Driven by the Numbers: Predicting retail price of cars

Abstract:
Recognizing the economic importance of owning a vehicle and its corresponding financial commitment, we examined which variables would best predict the price of a car. As a group, we compiled an unclean data set of 6,515 observations of specific car models and their corresponding specifications to predict the manufacturer's suggested retail price (MSRP). After applying a transformation and simplifying our model by removing predictors, we created a multiple linear regression model to predict the MSRP of the car given its horsepower, fuel type, transmission type, vehicle make, and vehicle popularity. From 2014-2017, our model indicated that the most underpriced cars were Porsches, Lincolns, and Volvos, while the most overpriced cars were Dodges and Kia. Utilizing the prediction interval of our model, we created a Shiny Web App that provides predicted intervals for MSRP: https://ignaciofeged.shinyapps.io/shiny/.

Main Content:

1. Background and Introduction

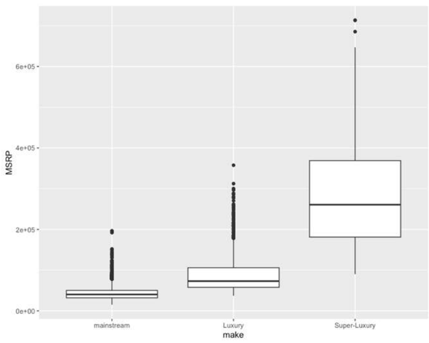
Our initial interest in the concept of car pricing began with our individual experiences with the automobile industry. As we have come of age to carry a driver’s license and thus operate a car, we have begun to realize how crucial personal vehicles are for transportation and lifestyle efficiency for us and society as a whole. Interestingly, “having continued car access makes it 2.2 times more likely that someone unemployed will have moved into employment two years later, compared to not having car access” (ISER 2). One of our group members has volunteered at a nonprofit that works to supply extremely affordable cars to individuals returning to society from prison or rehabilitation. Through that experience, he realized that in most parts of the country, owning a car is vital to securing and sustaining a reliable source of income. Due to their necessity, many individuals spend an enormous sum on older, less safe models. From there, we considered which aspects of the car affect the sale price the most. We assumed that the brand and reputation of that brand played an important role, but this was simply speculation. Furthermore, due to the ease of statistical analysis and calculations, we decided to focus strictly on new cars to predict a more consistent interval without considering many new major variables introduced when dealing with used cars. With this in mind, we worked to build a simple and effective multiple linear regression model that would predict as accurately as possible the price of a car based on its various aspects.

