UM-SJTU Joint Institute

Applied Regression Analysis
(STAT4130J)

MotiveMetrics: Predicting Car Prices In America

Project Report

Summer 2024

Group 6
Name: Tianyi Jiang Student ID: 520370910102
Name: Yi Zhou Student ID: 521010910003
Name: Yujia Gao Student ID: 520370910045
Date: July 28, 2024

1 Background

Car prices are affected by many factors. Understanding what factors are significant in predicting car prices can help car companies set reasonable price for cars with different features. We obtained a car price dataset from Kaggle [1] to build car price models based on several significant factors. This car price dataset is composed of 205 records of different car prices with corresponding influencing factors: `symboling`(insurance risk rating), car brand, fuel type, aspiration, number of doors, car body, drive wheel, engine location, wheelbase, car length, car width, car width, curb weight, cylinder number, type of engine, engine size, fuel system, bore ratio, stroke, compression ratio, horsepower, peak rpm, Mileage(city) and Mileage(highway).

2 Design and Analysis

2.1 Data Cleaning

Before exploring the data, we performed data cleaning, which included handling missing values, removing irrelevant variables, and transforming and correcting certain variables. Results showed no missing values in this dataset. Upon examining the original dataset, we removed the column `car_ID`, because it's merely the index of entries and therefore insignificant. Additionally, we transformed the data in the `CarName` column from containing numerous specific categories, such as "alfa-romero stelvio" or "audi 100 ls", to including only the brand names. This allows us to consider the car brand as a more integrated variable.

2.2 Data Exploration

2.2.1 Basic Visualization and Analysis

We began by examining the overall patterns and distributions of all variables. Their histograms are shown in Figure 1.

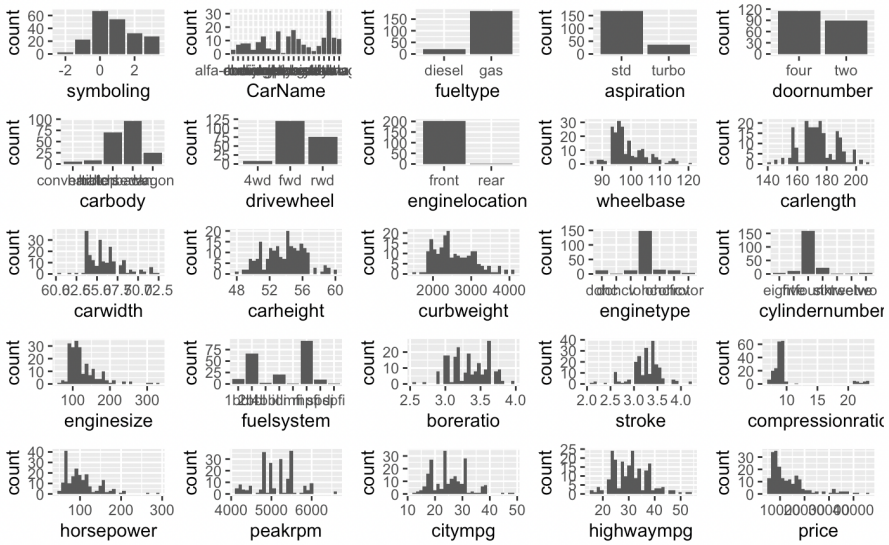


Figure 1: Histograms of all variables in CarPrice dataset

From the histograms, it is evident that many categorical variables, such as `fueltype` and `enginetype`, are unevenly distributed. Some numerical variables, like `compressionratio`, exhibit high concentration with peaks in certain regions while being sparse elsewhere. Based

on these observations, we may consider transforming certain variables to address these extremes. Outliers and influential points will also be evaluated or even be deleted from the dataset to enhance model performances.

In this project, we selected car **price** as our response variable and aimed to build models that best predict it. Its original and log-transformed distributions are shown in Figure 2. The transformation was applied due to the right-skewness of the original distribution. After transformation, it became more normally distributed, providing a solid qualitative foundation for model building.

Figure 2: Car Price distribution plot: original (left) and log-transformed (right)

2.2.2 Correlation Study

To better analyze the correlations between variables, we divided the cleaned dataset into numerical and categorical subsets. For numerical predictors, we performed a pairwise correlation analysis, as shown in Figure 3.

