Multiple Linear Regression Analysis on Car-sales' MSRP

Yifu Zheng
Master of Data Analytics
University of Western Ontario
London, Canada
yzhen487@uwo.ca

I. Introduction

The pricing of automobiles is influenced by a myriad of factors, ranging from brand reputation and technical specifications to consumer sentiment and market trends. Understanding the relationships between these factors and a car's Manufacturer's Suggested Retail Price (MSRP) is critical for stakeholders such as manufacturers, consumers, and market analysts. This project seeks to explore the dynamics between car features and pricing through a comprehensive regression analysis.

Utilizing a dataset from Kaggle that encompasses car attributes such as make, model, year, engine specifications, and social metrics like popularity, this analysis aims to uncover the key determinants of car prices. By investigating the effects of various features, the role of branding, and the potential for overpricing, the study aspires to deliver insights that could guide pricing strategies and consumer decisions. Additionally, this project will assess the interplay between price and market categories, shedding light on how different car types align with pricing trends.

The report presents the goals of this analysis, the methodology employed, the results obtained, and their implications, providing a detailed and data-driven perspective on the automotive market.

II. Data Analysis

A. Data Description

The dataset consists of 11,914 entries and 16 columns, containing information about cars, their features, and their corresponding Manufacturer's Suggested Retail Prices (MSRP). Below is an overview of the variables in the dataset:

1. Make: The manufacturer or brand of the car (e.g., BMW, Ford).

2. Model: The specific model of the car.

3. Year: The year the car model was manufactured.

4. Engine Fuel Type: The type of fuel the engine uses (e.g., premium unleaded, diesel).

5. Engine HP: The horsepower of the engine, indicating its power output.

6. Engine Cylinders: The number of cylinders in the engine.

7. Transmission Type: The type of transmission system (e.g., automatic, manual).

8. Driven_Wheels: The drivetrain configuration (e.g., rear-wheel drive, all-wheel drive).

	Make	Model	Year	Engine.Fuel.Type	Engine.HP	Engine.Cylinders	Transmission.Type	Driven_Wheels	Number.of.Doors	Market.Category	Vehicle.Size	Vehicle.St
1	BMW	1 Series M	2011	premium unleaded (required)	335	6	MANUAL	rear wheel drive	2	Factory Tuner,Luxury,High-Performance	Compact	Coupe
2	BMW	1 Series	2011	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Convertibl
3	BMW	1 Series	2011	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,High-Performance	Compact	Coupe
4	BMW	1 Series	2011	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Coupe
5	BMW	1 Series	2011	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury	Compact	Convertibl
6	BMW	1 Series	2012	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Coupe
7	BMW	1 Series	2012	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Convertibl
8	BMW	1 Series	2012	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,High-Performance	Compact	Coupe

Fig. 1: Showcase of Data Before Modifying

9. Number of Doors: The number of doors on the car.

10. Market Category: A categorization of the car based on market positioning (e.g., luxury, high-performance).

11. Vehicle Size: The size classification of the vehicle (e.g., compact, midsize).

12. Vehicle Style: The style or body type of the vehicle (e.g., sedan, SUV).

13. Highway MPG: The fuel efficiency of the car on the highway, measured in miles per gallon.

14. City MPG: The fuel efficiency of the car in city driving conditions, measured in miles per gallon.

15. Popularity: A numeric measure indicating the car's popularity.

16. MSRP: The Manufacturer's Suggested Retail Price, indicating the car's price.

The dataset includes a mix of numerical and categorical variables, with some columns containing missing values. The target variable for the regression analysis is MSRP, while the other columns provide features that can potentially influence the price.

B. Data Pre-treatment

Firstly, we do believe that even before any sort of training, there is the need to convert some of the messy datas. Shown in Fig 1.

We do believe before even fitting it into any specific model, we have to acknowledge that Market.Category would be potentially a very significant predictor to our model and making it into dummies would be a better indicator for the overall prediction. However, due to it being compacted into a single column we would need to split it up into different dummy variable columns that we would be able to use in the model. After this has been done, we also standardized columns so it would be on a similar scale when fitted. Also, we believe that removing make and model would extremely help with the result as including these would definitely create problem with

overfitting as these are way too ”good” of a predictor for the MSRP.

