**Price Prediction of Automobiles Using Multiple Linear Regression**

Armin Kavian  
Simon Fraser University

**Abstract**

In this report, we develop a multiple linear regression model to predict automobile prices using the core characteristics of the car such as its model, KM’s driven, mileage, engine power, etc. We establish what the most important factors in price prediction are and how they effect the price of the automobile.

1. **Introduction and Motivation**

Historically buying or selling an automobile has been done mostly through dealerships or friends. However, over the past 20 years, the Internet has become an especially important and powerful tool for buyers and sellers of automobiles. While this new market is highly effective at putting buyers and sellers in touch and increasing the accessibility of options for both sides, it could also be a very confusing market as this increase in options, leads to a mispricing of automobiles. This is the motivation behind this report as now more than ever, we need a model that can give an accurate price so that the buyers and sellers can have a better understanding of what their automobile is really worth.

This report is organized as follows. Section 2 presents how our data was acquired and methods used to clean this data. The visualization of our data has been moved to our methodology section to improve the flow of the report. Our methods in evaluating the best model as well as visualizations of the dataset is presented in section 3. Section 4 discusses our results and accuracy of the model. Finally, some concluding remarks and ways of improving the model are provided in our Conclusion section.

1. **Data Description**

Our data was the 9th data set from the LIONBRIDGE datasets for linear regression. The main difficulty with our data set was the missing values. Only 1 of the 3 files contained values, therefore we had to fill the other 2 data sets based on a few assumptions. The following are the assumptions made in completing this data set.

1. We matched the car models from the complete data set with the incomplete sets and assumed that cars with the same model and year will have the same core characteristics such as engine size, mileage, power, etc.
2. We matched the car models and years from the complete data set and the missing data sets. Then we would find the cars with the closest prices and assume they had the same number of previous owners. For example, if a BMW series 3 from year 2014 was worth $20,000 with 2 previous owners (this is from the complete data set), then we would find a BMW series 3 from year 2014 with a price as close to $20,000 as possible from the missing data sets and assign 2 previous owners to it.

These assumptions allowed us to evaluate our model, but they also decreased the accuracy of our model since the filled data sets are not actually observed but filled based on assumptions.

We also had to make certain adjustments to out data sets. The following are the adjustments made.

1. We removed any cars with a selling price less than $450. There were a few hundred observations with prices as low as $1 and we assumed that these were mistakes and not outliers.
2. In the mileage section, some of the units were KM/Liters whereas the others were KM/Kg, we adjusted the units so that all observations were Km/Liters.
3. The torque section included 3 different types of observations. Newton meters (Nm) @ x rpm, Newton meters (Nm) @ x-y rpm (a range of rpm instead of a single rpm), Kilogram @ x rpm. We adjusted the units so that everything was in Newton meters (Nm) @ x rpm (a single rpm). As for the Kg @ rpm, we simply multiplied the value of the kg parameter by 9.8 to change it to Newton meters. Lastly, as for the rpm’s that were a range instead of a single value, we simply took an average of the range.

Additional to the removal of observations with too low of a price, we also had to remove an observation with Fuel type Electric. This was due to this observation being the only electric car and that car model also being a one and only. It would not then make sense to include it in the model because then it is unclear where the price of that car is coming from, is it from it being electric or it is actually due to the model of the car.

We also divided our dataset into two sections. First section including 60% of our data set (observations picked at random) was used to develop our model, and later the remaining 40% was used to evaluate our prediction error and evaluate the accuracy of our model. However this lead to some issues in predictions later on because we only had a few observations for each car model and we would end up with many car models not being in the 60% which meant we could not make a prediction for the 40%. Therefore, we went back to using our entire dataset instead.