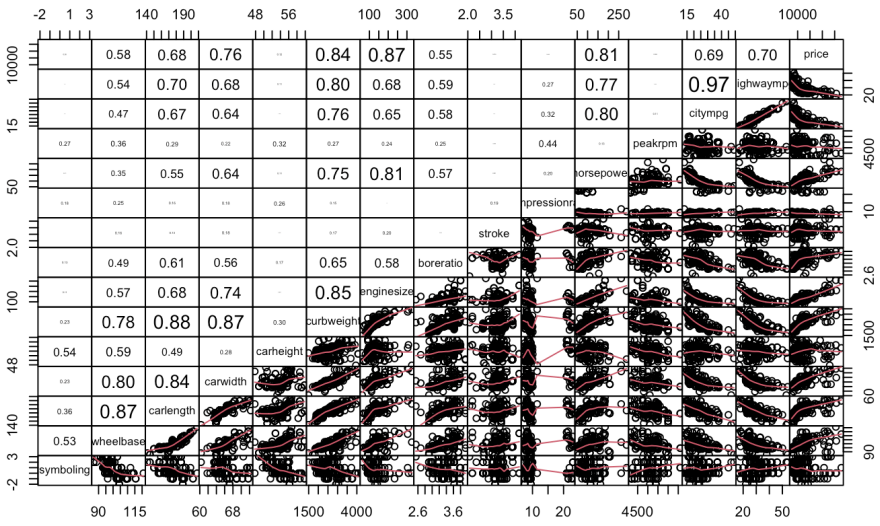


Figure 3: Pairwise and correlation plot of numerical variables

From the analysis, the response variable **price** shows strong correlations with `enginesize`, `curbweight`, `horsepower`, `carwidth` and `highwaympg`. It also has relatively high correlations with `citympg`, `carlength`, `wheelbase`, and `boreratio`. Additionally, three sets of predictor variables exhibit high collinearity, with the first set among `highwaympg`, `citympg`, `horsepower`, `curbweight`, `carlength`, `carwidth`, and `enginesize`; the second set includes `boreratio`, `curbweight`, and `carlength`, and the third set includes `curbweight`, `carlength`, `carwidth`, and `wheelbase`.

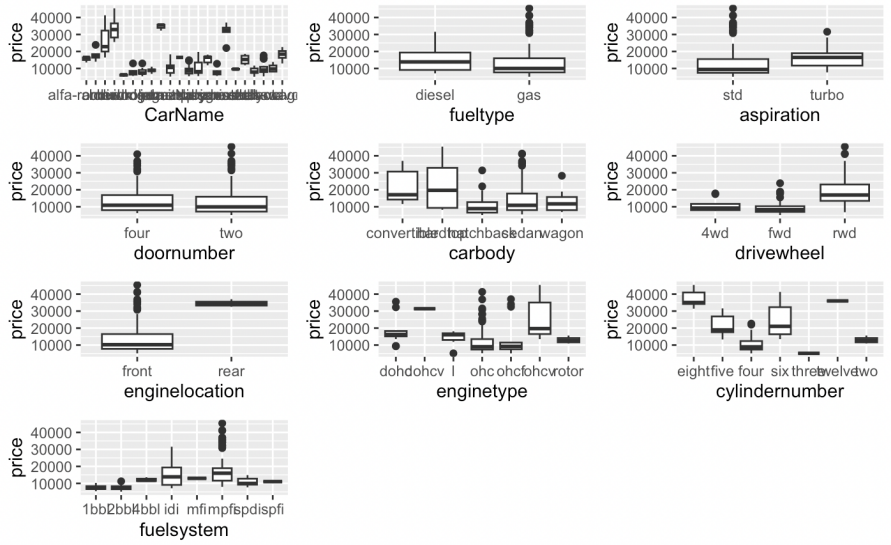


Figure 4: Boxplots of categorical variables

To find significant categorical predictors, boxplots were plotted with respect to the response variable, as shown in Figure 4.

From the boxplots, we observed notable patterns between the the response variable `price` and the following predictors: `CarName`, `enginetype`, `cylindernumber`, `fueltype`, `aspiration`, `carbody`, and `drivewheel`. For variables `CarName`, `carbody`, `enginetype`, and `cylindernumber`, they show significant price variations, likely due to differences in brands and design strategies. The `fueltype` and `aspiration` variables indicate that prices for "diesel" cars are generally higher than those for "gas" cars, while prices for "turbo" cars are higher compared to "std" cars respectively. Additionally, `drivewheel` variable shows that "rwd" cars tend to have higher prices. `enginelocation` are excluded from our consideration for future feature selection because the sample size for one of its subcategory, "rear" , is too small to be conclusive.

2.2.3 Interaction Study

To examine interactions between numerical and categorical variables, we created a multi-panel plot in Figure 5. We selected three numerical variables that are highly correlated with price from previous analysis — `highwaympg`, `horsepower`, and `carlength` — to explore their interactions with the seven previously chosen categorical variables. Due to space constraints, only a selection of plots are included in this report. However, the summarized results about the whole analysis are as follows: `highwaympg` may interact with `enginetype`, `drivewheel`, and `carbody`; `horsepower` may interact with `cylindernumber`, `aspiration`, and `carbody`; `carlength` may interact with `drivewheel`.

