**Stat 3340 Final Project-Group 38**

By Mackenzie Frolic-Smart, Samson Mifflin, and Dawei Yuan

**Abstract**

In this report we analyse a dataset of the specifications and selling prices of used cars in order to determine what factors play in to a used car’s selling price. Before we could analyse the data we had to clean it and transform it into usable metrics. Then, using multiple linear regression, residual analysis, stepwise comparison, and cross validation we created a model that not only showed us the most important aspects of a used car, but also gave us the power to predict the price of future used cars, based on their specifications. We found that a car’s age and transmission were the largest factors that contributed to a lowering of a car’s price. On the other had, a car being a test drive car and its power and engine were important specifications in raising a car’s price.

**Introduction**

Our group was assigned the vehicle data set posted on Lionbridge. This data set is a list of second-hand cars, along with their selling price, year, kilometers driven, and various other specifications. Our group decided that the most relevant analysis to do on this dataset would be to investigate how the selling price of a used car is determined. To do this we decided to apply linear regression to the data set, attempting to see if the various specifications listed in the data set could be used to predict the selling price of the car. An in depth regression analysis of this data will not only be able to tell us how a cars various features (such as number of seats, engine, max power, kilometers driven, etc.) impact its values on the secondary market, but also which of these features are even relevant at all. Cars are such complex machines with so many intricacies that it is unlikely a potential buyer will take them all into account when considering a new car, we hypothesise that there are only a few key features that a buyer will look at when considering a purchase. Linear regression is a perfect tool to identify what these important features are.

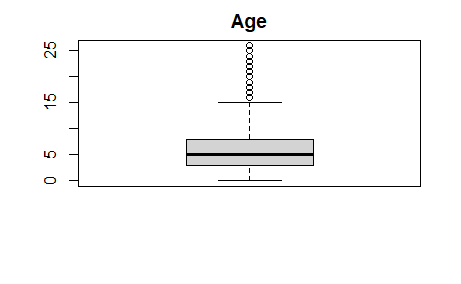
**Description of the Data**

Before analysing the data our group took several steps to make sure it was usable and coherent. The first thing we did was introduce the new data point. Given that the data set was about cars it made sense to choose a real car and use those specifications to fill in the table. The car we chose was the 2020 Kia Soul LX 4dr Wagon. This car was randomly selected from a google search returning the most popular cars of 2020. The full data entry for the 2020 Kia Soul LX 4dr Wagon is below:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| name | year | Sell price | Km driven | fuel | Seller type | transmission | owner | mileage | engine | Max power | Torque | seats |
| Kia Soul LX 4dr Wagon | 2020 | 9495 $ | 500000 | petrol | Dealer | automatic | First owner | 11.48 kmpl | 1,999 cc | 147 bhp | 132 lb.-ft. @ 4,500 rpm | 5 |

The search criteria of popular cars from 2020 was selected because our group decided that we wanted our data point to be something of an outlier, rather than simply having a data point that blended in with the others. We felt that the introduction of an outlier might make our analysis more interesting and make it stand out from the others. Using manufacturer information, we determined the name, year, fuel, transmission, mileage, engine, max power, and seats for the Kia Soul LX 4dr Wagon. In keeping with the goal of creating an outlier point a relatively low selling price was selected (approximately half of the manufacturer MSRP) and a large number of kilometers driven was selected just to make the data stranger. Seller type and owner were arbitrarily selected from the possible values in the data set.

With the new data point integrated into our data sheet we began preliminary analysis of the data. The first thing we did was look at the summaries and create box plots of all our numeric variables, in order get a sense of the distribution of the data, as well as check for any outliers.



|  |  |  |  |
| --- | --- | --- | --- |
| Min | Mean | Max | SD |
| 0 | 5 | 26 | 3.865 |

Chart, box and whisker chart

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Min | Mean | Max | SD |
| 9 | 19.43 | 42 | 3.902 |

Table

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Min | Mean | Max | SD |
| 4 | 5 | 14 | 0.962 |

From the box plots, the car age, number of seats, and mileage variables seem to be normal for most part with outliers at one end of the distribution Chart, diagram, box and whisker chart

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Min | Mean | Max | SD |
| 624 | 1248 | 3604 | 504.061 |

The engine variable has more significant outliers but again nothing too worrisome.

