Refining the Regression Model: Exploring the Automobile Dataset, Variable Selection, Transformation, Validation, and Car Price Prediction

Quang Duy Tran, Maiqi Zhang, Baixue Zhang

I. Introduction

A. Data Exploration

The Automobile Dataset [1], sourced from the 1985 *Ward's Automotive Yearbook*, was contributed to the UC Irvine Machine Learning Repository on August 15, 1987. The data repository provided 205 brand-new cars (each observation represents a specific car model) in 1985 with 26 variables: 10 categorical and 15 continuous. From these 26 variables, our project focuses on predicting car price as the response variable. Here is a brief description of all the variables in the dataset:

Description of categorical variables:

Symboling: insurance risk rating, ranging from -3 to 3 (-3 = safer, 3 = riskier)

Make: manufacturer of the car (summarized in **Figure 1a**)

Fuel type: type of fuel used (185 gas, 20 diesel)

Aspiration: type of induction; the process of introducing fuel and air into the vehicle's engine combustion chamber for ignition (168 standard, 37 turbo)

Number of doors: number of doors (114 four, 89 two)

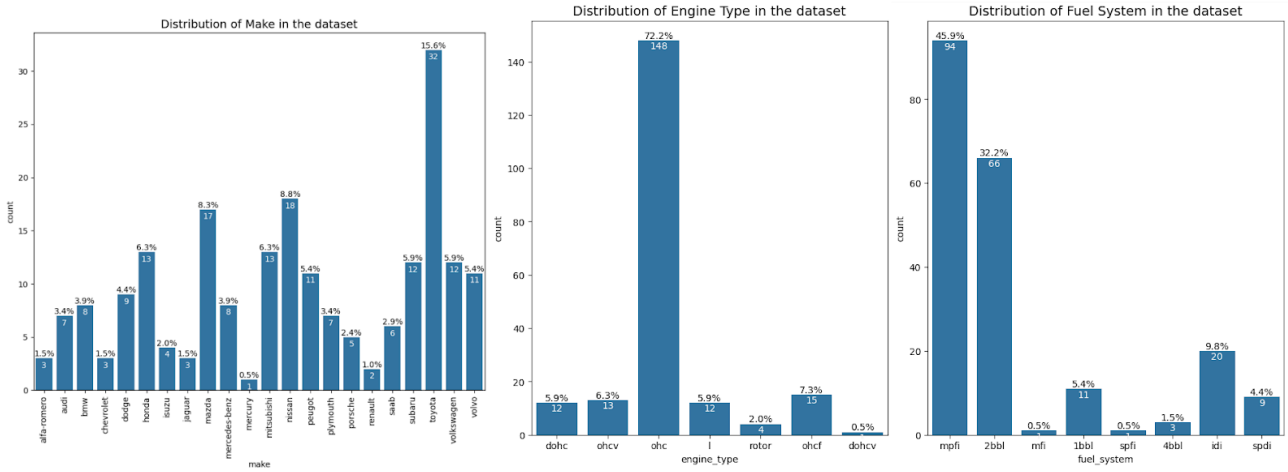
Body style: car design type (96 sedans, 70 hatchbacks, 25 wagons, 8 hardtops, 6 convertibles)

Drive wheels: type of drivetrain (120 front-wheel drive, 76 rear-wheel drive, 9 four-wheel drive)

Engine location: position of the engine (3 rear, 202 front)

Engine type: configuration of the engine (summarized in Figure 1b)

Fuel system: fuel injection system type (summarized in Figure 1c)



a. Distribution of Make **b. Distribution of Engine Type** **c. Distribution of Fuel System**

Figure 1: Distribution (in count and percentage) of **a.** Make; **b.** Engine Type; **c.** Fuel System

Description of continuous variables:

Normalized losses: Relative average loss payment per insured vehicle; values ranging from 65-256; unit: unitless

Wheelbase: distance between the front and rear axles; values ranging from 86.6 to 120.9; unit: inches

Length: overall length of the car; values ranging from 141.1 to 208.1; unit: inches