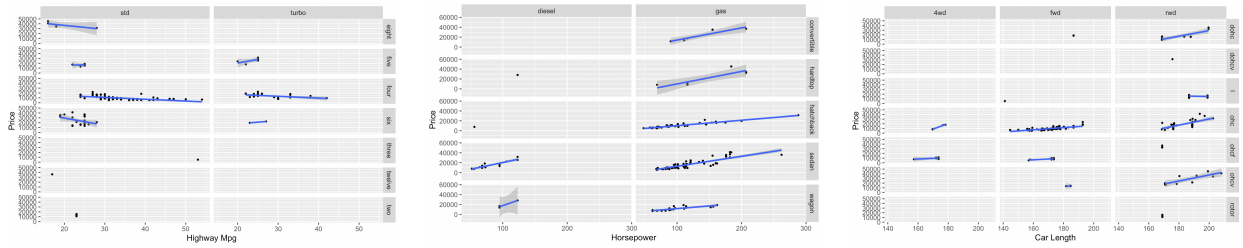


Figure 5: Interaction study of `highwaympg`, `horsepower`, `carwidth` and categorical variables

We also examined pairwise interactions among categorical variables using boxplots. Also

due to space limitations, only a subset of these boxplots is shown in Figure 6. Our analysis suggests possible interactions between `fueltype` and `carbody`, `fueltype` and `aspiration`, `carbody` and `drivewheel`, and lastly, `cylindernumber` and `fueltype`.

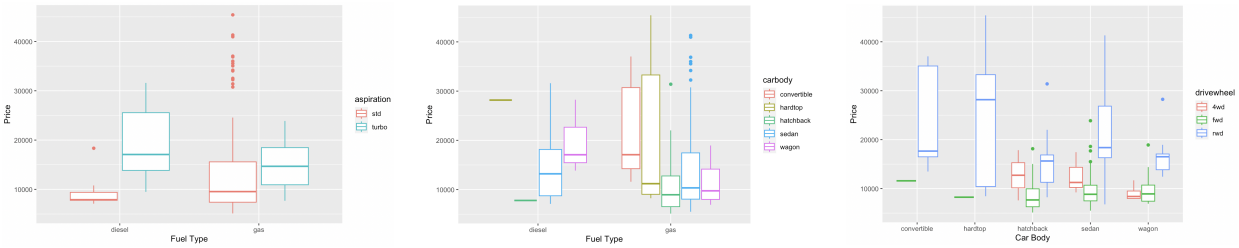


Figure 6: Interaction study among categorical variables

2.3 Model Building

To start, the dataset of 205 rows was divided into a training set (80%) and a test set (20%) for evaluation. Given the significant skewness observed in the response variable **price**, we used the Box-Cox method based on the following model to determine the appropriate power transformation needed.

```
fit <- lm(price~ curbweight + CarName + carheight + enginelocation + wheelbase +
    horsepower, data = car_price)
```

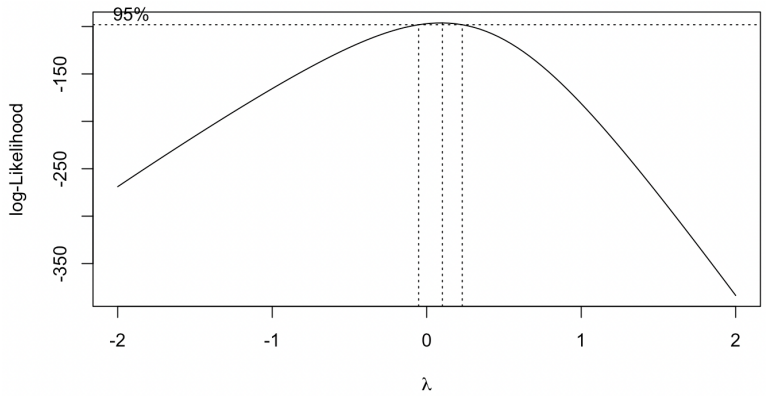


Figure 7: Box-Cox plot for CarPrice

As shown in Figure 7, we selected a log transformation for **price** with $\lambda = 0$. This transformation proved effective, as evidenced by Figure 8, where the residuals are more consistent and randomly distributed across the fitted values after the transformation. This confirms that our initial decision to apply a log transformation to **price** during data exploration was appropriate.

