Statistics for Business and Economics

Regression Modeling for Automobile Price Prediction

Sekai Kanamori

June 18th, 2025

Recommended Model

The most **optimal model** that was found through statistical and data analysis was:

	(11.592)	(8.288)	(2.770)	(11.826)	(2.255)
log(price) =	0.0002207**hp* +	0.00001578**weight* +	0.01276**rearwd* +	0.02908**luxury* +	0.009898**hybrid*

With the **Adjusted R^2 = 0.8448**.

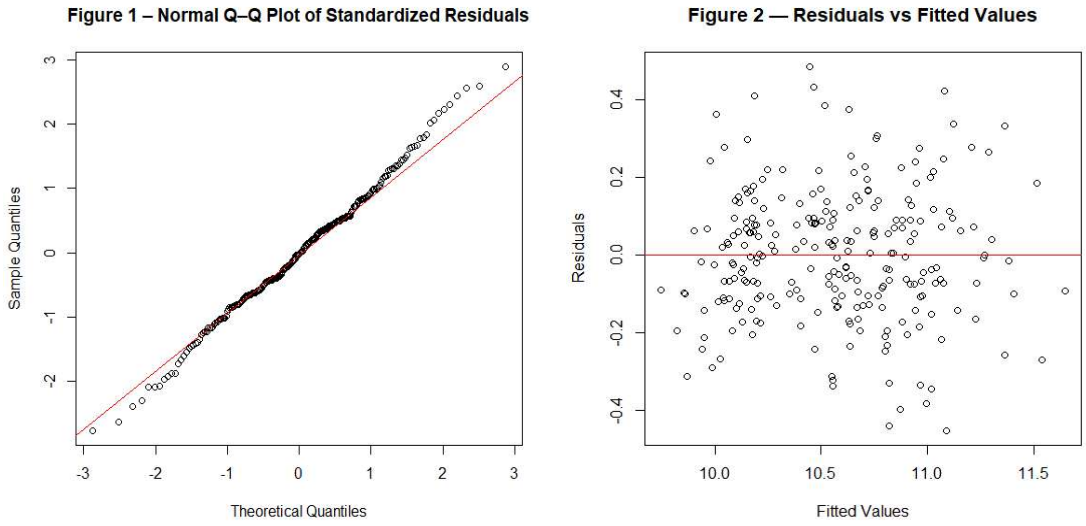

Figure 1 shows a **Q-Q plot** of the standardized residuals. The points lie almost precisely on the reference line, with only minor tail deviations, ensuring that the residuals are approximately normal, meaning that the model is a good fit for the data.

Figure 2 displays the **residuals vs. fitted values**. The residuals are randomly scattered around zero with no evident pattern or unequal variance, and no extreme outliers exist, which means the model is appropriately suitable for the data.

Three extreme brand-power outliers were excluded beforehand to focus on engineer-controllable factors; the remaining residuals fit well with normality and unequal variance assumptions. Furthermore, the missing values that were in the data have been removed manually prior to the analysis, ensuring there are no effects from them.

In summary, ***horsepower***, ***curb weight***, ***drivetrain configuration (rear-wheel drive)***, ***luxury options***, and ***hybrid technology*** explain 84.48% of the variation in car prices, demonstrating which traits are most strongly associated with price.

Introduction

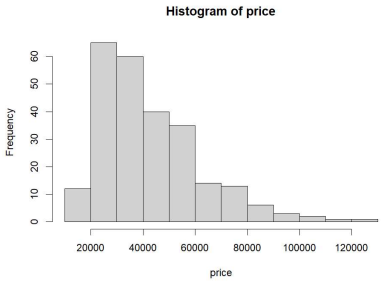

Figure 1.1: Histogram of price distribution

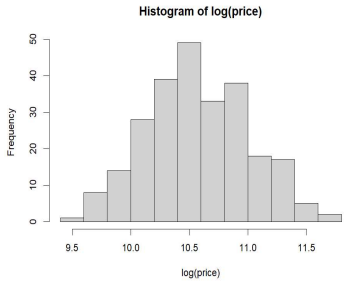

Figure 1.2: Histogram of log-transformed price distribution

Before starting anything, the first step was to render a histogram to determine the price distribution. This diagram will tell us if the price data needs to be log-transformed. This process is important because linear regression assumes that residuals are normally distributed and that the relationship between variables is linear. As shown in Figure 1.1, the distribution is skewed to the right, indicating that the data violates the normality assumption. Hence, we will need to take the logarithm of the data to normalize the distribution (Figure 1.2), stabilizing variance across values. From now on, the price will be referred to as the log(price), ensuring that our regression model meets the necessary statistical assumptions for valid analysis.