Diagram

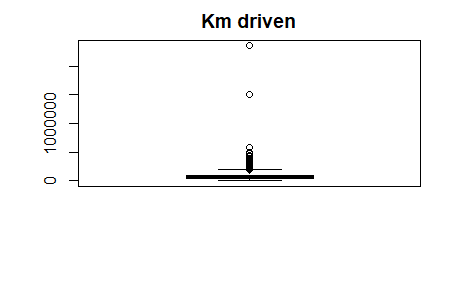
Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Min | Mean | Max | SD |
| 32.80 | 91.59 | 400 | 35.7732 |

Diagram

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Min | Mean | Max | SD |
| 9495 | 695000 | 10000000 | 817003.1 |



|  |  |  |  |
| --- | --- | --- | --- |
| Min | Mean | Max | SD |
| 1000 | 69181 | 2360457 | 357078.54 |

On the other hand, selling price and kilometers driven, and max power all had extreme outliers. To help offset these outliers we decided to run our regression using the logarithm of selling price, the logarithm of km driven, and the logarithm of max power.

Diagram, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

The box plots after the log transformation look much more reasonable than before as the transformation blunted the effect of the extreme outliers.

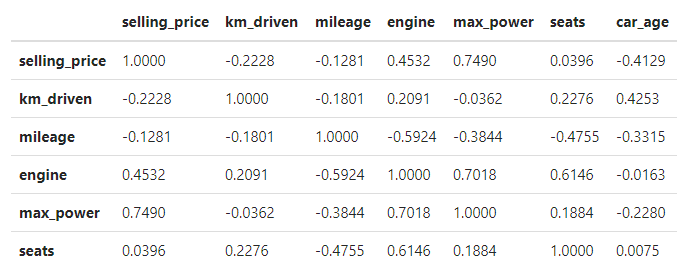
To get a sense of our data we also ran a pairs command comparing all the numeric variables to each other.

Diagram

Description automatically generated

Judging from these pairs plots selling price and both engine and max power seem to be positively linearly related, and selling price and car age appears to be negatively linearly related. Additionally, max power and engine appear to be linearly related to each other, indicating the possibility that they are collinear, something that we will have to investigate further.

The following table gives the pairwise correlations of the numerical variables.



In addition to the creating of a new data point our group also had to clean the data and make it usable. The first step was to drop the names category. The name variable is categorical, and we felt it would be too arduous to create indicator variables for every different car name, so we dropped the entire column. Similarly, we decided to drop the torque data column. The torque data did not have a consistent unit of measurement, and even if we were to select one unit we were unsure if we could convert one measurement of torque into another. Not only did some entries in the torque column have different units of measurement, some also were given as a range of values (for example, “*24@ 1900 – 2750 (kgm @pm)”*)

Next we considered the fuel column. The vast majority of the cars in this data set used either diesel or petrol fuel, however a few used LPG (liquid propane gas) or CNG (compressed natural gas) fuel. Of the 8129 cars in the original data set 38 used LPG and 57 used CNG, together representing approximately 0.01% of the data. Since there we so few data points for either of these categories we felt we would be unable to construct a useful regression model with these categories included, so we decided to drop all cars using CPG or LPG from the data set.

We also decided to transform the year data column. The year column represents the age of the vehicle being sold, however that is a little hard to understand if the column is just kept as the year the car was made, so instead we decided to subtract the year the car was made from 2020 thereby changing the year column into the age column. The data becomes more meaningful this way.

Some of observation rows had mileage recorded as 0. If this were true, this means the car represented by this observation runs 0 kmpl. This shows an error in recording the observation, so we deleted the rows from our data frame. We also deleted the rows that had missing entries.

We also noticed that the numerical variables are on different scales so we decided to do unit normal scaling on them. This also gives better meaning to the intercept term of the regression model. Without unit normal scaling, the intercept term will give the logarithm of the selling price of a car that has age 0, with 0 log(km driven), 0 kmpl mileage, 0 seats, 0 max power, and 0 CC engine. Doing unit normal scaling allows us to interpret the intercept term as the expected value of the logarithm of the selling price when the predictor values are set to the average values from our data set. Each regression coefficient will then give the predicted change on a standard deviation scale of the logarithm of the selling price associated with a change of one standard deviation of the predictor with everything else held constant.