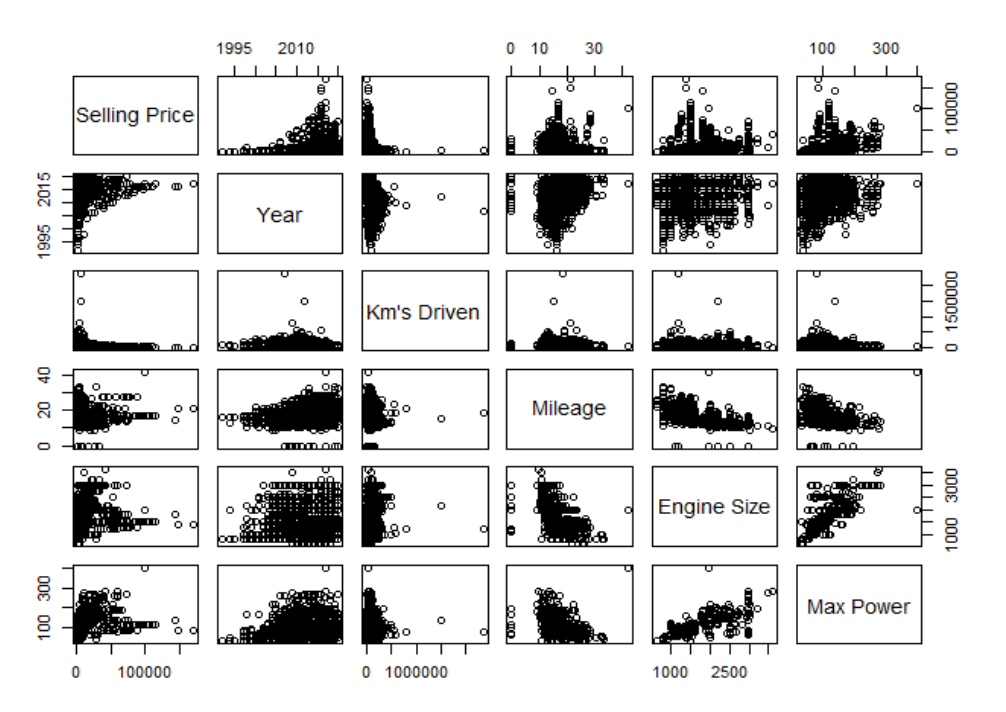
Lastly, we made a final adjustment to the data set by sorting all observations by the variable “Car Name” in alphabetical order.

**Additional observation:**

We decided to add a BMW x4, 2019, with the following characteristics : an unusually large Km’s driven, unexpectedly low selling price, Diesel fuel, first owner, manual. The remaining parameters were matched to a real observation. We did this to see how an extreme observation would affect our model. We will further discuss the effects of this observation in our results section.

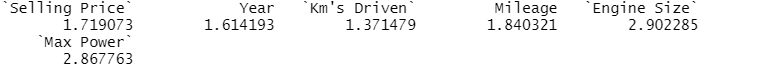
1. **General Methodology**

First, we looked at the behaviour of our non categorical variables and the correlation between these variables. figure 1 represents the pairs plot between the following variables: Selling Price, Year, KMs Driven, Mileage, Engine Size, Power. We did not include the torque as it is the same variable as Power. (Power = torque \* rpm / 5252). There are two notable mentions in this figure. First, there seems to be a strong positive correlation between the variable Engine size and Max power which is to be expected. The second, we expected a positive correlation between engine size and mileage however there is no evidence of such correlation according to figure 1.

Figure - Pairs Plot of the Non-Categorical Variables

We also calculated the VIF values for our non categorical variables to get a better understanding of multicollinearity between our variables. Figure 2 represents the VIF values for these variables. The VIF values in Figure 2 do not seem to support a strong multicollinearity.

Figure - VIF values for Non-Categorical Values



Since figure 2 does not support strong multicollinearity we decided to keep the max power variable in our model and later decide through AIC if it should be removed or not.

Next, we evaluated our categorical values.

1. For our Car Model Variable, we decided to have the base be Ambassador CLASSIC as it was different than Ambassador Classic but had a similar name, and we wanted to avoid confusion.
2. For the Owners variable, we set the base variable as “Fourth and Above”
3. For the Transmission variable, we set the base variable as “Manual”
4. For the Seller variable, we set the base variable as “individual”
5. For the Fuel variable, we set the base variable as “Petrol”