Width: overall width of the car; values ranging from 60.3 to 72.3; unit: inches

Height: overall height of the car; values ranging from 47.8 to 59.8; unit: inches

Curb weight: weight of the car without passengers or cargo; values ranging from 1488 to 4066; unit: pounds

Engine size: engine displacement; values ranging from 61 to 326; unit: cubic inches

Bore: diameter of the cylinder bore; values ranging from 2.54 to 3.94; unit: inches

Stroke: distance the piston travels in the cylinder; values ranging from 2.07 to 4.17; unit: inches

Compression ratio: ratio of the cylinder's volume at the bottom versus the top of the compression stroke; values ranging from 7 to 23; unit: unitless

Horsepower: engine power output; values ranging from 48 to 288; unit: horsepower

Peak-rpm: engine speed at peak horsepower; values ranging from 4150 to 6600; unit: revolutions per minute

City-mpg: fuel efficiency in city driving; values ranging from 13 to 49; unit: miles per gallon

Highway-mpg: fuel efficiency in highway driving; values ranging from 16 to 54; unit: miles per gallon

Price: the selling price of the car; values ranging from 5118 to 45400; unit: USD

After examining the dataset, we found no duplicate observations. However, 46 out of 205 observations (approximately 22%) contain NaN values. In which, 41 NaN for *Normalized Losses*, 2 NaN for *Number of doors*, 4 NaN for *Bore*, 4 NaN for *Stroke*, 2 NaN for *Horsepower*, 2 NaN for *Peak-rpm*, and 4 NaN for *Price*. Since our objective is to predict *Price*, the four observations with NaN values for this variable were removed. The NaN values in predictor variables were left unchanged, as these may be excluded during variable selection. Further details on handling NaN values are provided in the Variable Selection section.

To visualize the Price distribution, we grouped it into $1,000 bins. **Figure 2** presents a histogram and a box plot highlighting quartiles and the mean. Most cars are priced between $5,000 and $15,000. The distribution is right-skewed, with a long tail toward higher prices, suggesting a small number of expensive cars disproportionately impact the upper range.

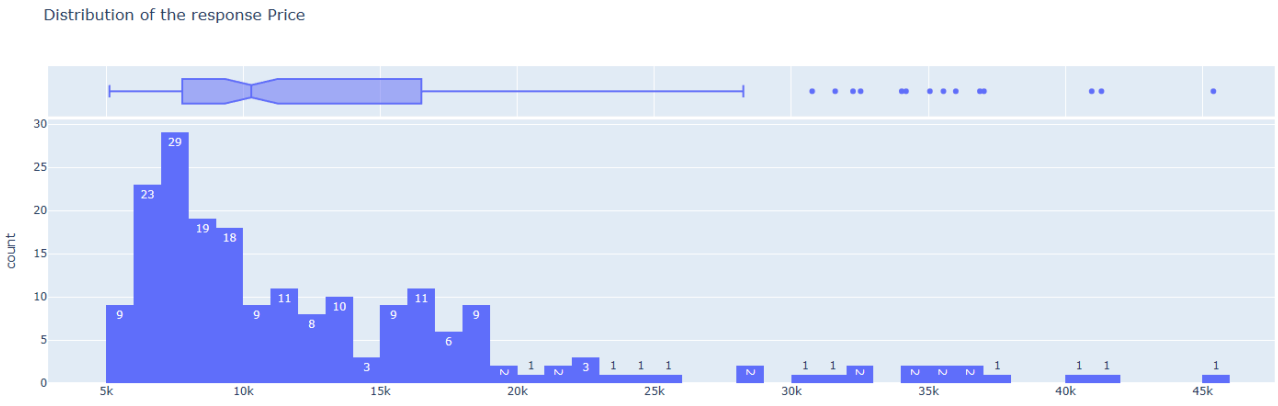


Figure 2: Distribution of the response Price; grouped into $1000 bins.