After cleaning the data and paring it down we were left with our dependent variable (selling price), and ten regressors. Of those ten, four were categorical and the rest quantitative. Of the categorical variables fuel had 2 categories (Petrol and Diesel), seller type had 3 categories (Individual, Dealer, Trustmark Dealer), transmission had 2 categories (Automatic and Manual) and owner had 5 categories (test drive car, first, second, third, and ,fourth & above)

Once we cleaned our data, we were ready to start analysing it. We analysed our data using multiple linear regression stepwise comparison, and cross validation.

**Methodes**

Before doing any regression or performing stepwise comparison for variable selection, we set aside 20 percent of our data set to serve as test data to validate our model. We will be performing stepwise variable selection using our training data comprised of the other 80 percent.

Multiple linear regression was used to not only tell us what the most important factors are in determining the selling of a used car but to also construct a model which we could use to predict the selling price of future used cars taking their various specifications into consideration. To that end, once we had our regression model we then conducted step a stepwise comparison to prune our model into one that more parsimonious while still having all statistically significant regressors. We used the stepAIC function with the direction set to "both".

The model obtained from stepAIC variable selection has all the predictor variables as significant. It has an adjusted-r2 of 0.8594, a condition number of 48.6 (using the kappa function) and an AIC of 5422.7 (using the AIC() function). Using the vif function gives a value of more than 5 for the engine variable.

lm(formula = selling\_price ~ km\_driven + fuel + seller\_type + transmission + owner + mileage + engine + max\_power + seats + car\_age, data = train\_data)

Residuals: Min 1Q Median 3Q Max -6.1221 -0.2252 0.0324 0.2376 2.0135

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.501265 0.020241 24.765 < 2e-16 \*\*\*

km\_driven -0.077669 0.006492 -11.964 < 2e-16 \*\*\*

fuelPetrol -0.163294 0.014350 -11.379 < 2e-16 \*\*\*

seller\_typeIndividual -0.137375 0.015252 -9.007 < 2e-16 \*\*\* seller\_typeTrustmark Dealer -0.090267 0.031007 -2.911 0.00361 \*\* transmissionManual -0.323164 0.017679 -18.279 < 2e-16 \*\*\*

ownerFourth & Above Owner -0.145141 0.035645 -4.072 4.72e-05 \*\*\*

ownerSecond Owner -0.077729 0.012236 -6.352 2.27e-10 \*\*\*

ownerTest Drive Car 0.784113 0.167597 4.679 2.95e-06 \*\*\*

ownerThird Owner -0.140397 0.021330 -6.582 5.01e-11 \*\*\*

mileage 0.079659 0.008362 9.527 < 2e-16 \*\*\*

engine 0.196488 0.010799 18.195 < 2e-16 \*\*\*

max\_power 0.395094 0.008060 49.020 < 2e-16 \*\*\*

seats 0.028440 0.007132 3.987 6.75e-05 \*\*\*

car\_age -0.456304 0.007408 -61.595 < 2e-16 \*\*\*

--- Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3731 on 6228 degrees of freedom Multiple R-squared: 0.8597, Adjusted R-squared: 0.8594 F-statistic: 2727 on 14 and 6228 DF, p-value: < 2.2e-16

Here, we note that R automatically constructs the factors of a categorical variable in alphabetical order. So, if fuelPetrol=0 then the fuel type is diesel while fuelPetrol = 1 means fuel type is petrol. For the seller\_type, the base level is Dealer. This means if seller\_typeIndividual=0 then we have a Dealer while seller\_typeIndividual=1 means seller type is Individual. For the owner, the base level is First Owner. This means if ownerFourth & Above Owner=0 then we have a First Owner while ownerFourth & Above Ownerl=1 means an owner that is at least the fourth owner. The car is automatic if transmissionManual=0 .

