# Car Price Linear Regression Analysis

Nancy Ruan, Monashree Sanil, Wei He

## 1. Introduction

Estimating the price of a car is a complicated process. Before launching a car into the market, a good business strategy opted by most automobile companies is to perform market analysis. One aspect of such market analysis is to understand what factors affect the price of the car. There are fixed and dynamic variables in play. In this project, we are trying to identify what fixed variables play a significant role in estimating the price of a car and how significant is the effect.

We have used a dataset from Kaggle. The dataset contains 15 features and the target variable we are interested in:

| Variable Name | Type | Description |
|---|---|---|
| Make | Categorical | Manufacturing Company |
| Model | Categorical | Vehicle Model Series |
| Year | Categorical | Launched Year |
| Engine Fuel Type | Categorical | Engine Fuel Type |
| Engine HP | Numerical | Engine Horsepower |
| Engine Cylinders | Categorical | Engine Cylinders |
| Transmission Type | Categorical | Transmission Type |
| Driven_Wheels | Categorical | Wheel Drive System |
| Number of Doors | Categorical | Number of Doors |
| Market Category | Categorical | Market Category |
| Vehicle Size | Categorical | Vehicle Size |
| Vehicle Style | Categorical | Vehicle Style |
| highway MPG | Numerical | Miles Per Gallon in Highway |
| city mpg | Numerical | Miles Per Gallon in City |
| Popularity | Numerical | Popularity |
| **MSRP** | **Numerical** | **Manufacturer's Suggested Retail Price** |

In the above table, ‘MSRP’ is our response variable. It contains manufacturer's suggested retail prices for different cars. The raw dataset has 11,914 observations to work with.

## 2. Exploratory Data Analysis

Firstly, we would be performing exploratory data analysis to get some insights from the data and understand it better. Using the analysis, we would also perform data cleansing.

### 2.1 Drop ‘Model’ feature

‘Model’ feature contains the name of the car model series. If we treat it as a categorical variable, it would contain 914 categories. Also, more than half of the car model series have less than 10 samples. Hence, this feature is not that important for our goal and we can choose to drop this variable.

### 2.2 Unnecessary Subcategories in ‘Engine Fuel Type’, ‘Vehicle Style’ and ‘Market Category’

‘Engine Fuel Type’ : There are 10 categories in this feature. We would just need the basic level of information about what kind of fuel the car requires. Also, there are not many samples for certain categories. Hence, we can club 'premium unleaded (required)', 'premium unleaded (recommended)' under one category 'premium'. Also, we can club all categories starting with flex-fuel under one category.

‘Vehicle Style’ : Certain categories like '4dr SUV', '2dr SUV', ‘ Extended Cab Pickup’ etc. are providing redundant information that is already captured by the 'Number of Doors' feature and ‘Vehicle Size' feature. This can lead to multicollinearity among the features. Hence, we can group such categories under one category so that they distinguish cars based on their type like ‘SUV’, ‘Convertible’ etc.

‘Market Category’ : We have overlapping categories in the ‘Market Category’ feature. Certain categories indicate attributes that are already captured by other features like 'Vehicle Style', 'Engine Fuel Type' etc. Hence, only categories that can add information is whether the car is luxury or exotic.

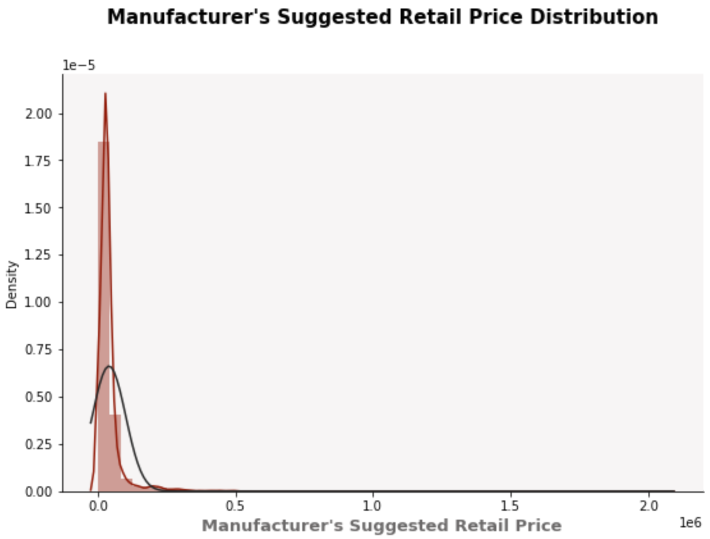
## 2.3 Missing data

Very few observations have missing values in 'Number of Doors', 'Engine Fuel Type', 'Engine Cylinders' and 'Engine HP'. We can choose to ignore these observations. For the 'Market Category' feature, we can replace missing values with a new category 'Unknown'.

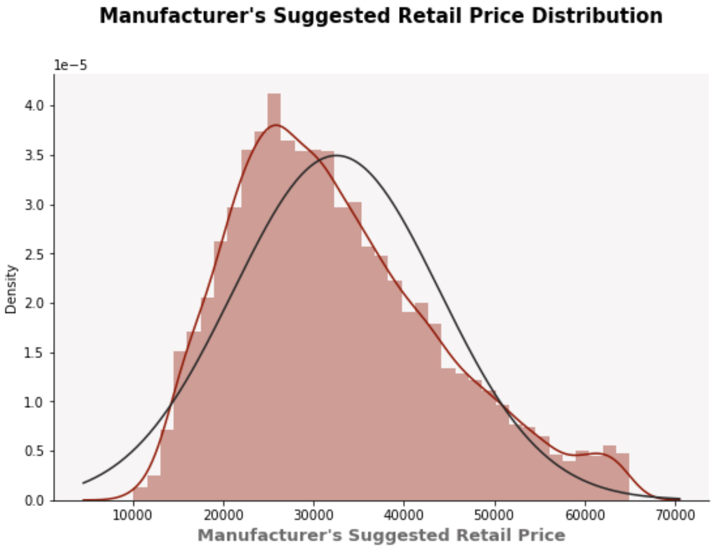
## 2.4 Transform ‘Year’

Intuitively, it makes more sense to include the number of years it has been since the car was launched rather than the year it was launched.

## 2.5 Distribution of MSRP

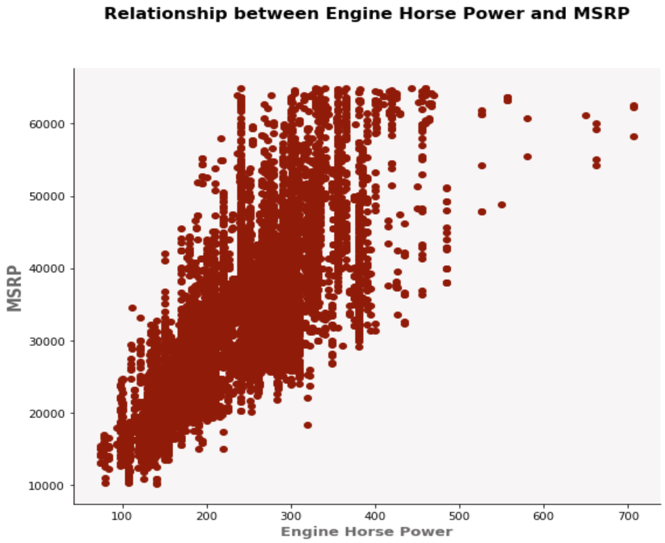


We can infer that the sale price values are strongly right-skewed. Some cars have very large values of MSRP. It would be difficult to infer the data for such a wide range of MSRP and regression analysis would not be reliable as it would violate the assumption of normality. Hence, using percentiles, we can determine that most of the cars lie in the range between 10000 and 65000. We can restrict MSRP to this range and carry on with our analysis.