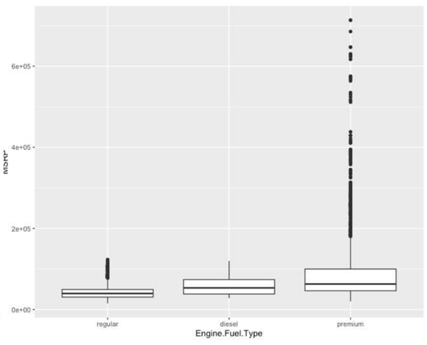
2. Data and Exploratory Analysis

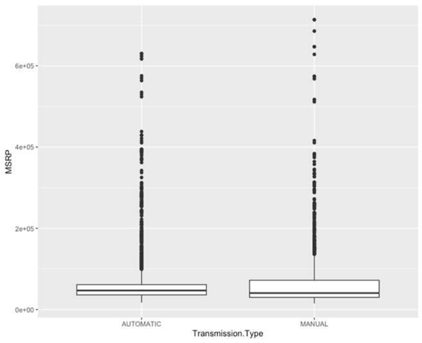
a. Data and Variables

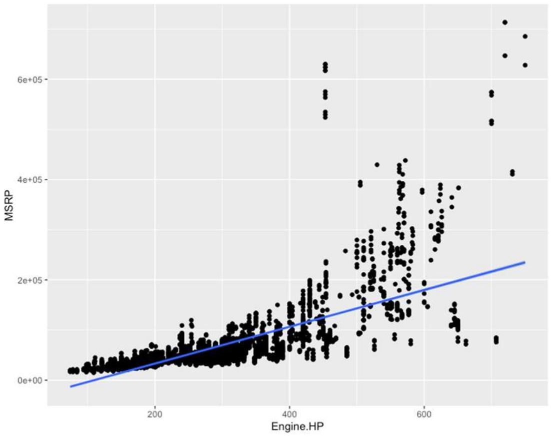
The primary data source for analysis (Cooper Union 1) was obtained via a reliable .csv file on the statistical database Kaggle, where our file was compiled from data found on edmunds.com, an automobile pricing website. The unclean file contained 6,515 observations of specific car models, along with many details about each car: year (2014-2017); fuel type (ie. regular, diesel); engine horsepower (range of 74-750); engine cylinders (0-12); transmission type (ie. automatic, manual); driven wheels (ie. all wheel, front wheel); the number of doors (2-4); vehicle size (ie. compact, midsize); vehicle style (ie. sedan, SUV); highway miles per gallon (13-354); city miles per gallon (10-137). Another variable was price, which was provided through MSRP (Manufacturer Suggested Retail Price) for the new car, which we adjusted to 2023 dollars using the Consumer Price Index measure of inflation. Ultimately, we use MSRP as our response variable. Edmunds.com also scores each car with a popularity rating, corresponding to how much traffic that car got on its website. We use this as a price predictor in our final model. The dataset also had a predictor labeled make, where it gave us the car brand, but with over 30 different brands, far too many for a more stable categorical predictor variable. In response, we shrunk these categories into three categories: super-luxury (ie. Ferrari, Lamborghini, Rolls-Royce), luxury (ie. BMW, Mercedes-Benz, Lexus), and mainstream (ie. Ford, Nissan, Chevrolet). This resulted in a reliable categorical predictor variable that we ended up using in our final model. A final categorical variable, the model, had even more factors (416), with some factors only having one or two observations.

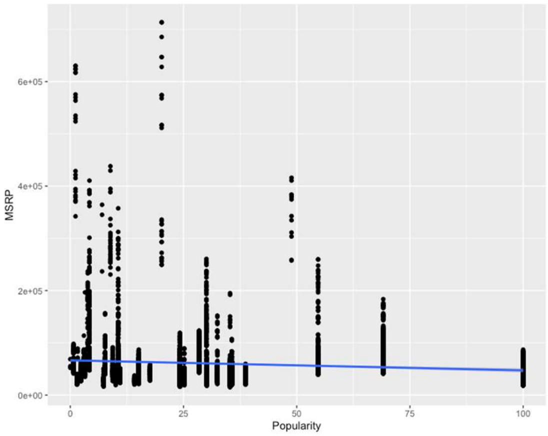
We were amazed when we first found this data source; it perfectly represented our target population for our research question. This was a shifting point in our research process because most other sources contained used and new cars or did not provide enough price predictors.











b. Exploratory Data Analysis

After some exploratory data analysis, we cut down our predictors significantly due to confounding variables, the complexity of the model, and relevance. We discuss the reasoning for these cuts later but above are five graphs of the variables we included in our final model. The top row shows boxplots of categorical variables make, engine fuel type, and transmission type. Our make predictor shows that cars in the super-luxury category tend to have the highest median price, followed by luxury cars, while mainstream designated cars possess the lowest median MSRP. Engine fuel type shows that the median price is lowest for regular gas, then diesel is in the middle, and premium gas cars have the highest median price, but this category seems to have a substantial quantity of outliers. Next, transmission type shows us that automatic and manual cars possess similar relationships with price, with the manual's interquartile range being slightly greater than the automatic's interquartile range. In the second row, we have scatterplots comparing engine horsepower and popularity to MSRP, respectively. Engine horsepower (in our opinion, our strongest quantitative predictor variable) possesses a strong, positive association with MSRP with an correlation coefficient of 0.71, which makes sense as society usually puts more money towards more powerful engines. Finally, popularity vs. MSRP shows us an interesting negative, fairly weak correlation with a correlation coefficient of -0.08. This shows that the most popular cars on Edmonds.com are cheaper and more affordable.