Satisfied with our model, we then conducted a stepwise comparison of our regression model. The Stepwise comparison was used in order to prune our model down to only its relevant regressors, as well as make our model more parsimonious. Our stepwise regression returned the following VIFs:

|  |  |
| --- | --- |
| Variable | VIF |
| km\_driven | 1.914041 |
| seller\_typeTrustmark Dealer | 1.233285 |
| ownerSecond Owner | 1.286511 |
| mileage | 3.145929 |
| Seats | 2.271987 |
| fuelPetrol | 2.290148 |
| transmissionManual | 1.592361 |
| ownerTest Drive Car | 1.008186 |
| engine | 5.200269 |
| car\_age | 2.491215 |
| seller\_typeIndividual | 1.467964 |
| ownerFourth & Above Owner | 1.095945 |
| ownerThird Owner | 1.205094 |
| max\_power | 2.911937 |

Surprisingly, our stepwise comparison found that all regressors in our model contributed to it’s predictive capabilities and were worth keeping. Looking at the VIF’s of all our regressors the majority are satisfactorily small, indicating that they are not colinear with any other regressors. The notable exception to this is the engine variable. Engine has a VIF of 5.2, significantly higher than any other regressor, and hinting at the possibility of it being coliniear with another regressor. This is not surprising, as we saw the engine and max power were likely to be collinear in the pairs plots.

Regarding the engine variable with VIF of above 5, we did regression to fit a model without the engine variable. However, the model we obtained has an adjusted-r2 of 0.852 and has mileage variable and the seller\_typeTrustmark Dealer factor marked as not significant.

lm(formula = selling\_price ~ km\_driven + fuel + seller\_type + transmission + owner + mileage + max\_power + seats + car\_age, data = train\_data)

Residuals: Min 1Q Median 3Q Max

-6.1421 -0.2403 0.0275 0.2393 1.8113

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.595494 0.020079 29.658 < 2e-16 \*\*\*

km\_driven -0.068861 0.006643 -10.366 < 2e-16 \*\*\*

fuelPetrol -0.281021 0.013144 -21.381 < 2e-16 \*\*\*

seller\_typeIndividual -0.143453 0.015648 -9.168 < 2e-16 \*\*\* seller\_typeTrustmark Dealer -0.036336 0.031672 -1.147 0.2513 transmissionManual -0.363618 0.017997 -20.204 < 2e-16 \*\*\*

ownerFourth & Above Owner -0.162841 0.036564 -4.454 8.59e-06 \*\*\*

ownerSecond Owner -0.081105 0.012555 -6.460 1.13e-10 \*\*\*

ownerTest Drive Car 0.761109 0.171975 4.426 9.78e-06 \*\*\*

ownerThird Owner -0.153987 0.021874 -7.040 2.13e-12 \*\*\*

mileage 0.012854 0.007709 1.667 0.0955 .

max\_power 0.470488 0.007094 66.323 < 2e-16 \*\*\*

seats 0.080382 0.006707 11.985 < 2e-16 \*\*\*

car\_age -0.457155 0.007602 -60.138 < 2e-16 \*\*\*

--- Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3828 on 6229 degrees of freedom Multiple R-squared: 0.8523, Adjusted R-squared: 0.852 F-statistic: 2765 on 13 and 6229 DF, p-value: < 2.2e-16

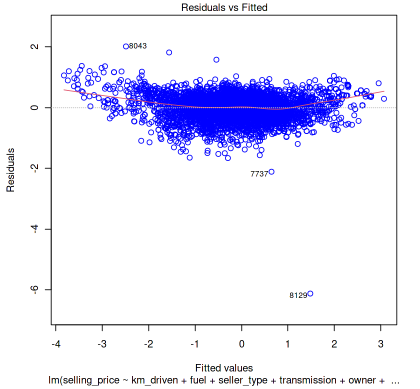
We compared these two models using anova and got a p-value of 4.0809e-72, indicating a significantly improved fit when the engine variable joins all the other variables in the model.

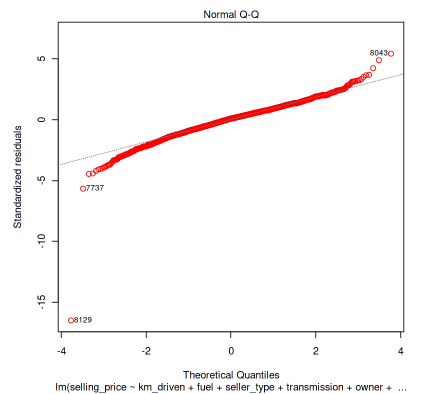
Res.Df RSS Df Sum ofSq F Pr(>F)