We will then consider to fit it into a MLR model, with MSRP being the dependent variable and test using AIC with stepwise to gain the best model. We will first start with all of the predictors first and continue from there to see if we would also need interation terms.

III. Model Fitting and Optimization

In this step, before anything and the aforementioned data alterations, we have randomly selected 600 entries from the dataset, using the random seed number 88(for reproducibility). We then transformed the 600 entries together with the following step:

1. Add a temporary Row_ID to track rows
2. Extract Market.Category into a separate DataFrame
3. Split and preprocess Market.Category
4. Create one-hot encoding for categories
5. Aggregate One-Hot Encoded Data
6. Merge Back with subset_data
7. Merge the one-hot encoded data back into subset_data
8. Remove NA columns and Split the 600 rows into training (500) and testing (100)

A. Model Definition

We believe that for this type of task given, the optimal choice is to assume a Multiple-Linear Regression Model to start first. Our model is to fit it as followed shown in Fig 2. As we see here, the three assumptions of a model is violated:

1. Equal Variance: We see a compiled bunch in the very beginning from the Residual Plot and large spread at the very end, which means Equal Variance is definitely violated.

2. Normality: In the normal QQ plot, since the values at both end are far from the linear line, it seems that the normal assumption is violated.

3. Linearity: As shown in the Residual plot, the average is roughly around 0, and linearity does seemed to hold.

This lead to us to consider a transformation for $\hat{Y}$ in order to fit the violation again equal variance and normality. And since the Var(e) is seemed to be a quadratic function of $\hat{Y}$, we will start with a log transformation on the dependant variable MSRP.

B. Transformed Model with Feature Selection

Recall to our observation on the model summary shown in Fig 2-4. Even though the p-value of the predictor are being shown to be significant, we believe the first step is to fit the assumptions and then a AIC predictor reduction. As Var(e) is showing us a quadratic function of $\hat{Y}$ so we'll now fit it with a log transformation. The residual plot and QQ plot of the model after log transformation are shown in Fig 5-6.

With this analysis, we check the 3 assumptions again:

1. Equal Variance: Compared to the initial model, heteroscedasticity has improved. However, it can be seen from residual plot that the distribution of residuals is not constant. In the initial range of fitted values, the residual distribution is tighter, while in the higher range of fitted values, the fluctuation of the residuals increases significantly. This pattern indicates a clear violation of the equal variance assumption.

2. Normality: The Q-Q plot reveals deviations from the theoretical normal distribution, particularly at both tails. Residuals at the extreme ends diverge significantly from the reference line, suggesting that the residuals do not follow a normal

```
Call:
lm(formula = MSRP ~ Year + Engine.HP + Engine.Cylinders + Number.of.Doors +
    highway.MPG + city.mpg + Popularity + Crossover + `Factory Tuner` +
    Performance + `High-Performance` + Luxury + `Flex Fuel` +
    Exotic + Hatchback + Diesel + Hybrid, data = training_data)

Residuals:
    Min      1Q  Median      3Q     Max
-122389   -6476     895    6623  117412

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.420e+06  3.871e+05  -3.668 0.000271 ***
Year               6.908e+02  1.947e+02   3.548 0.000426 ***
Engine.HP          1.622e+02  2.436e+01   6.660 7.48e-11 ***
Engine.Cylinders   2.853e+03  1.146e+03   2.490 0.013115 *
Number.of.Doors   -5.120e+02  1.179e+03  -0.434 0.664254
highway.MPG        1.950e+02  4.429e+02   0.440 0.659858
city.mpg           2.104e+02  4.116e+02   0.511 0.609353
Popularity        -1.148e-01  6.153e-01  -0.187 0.852066
Crossover          1.509e+03  2.827e+03   0.534 0.593811
`Factory Tuner`   -1.733e+03  4.438e+03  -0.391 0.696298
Performance       -3.118e+03  2.842e+03  -1.097 0.273239
`High-Performance` -4.758e+03  4.774e+03  -0.997 0.319403
Luxury             1.054e+04  2.372e+03   4.445 1.09e-05 ***
`Flex Fuel`       -4.916e+03  3.409e+03  -1.442 0.150021
Exotic             1.337e+05  5.230e+03  25.572  < 2e-16 ***
Hatchback          1.414e+03  3.564e+03   0.397 0.691730
Diesel             6.595e+03  6.220e+03   1.060 0.289516
Hybrid             1.001e+03  6.640e+03   0.151 0.880192
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20130 on 482 degrees of freedom
Multiple R-squared:  0.8367,    Adjusted R-squared:  0.831
F-statistic: 145.3 on 17 and 482 DF,  p-value: < 2.2e-16
```