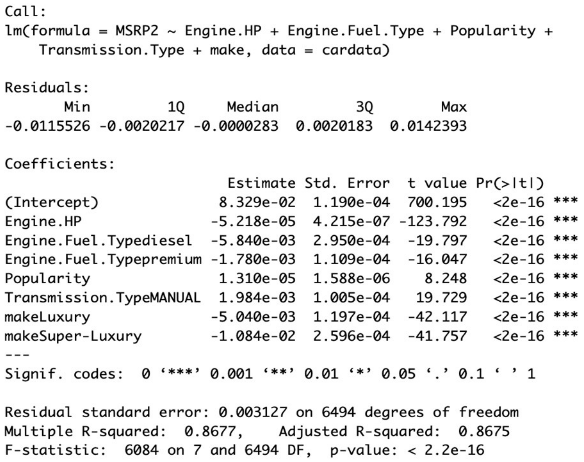
3. Model and Results

a. Analytic Methods

We decided to use a multiple linear regression model to predict price. When examining the model diagnostics, our initial model violated the constant variance and normality conditions, and box-cox transformation suggested that we transform MSRP into MSRP^(-¼). After applying the transformation, all model assumptions are reasonably satisfied. Continuing to the variable section, the best subset and stepwise selection (forward, backward, and both directions) suggest that the full model is the best model. Concerns arose regarding the complexity of our model as well as multicollinearity. We examined the variance inflation factor, revealing that highway MPG and Vehicle Style: Sports Cars suffered from extreme multicollinearity with other predictors in our model. Still, our model remained extremely complex, so we sacrificed a very small loss in adjusted R-squared for a more interpretable model and cut our final model to contain just the four predictor variables detailed above (make, fuel type, transmission type, and engine horsepower). The 'make' variable contains indicator variables *luxury* and *super-luxury*, while 'fuel type' brings in indicators of *diesel* and *premium* in addition to the transmission type's *manual* indicator variable. Engine horsepower and popularity are quantitative variables. By omitting interaction terms, we assume the presence of parallel lines; however, our plots indicate that the lines are not perfectly parallel. We chose to expect this limitation in our findings to maintain the model's relative simplicity.

$$\widehat{MSRP}^{-1/4} = 0.08329 - 0.00005218\beta_{Engine.HP} - 0.005840\beta_{Engine.Fuel.Typediesel} - 0.001780\beta_{Engine.Fuel.Typepremium}$$

$$+ 0.00001310\beta_{Popularity} + 0.001984\beta_{Transmission.TypeMANUAL} - 0.00504\beta_{makeLuxury} - 0.01084\beta_{makeSuper-Luxury}$$



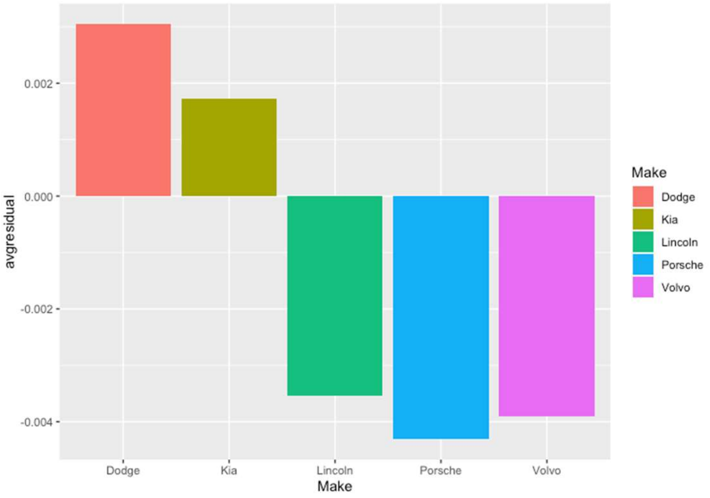
```
Call:
lm(formula = MSRP2 ~ Engine.HP + Engine.Fuel.Type + Popularity +
    Transmission.Type + make, data = cardata)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0115526 -0.0020217 -0.0000283  0.0020183  0.0142393

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              8.329e-02  1.190e-04  700.195   <2e-16 ***
Engine.HP               -5.218e-05  4.215e-07 -123.792   <2e-16 ***
Engine.Fuel.Typediesel  -5.840e-03  2.950e-04  -19.797   <2e-16 ***
Engine.Fuel.Typepremium -1.780e-03  1.109e-04  -16.047   <2e-16 ***
Popularity               1.310e-05  1.588e-06    8.248   <2e-16 ***
Transmission.TypeMANUAL  1.984e-03  1.005e-04   19.729   <2e-16 ***
makeLuxury              -5.040e-03  1.197e-04  -42.117   <2e-16 ***
makeSuper-Luxury        -1.084e-02  2.596e-04  -41.757   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003127 on 6494 degrees of freedom
Multiple R-squared:  0.8677,    Adjusted R-squared:  0.8675
F-statistic:  6084 on 7 and 6494 DF,  p-value: < 2.2e-16
```