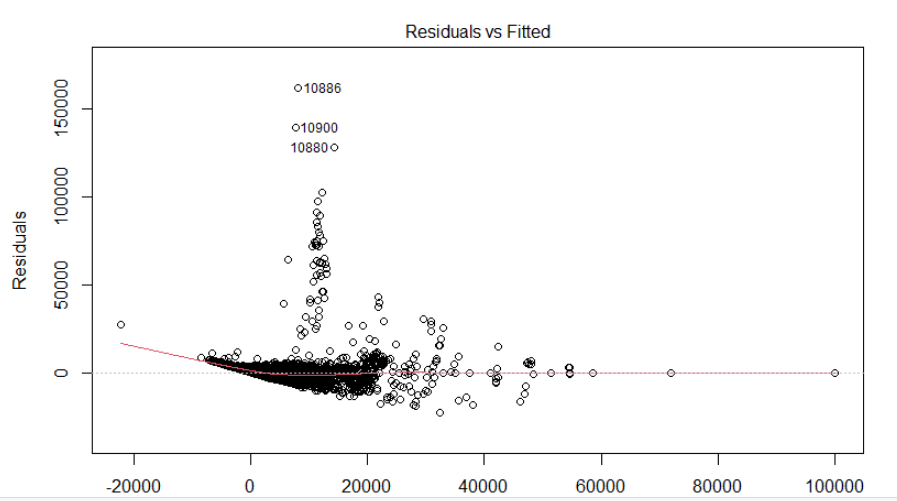
Figure 3-A is the residuals vs fitted values plot for a full linear model including all mentioned variables. It is clearly seen that the residuals are spreading in a fan shaped manner. Figure 3-B is the normal QQ plot for a linear model including all mentioned variables. The plot suggests heavy tails as well as right skewness.

Figure -A Residuals Vs Fitted Values Plot. Visible fan shaped increase in residuals

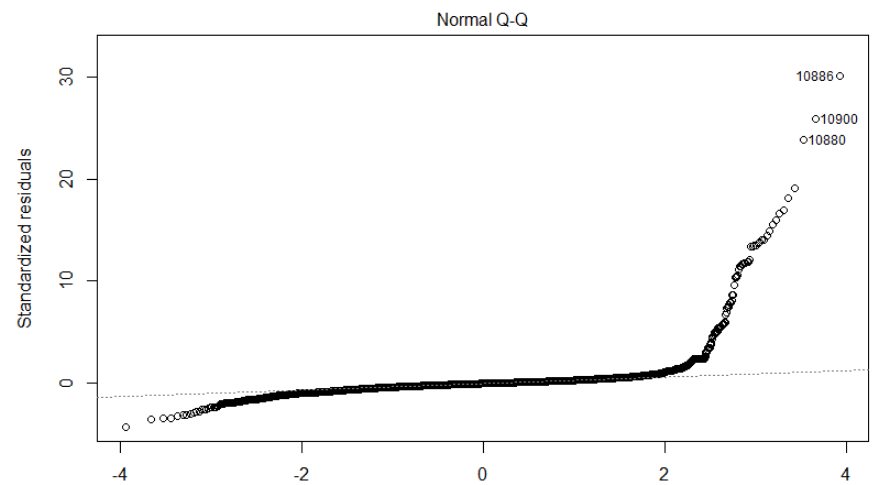


Figure 3-B Normal Q-Q Plot

In order to fix the fan shaped increase in the residuals, we re-evaluated our model using the Log of the selling prices. Figure 4-A shows a clear improvement and we no longer see a clear pattern in residuals. Figure 4-B is the upgraded Normal Q-Q plot. It shows a clear improvement in our normality assumption, however the fat tails and right skewness is still visible.

As clearly seen in Figures 4-A and 4-B, there seem to be certain values that have remarkably high residuals and are also far off in tails causing the skewness. In order to better understand these observations, we performed a leverage analysis through partial regression analysis as well as looking at extreme values on the diagonal of our hat matrix.

Unfortunately, even though this analysis gives us a good idea of the influential points in our data set, we cannot detect any points as an outlier. This is due to the nature of our data sets where just by changing the color of a car or adding a leather or ABS option the price of the car drastically changes.