B. Research Objectives

Understanding the key factors influencing car prices helps consumers make informed purchasing decisions within their budget. A model that explains this data can save customers time, effort, and money. For manufacturers and dealerships, this information enables more effective pricing

strategies to stay competitive and maximize profitability. Based on these insights, our project aims to address the following two questions:

1. Which predictors contribute significantly to the price of a brand-new car?
2. How well can we predict the price of a brand-new car on the smaller subset of predictors?

To answer these questions, we will use forward, backward, and stepwise variable selection methods to identify significant predictors of a new car's price. We will compare R^2, Adjusted R^2, MSRes, and Mallows's Cp to assess model performance. Additionally, we will apply Box-Cox transformation to optimize the response variable (price) and adjust the regression models accordingly. Finally, the best model will undergo evaluation through double cross-validation.

II. Methodology

A. Variable Selection

Using all 25 predictor variables can be impractical since not all of them are significant. Our first step in handling the dataset is to identify a smaller subset of predictors that effectively model the response variable—addressing the variable selection problem.

This process consists of the following steps:

1. Visual Inspection: Identify missing response variables (y), categorical variables with only one level, and variables with a significant amount of missing data.
2. Address Multicollinearity: Identify multicollinearity issues and determine which variables are highly correlated and select representative variables to mitigate this issue.
3. Perform variable selection in R using two significance levels, 0.05 and 0.1, compare the resulting models. Apply forward selection, backward selection, and stepwise selection methods to evaluate and find the optimal predictors.
4. Examine Exhaustive Selection: skipped due to the large number of possible subsets
5. Final Variables: Process a stepwise variable selection again by R. Compare the results and finalize the selection of variables.

B. Residual Analysis

1. Model Normality and Homoscedasticity Check

A Q-Q (quantile-quantile) plot is a tool to evaluate if a dataset fits normal distribution. A residual plot is a graphical tool for identifying potential problematic points and checking homoscedasticity in the model (chapter 4.2) [2].

2. Outlier detection

Outliers in a dataset are extreme observations that deviate significantly from other data points and can greatly affect the outcomes of the regression model. Standardized residuals are used to identify outliers. Points with standardized residuals exceeding 3 are regarded as potential outliers (Chapter 4.2) [2]. After identifying outliers, additional tests will be conducted to determine how their inclusion or exclusion affects the model's price prediction performance.

C. Leverage and Influential

1. Leverage point identification

A leverage point is an observation with an unusual value in the predictor variable space. The influence of an observation on the model is assessed using the hat matrix (H). The diagonal elements (h_{ii}) of the hat matrix, derived from the predictor matrix X, measure the leverage.

These diagonal values are standardized measures that indicate how distant an observation is from the centroid of the x-values. Observations whose leverage value (h_ii) exceeds twice the average leverage are considered significant leverage points (chapter 6.2) [2].

2. Influential point detection

Influential points are observations that cause large changes in the model's coefficient betas when being removed from the dataset. Cook's distance is a metric to identify these points. Cook's distance greater than 1 will be regarded as influential, because it shows that the point could move the estimated coefficients outside a 50% confidence region for the parameters, based on the complete dataset (chapter 6.4) [2].

D. Transformation Implementation

To identify proper transformations, Box-Cox and visual inspections will be applied. Box-Cox is to identify the optimal lambda for normalizing residuals and stabilizing variance, Visual inspection is to visually check residual plots for patterns. Further, chosen transformations will be assessed based on improvements in Adjusted R-squared and reductions in the Standard Error of Estimates. QQ plots and residual plots will also be used to check normality after transformation.

E. Data Validation

To ensure the fitted model accurately represents the true relationship, we perform model validation. Specifically, we use a data-splitting technique to reserve a portion of the data for testing the model's predictive performance. This approach, a form of cross-validation, involves repeatedly splitting the data into estimation and prediction sets. We repeat this process 1,000 times, calculating the mean squared prediction error (MSP) for each iteration. The coefficients and MSP values from all iterations are stored and visualized as histograms. Ideally, the distribution of mean square error should approximate normality. And most of the MSP values should be close to zero.