b. Final Model and Results

Our final model, as shown by the two-lined equation above, predicts the manufacturer's suggested retail price (MSRP) of a new car to the power of -¼ using variables engine horsepower, fuel type, popularity, transmission type, and make. Our model has a adjusted R-squared of 0.8675, meaning that our model is able to account for 86.75% of the variability of car MSRP, and a p-value associated with the F-statistic of < 2.2e-16, confirming that our model is more useful than an intercept only model at predicting MSRP. Additionally, each variable in our model has a p-value of < 2.2e-16, suggesting that they are each statistically significant in predicting MSRP given other variables in the model. When looking at the coefficients of each variable, however, we can notice that engine horsepower has a negative coefficient in this conditional relationship to MSRP, whereas in the marginal relationship between engine horsepower and MSRP, the coefficient is positive. Therefore, the effect on predicted MSRP when changing engine horsepower and holding all other variables constant does not necessarily make the most sense in the context of this problem. When interpreting the model, this is something to consider.

4. Discussion and Conclusion

Returning to our research question, the horsepower, fuel type, popularity, transmission type, and make of a vehicle, when evaluated together, are the best predictors of MSRP. Examining the observations with the largest positive and negative residuals, according to our model, the most underpriced cars between 2014-17 were Porsches, Lincolns, and Volvos, while the most overpriced cars were Dodges and Kias. While these observations are subject to our model's limitations, our data suggests that, pound-for-pound, these are the best brands for consumers. Using our model’s 95% prediction interval, we created an interactive Shiny App that allows users to see our model’s predicted price for their vehicle and others. While these predictions are limited, considering that our data ends in 2017, we hope it may still serve as a starting point for consumers.



References (MLA 9)

Social and Economic Research (ISER) at the University of Essex. “How Owning a Car Impacts on Our Life Opportunities.” *Understanding Society*, 4 Dec. 2019, www.understandingsociety.ac.uk/2019/12/04/how-owning-a-car-impacts-on-our-life-opportunities.

CooperUnion. “Car Features and MSRP.” *Kaggle*, 21 Dec. 2016, www.kaggle.com/datasets/CooperUnion/cardataset/data.

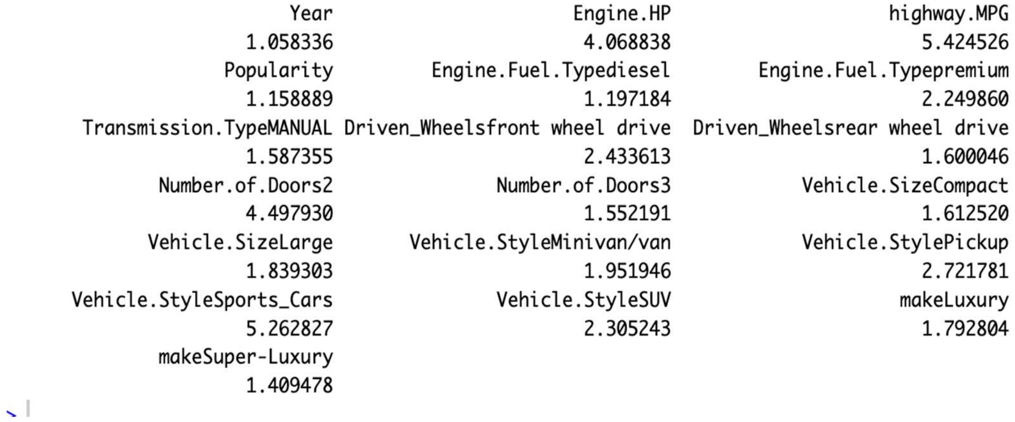
“New Cars, Used Cars, Car Reviews and Pricing.” Edmunds, www.edmunds.com/ . Accessed 11 Dec. 2023.

APPENDIX

Shiny app (for your own car price inquiries): https://ignaciofeged.shinyapps.io/shiny/