Figure 2.1 (see next page) shows a Scatterplot matrix. This plot shows pairwise relationships between variables using scatterplots. For binary variables (***7over, cvt***, **AWD, *4WD_dummy***, ***rear***, ***SUV***, ***Pickup***, ***Minivan***, ***Sports***, ***Luxuary*** and ***Hybrid***), jitter was used to spread the points and show differences between groups. This matrix helps identify which variables have strong linear relationships and whether some may overlap.

Once again, three extreme brand-power outliers were excluded beforehand to focus on engineer-controllable factors

Model 1

When purchasing a vehicle for transportation, consumers will typically pay attention to the size of the vehicle first because the size of the car can determine the number of passengers it can accommodate, affect fuel efficiency, and generally offer better safety in crashes. In this data, ***length(inch)***, and ***Curb Weight(lb)*** are variables responsible for depicting the size of the car. The correlation between these two variables is reasonably high (you can see this in Figure 1.2).

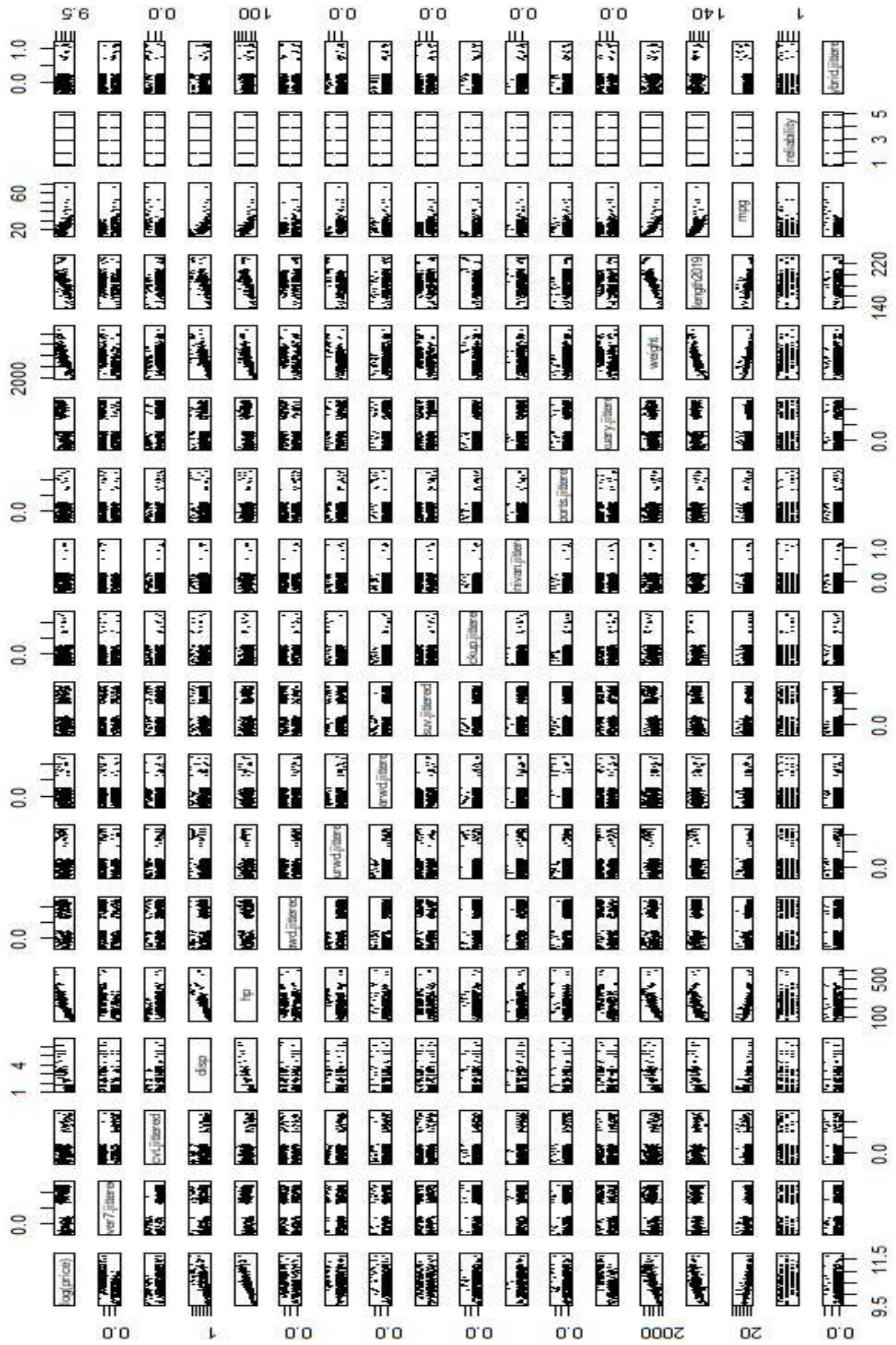

Figure 2.1: Scatterplot matrix of the variables.

After analyzing the adjusted scatterplot (***log(price)*** adjusted for ***Curb Weight(lb)***, and ***length(inch)*** adjusted for ***Curb Weight(lb)***, which is a scatterplot between residuals with the influence of ***Curb Weight(lb)*** removed from ***log(price)*** and ***length(inch)***) with a line that shows the trend between these two variables, the scatter plot showed that without the influence of weight, length of the car had nothing to do with the price of the car. This means that the length provides almost nothing beyond what the curb weight already explains. The regression analysis between these residuals shows that length has a t-value (which illustrates how confident we can be that a variable truly matters for predicting the outcome rather than just being due to random chance) of 0.283 and an estimate of 5.758e-04. While the estimate's positive sign is logical (longer cars should cost more), the t-value of 0.283 is