III. Results

A. Variable selection

1. Visual Inspection

By looking and searching through the dataset, we detect the following issues in our dataset:

1. There are 4 observations with missing values for the dependent variable price, in rows 10, 45, 46, and 130. Since these observations do not contribute to our goal of building a model with price as the dependent variable, we will exclude them from the dataset.
2. The variable engine_location has 3 observations classified as rear and over 190 classified as front. Notably, the 3 cars with rear engine locations have relatively high prices: $32,528, $34,028, and $37,028, compared to the overall mean car price of $13,207.13.
3. The variable normalized_losses has 20% missing values, which could present challenges during the variable selection process.

2. Address Multicollinearity

To build a robust and interpretable model, addressing multicollinearity among the predictor variables is crucial. This was done by calculating the Variance Inflation Factors (VIFs) and Pearson correlation coefficients.

a. A Variance Inflation Factor (VIF) analysis was conducted on all numerical predictor variables to identify those with high correlations. The analysis reveals extremely high VIF values for several variables: curb_weight (16.05), city_mpg (26.42), and hwy_mpg

(24.43). These elevated VIF values suggest that these variables are strongly correlated with one or more other predictors in the dataset.

b. There are two distinct groups of variables that exhibit high correlations. The first group describes the size of a car and includes variables length, width, wheel_base (distance between front and rear wheels), and curb_weight. These variables are strongly interconnected, indicating that they collectively describe the overall dimensions and weight of a vehicle. The second group relates to engine attributes and consists of engine_size, horsepower, highway_mpg, and curb_weight. Notably, curb_weight appears in both groups. To avoid multicollinearity, we exclude curb_weight. Subsequently, we select one representative variable from each group based on the result of forward, backward, and stepwise selection methods in R.

3. Variable selection by R

When fitting the model with all predictor variables, an issue arose: engine_location and normalized_losses could not be included in the same model. This problem occurs because engine_location has only three observations labeled as "rear," while over 190 are labeled as "front." Moreover, the three "rear" observations have missing normalized_losses values. As a result, when both variables are included, R treats engine_location as having only one level and cannot process the model.

To ensure R cooperates, this process was separated into two cases:

1. Case 1: Include engine_location and exclude normalized_losses. Later, include normalized_losses if engine_location is excluded. However, since engine_location was never dropped, normalized_losses could not be included.
2. Case 2: Include normalized_losses and exclude engine_location. Later, include engine_location if normalized_losses is excluded. Eventually, R excludes normalized_losses, allowing us to include engine_location.

This approach ensures the model can be processed without errors while addressing the conflicting variables.

In both cases, a forward variable selection, backward variable selection, and a stepwise variable selection were processed. The variables that were excluded by R backward selection and not picked by R forward selection were excluded.

After conducting variable selection in R, the result in Case 1 is: number_of_doors, city_mpg, horsepower, drive_wheels, and symboling were excluded in that order at a 0.1 significance level, and extra variables hwy_mpg, Compression_ratio, Fuel_system, fuel_type were excluded at a 0.05 significance level. Among the highly correlated variables, width was selected to represent the car's size (length, wheel_base), as it was the fifth variable chosen by R's forward selection and was never excluded during backward selection. Additionally, a clear linear relationship was observed between width and other car size-related variables through plots. For engine attributes(engine_size, horsepower, highway_mpg), engine_size was selected since it was the first variable picked by R's forward selection. Similarly, in Case 2, the excluded variables are symboling, stroke, peak_rpm, normalized_losses, city_mpg, hwy_mpg, fuel_system, bore, and fuel_type at a 0.1 significance level, at extra variables fuel_system and num_of_doors at a 0.05 significance level. In this case, wheel_base and horsepower were chosen to represent car size and engine attributes, respectively based on the sequence of R forward and backward variable selection.