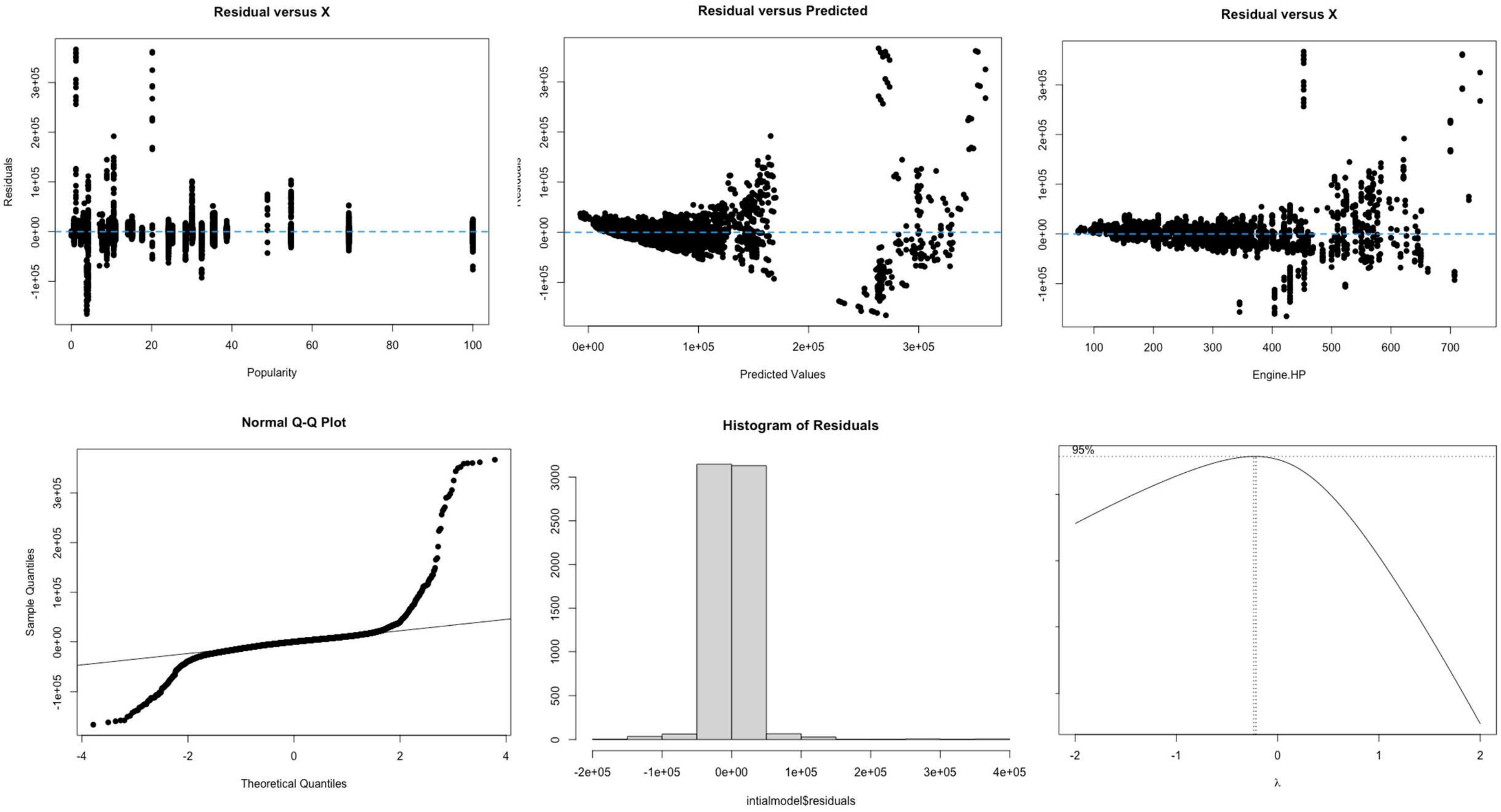
Variable Name	Variable Label (description)	Variable Range/Code
Make	The car’s brand; categorized	Categorical, ie. luxury, super-luxury, and mainstream (3 factors)
Engine.Fuel.Type	Type of fuel for the car	Categorical, ie. regular, premium, and diesel (3 factors)
Engine.HP	The Horsepower of the car’s engine	Quantitative, from 74 to 750 hp
Transmission.Type	The car’s transmission	Categorical, ie. automatic, manual (2 factors)
Popularity	The Edmunds.com popularity rating	Quantitative, from 0 to 100
MSRP	The car’s list price (in 2023 USD)	Quantitative, from $11,990 to $548,800

Initial Model’s VIF values:

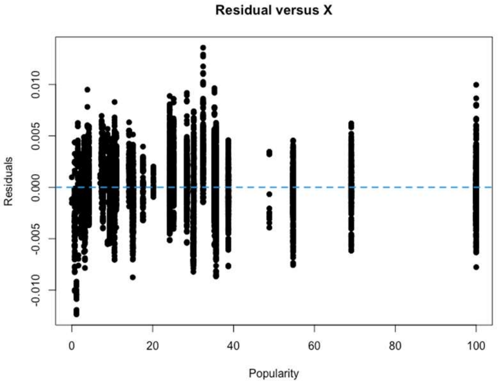


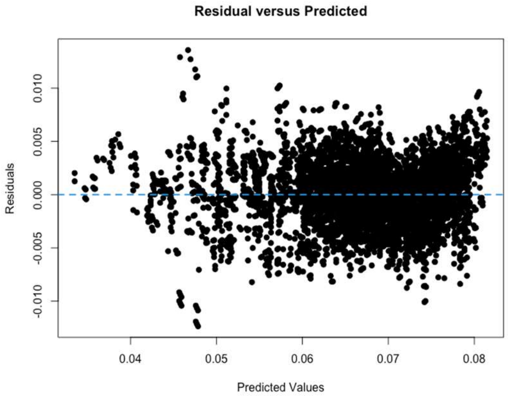
```
                          Year                      Engine.HP                    highway.MPG 
                      1.058336                       4.068838                       5.424526 
                    Popularity         Engine.Fuel.Typediesel        Engine.Fuel.Typepremium 
                      1.158889                       1.197184                       2.249860 
       Transmission.TypeMANUAL Driven_Wheelsfront wheel drive  Driven_Wheelsrear wheel drive 
                      1.587355                       2.433613                       1.600046 
              Number.of.Doors2               Number.of.Doors3            Vehicle.SizeCompact 
                      4.497930                       1.552191                       1.612520 
             Vehicle.SizeLarge       Vehicle.StyleMinivan/van            Vehicle.StylePickup 
                      1.839303                       1.951946                       2.721781 
      Vehicle.StyleSports_Cars               Vehicle.StyleSUV                     makeLuxury 
                      5.262827                       2.305243                       1.792804 
              makeSuper-Luxury 
                      1.409478 
```

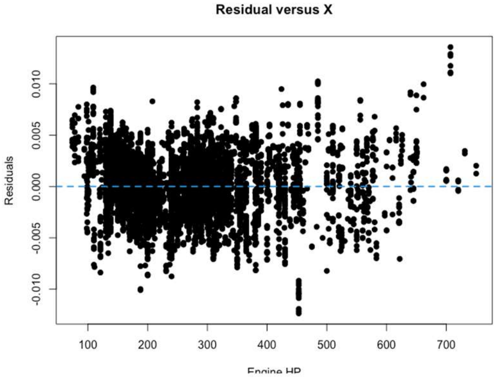
Diagnostic Plots Before Transformation:

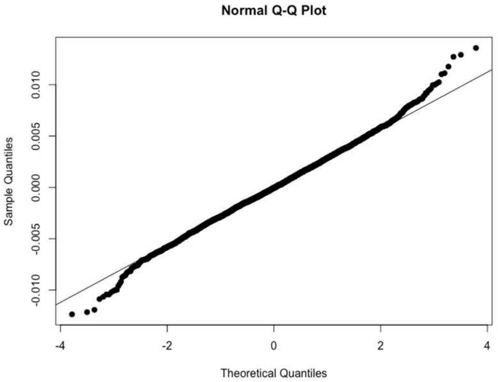


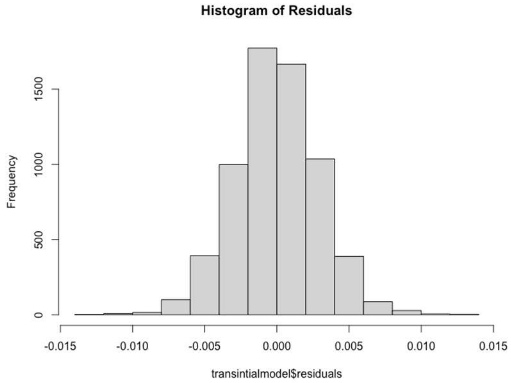
Diagnostic Plots after (^-¼) Transformation:

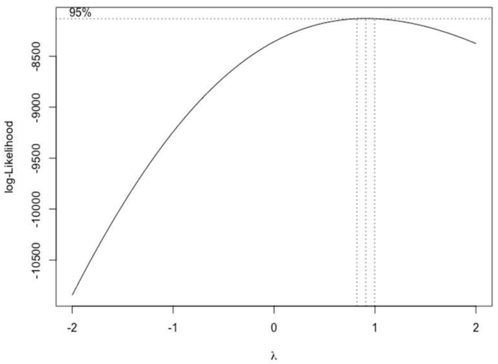

Residual versus X
Residuals
Popularity


Residual versus Predicted
Residuals
Predicted Values


Residual versus X
Residuals
Engine HP


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


Histogram of Residuals
Frequency
transintialmodel$residuals


log-Likelihood
λ