Fig. 2: Model Summary

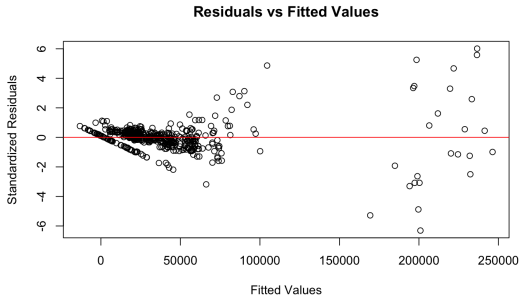

Fig. 3: Residual Plot

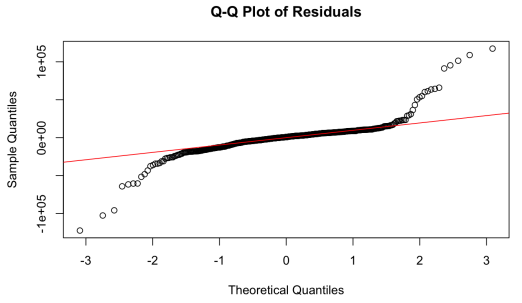

Fig. 4: QQ Plot

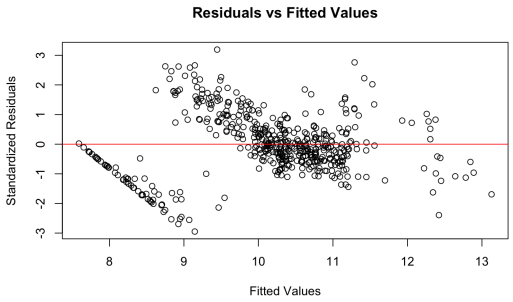

Fig. 5: Residual Plot after Y transformation

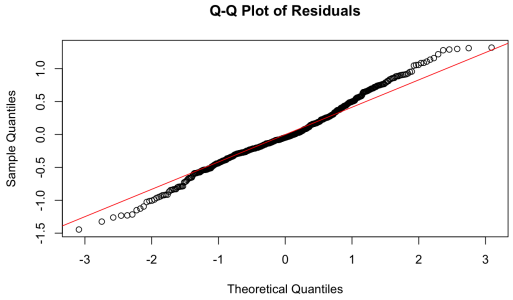

Fig. 6: QQ Plot after transformation

distribution, thus violating the normality assumption.

3. Linearity: As observed in the residuals plot, the residuals are centered around zero across the range of fitted values. This indicates that the linearity assumption holds, as there is no systematic pattern in the residuals that would suggest a non-linear relationship.

With this in mind, we considered adding in interaction terms, and also box-cox transformation. However, both approaches lead us to worse results. We will be able to see the BP test and Shapiro-Wilk normality test result in Fig 7 and 8.

