between the response variable price and the predictor variables, it was found that the linear model assumption of constant variance of errors had been broken. One of the primary methods to obtain constant variance is to transform the response variable, thus, equalizing the variance across all levels of the explanatory variables. A box-cox function was used to identify the optimal transformation that could be applied to avoid breaking the constant variance assumption. The box-cox plot identified a λ value of approximately 0, with this value identifying the log transformation as the optimal response variable transformation.

3.2 Fitting a Linear Regression Model

The above section explored the possibility of relationships between the price of a car, and the various car characteristics included in our data set. Henceforward, linear regression models will be fitted between the remaining explanatory variables mileage, make, type, liter, cruise, sound and leather and the response variable price with the hope of distinguishing the significance of each explanatory variable.

As mentioned above a linear regression model was fitted between the response variable price and all of the remaining explanatory variables. However, it was clear when examining the Scale-Location diagnostic plot that the assumption of constant variance across the residuals had been broken. The residuals also did not appear to be normally distributed when examining the QQ-plot. This can be seen in the diagnostic plots shown in Figure 5 below.

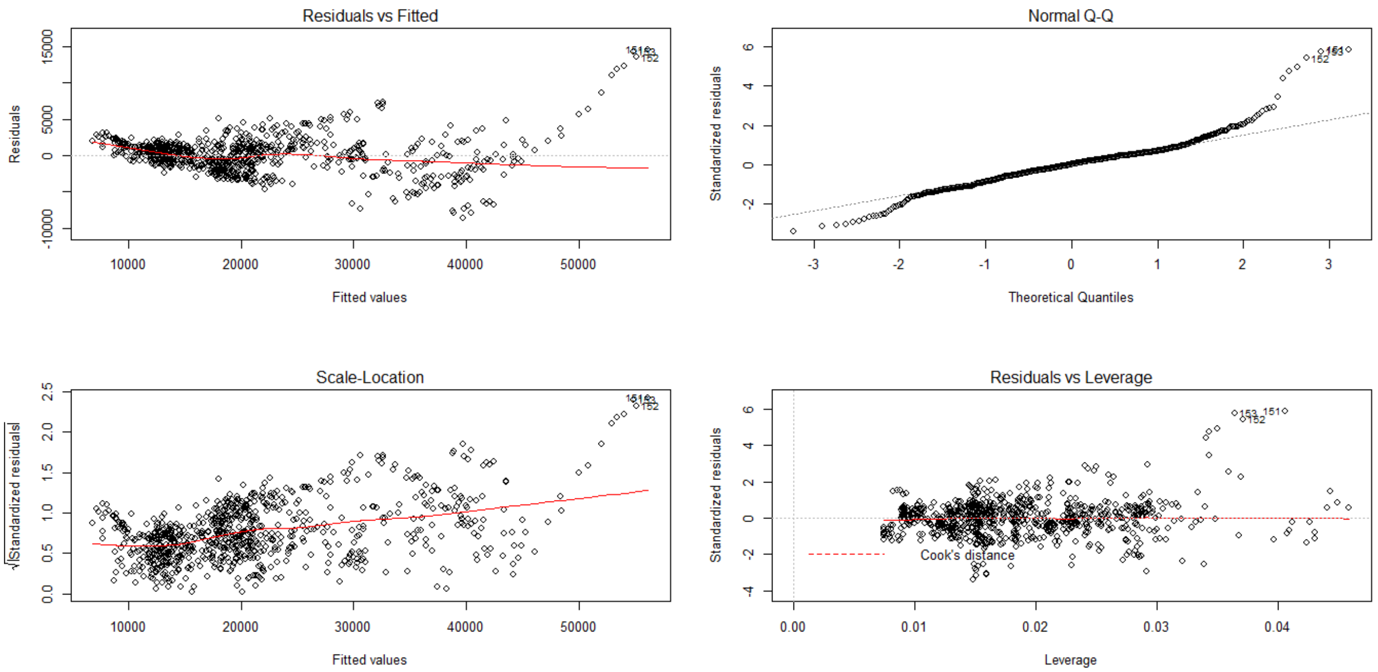

Figure 5: Diagnostic plots of the full model

Thus, a log transformation was used to transform the response variable price. A linear model was then fitted between the response log(price) and all of the remaining explanatory variables. Examining the diagnostic plots all of the linear regression assumptions appeared to be satisfied. For this model an adjusted R^2 value of 0.952 was obtained. Thus, indicating that over 95% of the variation present within the log of the response variable price is accounted for by the 7 explanatory variables.