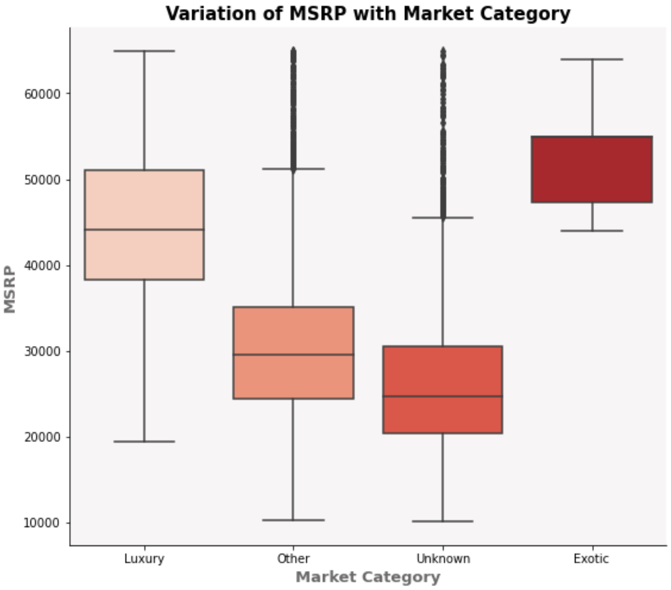
Now, we can see that the MSRP has approximately normal distribution. The data is now suitable for regression analysis.

## 2.6 Engine Horsepower has a strong positive linear relationship with MSRP



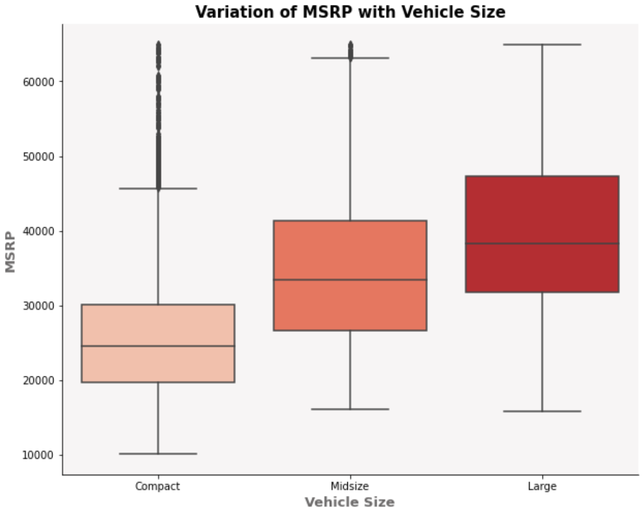
As can be observed, the horsepower of the engine has a positive correlation with the MSRP of the car. There are some outliers which do not follow the trend but they are not significantly out of trend and may not be significantly influential.

## 2.7 Luxury and exotic cars have high MSRP as compared to others.

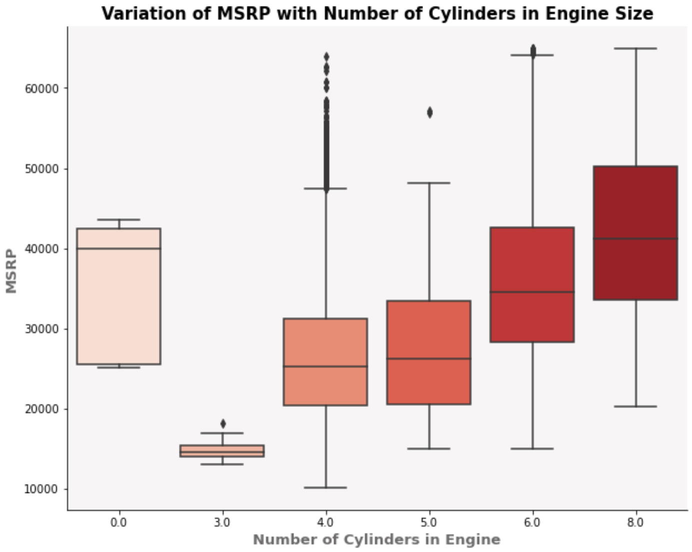


As expected, luxury and exotic cars have high prices. The range for luxury cars is wide but exotic cars have exceptionally high prices.

## 2.8 MSRP increase as vehicle size increases

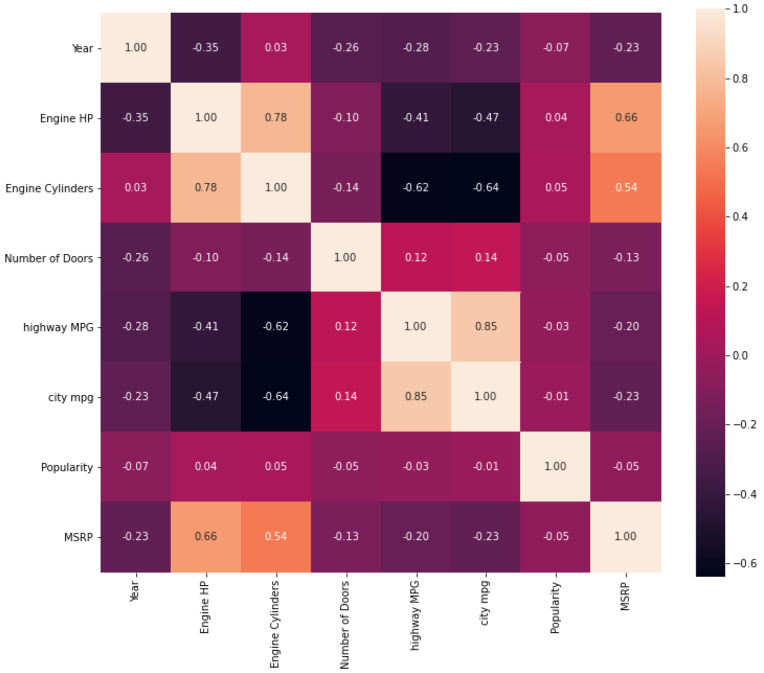


## 2.9 MSRP increase as number of cylinders increases



MSRP increases with the number of cylinders. However, there is an opposite trend for cars with cylinders less than 3. This may be missing data. The measurements for these observations may be missing.

## 2.10 Correlation Matrix for numerical features



The heatmap indicates that engine horsepower and number of cylinders in the engine are strongly and positively correlated to MSRP which is in line with our above analysis. In addition, we can see that MPG on highways is strongly correlated to MPG in the city. These columns may induce multicollinearity and add redundant information. Hence, it is better to remove one of them.

## 3. Type of Regression - Multiple Linear Regression

In this research, we are going to use Multiple Linear Regression(MLR) models to do model fitting and model analysis. Because MLR is suitable for predicting the possible outcome based on a number of variables. We would have an understanding of the relationship between these independent variables and the dependent variable. In our case, MSRP would be considered as our response variable and possible predictors after exploratory analysis would include: Make, Year, Engine Fuel Type, Engine HP, Engine Cylinders, Transmission Type, Driven Wheels, Number of Doors, Market Category, Vehicle Size, Vehicle Style, highway MPG, and Popularity.

We would firstly test multicollinearity before fitting our initial model. After that, we would go through initial model 1 fitting and diagnostics, initial model 2 fitting and diagnostics, and finally model selection. For model diagnostics, some of the potential problems we are going to look at would include influential points, heteroscedasticity problems, and normality problems.

With all the steps being taken, we wish to select an MLR model with relatively high trustworthy performance in predicting the manufacturers’ suggested retail price of cars.