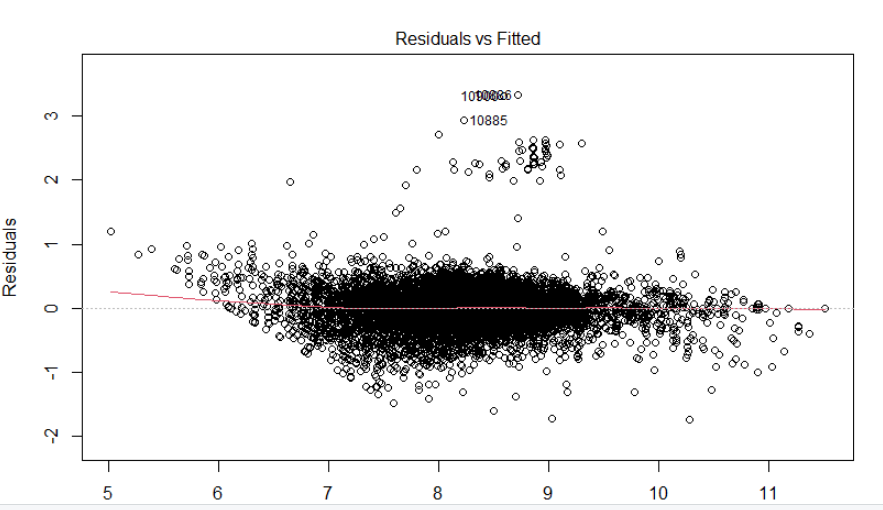


Figure -A Residuals vs Fitted Values calculated with Log(selling Prices) Model

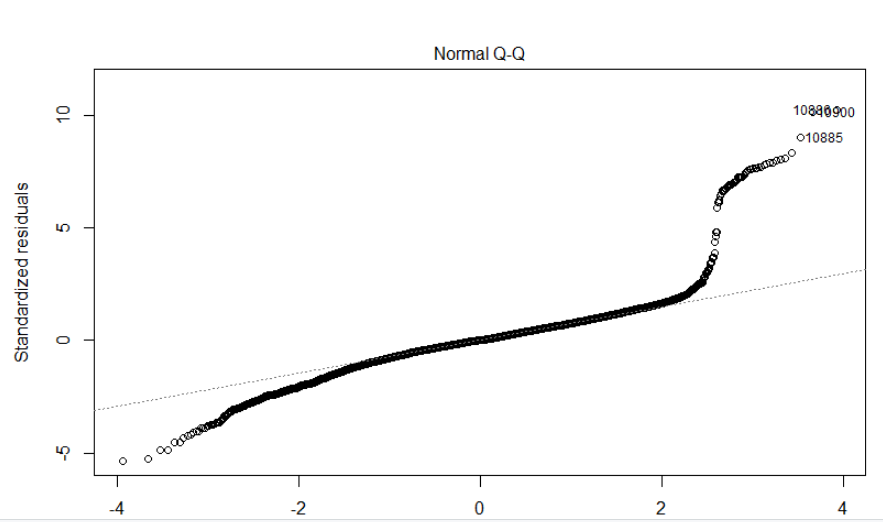
Since we do not have access to these additional variables in our data sets, we will have to include all these observations in our analysis.

Figure -B Normal Q-Q plot for Log(Selling Prices) Model

Lastly, we will develop the best model by AIC in a stepwise algorithm and evaluate the most important parameters through standardized inputs. In the results section we will discuss and interpret the model itself.

1. **Results**

In this section we will discuss the parameters, their importance and why certain parameters are omitted. The summary of the actual model (without the added variable) as well as the summary of the model including the added variable are both available at Appendix1.

* 1. **The Missing Variables and the intuition behind it**

The non-categorical variables present in the final model are: Year of the car, Log(Kms’ Driven), Car Mileage, and Max Power. We notice that the number of seating is not present in the model. This was somewhat expected as the number of seats is mostly constant for any car model, and we can expect this to be a constant included with our categorical variable “Car Model”.

The Categorical Variables present in the final model are: Type of the Car (The brand and the model), Ownership of the car (Only for test Drive cars, First owner, And Second Owner), Fuel type ( Only Diesel). We notice that the following categorical variables such as Third Owner, Fuel Type CNG and LPG, as well as the Seller Type are missing. Third owner was statistically not significant than the “Fourth Owners and Above” which was our base for Number of owners’ parameter, which explains why it is not in the model. Similarly, Fuel types CNG and LPG are not significantly different than the petrol which explains why they are not in the model either. Lastly the Seller parameter is entirely missing. This was due this parameter being a perfect linear combination of the Owners parameter.