Alpha = 0.05

```
Model 1: price ~ engine_size + make + width + engine_location + peak_rpm +
    aspiration + engine_type + stroke + body_style + height +
    bore
Model 2: price ~ engine_size + make + width + engine_location + peak_rpm +
    aspiration + engine_type + stroke + body_style + height +
    bore + compression_ratio + fuel_system + fuel_type
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    159 589523454
2    152 546850551  7  42672902 1.6945 0.1143
```

Alpha = 0.1

```
Model 1: price ~ make + aspiration + body_style + drive_wheels + engine_location +
    wheel_base + height + engine_type + num_of_cylinders + compression_ratio +
    horsepower
Model 2: price ~ make + aspiration + num_of_doors + body_style + drive_wheels +
    engine_location + wheel_base + height + engine_type + num_of_cylinders +
    fuel_system + compression_ratio + horsepower
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    156 605455638
2    148 576265522  8  29190116 0.9371 0.4879
```

Figure 3: ANOVA comparison for significance levels of 0.05 and 0.1 in both Case 1 and Case 2.

At the end of this step, each case results in two models, with additional predictors included at a 0.1 significance level. However, the model comparison in R (Figure 3) shows that these extra predictors are not significant, so we proceed with the variables picked with a 0.05 significance level.

4. Final variables

The next step involves performing a final variable selection using a second round of stepwise selection in R. In both cases, the process yields two models, and a comparison in R (Figure 4) indicated that the additional predictor(s) were not statistically significant. Therefore, we proceeded with a 0.05 significance level.

```
Model 1: price ~ make + body_style + engine_location + wheel_base + engine_type +
    num_of_cylinders + compression_ratio + horsepower
Model 2: price ~ make + body_style + drive_wheels + engine_location +
    wheel_base + height + engine_type + num_of_cylinders + compression_ratio +
    horsepower
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    162 633465830
2    159 606324718  3  27141112 2.3725 0.0724 .
```

```
Model 1: price ~ engine_size + make + width + engine_location + peak_rpm +
    aspiration + engine_type + body_style + bore
Model 2: price ~ engine_size + make + width + engine_location + peak_rpm +
    aspiration + body_style + bore
  Res.Df       RSS Df Sum of Sq      F  Pr(>F)
1    161 591024164
2    164 619087193 -3 -28063028 2.5482 0.05777 .
---
```

Figure 4: ANOVA comparison for significance levels of 0.05 and 0.1 in both Case 1 and Case 2.

Our final variable selection models are:

Model Start with including engine_location and exclude normalized_losses:

Model 1: price ~ engine_size + make + width + engine_location + peak_rpm + aspiration + body_style + bore

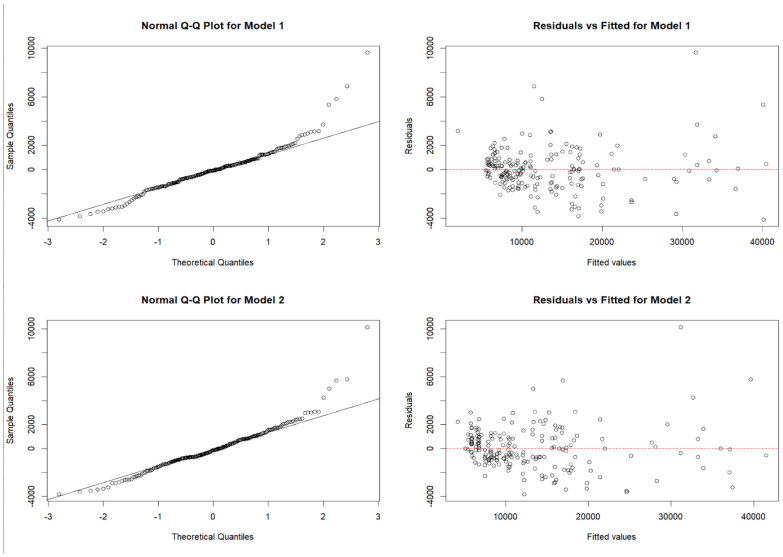
Model Start with including normalized_losses and exclude engine_location:

Model2: price ~ make + body_style + engine_location + wheel_base + engine_type + num_of_cylinders + compression_ratio + horsepower

B. Residual Analysis

1. Q-Q plots and Residual plots

In both model 1 and model 2, the Q-Q plots show that the points mostly adhere to the line, indicating that the residuals are approximately normally distributed. There are tails on both sides. These deviations can be critical as they often indicate the presence of potential outliers.