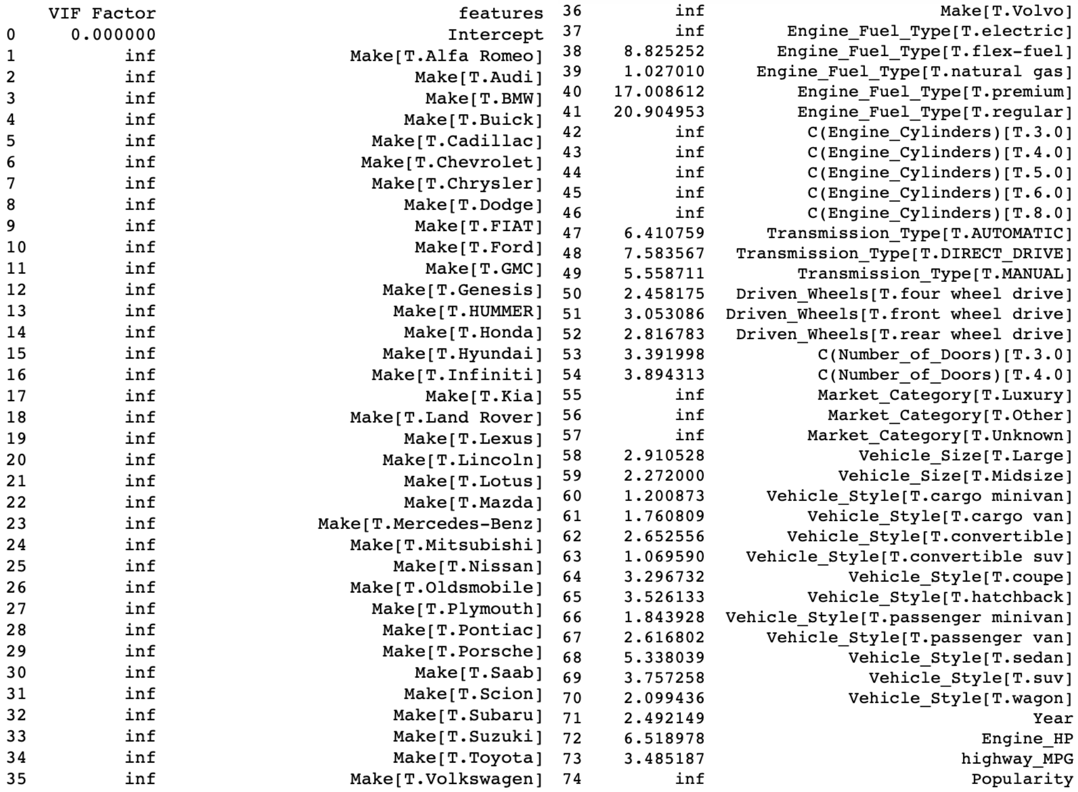
## 4. Before Model Fitting - Multicollinearity

Now we have a cleaned dataset for fitting an initial model. However, before fitting the model, we need to identify potential modeling problems that might potentially hurt the model performance and inferences through model diagnostics.

One of the model diagnostics that we conduct before initial modeling is checking multicollinearity. If critical multicollinearity problems exist in this dataset, then some of the predictors are highly correlated with each other, suggesting very sensitive coefficient estimates, and thus the test result would become unreliable. In our research, we use Variance Inflation

Factors (VIF) to measure how much the variance is inflated in the coefficient variables and detect predictors that are causing multicollinearity problems.

Referring to the VIF results (fig. 4.1), we identify 'Engine_Fuel_Type', 'Engine_Cylinders', 'Market_Category', 'Make', and 'Popularity' as variables that are causing the multicollinearity problem. Since each of them has a VIF value higher than 10, we consider these predictors are highly impacted by multicollinearity and decide to drop them out of the initial model.

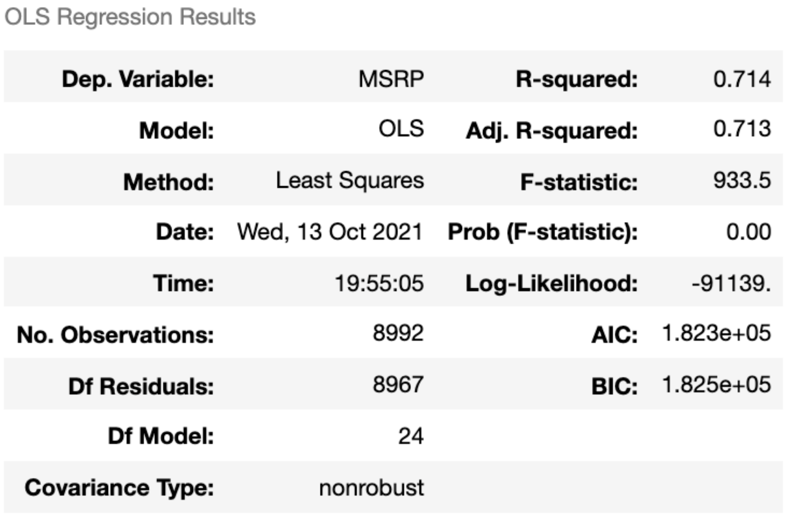


```
    VIF Factor                    features
0     0.000000                   Intercept
1          inf         Make[T.Alfa Romeo]
2          inf               Make[T.Audi]
3          inf                Make[T.BMW]
4          inf              Make[T.Buick]
5          inf           Make[T.Cadillac]
6          inf          Make[T.Chevrolet]
7          inf           Make[T.Chrysler]
8          inf              Make[T.Dodge]
9          inf               Make[T.FIAT]
10         inf               Make[T.Ford]
11         inf                Make[T.GMC]
12         inf            Make[T.Genesis]
13         inf             Make[T.HUMMER]
14         inf              Make[T.Honda]
15         inf            Make[T.Hyundai]
16         inf           Make[T.Infiniti]
17         inf                Make[T.Kia]
18         inf         Make[T.Land Rover]
19         inf              Make[T.Lexus]
20         inf            Make[T.Lincoln]
21         inf              Make[T.Lotus]
22         inf              Make[T.Mazda]
23         inf      Make[T.Mercedes-Benz]
24         inf         Make[T.Mitsubishi]
25         inf             Make[T.Nissan]
26         inf         Make[T.Oldsmobile]
27         inf           Make[T.Plymouth]
28         inf            Make[T.Pontiac]
29         inf            Make[T.Porsche]
30         inf               Make[T.Saab]
31         inf              Make[T.Scion]
32         inf             Make[T.Subaru]
33         inf             Make[T.Suzuki]
34         inf             Make[T.Toyota]
35         inf         Make[T.Volkswagen]
36         inf                          Make[T.Volvo]
37         inf           Engine_Fuel_Type[T.electric]
38    8.825252          Engine_Fuel_Type[T.flex-fuel]
39    1.027010        Engine_Fuel_Type[T.natural gas]
40   17.008612            Engine_Fuel_Type[T.premium]
41   20.904953            Engine_Fuel_Type[T.regular]
42         inf            C(Engine_Cylinders)[T.3.0]
43         inf            C(Engine_Cylinders)[T.4.0]
44         inf            C(Engine_Cylinders)[T.5.0]
45         inf            C(Engine_Cylinders)[T.6.0]
46         inf            C(Engine_Cylinders)[T.8.0]
47    6.410759       Transmission_Type[T.AUTOMATIC]
48    7.583567    Transmission_Type[T.DIRECT_DRIVE]
49    5.558711          Transmission_Type[T.MANUAL]
50    2.458175   Driven_Wheels[T.four wheel drive]
51    3.053086  Driven_Wheels[T.front wheel drive]
52    2.816783   Driven_Wheels[T.rear wheel drive]
53    3.391998            C(Number_of_Doors)[T.3.0]
54    3.894313            C(Number_of_Doors)[T.4.0]
55         inf             Market_Category[T.Luxury]
56         inf              Market_Category[T.Other]
57         inf            Market_Category[T.Unknown]
58    2.910528                Vehicle_Size[T.Large]
59    2.272000              Vehicle_Size[T.Midsize]
60    1.200873       Vehicle_Style[T.cargo minivan]
61    1.760809           Vehicle_Style[T.cargo van]
62    2.652556         Vehicle_Style[T.convertible]
63    1.069590     Vehicle_Style[T.convertible suv]
64    3.296732               Vehicle_Style[T.coupe]
65    3.526133           Vehicle_Style[T.hatchback]
66    1.843928   Vehicle_Style[T.passenger minivan]
67    2.616802       Vehicle_Style[T.passenger van]
68    5.338039               Vehicle_Style[T.sedan]
69    3.757258                 Vehicle_Style[T.suv]
70    2.099436               Vehicle_Style[T.wagon]
71    2.492149                                Year
72    6.518978                           Engine_HP
73    3.485187                         highway_MPG
74         inf                          Popularity
```

(fig. 4.1 VIF results)

## 5. Initial Model 1 and Model Diagnostics

Thus, after deleting the predictors we mentioned above, we have our initial MLR model 1: Manufacturer's Suggested Retail Price ~ Year + Engine HP + Transmission Type + Driven Wheels + Number of Doors + Vehicle Size + Vehicle Style + highway MPG. Next step, we fit the initial model 1 (fig. 5).