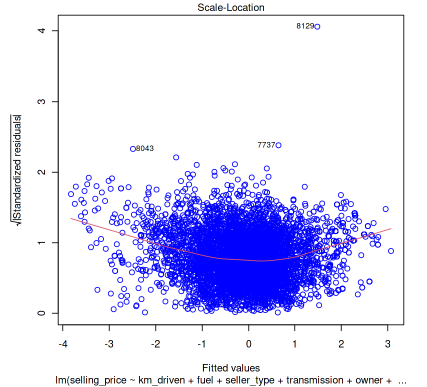
6229 912.8842 NA NA NA NA

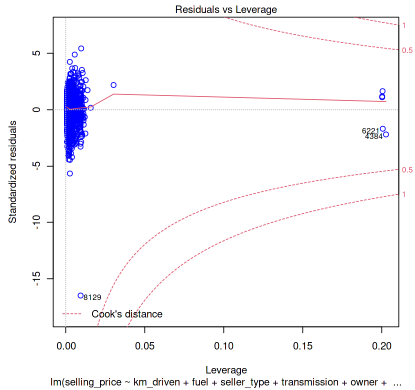
6228 866.8089 1 46.07533 331.0501 4.080973e-72

The following residual plots from the plot(lm) function show that observation 8129 in cardata is an outlier in all four of the plots.



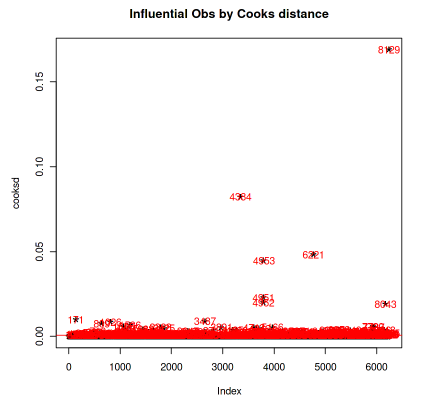






If we ignore the observation 8129 in the residual plots, the Residuals vs Fitted plot shows that the residuals are spread randomly around the 0 line and form an approximate horizontal band around the 0 line. This indicates a linear relationship and homogeneity of error variance. The Normal Q-Q plot shows that an approximately normal distribution with some heavy tails. The Scale-Location plot shows a random spread above and below and along the line. The Residuals vs Leverage plot shows observation 8129 near the red line that marks Cook’s distance.

We also plot Cook’s distance and see that observation 8129 is clearly an outlier. Its Cook’s distance is approximately 0.17 while the traditional cut-off value 4/n is 0.0006 and the average Cook’s distance is 0.0002.



We checked what is in observation 8129. It turns out that it is the synthetic data point we introduced, the data for the 2020 Kia Soul LX 4dr Wagon. We also checked other observations with large Cook’s distances and didn’t see anything wrong with the data contained in them.

We then re-did the regression without our synthetic data point and get a higher adjusted-r2 of 0.8651. This is expected since we removed an influential point with a poor fit. This model had all the predictor variables as statistically significant. We also obtained a condition number of 48.59 and an AIC of 5144.3. The VIF of the engine variable is 5.2 while the others are less than 4.

lm(formula = selling\_price ~ km\_driven + fuel + seller\_type + transmission + owner + mileage + engine + max\_power + seats + car\_age, data = train\_data, subset = 1:6242)

Residuals: Min 1Q Median 3Q Max

-2.11612 -0.22578 0.03298 0.23476 2.03263

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.513172 0.019808 25.907 < 2e-16 \*\*\*

km\_driven -0.069315 0.006368 -10.884 < 2e-16 \*\*\*

fuelPetrol -0.159324 0.014037 -11.350 < 2e-16 \*\*\*

seller\_typeIndividual -0.144405 0.014923 -9.677 < 2e-16 \*\*\* seller\_typeTrustmark Dealer -0.100311 0.030331 -3.307 0.000948 \*\*\* transmissionManual -0.330371 0.017296 -19.101 < 2e-16 \*\*\*

ownerFourth & Above Owner -0.143955 0.034861 -4.129 3.69e-05 \*\*\*

ownerSecond Owner -0.078757 0.011967 -6.581 5.05e-11 \*\*\*

ownerTest Drive Car 0.773642 0.163913 4.720 2.41e-06 \*\*\*

ownerThird Owner -0.140788 0.020861 -6.749 1.62e-11 \*\*\*

mileage 0.074565 0.008183 9.112 < 2e-16 \*\*\*

engine 0.195960 0.010562 18.554 < 2e-16 \*\*\*

max\_power 0.393450 0.007883 49.909 < 2e-16 \*\*\*

seats 0.025527 0.006978 3.658 0.000256 \*\*\*

car\_age -0.463307 0.007257 -63.841 < 2e-16 \*\*\*

--- Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3649 on 6227 degrees of freedom Multiple R-squared: 0.8654, Adjusted R-squared: 0.8651 F-statistic: 2860 on 14 and 6227 DF, p-value: < 2.2e-16