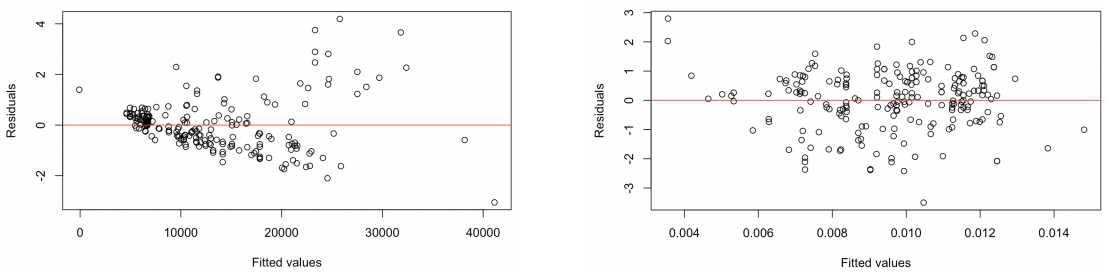


Figure 8: Residual plots before (left) and after (right) log transformation

2.3.1 BIC Model

BIC model was obtained through the built-in step() function in R to identify the optimal combination of predictors that minimized the BIC value without including interaction terms. We chose BIC as the evaluation criterion, to find a simple yet effective model given different kinds of potential predictors. The selected predictor combination is `curbweight`, `CarName`, `carheight`, `enginelocation`, `wheelbase` and `horsepower`. The model takes the form of:

```
model1 <- lm(log(price)~ curbweight + CarName + carheight + enginelocation +
    wheelbase + horsepower, data = train_data)
```

Additionally, in this model, we identified an outlier at the 44^{th} observation. To decide whether to retain or exclude this outlier, we assessed its influence by examining the DFFITS plot, which shows the impact on the mean response when the i^{th} observation is removed.

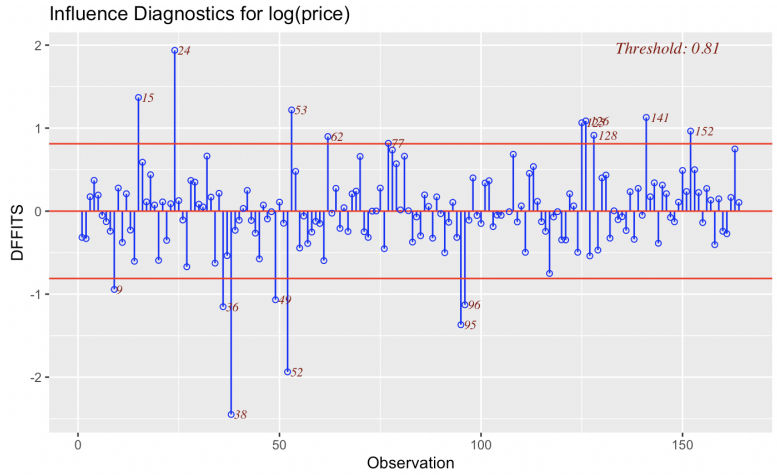


Figure 9: DFFITS plot for model 1

From Figure 9, labelled datapoints are beyond the threshold, meaning they are more or less influential to the model. Since datapoint with index 44 is not an influential point, even as an outlier, it will not affect the regression of the model. Therefore, we kept it in model building.

2.3.2 Self-Chosen Model

After conducting the aforementioned analysis on the dataset, we also selected our own predictors to build models based on previous data exploration analysis. Since on the above model, car brands (`CarName`) explain a lot about car prices and we would like to know more about different product features that influence car prices. Therefore, our chosen predictors include three numerical variables: `highwaympg`, `horsepower`, and `carlength`, as well as two categorical variables: `fueltype` and `carbody`. These predictors have strong correlations with the response variable **price**. The fitted model takes the form of:

```
model2_orig <- lm(log(price)~ highwaympg + horsepower + carlength + fueltype +
    carbody, data=train_data)
```

2.3.3 Self-Chosen Model with Parameter Transformation

From the component plus residual plot of the Self-Chosen model shown in Figure 12, we identified a slight non-linear trend in `highwaympg` and `carlength`. To better align the model with the data's actual moving average, we reduced the power of `highwaympg` and added a polynomial term for `carlength`. This gave us a new model taking the following form:

```
    model2_tran <- lm(log(price)~ sqrt(highwaympg) + horsepower + poly(carlength,
        2) + fueltype + carbody, data=train_data)
```

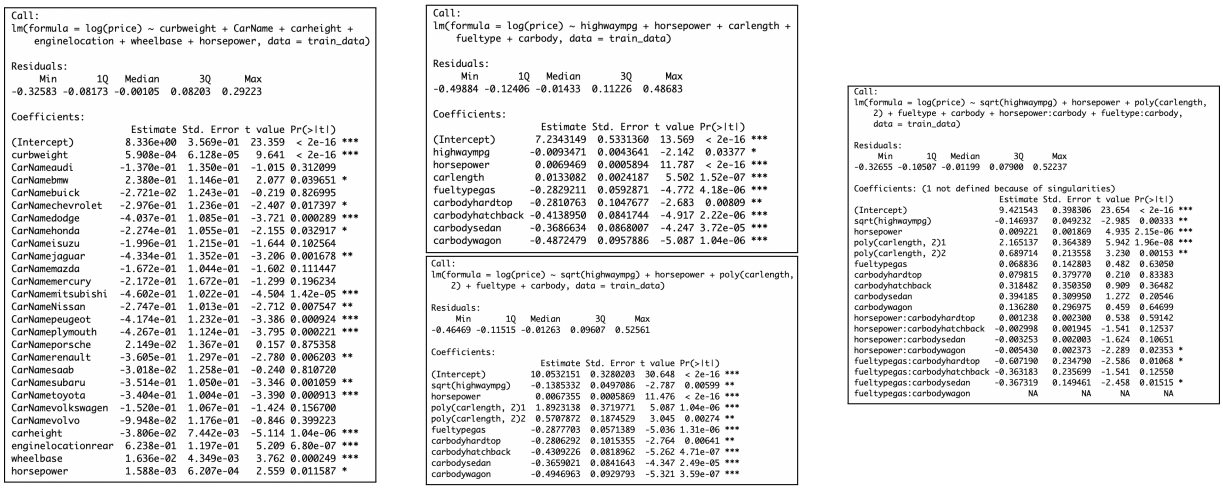
2.3.4 Self-Chosen Model with Interaction