* 1. **Adjusted R Squared and accuracy of the model**

With a model as large as ours with over 200 parameters one would expect a rather low adjusted R squared. However, our model has an extremely high adjusted R squared value of 0.899. I believe it is possible to further improve this model by adding certain options such as color, ABS, leather, and other luxuries. Although that model should still be tested as the introduction of these new parameters might not add enough to our adjusted R Squared to justify the extra parameters.

**4.3 the Additional Observation**

As it turns out, we made a remarkably interesting choice with our additional observation. We picked a BMW X3, a car model that was originally a petrol base car associated with a high selling price and changed it to a low selling price, diesel-based car. This added observation also happened to be a high leverage point due to its remarkably high Km’s driven. The most notable changes to the model were to the coefficients for Km’s Driven as well as the baseline selling price for a BMW X3 model. Since the selling price of our added observation was much lower than the average BMW X3 we expected a decrease in the coefficient for this variable however this coefficient increased from 0.77 to 1.1. This change will make more sense after we explain the changes to the Km’s driven coefficient. The most dramatic change was done to the Km’s driven coefficient as it went from -5.3 e-2 to -6.7 e-7. This does make sense as we used the log of Km’s driven to develop our model and since our added variables Km’s driven was an order of magnitude larger than any other observations, the model took away from the importance of this variable and adjusted the baselines instead. This baseline adjustment is also present in all the other car models however they seem to be rather small.

Lastly, not to our surprise, the addition of this observation slightly lowered our adjusted R squared value to 0.898 from 0.899. this change is again rather small and that is mostly due to the large number of observations in our data set which stops one bad observation from destroying our model.

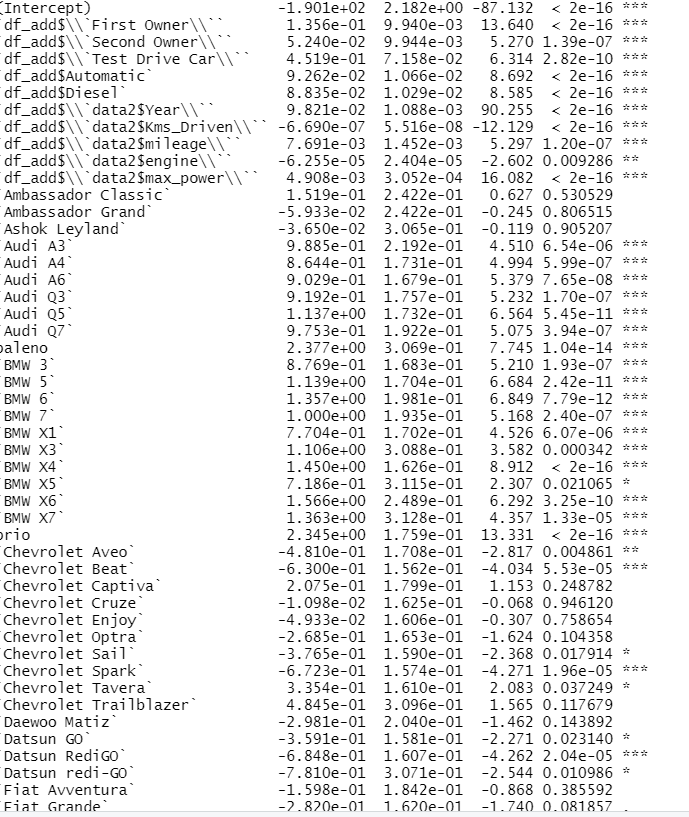
Based on the changes mentioned above we can conclude that our added observation was an influential point since including it in the model drastically changes the slope of our parameters.