We also fit a model where observation 8129 and the engine variable are both removed. This model has an adjusted-r2 of 0.8577.

lm(formula = selling\_price ~ km\_driven + fuel + seller\_type + transmission + owner + mileage + max\_power + seats + car\_age, data = train\_data, subset = 1:6242)

Residuals: Min 1Q Median 3Q Max

-2.17568 -0.23996 0.02714 0.23829 1.83094

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.607185 0.019670 30.869 < 2e-16 \*\*\*

km\_driven -0.060503 0.006523 -9.275 < 2e-16 \*\*\*

fuelPetrol -0.276721 0.012870 -21.501 < 2e-16 \*\*\*

seller\_typeIndividual -0.150490 0.015325 -9.820 < 2e-16 \*\*\* seller\_typeTrustmark Dealer -0.046558 0.031013 -1.501 0.133 transmissionManual -0.370739 0.017625 -21.035 < 2e-16 \*\*\*

ownerFourth & Above Owner -0.161604 0.035796 -4.515 6.46e-06 \*\*\*

ownerSecond Owner -0.082126 0.012291 -6.682 2.57e-11 \*\*\*

ownerTest Drive Car 0.750666 0.168365 4.459 8.40e-06 \*\*\*

ownerThird Owner -0.154342 0.021415 -7.207 6.38e-13 \*\*\*

mileage 0.007924 0.007553 1.049 0.294

max\_power 0.468635 0.006946 67.470 < 2e-16 \*\*\*

seats 0.077320 0.006569 11.771 < 2e-16 \*\*\*

car\_age -0.464179 0.007454 -62.270 < 2e-16 \*\*\*

--- Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3748 on 6228 degrees of freedom Multiple R-squared: 0.858, Adjusted R-squared: 0.8577 F-statistic: 2894 on 13 and 6228 DF, p-value: < 2.2e-16

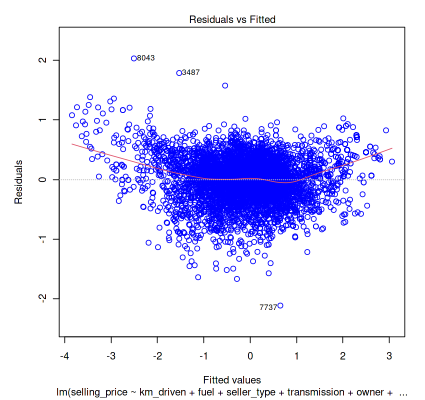
Comparing the two models without observation 8129 using anova, we obtained a p-value of 7.700e-75, indicating a significantly improved fit when the engine variable joins all the other variables in the model.

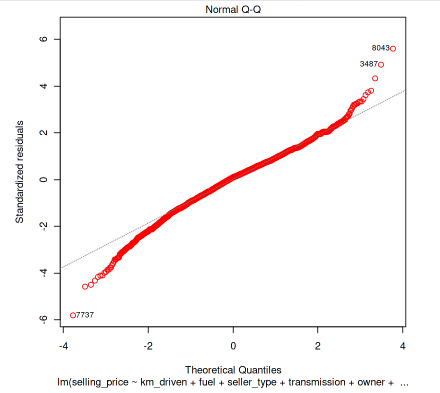
Res.Df RSS Df Sum ofSq F Pr(>F)