We were also interested in whether including interactions would give better predictions. To identify the relevant interaction pairs, we performed forward selection on the transformed model. From the results, additional interaction terms `horsepower:carbody` and `fueltype:carbody` were selected from the feature selection function, which also aligned with our previous conclusion in interaction study. The fitted model takes the following form:

```
model2_inter <- lm(log(price)~ sqrt(highwaympg) + horsepower + poly(carlength
    ,2) + fueltype + carbody + horsepower:carbody + fueltype:carbody, data=
    train_data)
```

Overall, we obtained four models, with their summary tables shown in Figure 10. Each table includes the predictor's name, its coefficient, and t-test results indicating the significance of each predictor. The null hypothesis for the t-test states that the predictor's coefficient is zero (it should not be included then), while the alternative hypothesis is that the coefficient is not zero. A p-value smaller than 0.05 indicates the predictor is significant.

In the BIC model, most predictors are significant, indicating their importance in explaining the response variable. Some sub-categories of `CarName` are insignificant, which could be due to the small number of entries in these sub-categories or their irrelevance to the model. For the Self-Chosen model and Self-Chosen Transformed model, all predictors are significant, suggesting good model construction and weak collinearity among predictors. In the Self-Chosen Transformed model with Interactions, some predictors became insignificant on their own, though their interaction terms were significant. This occurs because interaction terms assume that the effect of some predictors depends on the levels of others, capturing additional variation that single-variable models could not explain previously. This redistribution of explanatory power causes the significance of some independent effects to decrease. Additionally, we found no outliers in the three self-chosen models, thus eliminating the need to remove outliers and re-fit the models.



```
Call:
lm(formula = log(price) ~ curbweight + CarName + carheight +
    enginelocation + wheelbase + horsepower, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32583 -0.08173 -0.00105  0.08203  0.29223

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.336e+00  3.569e-01  23.359  < 2e-16 ***
curbweight         5.908e-04  6.128e-05   9.641  < 2e-16 ***
CarNameaudi       -1.370e-01  1.350e-01  -1.015 0.312099
CarNamebmw         2.380e-01  1.146e-01   2.077 0.039651 *
CarNamebuick      -2.721e-02  1.243e-01  -0.219 0.826995
CarNamechevrolet  -2.976e-01  1.236e-01  -2.407 0.017397 *
CarNamedodge      -4.037e-01  1.085e-01  -3.721 0.000289 ***
CarNamehonda      -2.274e-01  1.055e-01  -2.155 0.032917 *
CarNameisuzu      -1.996e-01  1.215e-01  -1.644 0.102564
CarNamejaguar     -4.334e-01  1.352e-01  -3.206 0.001678 **
CarNamemazda      -1.672e-01  1.044e-01  -1.602 0.111447
CarNamemercury    -2.172e-01  1.672e-01  -1.299 0.196234
CarNamemitsubishi -4.602e-01  1.022e-01  -4.504 1.42e-05 ***
CarNameNissan     -2.747e-01  1.013e-01  -2.712 0.007547 **
CarNamepeugeot    -4.174e-01  1.232e-01  -3.386 0.000924 ***
CarNameplymouth   -4.267e-01  1.124e-01  -3.795 0.000221 ***
CarNameporsche     2.149e-02  1.367e-01   0.157 0.875358
CarNamerenault    -3.605e-01  1.297e-01  -2.780 0.006203 **
CarNamesaab       -3.018e-02  1.258e-01  -0.240 0.810720
CarNamesubaru     -3.514e-01  1.050e-01  -3.346 0.001059 **
CarNametoyota     -3.404e-01  1.004e-01  -3.390 0.000913 ***
CarNamevolkswagen -1.520e-01  1.067e-01  -1.424 0.156700
CarNamevolvo      -9.948e-02  1.176e-01  -0.846 0.399223
carheight         -3.806e-02  7.442e-03  -5.114 1.04e-06 ***
enginelocationrear  6.238e-01  1.197e-01   5.209 6.80e-07 ***
wheelbase          1.636e-02  4.349e-03   3.762 0.000249 ***
horsepower         1.588e-03  6.207e-04   2.559 0.011587 *
```

```
Call:
lm(formula = log(price) ~ highwaympg + horsepower + carlength +