In model 1, the spread of residuals does not look constant, because there is a varying density of points across the fitted values that forms a funnel pattern. This shows potential issues with heteroscedasticity. A few points lie significantly away from the central cluster, which means they can be potential outliers. In model 2, the variance of the residuals in Model 2 is more consistent across the range of fitted values compared to Model 1, but it still has the funnel pattern, which shows heteroscedasticity. Model 2 also shows several points that deviate significantly from the rest, which could be potential outliers.

2. Outliers

Model 1: 4 data points (ID = 16, 59, 60, 68) are identified as outliers. For # 68, R^2 shows the most noticeable increase from upon removal, and MS_Res decreases substantially. This shows that #68 is affecting the model fit more than the others. Overall, the model appears to be robust. The outlier #68 shows the most substantial impact on both R^2 and MS_Res, showing a closer examination of this data point might be needed. However, none of the outliers cause a significant decrease in the standard errors of the regression coefficients. This shows that these outliers do not largely influence the regression estimates' precision, and the model's parameter estimates are stable with or without these points.

	Adjusted R^2 with the point	Adjusted R^2 without the point	MS_Res with the point	MS_Res without the point	Any Slope that has obvious change on Standard Errors
#16	0.9418388	0.9403236	3174806	3191119	No
#59	0.9418388	0.9418388	3174806	3174806	No

#60	0.9418388	0.9422036	3174806	3163245	No
#68	0.9418388	0.9426322	3174806	3106651	No

Model 2: 3 data points (ID = 16, 72, 199) are considered as outliers. For #199, there is an increase in R^2, and for both #72 and #199, there is a decrease in MS_Res. This shows that #199's removal does lead to an enhancement of the model. Overall, the model is relatively robust, but further examination of outlier #199 could be useful to enhance model performance.

	Adjusted R^2 with the point	Adjusted R^2 without the point	MS_Res with the point	MS_Res without the point	The Slope that has obvious decrease on Standard Errors
#16	0.9385753	0.9369813	3183245	3198432	No
#72	0.9385753	0.9377706	3183245	3124377	No
#199	0.9385753	0.9387636	3183245	3179110	No

C. Leverage and Influential

1. Leverage

Model 1: The data points identified as leverage points in model 1 are #1, #2, #3, #18, #19, #20, #43, #44, #47, #66, #69, #119, #120, #121, #122, #125. Among all the leverage points, #66 stands out as having the biggest impact on the model. With the removal of this point, the adjusted R^2 changes by approximately 0.369%, which is the most significant increase among all leverage points. Additionally, the beta value for the predictor makemercury changes by -174.98%, showing a change in the coefficient when this outlier is excluded in model 1. Overall, #66 is a critical leverage point that heavily influences the model's statistics and regression coefficients. It is necessary to do further analysis of this point for better price prediction performance.

	AdjR2 with the Point	AdjR2 without the point	Predictor Estimator that is Changed the Most	AdjR2 Change Percentage (%)	Estimator Change Percentage (%)
#1	0.9418388	0.9421302	makemercury	0.03	392.39
#2	0.9418388	0.9420686	makepeugot	0.02	-110.32
#19	0.9418388	0.9429675	makechevrolet	0.12	245.46
#20	0.9418388	0.9418621	makechevrolet	0.002	-117.15
#66	0.9418388	0.9453205	makemercury	0.369	-174.98

Model 2: The data points identified as leverage points in model 2 are #1, #2, #3, #4, #18, #19, #20, #43, #44, #47, #70, #73, #123, #124, #125, #126, #186. Among all the leverage points, #2 stands out as having the biggest impact on the model. With the removal of this point, the adjusted R^2 changes by approximately 7.73%, which is the largest increase observed among all leverage points. Additionally, the beta value for the predictor makeaudi changes by -10237.97%, indicating a drastic change in the regression coefficient when this outlier is excluded. Overall, #73 is a critical leverage point, having an unusual value in the predictor space that heavily

influences the model's statistics. It is important to do further analysis of this point for better predictive performance.