It was felt that the above model contained too many variables, thus, the next step was to use a stepwise regression procedure to reduce the number of variables present in the model and to find a more parsimonious model. Running a bidirectional stepwise regression procedure using AIC as its information criterion, no variables were removed from the model. The full model remained, which can be seen below:

$$log(Price) \sim Make + Liter + Type + Mileage + Leather + Cruise + Sound$$

However, it was decided to use the stricter information criterion, BIC, to see if a different model would be achieved. Again, a bidirectional stepwise regression procedure was conducted using the Bayesian Information criterion. This time 3 of the variables were removed. These variables were leather, sound and cruise. The following model was obtained:

$$log(Price) \sim Mileage + Make + Type + Liter$$

Next, this models diagnostic plots were examined to ensure all of the linear regression assumptions were satisfied. The diagnostic plots of the final model are shown below in Figure 6.

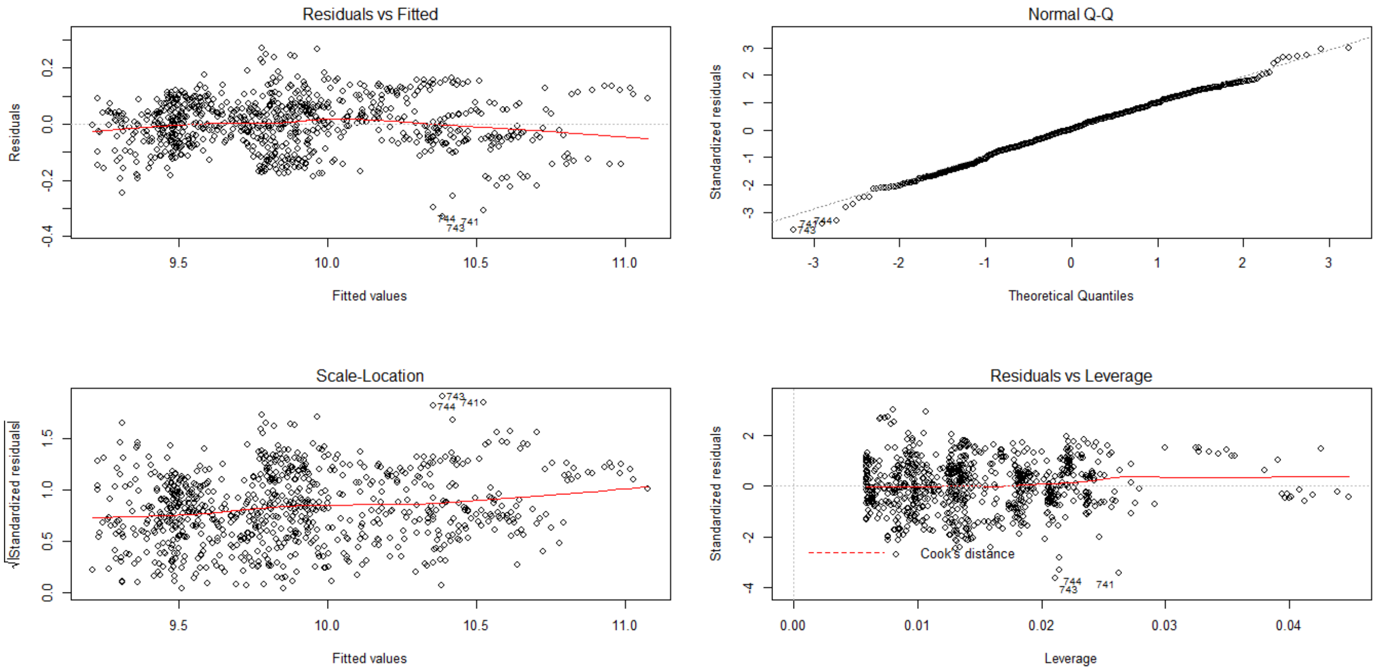

Figure 6: Diagnostic Plots of the Final Model

It is clear from the QQ-plot that the residuals are normally distributed. The plot of the residuals against the fitted values shows how the residuals are evenly scattered above and below 0 with the straight line of best fit indicating how nearly no structure is present in the residuals as all structure has been captured within the model. The Scale-Location plot which indicates the presence or lack of constant variance among the residuals, clearly shows no change in the variance across all the residuals.

The above model obtained an adjusted R^2 value of 0.9514, indicating how the 4 explanatory variables explained over 95% of the variation present in the log of the price response variable. There is only a slight difference in the adjusted R^2 values of the full and reduced models. This would indicate how the variables sound, leather and cruise added little or no additional information than the variables already included in the reduced model. All of the remaining variables were statistically significant at the 5% level of significance. The final models equation is given as follows:

$$\begin{aligned} log(Price) = 9.610 - 0.000008x_{Mileage} + 0.446x_{MakeCadillac} - 0.125x_{MakeChevrolet} - 0.098x_{MakePontiac} \\ + 0.560x_{MakeSAAB} - 0.105x_{MakeSaturn} - 0.312x_{TypeCoupe} - 0.368x_{TypeHatchback} - 0.325x_{TypeSedan} \\ - 0.165x_{TypeWagon} + 0.223x_{Liter} \end{aligned} \tag{2}$$

3.3 Model Interpretation