    fueltype + carbody, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49884 -0.12406 -0.01433  0.11226  0.48683

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.2343149  0.5331360  13.569  < 2e-16 ***
highwaympg       -0.0093471  0.0043641  -2.142  0.03377 *
horsepower        0.0069469  0.0005894  11.787  < 2e-16 ***
carlength         0.0133082  0.0024187   5.502 1.52e-07 ***
fueltypegas      -0.2829211  0.0592871  -4.772 4.18e-06 ***
carbodyhardtop   -0.2810763  0.1047677  -2.683  0.00809 **
carbodyhatchback -0.4138950  0.0841744  -4.917 2.22e-06 ***
carbodysedan     -0.3686634  0.0868007  -4.247 3.72e-05 ***
carbodywagon     -0.4872479  0.0957886  -5.087 1.04e-06 ***
```

```
Call:
lm(formula = log(price) ~ sqrt(highwaympg) + horsepower + poly(carlength,
    2) + fueltype + carbody, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46469 -0.11515 -0.01263  0.09607  0.52561

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       10.0532151  0.3280203  30.648  < 2e-16 ***
sqrt(highwaympg)  -0.1385332  0.0497086  -2.787  0.00599 **
horsepower         0.0067355  0.0005869  11.476  < 2e-16 ***
poly(carlength, 2)1  1.8923138  0.3719771   5.087 1.04e-06 ***
poly(carlength, 2)2  0.5707872  0.1874529   3.045  0.00274 **
fueltypegas       -0.2877703  0.0571389  -5.036 1.31e-06 ***
carbodyhardtop    -0.2806292  0.1015355  -2.764  0.00641 **
carbodyhatchback  -0.4309226  0.0818962  -5.262 4.71e-07 ***
carbodysedan      -0.3659021  0.0841643  -4.347 2.49e-05 ***
carbodywagon      -0.4946963  0.0929793  -5.321 3.59e-07 ***
```

```
Call:
lm(formula = log(price) ~ sqrt(highwaympg) + horsepower + poly(carlength,
    2) + fueltype + carbody + horsepower:carbody + fueltype:carbody,
    data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32655 -0.10507 -0.01199  0.07900  0.52237

Coefficients: (1 not defined because of singularities)
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     9.4215431  0.3983806  23.654  < 2e-16 ***
sqrt(highwaympg)               -0.1469370  0.0492320  -2.985  0.00333 **
horsepower                      0.0092210  0.0018690   4.935 2.15e-06 ***
poly(carlength, 2)1             2.1651370  0.3643890   5.942 1.96e-08 ***
poly(carlength, 2)2             0.6897140  0.2135580   3.230  0.00153 **
fueltypegas                     0.0688360  0.1428030   0.482  0.63050
carbodyhardtop                  0.0798150  0.3797700   0.210  0.83383
carbodyhatchback                0.3184820  0.3503500   0.909  0.36482
carbodysedan                    0.3941850  0.3099500   1.272  0.20546
carbodywagon                    0.1362800  0.2969750   0.459  0.64699
horsepower:carbodyhardtop       0.0012380  0.0023000   0.538  0.59142
horsepower:carbodyhatchback    -0.0029980  0.0019450  -1.541  0.12537
horsepower:carbodysedan        -0.0032530  0.0020030  -1.624  0.10651
horsepower:carbodywagon        -0.0054300  0.0023730  -2.289  0.02353 *
fueltypegas:carbodyhardtop     -0.6071900  0.2347900  -2.586  0.01068 *
fueltypegas:carbodyhatchback   -0.3631830  0.2356990  -1.541  0.12550
fueltypegas:carbodysedan       -0.3673190  0.1494610  -2.458  0.01515 *
fueltypegas:carbodywagon               NA         NA      NA       NA
```

Figure 10: (Partial) summary tables of BIC model (left), self-chosen model (middle top), self-chosen model with parameter transformation (middle bottom), and self-chosen model with interaction (right)

3 Results

3.1 Diagnostic Analysis

For each model, we conducted tests for linearity, normality, and collinearity among predictors to ensure our models meet the assumptions of linear regression. Component-residual plots assessed the relationship between the response variable and each predictor. In this plot, a closer alignment between the fitted line (blue dashed) and the moving average (pink solid) indicates a more linear relationship. Q-Q plots evaluated the normality of residuals; residuals following a straight line suggest they are approximately normally distributed. Finally, collinearity among predictors was checked using the Variance Inflation Factor (VIF), with a threshold of 5 indicating high collinearity.