	AdjR2 with the Point	AdjR2 without the point	Predictor Estimator that is Changed the Most	AdjR2 Change Percentage (%)	Estimator Change Percentage (%)
#2	0.9385753	0.9393009	makeaudi	7.73	-10237.97
#3	0.9385753	0.9390276	makeaudi	4.82	17289.00
#4	0.9385753	0.9386148	makeaudi	4.21	-9122.59
#19	0.9385753	0.9385647	horsepower	-0.11	6800.19
#73	0.9385753	0.9368719	makeaudi	-18.14904103	-6541.27

2. Influential Points Detection

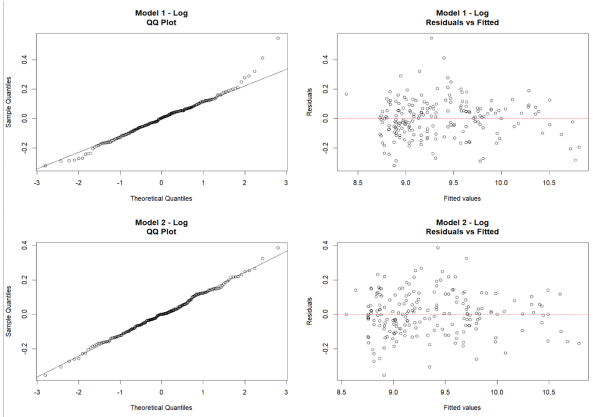
No observations are identified with a Cook's distance greater than 1. This result suggests that the model's coefficients are relatively stable and not overly sensitive to the removal of any single observation in terms of Cook's distance.

D. Transformation

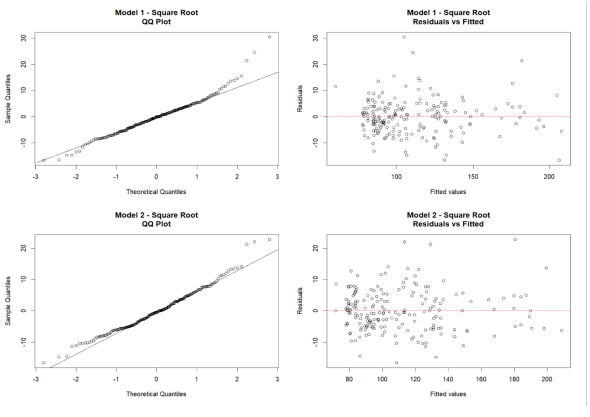
The estimated optimal lambda value from the boxcox plot is 0.303 for Model 1 and 0.2222 for Model 2. We choose 0 for both models by rounding it to the nearest integer, which suggests log transformation. Additionally, the residual plot shows a funnel pattern from both models, which means the variance of the residuals is not constant but increases with the price. This pattern is observed when y follows a distribution where the variance is a function of the mean (Poisson-distributed data). To solve this issue, square root transformation became another reasonable choice, because it equalizes the variance of the residuals across different levels of the response, effectively reducing heteroscedasticity (Chapter 5.2) [2]. So both log transformation and square root transformation will be applied.

	Model	Transformation	R_squared	Adj_R_squared	SE_Original_Scale
1	Model 1	Original	0.9414	0.9235	1443.658
2	Model 1	Square Root	0.9519	0.9431	1690.497
3	Model 1	Log	0.9364	0.9247	1843.631
4	Model 2	Original	0.9337	0.9122	1546.282
5	Model 2	Square Root	0.9532	0.9428	1677.584
6	Model 2	Log	0.9420	0.9291	1769.988