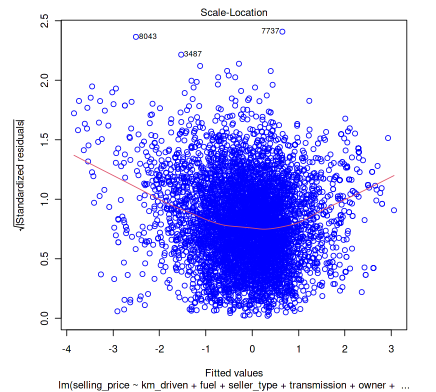
6228 874.8067 NA NA NA NA

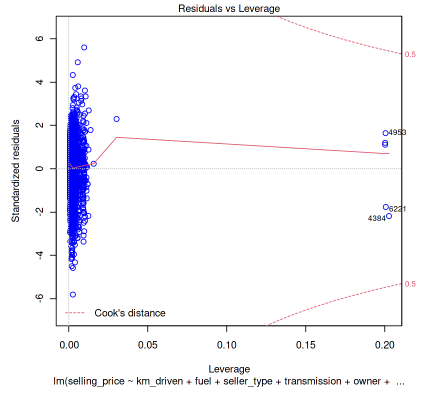
6227 828.9794 1 45.82731 344.2386 7.700291e-75

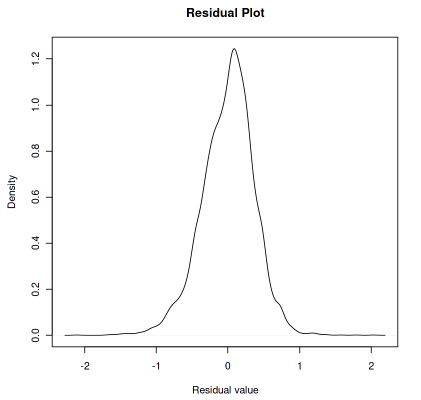
The following are the residual plots of the full model when observation 8129 is removed.











The Residuals vs Fitted plot gives a good indication that there are no nonlinear relationships since it seems like the residuals are randomly spread around the horizontal line. The Scale-Location plot also shows a somewhat random spread of points. The Q-Q plot suggests heavy tails in the distribution and the Density-Residual plot shows that a normal distribution is a good approximation for the distribution of the residuals.

Finally, we tested the predictive capabilities of our model using cross validation. We also decided to compare the RMSE of the no engine model to the full model, to again check if we should keep the engine variable in our data. First we cross validated the model with engine removed, comparing the models predicted values to the actual values present in the isolated training data. The cross validation returned an RMSE of 0.3768. This indicates that the no engine model is a fairly good predictor of the actual data. This can be seen when the values predicted by the model are graphed against those that were found in the training data.

Chart, scatter chart

Description automatically generated

The cross validation of our full model returned an RMSE of 0.3670255. While very similar, the RMSE of the full model is slightly smaller than that of the no engine model, indicating that the full model a more powerful predictor than the engine model. While this may seem like a pretty minor change in RMSE, it is apparent when you look at the graph of the full model.

Chart, scatter chart

Description automatically generated

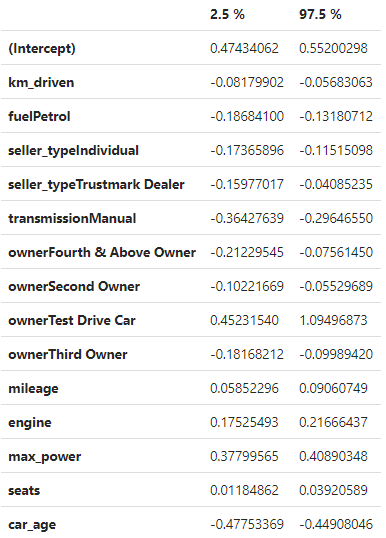
As you can see, the distribution of the points in the full model is just a little tighter than those of the no engine model. Because the full model had a lower RMSE than the no engine model we decided to keep engine in our model, despite its higher than usual VIF.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

We give the following 95% confidence interval for the coefficients of our full model:



**Results**

Our full linear regression model is as follows:

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 0.513172 0.019808 25.907 < 2e-16 \*\*\***