OLS Regression Results

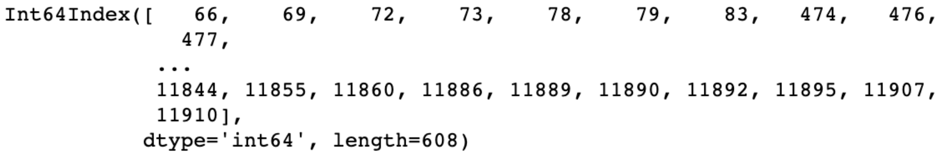
| | | | |
|---|---|---|---|
| **Dep. Variable:** | MSRP | **R-squared:** | 0.714 |
| **Model:** | OLS | **Adj. R-squared:** | 0.713 |
| **Method:** | Least Squares | **F-statistic:** | 933.5 |
| **Date:** | Wed, 13 Oct 2021 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 19:55:05 | **Log-Likelihood:** | -91139. |
| **No. Observations:** | 8992 | **AIC:** | 1.823e+05 |
| **Df Residuals:** | 8967 | **BIC:** | 1.825e+05 |
| **Df Model:** | 24 | | |
| **Covariance Type:** | nonrobust | | |

(fig. 5 Initial Model 1)

Before we draw any conclusions after model fitting, we need to run model diagnostics to detect model issues. There might be data structural problems like influential points and model assumption violations including heteroscedasticity, non-normality residuals, false assumption of linearity between MSRP and predictors, etc.

## 5.1 influential points

We use threshold externally studentized residuals and Cook's distance together to find points that are potentially influential to the model. In the studentized residual method, we use alpha = 0.05 to evaluate data points. When calculating Cook's distance, we use one-fourth of the number of observations (8992/4) as the threshold to choose influential points. By unioning the index results we found in both methods together, we identify 608 influential points in total out of 8992 observations (fig. 5.1.1).



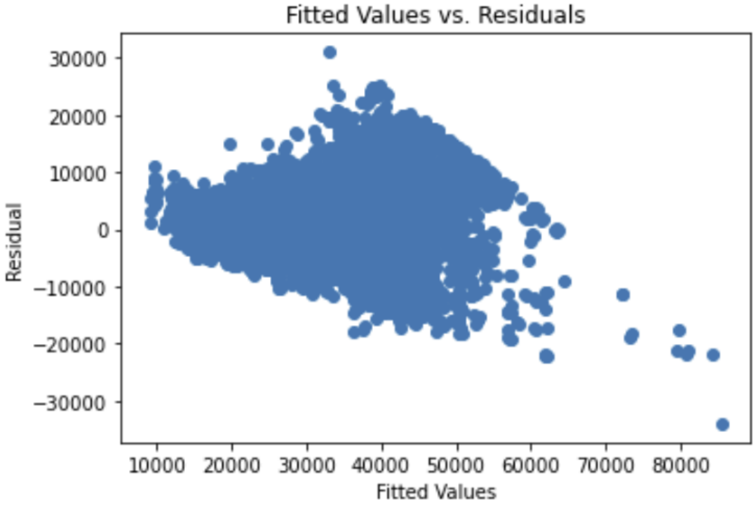
```
Int64Index([   66,    69,    72,    73,    78,    79,    83,   474,   476,
              477,
            ...
            11844, 11855, 11860, 11886, 11889, 11890, 11892, 11895, 11907,
            11910],
           dtype='int64', length=608)
```

(fig. 5.1.1 Initial Model 1 - Influential Points)

It is no good to simply delete those influential points. We would look at the model result with and without those influential points.

## 5.2 Heteroscedasticity Problem

Heteroscedasticity occurs when the variances of errors are non-constant. With heteroscedasticity problems, the OLSE estimates of coefficients are still linear and unbiased, but not the “best” anymore. Smaller variance estimates, misleading t-test, and confidence interval results might be harmful to model fitting. Thus, with influential points included in the model, we draw a scatter plot of “Fitted Values vs. Residuals Plot” to detect this issue(fig. 5.2.1).

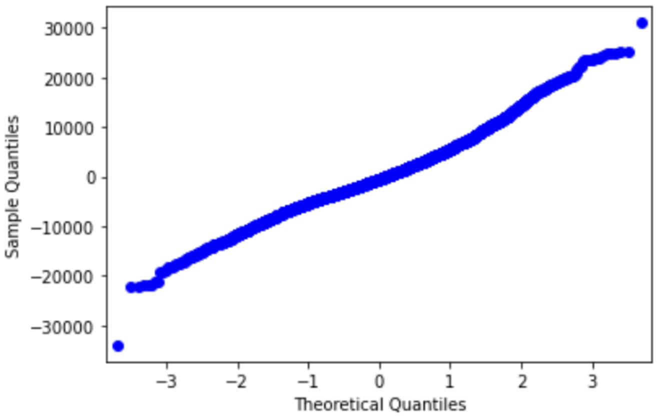


(fig. 5.2.1 Fitted Value v.s. Residuals)

We could see that there is an obvious change in the bandwidth with fitted values of MSRP, suggesting the existence of heteroscedasticity problems. Breusch-Pagan test is also used in detecting heteroscedasticity problems. The test result with statistics = 1691.70 and p-value = 0.0 suggest the existence of a significant heteroscedasticity problem (alpha = 0.05). In order to improve our model, we would use natural-log transformation on MSRP in future steps.

## 5.3 Normality Problem

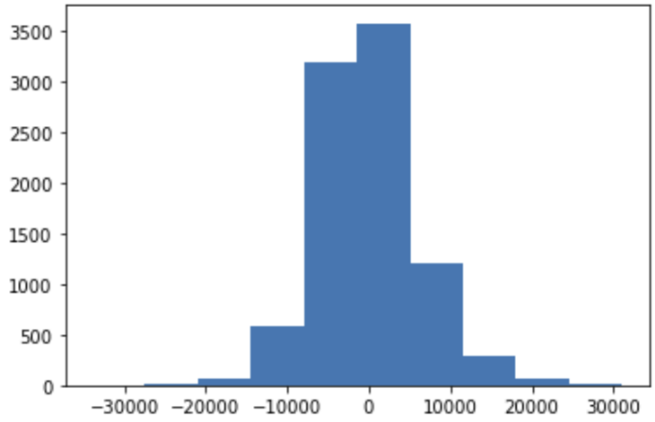
Normality assumptions should be tested for our initial model, too. When there’s a violation of normality, it means that the error term does not follow the normal distribution, which may lead to problems when we compute t-test, confidence intervals, and ANOVA test for our model. In this research, we use the QQ plot to identify the potential normality problem.



(fig 5.3.1 QQ plot)

Based on the results of the QQ plot(fig. 5.3.1), we see the graphs are approximately diagonal, showing that the distribution of the data might not exist as a normality problem. It is normally distributed. However, we identify weird points on both sides of the QQ plot. Further improvements on model fitting could be done through natural-log transformation.