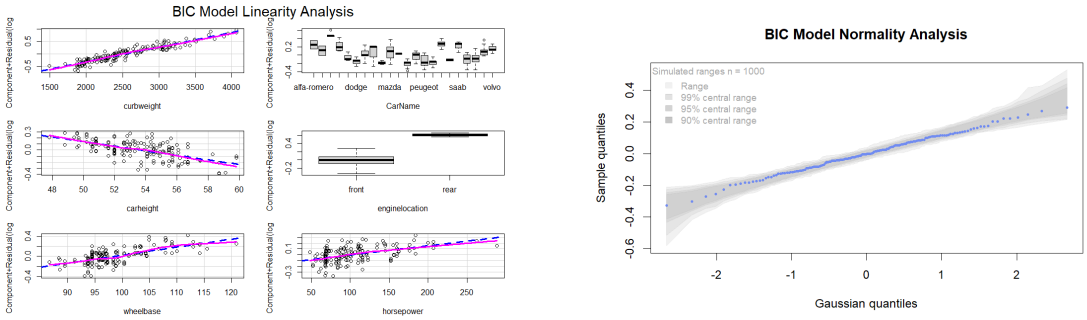


Figure 11: Diagnostic plots for BIC model. crplot checking linearity (left) and Q-Q plot for normality (right).

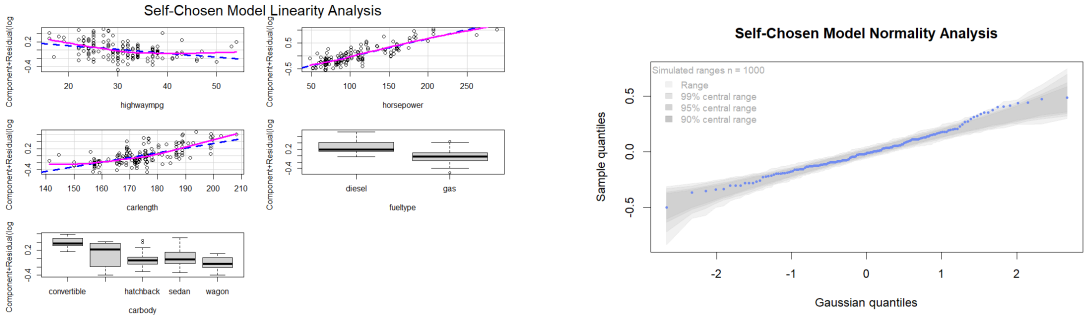


Figure 12: Diagnostic plots for Self-Chosen model. crplot checking linearity (left) and Q-Q plot for normality (right).

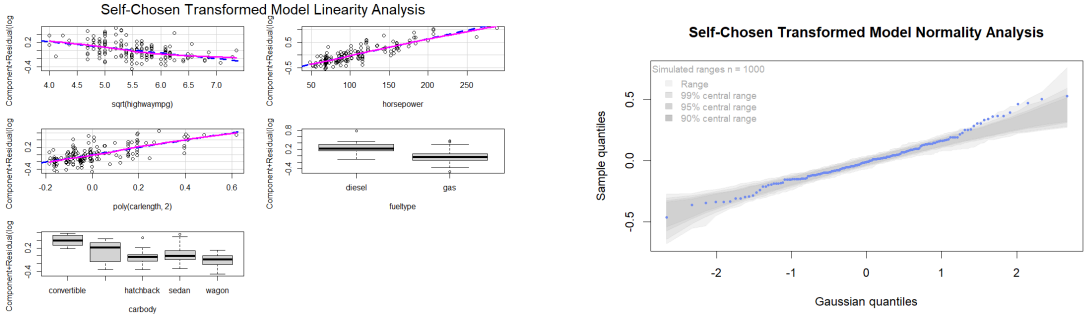


Figure 13: Diagnostic plots for Self-Chosen Transformed model. crplot checking linearity (left) and Q-Q plot for normality (right).

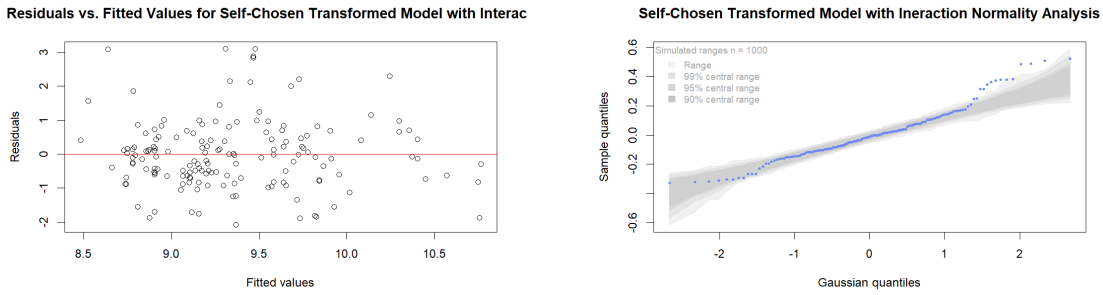


Figure 14: Diagnostic plots for Self-Chosen Transformed Model with Interaction. Residual plot checking linearity (left) and Q-Q plot for normality (right).