C. Final Model

After all approaches, we believe the initial model after log transformation on dependent variable's performance was the best, and an acceptable model assumption to continue. We then use the stepwise regression method to select important variables in the model and diagnose the optimized model.

```
        studentized Breusch-Pagan test

data:  model_interaction_step
BP = 193.74, df = 8, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  residuals(model_interaction_step)
W = 0.97492, p-value = 1.48e-07
```

Fig. 7: Model Assumption Result After Adding in Interactions

```
        studentized Breusch-Pagan test

data:  model_boxcox
BP = 206.38, df = 17, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  residuals(model_boxcox)
W = 0.5567, p-value < 2.2e-16
```

Fig. 8: Model Assumption Result After Utilizing Box-Cox

```
lm(formula = log(MSRP) ~ Year + Engine.HP + Engine.Cylinders +
    highway.MPG + city.mpg + Performance + Luxury + Exotic, data = training_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.41114 -0.29174 -0.04479  0.28724  1.35798

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.024e+02  8.317e+00 -24.339  < 2e-16 ***
Year              1.054e-01  4.170e-03  25.279  < 2e-16 ***
Engine.HP         1.310e-03  4.559e-04   2.874  0.00423 **
Engine.Cylinders  8.413e-02  2.567e-02   3.278  0.00112 **
highway.MPG      -2.265e-02  9.660e-03  -2.345  0.01944 *
city.mpg          1.368e-02  8.925e-03   1.532  0.12606
Performance       1.181e-01  6.143e-02   1.923  0.05502 .
Luxury            2.970e-01  5.637e-02   5.269 2.06e-07 ***
Exotic            1.350e+00  1.217e-01  11.092  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4978 on 491 degrees of freedom
Multiple R-squared:  0.804,     Adjusted R-squared:  0.8008
F-statistic: 251.8 on 8 and 491 DF,  p-value: < 2.2e-16
```

Fig. 9: Final Model Summary

Stepwise regression was performed on the log transformed model to optimize the model. By combining forward selection and backward elimination methods, stepwise regression automatically identifies and eliminates variables that contribute little or are not significant to the model, while retaining significant variables to improve model simplicity and predictive performance. The AIC of the model after stepwise regression is 732.2841, which is significantly lower than the initial model's 743.4987, indicating that stepwise regression improves the model's predictive ability and goodness of fit while reducing model complexity.

Now, we will head back to the model assumption:

1. Equal Variance: Compared to the initial model, the updated model shows improvements in heteroscedasticity, as the residuals in the residuals plot are slightly more evenly spread. However, there is still a noticeable pattern where residuals at lower fitted values are tightly clustered, while residuals at higher fitted values exhibit increased variability. This indicates a remaining violation of the equal variance

	df <dbl>	AIC <dbl>
model	19	743.4987
model_step	10	732.2841

Fig. 10: AIC Comparison Before and After

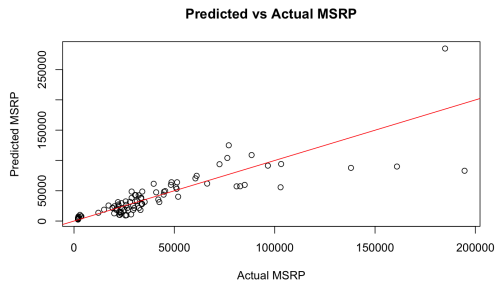

Fig. 11: Test Result

assumption. The unequal spread of residuals may stem from omitted interactions or non-linear relationships that are not captured by the current model.

2. Normality: The Q-Q plot for the updated model shows some improvement in terms of the overall alignment of residuals with the reference line, particularly in the middle range. However, the residuals at both tails continue to deviate significantly from the line, suggesting that the normality assumption is still violated. This could be due to the presence of outliers or the influence of extreme values that distort the residual distribution, or potentially due to skewed predictor variables.

3. Linearity: The residuals plot for the updated model indicates that the linearity assumption holds, as the residuals are still centered around zero and exhibit no systematic patterns across the range of fitted values.

The updated model shows significant progress in reducing heteroscedasticity and improving the normality assumption. These improvements are due to the screening of variables through stepwise regression, eliminating irrelevant variables, thereby simplifying the model structure. Year, Engine.HP, Engine.Cylinders, Luxury and Exotic are the independent variables that have the most significant impact on log(MSRP). These factors usually reflect the performance, brand positioning and market positioning of the car.The overall model performed well, adjusted R-squared is 0.8008: the model is able to explain approximately 80% of the variability in log(MSRP). The residual standard error is 0.4978, indicating that the model deviates less from the actual data. However, there are still slight heteroscedasticity and normality problems in the residuals in the model, which may be caused by the presence of high leverage points in the model and inappropriate functional forms of some predictor variables. All results are being shown in Fig 8-10.