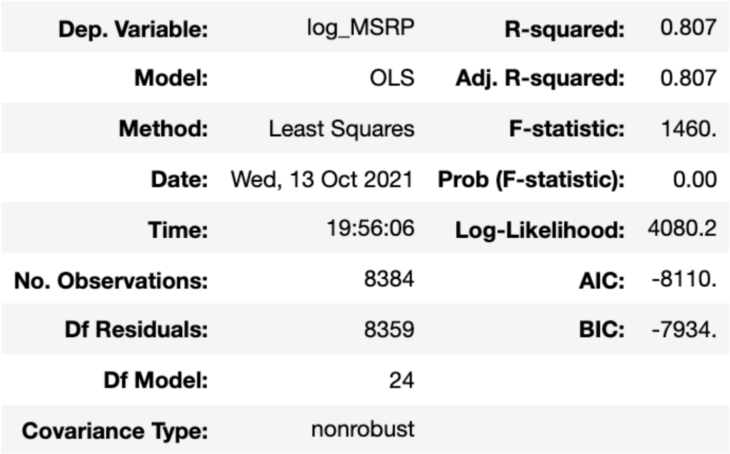
The histogram of residuals and a Jarque-Bera test are also used in identifying normality problems. The histogram of residuals(fig. 5.3.2) shows that the skewness of residuals is approximately 0, showing approximately symmetry. While for the Jarque-Bera test, with alpha = 0.05, we have JB statistics = 819.57 and p-value = 0.0. There's a skewness (normality) problem for this dataset and we could use natural-log transformation on MSRP to improve this problem. However, we need to note that since Jarque-Bera is sensitive to outliers, the outliers in the dataset might cause this rejection.



(fig. 5.3.2 Histogram of Residuals)

## 6. Initial Model 2 and Model Diagnostics

Now we are going to drop those influential outliers out of the initial dataset and perform a natural-log transformation on MSRP to have a cleaned dataset. We still use the predictors we selected in initial model 1 to fit the model, but the dependent variable changed from MSRP to natural-logged MSRP. Also, the number of observations changed from 8992 to 8384 (fig. 6).

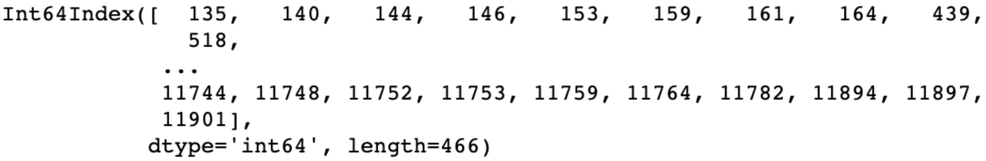


| | | | |
|---|---|---|---|
| **Dep. Variable:** | log_MSRP | **R-squared:** | 0.807 |
| **Model:** | OLS | **Adj. R-squared:** | 0.807 |
| **Method:** | Least Squares | **F-statistic:** | 1460. |
| **Date:** | Wed, 13 Oct 2021 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 19:56:06 | **Log-Likelihood:** | 4080.2 |
| **No. Observations:** | 8384 | **AIC:** | -8110. |
| **Df Residuals:** | 8359 | **BIC:** | -7934. |
| **Df Model:** | 24 | | |
| **Covariance Type:** | nonrobust | | |

(fig. 6 Initial Model 2)

### 6.1 Influential Points

We still want to run model diagnostics once again after fitting model 2. For influential points, we used the threshold externally studentized residuals methods (alpha = 0.05) and Cook's distance (threshold = 8384/4). The index of 466 influential points is identified when we combine the results together(fig. 6.1.1).

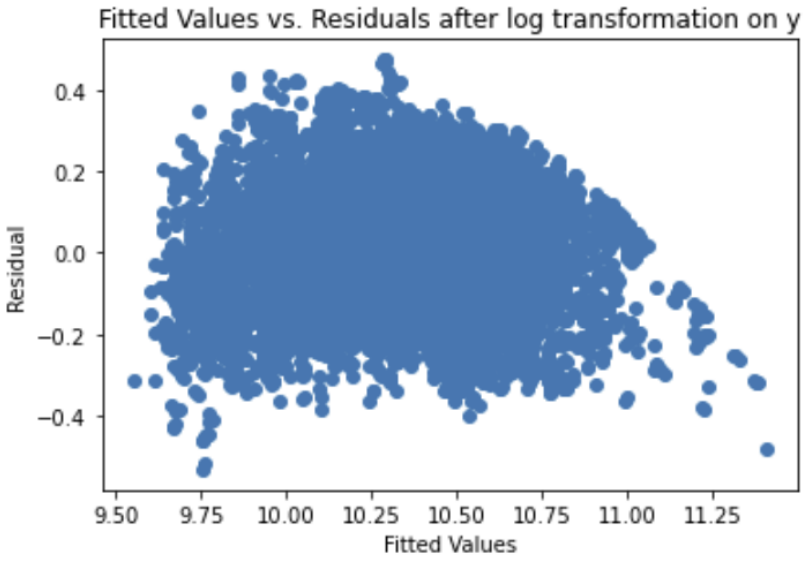


```
Int64Index([  135,   140,   144,   146,   153,   159,   161,   164,   439,
              518,
            ...
            11744, 11748, 11752, 11753, 11759, 11764, 11782, 11894, 11897,
            11901],
           dtype='int64', length=466)
```

(fig. 6.1.1 Initial Model 2 - Influential Point Index)

### 6.2 Heteroscedasticity Problem

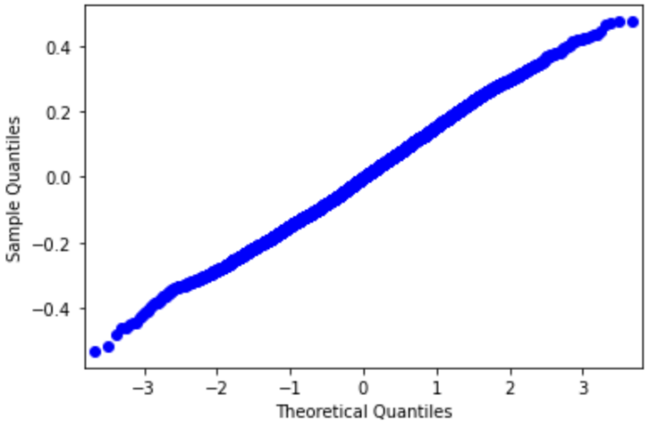
The heteroscedasticity problems are being detected again with the scatter plot of "Fitted Values v.s. Residuals after log transformation on MSRP" (fig. 6.2.1). We could see that the heteroscedasticity problem has been improved as the distribution of those plotted points shows similar bandwidth. On the other hand, the Breusch-Pagan test still suggests a significant heteroscedasticity problem with BP statistics = 312.74 and p-value = 4.524654899906637e-52 (alpha= 0.05). However, we need to note that the Breusch-Pagan test is very sensitive to outliers. This rejection of hypothesis testing might be caused by the remaining outliers.



(fig. 6.2.1 Fitted Values v.s. Residuals after log transformation on MSRP)

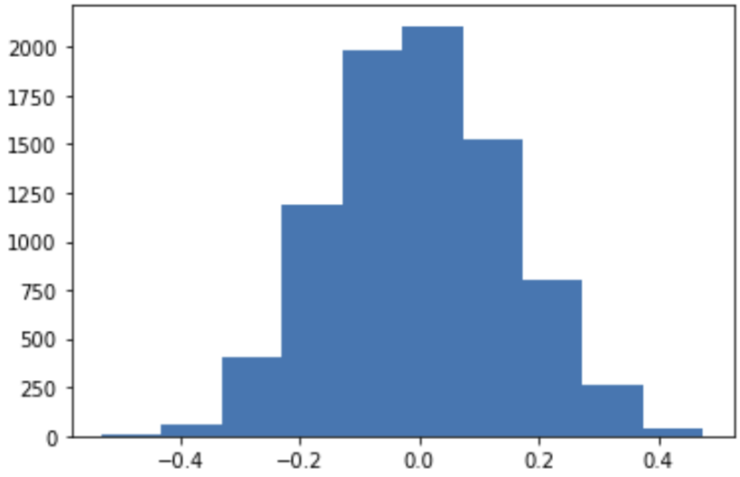
## 6.3 Normality Problem

For normality problems, we obtained a better QQ plot (fig. 6.3.1) with approximately perfect diagonal distribution, showing that the distribution of the data might not have the normality problem.



(fig. 6.3.1 QQ plot)

The histogram of residuals (fig. 6.3.2) also shows an approximately normal distribution, suggesting zero skewness and no normality problems.