**km\_driven -0.069315 0.006368 -10.884 < 2e-16 \*\*\***

**fuelPetrol -0.159324 0.014037 -11.350 < 2e-16 \*\*\***

**seller\_typeIndividual -0.144405 0.014923 -9.677 < 2e-16 \*\*\***

**seller\_typeTrustmark Dealer -0.100311 0.030331 -3.307 0.000948 \*\*\***

**transmissionManual -0.330371 0.017296 -19.101 < 2e-16 \*\*\***

**ownerFourth & Above Owner -0.143955 0.034861 -4.129 3.69e-05 \*\*\***

**ownerSecond Owner -0.078757 0.011967 -6.581 5.05e-11 \*\*\***

**ownerTest Drive Car 0.773642 0.163913 4.720 2.41e-06 \*\*\***

**ownerThird Owner -0.140788 0.020861 -6.749 1.62e-11 \*\*\***

**mileage 0.074565 0.008183 9.112 < 2e-16 \*\*\***

**engine 0.195960 0.010562 18.554 < 2e-16 \*\*\***

**max\_power 0.393450 0.007883 49.909 < 2e-16 \*\*\***

**seats 0.025527 0.006978 3.658 0.000256 \*\*\***

**car\_age -0.463307 0.007257 -63.841 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1**

**Residual standard error: 0.3649 on 6227 degrees of freedom**

**Multiple R-squared: 0.8654, Adjusted R-squared: 0.8651**

**F-statistic: 2860 on 14 and 6227 DF, p-value: < 2.2e-16**

The regression coefficients of the full model indicate what the most important aspects of a used car are to potential buyers. This suggests that the factor that increases the price of a used car the most is whether or not the car was a test drive car, as owner test drive car has the largest positive regression coefficient of 0.773642 . The next two variables with the largest positive regression coefficients the maximum power of the car, with a regression coefficient of 0.393450 and the displacement of the car’s engine, with a regression coefficient of 0.195960 . Inversely, a car’s age seemed to be the largest thing detracting from its price, having the largest negative regression coefficient of -0.463307. The next two variables with the largest negative regression coefficients were transmission manual with -0.330371, and fuel petrol with -0.159324.

These findings are mostly unsurprising. It makes sense that being a test drive car would make a second hand car more valuable because while the test drive car is still used it is likely to still be in excellent condition, as it was previously used and maintained by a car dealership. Max power and engine displacement being the second and third most important factors to the increasing a used car’s selling price is a little surprising, but it also makes sense as they are both an indicator of how powerful and useful a car is.

Similarly, the most prominent negative regression coefficients are also mostly unsurprising. Age being the largest negative regressors is logical, as the more the older the car is the less likely it is to still be in good condition, and les less time it is likely to be usable. It is also understandable than having a manual transmission dampens the price of a used car because not everyone can drive a manual transmission. Petrol fuel being the third most negative regressor is a little surprising, but a negative regressor for petrol means that cars with petrol engines are less valuable than those with diesel engines, which makes sense.

It is also interesting to note that the outlier that was identified in our residual analysis was actually our added data point. This is unsurprising, as we purposefully tried to create a datapoint that was an outlier so as to set our data apart from the data found on Lionsgate.

To test our model’s predictive powers we used cross validation to compare our model’s predicted answers for data points to the actual points in our training data. The cross-validation calculation returned an RMSE of 0.2995806, indicating that our model has fairly good predictive properties, on average only being off of the true selling price by 0.299.

The intercept value of 0.513172 means that when the factors and scaled variables are 0, the selling price is 0.513172 standard deviations above the arithmetic mean of log(selling\_price). This means the natural logarithm of the selling price is 13.43040219147 and it corresponds to a selling price of 680376.90. By interpreting the 0 levels of the factors and scaled variables, a used diesel automatic five-seater first-owned car sold by a dealer with a 1463cc engine, 91.94 bhp max power and mileage of 19.43kmpl on 69191 km driven has an expected selling price of 680376.90.

We try to predict the price of our synthetic data point by using the predict() function. The fitted value is 1.538633, which gives the natural logarithm of the selling price to be 14.2779876. Thus we have a predicted selling price of 1588002.66, with the 95% prediction interval corresponding to 877461.41 – 2888250.01.

**Conclusion**

All specifications recorded in the Lionbridge data set, except for torque and car names, are relevant for determining the price of a used car. We obtained a multiple linear regression model, which involved log transformation of the selling price and most of the numerical variables, that allows us to predict the price of used cars based on their specifications. We found that a car being a test drive car and its power as the largest contributors in raising a car’s price. On the other hand, the specifications that are most responsible for lowering a car’s price is its age and if it has a manual transmission.