IV. Output and Results

With the previous set aside test_data, we've gain the result:
Mean Absolute Error (MAE): 12176.43
Mean Squared Error (MSE): 471190437
Root Mean Squared Error (RMSE): 21706.92
R-squared (R^2): 0.6521
The overall fit has also been shown in Fig 11.

To further evaluate the performance of the model, a row of data was randomly selected from the test data set based on the model we finally selected, and the model was used to predict the log(MSRP) of the row. The predicted log(MSRP) is then converted back to the original MSRP value for interpretation. The predicted MSRP was 42,465.15, while the actual MSRP for this sample was 32,210. The absolute error between the predicted value and the actual value is 10255.15, and the percentage error is 31.84%.

These results indicate that the model's predictions on this test sample deviate significantly from the actual values, and the higher percentage error reflects the poor performance of the model on this sample. Possible reasons for this deviation include:

1. Model Limitations: The model may not fully capture the underlying relationship between the predictors and the MSRP, potentially due to omitted variables or non-linear effects.

2. Outlier Influence: The selected sample might be an outlier, with features or characteristics that differ significantly from the majority of the training data.

3. Test Sample Complexity: The test sample could belong to a category (e.g., a rare or highly specific vehicle type) that the model was not adequately trained to predict.

VI. Discussion of the findings and inferences

This article analyzes a car pricing model with log(MSRP) as the dependent variable and summarizes the main findings and model limitations obtained through initial modeling, transformation, and optimization. The following is a detailed discussion:

1. Improvement of the initial model by logarithmic transformation revealed that the model violated the assumptions of equal variance and normality. By logarithmically transforming the dependent variable MSRP, the model performance is improved.

2. By using the stepwise regression method to filter the variables of the model, and by combining forward selection and backward elimination, variables that have a significant impact on log(MSRP) are retained and irrelevant variables are eliminated. The final model has an improved goodness of fit and is more complex. degree decreased.

3. The optimized model achieved an adjusted R^2 of 0.8008, explaining approximately 80% of the variability in log(MSRP). The residual standard error is 0.4978, indicating that the model fits the data well. However, the prediction results on the test samples showed significant errors, possibly because the characteristics of the test samples (such as rare vehicle types) were not fully learned or captured by the model.

V. Limitations and Further Questions

A. Research Limitations

1. Although the model has been improved to a certain extent after optimization through logarithmic transformation and stepwise regression, the model still has slight heteroscedasticity, and the normality assumption is not fully established. These issues reflect structural limitations of current linear regression model frameworks.

2. High absolute and percentage errors in randomly selected test samples indicate that the model has limited ability to generalize to new data. This problem is particularly evident in certain categories, such as rare or special vehicle types, which may be underrepresented in the training data.

3. The data selected may have limitations because there are many factors that affect MSRP in real life, and the data selected for this study do not yet include all factors, such as regional price differences, fluctuations in consumer demand, and more detailed car specifications.

B. Suggestion For Future Studies

1. How can the model be enhanced to better address heteroscedasticity and outliers?

The persistence of heteroskedasticity and deviations from normality suggests possible deficiencies in current linear modeling approaches. Future research can explore the use of advanced statistical techniques such as weighted least squares or robust regression, or methods based on machine learning and deep learning (such as random forests or gradient boosting) to build models to ensure that the assumptions of the model can be used to the greatest extent possible. be satisfied.

2. How robust is the model to evolving datasets and market conditions?

With the rapid development of the automotive industry, market conditions and data characteristics are constantly changing. For example, the popularity of new energy electric vehicles, the development of autonomous driving technology, and consumer preferences for environmentally friendly and intelligent vehicles have significantly changed the structure and dynamics of the automobile market. These changes may lead to a decrease in the predictive performance of existing models, as the models may not be able to capture important features of emerging trends or changes. Therefore, future research needs to evaluate the robustness of the model in the face of these changes, that is, the stability and applicability of the model under different data sets and market conditions.