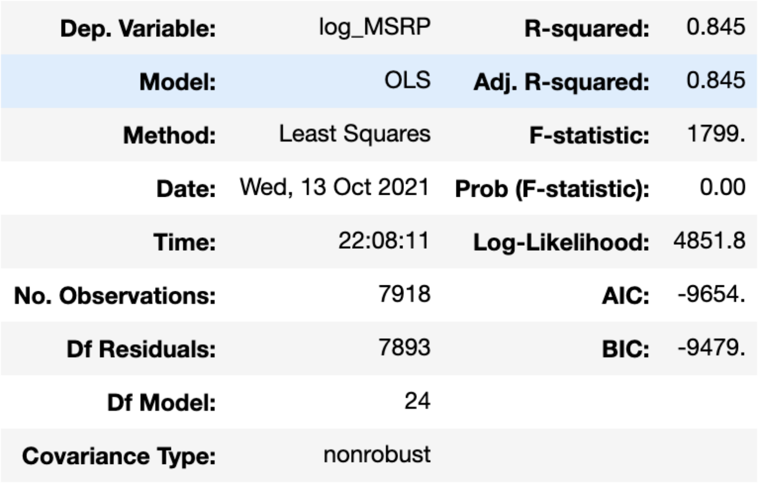
(fig. 6.3.2 Histogram of Residuals)

While QQ plot and histogram of residuals existence of normality problems, the Jarque-Bera test suggests a different result. With alpha = 0.05, JB statistics = 41.17 and p-value = 1.1459957427462086e-09, Jarque-Bera test shows a significant normality problem for initial model 2. We would note here, again, that the Jarque-Bera test is very sensitive to outliers. So the rejection might be influenced by those data points.

# 7. Model Selection

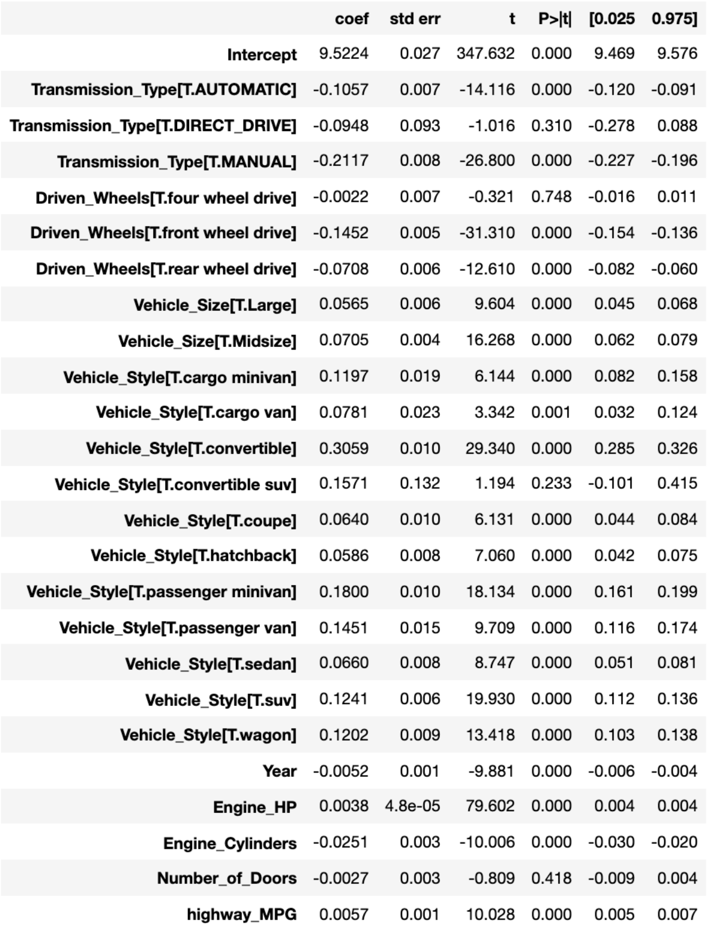
## 7.1 Comparison between models with or without influential points

After performing the log transformation on the MSRP, there are still some influential points existing. We fit model 3 to see whether the model performs better without these influential points. Below is the summary table of model 3:



| | | | |
|---|---|---|---|
| **Dep. Variable:** | log_MSRP | **R-squared:** | 0.845 |
| **Model:** | OLS | **Adj. R-squared:** | 0.845 |
| **Method:** | Least Squares | **F-statistic:** | 1799. |
| **Date:** | Wed, 13 Oct 2021 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 22:08:11 | **Log-Likelihood:** | 4851.8 |
| **No. Observations:** | 7918 | **AIC:** | -9654. |
| **Df Residuals:** | 7893 | **BIC:** | -9479. |
| **Df Model:** | 24 | | |
| **Covariance Type:** | nonrobust | | |

(fig. 7.1.1 Summary table of model 3)

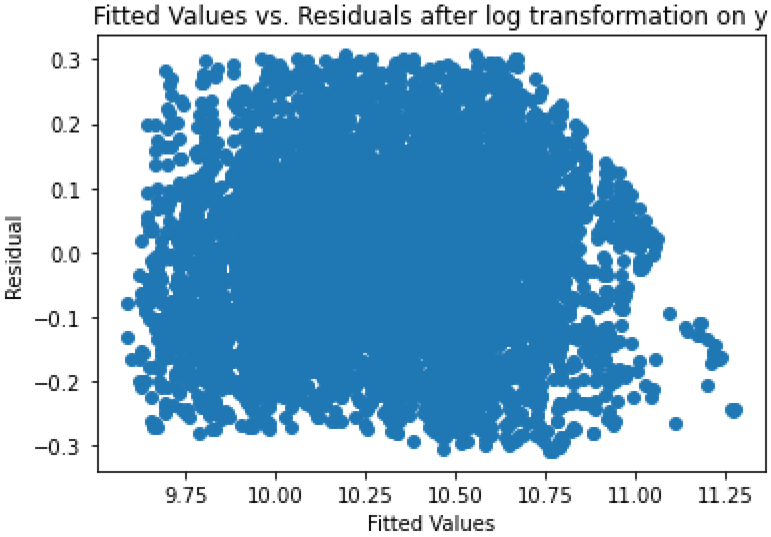


| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 9.5224 | 0.027 | 347.632 | 0.000 | 9.469 | 9.576 |
| **Transmission_Type[T.AUTOMATIC]** | -0.1057 | 0.007 | -14.116 | 0.000 | -0.120 | -0.091 |
| **Transmission_Type[T.DIRECT_DRIVE]** | -0.0948 | 0.093 | -1.016 | 0.310 | -0.278 | 0.088 |
| **Transmission_Type[T.MANUAL]** | -0.2117 | 0.008 | -26.800 | 0.000 | -0.227 | -0.196 |
| **Driven_Wheels[T.four wheel drive]** | -0.0022 | 0.007 | -0.321 | 0.748 | -0.016 | 0.011 |
| **Driven_Wheels[T.front wheel drive]** | -0.1452 | 0.005 | -31.310 | 0.000 | -0.154 | -0.136 |
| **Driven_Wheels[T.rear wheel drive]** | -0.0708 | 0.006 | -12.610 | 0.000 | -0.082 | -0.060 |
| **Vehicle_Size[T.Large]** | 0.0565 | 0.006 | 9.604 | 0.000 | 0.045 | 0.068 |
| **Vehicle_Size[T.Midsize]** | 0.0705 | 0.004 | 16.268 | 0.000 | 0.062 | 0.079 |
| **Vehicle_Style[T.cargo minivan]** | 0.1197 | 0.019 | 6.144 | 0.000 | 0.082 | 0.158 |
| **Vehicle_Style[T.cargo van]** | 0.0781 | 0.023 | 3.342 | 0.001 | 0.032 | 0.124 |
| **Vehicle_Style[T.convertible]** | 0.3059 | 0.010 | 29.340 | 0.000 | 0.285 | 0.326 |
| **Vehicle_Style[T.convertible suv]** | 0.1571 | 0.132 | 1.194 | 0.233 | -0.101 | 0.415 |
| **Vehicle_Style[T.coupe]** | 0.0640 | 0.010 | 6.131 | 0.000 | 0.044 | 0.084 |
| **Vehicle_Style[T.hatchback]** | 0.0586 | 0.008 | 7.060 | 0.000 | 0.042 | 0.075 |
| **Vehicle_Style[T.passenger minivan]** | 0.1800 | 0.010 | 18.134 | 0.000 | 0.161 | 0.199 |
| **Vehicle_Style[T.passenger van]** | 0.1451 | 0.015 | 9.709 | 0.000 | 0.116 | 0.174 |
| **Vehicle_Style[T.sedan]** | 0.0660 | 0.008 | 8.747 | 0.000 | 0.051 | 0.081 |
| **Vehicle_Style[T.suv]** | 0.1241 | 0.006 | 19.930 | 0.000 | 0.112 | 0.136 |
| **Vehicle_Style[T.wagon]** | 0.1202 | 0.009 | 13.418 | 0.000 | 0.103 | 0.138 |
| **Year** | -0.0052 | 0.001 | -9.881 | 0.000 | -0.006 | -0.004 |
| **Engine_HP** | 0.0038 | 4.8e-05 | 79.602 | 0.000 | 0.004 | 0.004 |
| **Engine_Cylinders** | -0.0251 | 0.003 | -10.006 | 0.000 | -0.030 | -0.020 |
| **Number_of_Doors** | -0.0027 | 0.003 | -0.809 | 0.418 | -0.009 | 0.004 |
| **highway_MPG** | 0.0057 | 0.001 | 10.028 | 0.000 | 0.005 | 0.007 |