Figures 11 to 14 summarize the diagnostic plots for all four models. The BIC model demonstrates good linearity, with the fitted line closely aligning with the moving average of the four numerical predictors. Its Q-Q plot is also a straight line. The Shapiro-Wilk test further validates this interpretation, with a p-value of 0.97 confirming good normality. Additionally, all predictors have VIF statistics below 5, indicating no significant collinearity.

For the three self-chosen models, the initial component-residual plots were unsatisfactory, but the Q-Q plot indicated normality. After transforming two variables, the model, shown in Figure 13, became much more aligned, with an improved Q-Q plot indicating relatively good normality. In the final model with interaction terms, we used a residual plot instead of `crPlot` to assess linearity. As shown in Figure 14, the residuals are well-scattered, indicating good condition, but the normality assumption is violated, with clear bending at the tails suggesting skewness. VIF tests for these three models still show no significant collinearity.

3.2 Validation and ANOVA Test Analysis

The R^2 statistics were used to evaluate the performance of these four models. As shown in Table 1 column 2, all models explain the actual data well, indicated by their high R^2 values. However, even though BIC was used as the evaluation criterion during the feature selection process—penalizing for the number of included predictors—we wanted to ensure these good performances were not due to overfitting. Therefore, we used these models to predict `price` on the test dataset, with the corresponding R^2 statistics included in Table 1. Given the consistently high R^2 values, we conclude that none of these models suffer from overfitting.

Model	R-squared(trained)	R-squared(tested)	p-value for F test
BIC	0.9498	0.9345	<2.2e-16
Self-chosen	0.8698	0.8353	<2.2e-16
Self-chosen transformed	0.8788	0.8817	<2.2e-16
Self-chosen transformed with interaction	0.8943	0.8197	<2.2e-16

Table 1: Summary of four models

We also performed ANOVA tests to compare the three self-chosen models, as they are nested models. This allowed us to check if including predictor transformations and interactions would improve the models. The result comparing the `self-chosen model` and the `self-chosen transformed model` is shown below, with H_0 : the self-chosen model is preferred and H_A : the self-chosen transformed model is preferred.

```
1 Res.Df RSS Df Sum of Sq F Pr(>F)
2 1 155 5.6478
```

```
3  2 154 5.2550 1 0.39275 11.51 0.0008801 ***
```

Given the very small p-value, we have sufficient evidence to reject the null hypothesis and conclude that adding transformed predictor terms improves the model. ANOVA tests were also used to check whether including interaction terms was significant, and the results confirmed that this inclusion was significant as well. Therefore, while all three models perform well on their own, we indeed find their performances increase with these enhancements.

```
1    Res.Df RSS Df Sum of Sq F Pr(>F)
2  1 154 5.2550
3  2 147 4.5854 7 0.66969 3.067 0.004815 **
```

4 Discussion and Conclusions

In conclusion, all four models explain car prices well using various predictors. The BIC model offers a comprehensive explanation of car prices with the most significant predictors, accounting for 94.98% of the variation in the response variable. The self-chosen model provides a strong explanation of the relationship between car price and key features. The self-chosen transformed model improves the parameters through transformation, and the self-chosen model with interaction terms incorporates interactions between predictors. These three self-chosen models show an increase in R^2 statistics, indicating that their performance improved with the adjustments made to their simpler forms. ANOVA tests further validate this conclusion.

However, some possible improvements and problems still exist. Firstly, about normality of the Self-Chosen model with Interaction, we may consider how to change the model's interaction terms to better align with normality assumptions, potentially through data transformation or more sophisticated modeling techniques. Secondly, the relatively small size of our test dataset poses a risk to the validity of our model evaluations. The R-squared statistic, while informative, may not fully capture the model's predictive power due to the limited sample size. Thirdly, although the dataset has many variables, some predictor categories are underrepresented because of few data points in certain categories. For example, fewer data with "dohcv" engine type and "rwd" drive wheel. and This could skew the conclusions drawn from the model. Additional data collection or alternative modeling strategies need to be considered to ensure reliable outcomes.

References

[1] Car Price Prediction Multiple Linear Regression, *Kaggle*, Available at https://www.kaggle.com/datasets/hellbuoy/car-price-prediction. Accessed on July 15, 2024.

[2] Ailin Zhang, "Regression-Diagnostics", "Collinearity-Diagnostics", "Addressing-Violations of Assumptions", "Feature-Selection", STAT4130J Lecture Slides.