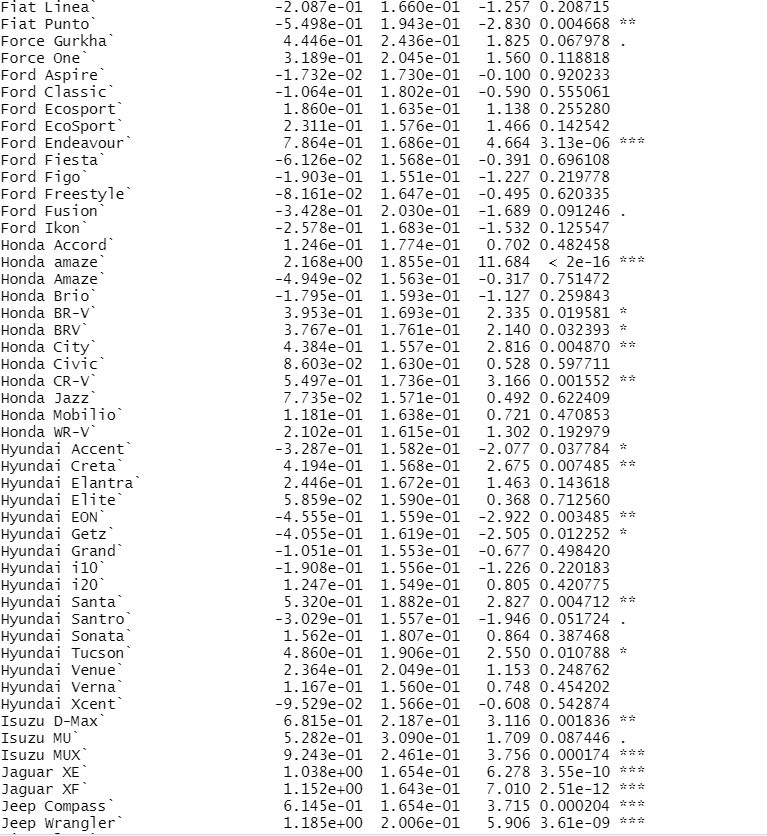
1. **Concluding Remarks**

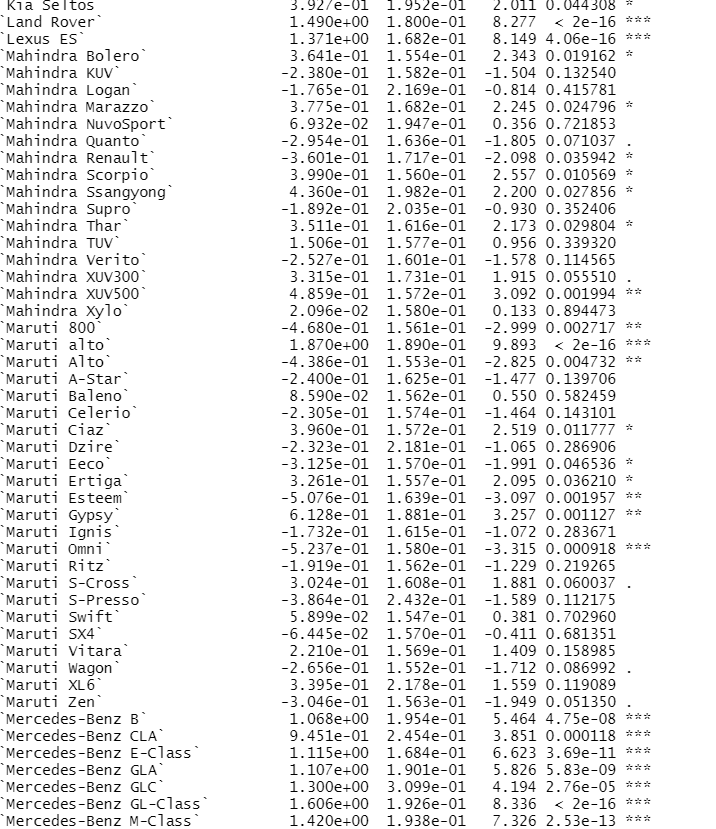
In this report we developed a model to predict prices of cars based on the model and core characteristics of these cars. We were unable to generate a prediction error due to the nature of our data in which we only had limited data for each car model, often 1-3 which we needed to use to develop the model, and we were left with no extra observations to test our model.