(fig. 7.1.2  t test results of model 3**)**

We can see compared with model 2, the Adj. R-squared value becomes higher which is 0.845 now. And the F-test value also increased from 1460 to 1799, which means the model gets a more significant result.

We draw the plot of fitted values v.s. residuals as below to check whether the heteroscedasticity is improved:

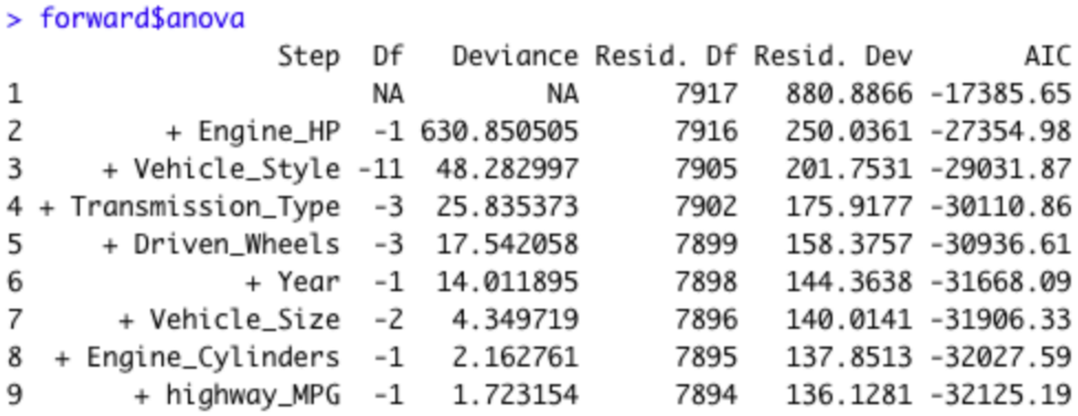


(fig. 7.1.2 Fitted values vs residuals)

This plot shows the residuals are all in the range of -0.3 to 0.3. The whole plot seems improved a lot. But from the Breusch-Pagan test, we find the  LM-Test p-value is 5.56e-32, which suggests that there is still heteroscedasticity existing in the data. We think it is due to the BP being too sensitive to the points on the bottom right. Therefore, we move forward to the stepwise approach of model selection.

### 7.2 Stepwise Model Selection

Since there are 9 candidates to choose, considering the computational burden, we decide to use the stepwise model regression approach to find out the best model. We first define the intercept-only model. Then set up the model with all predictors that fit natural-logged MSRP against Year, Engine HP, Engine Cylinders, Transmission Type, Driven Wheels, Number of Doors, Vehicle Size, Vehicle Style, and highway MPG. After performing forward stepwise regression based on these two models, we get the following results:



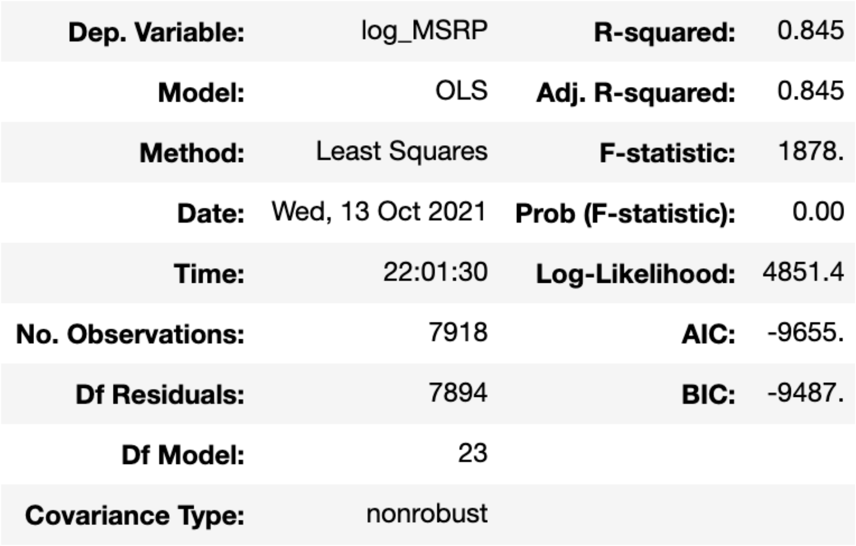
```
> forward$anova
                 Step  Df   Deviance Resid. Df Resid. Dev       AIC
1                      NA         NA      7917   880.8866 -17385.65
2         + Engine_HP  -1 630.850505      7916   250.0361 -27354.98
3     + Vehicle_Style -11  48.282997      7905   201.7531 -29031.87
4 + Transmission_Type  -3  25.835373      7902   175.9177 -30110.86
5     + Driven_Wheels  -3  17.542058      7899   158.3757 -30936.61
6              + Year  -1  14.011895      7898   144.3638 -31668.09
7      + Vehicle_Size  -2   4.349719      7896   140.0141 -31906.33
8  + Engine_Cylinders  -1   2.162761      7895   137.8513 -32027.59
9       + highway_MPG  -1   1.723154      7894   136.1281 -32125.19
```

(fig. 7.2 The forward stepwise regression output)

From the output of stepwise regression, we can see the model with 8 predictors has the lowest AIC value and  'Number of Doors' is not chosen. It is consistent with the Anova test of Model 3, in which the p-value of t-test of 'Number of Doors' is 0.418. Therefore, we can fit the model of log MSRP against the 8 chosen variables.