Log Transformation



Square Root Transformation



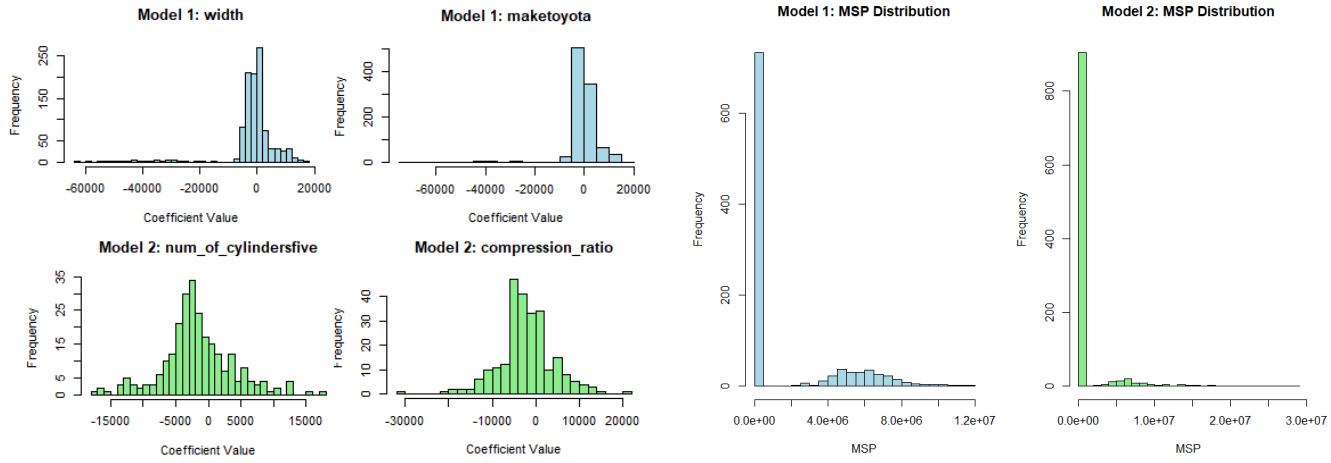
From the QQ plots and residual plots, both transformations demonstrate improvements in terms of normality and constant variance. The log transformation provides a better QQ plot, with data better aligned along the fitted line, showing better normality. Conversely, the square root transformation shows more obvious outliers on its QQ plot. The residual plots for both transformations show increased dispersion and a more random spread, showing reduced heteroscedasticity.

The square root transformation significantly improves the R^2 to 0.9532 from 0.9337, while the log transformation shows a slight improvement to 0.9420, showing a better fit than the original but not as strong as the square root transformation. Similar to R^2, the adjusted R^2 for the square root transformation is the highest at 0.9428. The log transformation shows small improvement. Both transformations in Model 2 show an increase in standard error, with the log transformation showing a slight decrease compared to the square root, yet still higher than the original. This can be caused by potential outliers, because there still appears to be some potential outliers from the QQ plot after transformation especially for square root transformation.

Conclusively, while both transformations mitigate certain levels of model inadequacies, the square root transformation might be preferable due to its better R^2, adjusted R^2, and residual plot.

E. Evaluation

The figures below summarize the coefficients of two betas from each model, while the distribution of prediction mean squared error (MSE) appears normal. Model 1 shows slight skewness, influenced by the inclusion of extreme observations in either the estimation or prediction set. The MSP histogram indicates that model 2 produces more MSP values near zero compared to model 1. This suggests that model 2 outperforms model 1 in predictive accuracy.



IV. Conclusion

The group of factors that influence the prediction of car prices in this dataset are car brand, body style, engine location, wheelbase, engine type, number of cylinders, compression ratio, and horsepower. Some relevant factors were excluded due to multicollinearity. The square root transformation model outperforms the non-transformation model in robustness, with potential improvements achievable by addressing outliers. Overall, the fitted model effectively supports price prediction, aiding consumer decision-making.

References

[1] Schlimmer, Jeffrey. "Automobile." UCI Machine Learning Repository, 1985, https://doi.org/10.24432/C5B01C.

[2] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. 5th ed., Wiley, April 2012.