As this is the final model, the model parameters and coefficients will be fully explained and interpreted below. It is worth noting that for a simple non-transformed linear model, ordinary least squares regression estimates the expected arithmetic mean of the response variable y. However, when the response variable is log transformed in a linear model, ordinary least squares regression estimates the expected geometric mean of the response variable y.

- **Intercept:** The intercept coefficient can be interpreted as the price of a Buick Convertible with 0 mileage and a 0 liter capacity engine. The geometric average cost of such a car is \$14907.21.

- **Mileage:** A one unit increase in the mileage variable decreases the geometric average price by 0.000823%, when all the other variables remain constant. It is also worth noting that a 10000 unit increase in mileage decreases the geometric mean price by 7.89%, keeping all other variables constant.
- **Make:** When all the other variables are held constant and the car make is either a Cadillac, a Chevrolet, a Pontiac, a SAAB or a Saturn; then the geometric average price would increase by 56.2%, decrease by 11.7%, decrease by 9.31%, increase by 75.08% and decrease by 9.97%, respectively.
- **Type:** When all the other variables are held constant and the model of the car is either a Coupé, a Hatchback, a Sedan or a Wagon then the geometric average price would decrease by 26.82%, 30.81%, 27.71% or 15.2%, respectively.
- **Liter:** Finally, a one unit increase in the liter variable causes the geometric average price of a car to increase by 24.98%, provided all other variables are held constant.

3.4 Further Discussion

Although the final model above obtained an adjusted R^2 value of 0.9514 it would be interesting to see if there were any other explanatory variables that might further improve this accuracy. Some possible variables that might prove to be a good indicator of price may include:

- **Fuel Type:** The fuel type of the car i.e. is a car a petrol or diesel car.
- **Manual:** Is the car a manual or automatic car.

4 Conclusion

The main aim of this investigation was to find the explanatory variables which best modelled and explained the price response variable. By fitting a linear regression model to the data it was possible to conduct analysis and determine that mileage, make, type and liter were the best variables for explaining the log of car price response variable. The final model satisfies the linear regression assumptions while also containing a small number of variables. Thus, from this investigation it was found that when modelling the log of price of a car, that the engine size of the car as well as its make, type and mileage are key indicators of the final price of such a car.