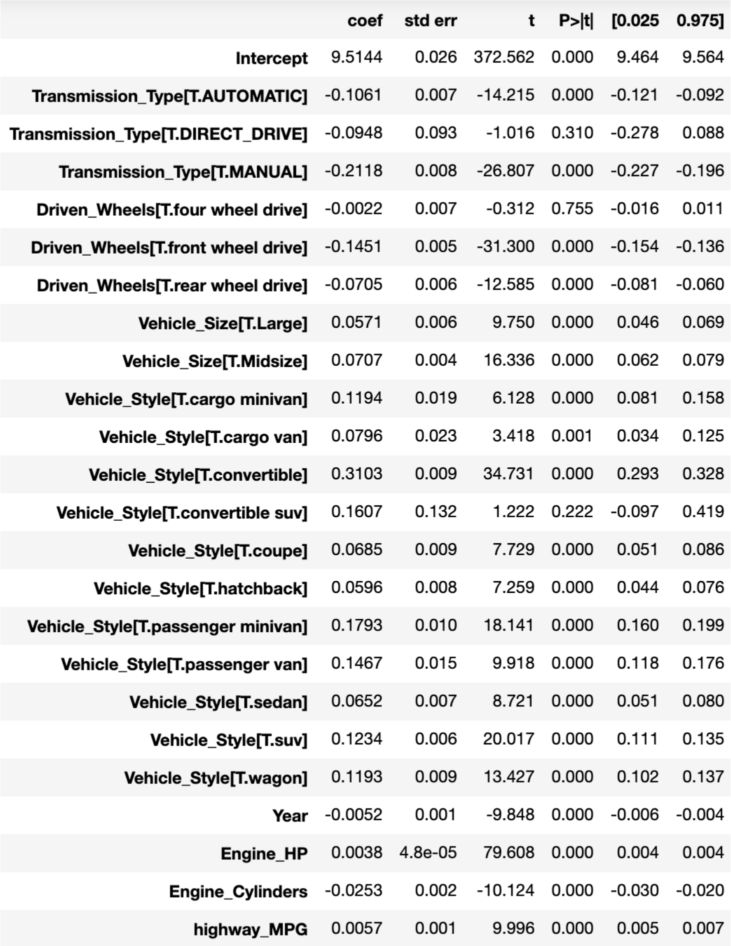
### 7.3 Fit the final model

Based on all the EDA, Model Diagnostics, and Model selection above, we fit the model using log MSRP against Year, Engine HP, Engine Cylinders, Transmission Type, Driven Wheels, Vehicle Size, Vehicle Style, and highway MPG.



| | | | |
|---|---|---|---|
| **Dep. Variable:** | log_MSRP | **R-squared:** | 0.845 |
| **Model:** | OLS | **Adj. R-squared:** | 0.845 |
| **Method:** | Least Squares | **F-statistic:** | 1878. |
| **Date:** | Wed, 13 Oct 2021 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 22:01:30 | **Log-Likelihood:** | 4851.4 |
| **No. Observations:** | 7918 | **AIC:** | -9655. |
| **Df Residuals:** | 7894 | **BIC:** | -9487. |
| **Df Model:** | 23 | | |
| **Covariance Type:** | nonrobust | | |

(fig. 7.3.1 The summary table of the final model)



| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 9.5144 | 0.026 | 372.562 | 0.000 | 9.464 | 9.564 |
| Transmission_Type[T.AUTOMATIC] | -0.1061 | 0.007 | -14.215 | 0.000 | -0.121 | -0.092 |
| Transmission_Type[T.DIRECT_DRIVE] | -0.0948 | 0.093 | -1.016 | 0.310 | -0.278 | 0.088 |
| Transmission_Type[T.MANUAL] | -0.2118 | 0.008 | -26.807 | 0.000 | -0.227 | -0.196 |
| Driven_Wheels[T.four wheel drive] | -0.0022 | 0.007 | -0.312 | 0.755 | -0.016 | 0.011 |
| Driven_Wheels[T.front wheel drive] | -0.1451 | 0.005 | -31.300 | 0.000 | -0.154 | -0.136 |
| Driven_Wheels[T.rear wheel drive] | -0.0705 | 0.006 | -12.585 | 0.000 | -0.081 | -0.060 |
| Vehicle_Size[T.Large] | 0.0571 | 0.006 | 9.750 | 0.000 | 0.046 | 0.069 |
| Vehicle_Size[T.Midsize] | 0.0707 | 0.004 | 16.336 | 0.000 | 0.062 | 0.079 |
| Vehicle_Style[T.cargo minivan] | 0.1194 | 0.019 | 6.128 | 0.000 | 0.081 | 0.158 |
| Vehicle_Style[T.cargo van] | 0.0796 | 0.023 | 3.418 | 0.001 | 0.034 | 0.125 |
| Vehicle_Style[T.convertible] | 0.3103 | 0.009 | 34.731 | 0.000 | 0.293 | 0.328 |
| Vehicle_Style[T.convertible suv] | 0.1607 | 0.132 | 1.222 | 0.222 | -0.097 | 0.419 |
| Vehicle_Style[T.coupe] | 0.0685 | 0.009 | 7.729 | 0.000 | 0.051 | 0.086 |
| Vehicle_Style[T.hatchback] | 0.0596 | 0.008 | 7.259 | 0.000 | 0.044 | 0.076 |
| Vehicle_Style[T.passenger minivan] | 0.1793 | 0.010 | 18.141 | 0.000 | 0.160 | 0.199 |
| Vehicle_Style[T.passenger van] | 0.1467 | 0.015 | 9.918 | 0.000 | 0.118 | 0.176 |
| Vehicle_Style[T.sedan] | 0.0652 | 0.007 | 8.721 | 0.000 | 0.051 | 0.080 |
| Vehicle_Style[T.suv] | 0.1234 | 0.006 | 20.017 | 0.000 | 0.111 | 0.135 |
| Vehicle_Style[T.wagon] | 0.1193 | 0.009 | 13.427 | 0.000 | 0.102 | 0.137 |
| Year | -0.0052 | 0.001 | -9.848 | 0.000 | -0.006 | -0.004 |
| Engine_HP | 0.0038 | 4.8e-05 | 79.608 | 0.000 | 0.004 | 0.004 |
| Engine_Cylinders | -0.0253 | 0.002 | -10.124 | 0.000 | -0.030 | -0.020 |
| highway_MPG | 0.0057 | 0.001 | 9.996 | 0.000 | 0.005 | 0.007 |

(fig. 7.3.2  t test results of the predictors**)**

In this final model, we drop the predictor "Number of doors". Adj. R-squared is still 0.845, while the F-statistic becomes bigger. P-value of the t-test of all the other variables are all much smaller than 0.05(For the category variables, each of them at least has 2 significant dummy variables). Therefore, we think this model performs well.

## 8. Summary of findings

Beginning with the original data, we want to find out the impactive predictors of MSRP within 16 variables. In EDA, we integrate the unnecessary categories of categorical variables, replace missing values, check the linear relationships between variables,  and constrain our target data between a reasonable range. In the model Diagnostics, we check whether there are multicollinearity, influential points, Heteroscedasticity, and non-normality problems existing in the data set. After eliminating all these problems, we choose 9 variables to be our final candidates and fit the model.

In the model selection, we firstly compare the models with or without the influential points, and conclude that by deleting the influential points, we can much improve the heteroscedasticity, get higher Adj-R square value, and let the F test result of the model become more significant. Then we utilize stepwise model regression to choose the best model and find that Number of doors does not have a significant impact on MSRP.

Finally, we set up the model by fitting MSRP against Year, Engine HP, Engine Cylinders, Transmission Type, Driven Wheels, Number of Doors, Vehicle Size, Vehicle Style, and highway MPG. There are four numerical variables and four categorical variables. The t test result shows that all these predictors have significant impacts on MSRP. Adj.R squared value is 0.845, which means compared with model 1, model 2, and model 3, this model fits very well.

Based on our final model, the conclusion is that Year and Engine Cylinders have significant negative impact on MSRP while Engine HP and  highway MPG have significant positive impact on MSRP. For the categorical variable Transmission Type, AUTOMATIC or MANUAL have a significant negative impact on MSRP compared with AUTOMATED MANUAL. For the categorical variable Driven Wheels, compared with all-wheel drive, cars only have front wheel drive or rear wheel drive will have lower MSRP. But there is no significant difference between the price of all-wheel drive and four wheel drive. For the categorical variable Vehicle Size, midsize cars and large size cars both have a significantly higher MSRP. At last, for the categorical variable vehicle style, apart from convertible suv, all the other vehicle styles of cars have significantly higher MSRP prices compared with  cab pickup.

## 9. Potential problems

Throughout our analysis, we think there are mainly two potential problems existing. Firstly, the Breusch-Pagan test result still shows that there is heteroscedasticity existing in the data. Although it is reasonable for us to speculate that the reason is BP test is too sensitive to the data points on the bottom right of fitted values vs residual plot, but we are not completely sure.

Secondly, considering the computational burden, we use the stepwise model regression approach to find out the best model. But the stepwise model regression approach may miss the true best model.

In future potential model improvements, we would like to focus on these two potential problems and make some adjustments.