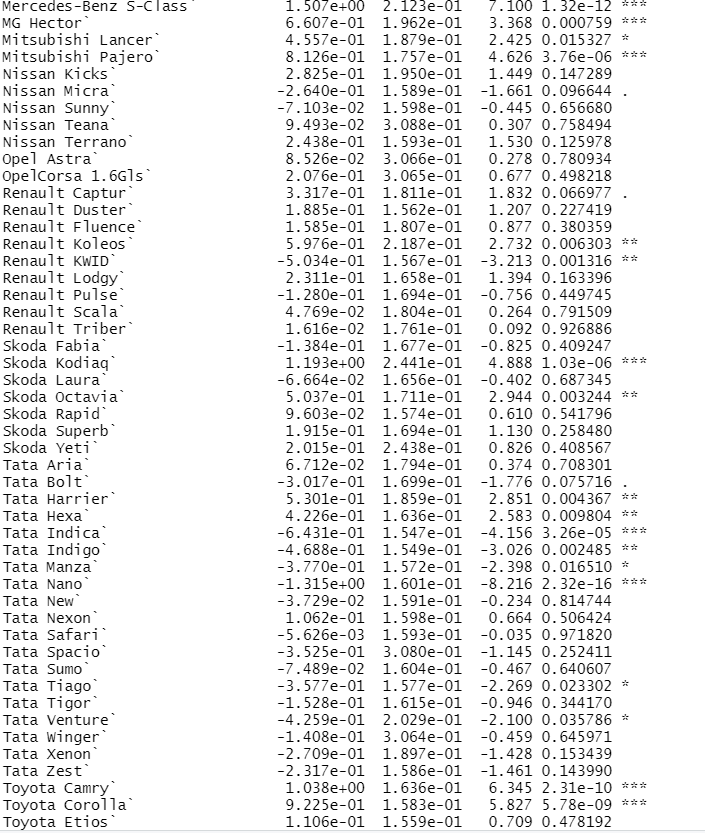
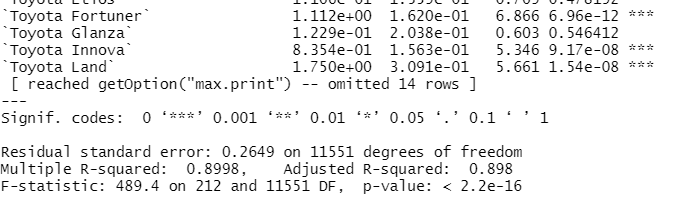
Lastly, we believe this model could be further improved on in 2 ways. First, by having more observations per car model, but also more accurate information since as mentioned in our data section, we had to fill many missing variables which hurt our model’s accuracy. Second, by introducing new luxury variables such as color of the car, leather option, etc. once such data is available this should be the topic of future reports.

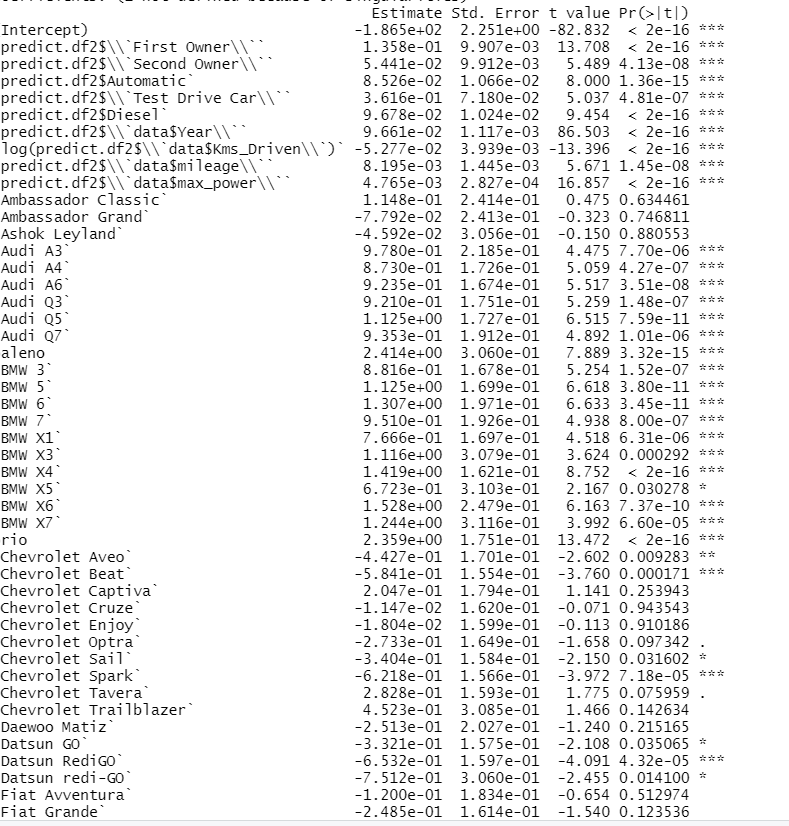
Appendix1.

Model with the additional observation







Model Without the Additional Observation