well below the significance threshold of |2|, indicating the relationship is statistically insignificant. On the other hand, weight has an estimate of 3.492e-04 and a t-value of 9.427, which exceeds the significance threshold. These results suggest using weight instead of length in our model due to the weight's greater statistical significance.

Another factor that consumers will look at is the engine size. In this data, ***Disp*** and ***Hp*** are variables responsible for defining the size of the engine. Once again, the correlation between these two variables is reasonably high (again, you can see this in Figure 1.2). After observing the adjusted scatter plot (***log(price)*** adjusted for ***Hp*** and ***Disp*** adjusted for ***Hp***), the results suggest that including both ***Hp*** and ***Disp*** in the regression equation may be problematic. This observation can be affirmed with numerical values; ***Disp*** has a t-value of -2.762 and an estimate of -5.542e-02; the t-value is significant, but the estimate is not logical. In contrast, ***Hp*** has a t-value of 18.353, well above the significance threshold, and an estimate of 4.4813e-03, which makes sense. This illogical behavior of the estimate only occurs when ***Disp*** and ***Hp*** are used concurrently because these variables suffer from multicollinearity; they both measure aspects of engine size and power. To summarize, ***Disp*** becomes negative when used with ***Hp***; hence, we now know it is a bad idea to use ***Disp*** and ***Hp*** simultaneously. Since HP shows stronger statistical significance and logical coefficient behavior, we will use ***Hp*** instead of ***Disp*** for our model.

Now, we have the base for our model, the weight, and hp, which exhibit strong t-values and logical estimates.

Another factor is the ***Pickup*** variable. As we know, pickup trucks are undoubtedly popular in North America and control a significant portion of the market. Creating an adjusted scatter plot helps determine if this variable correlates with the ***log(price).*** As a result, a trend is observable: pickup trucks tend to be cheaper than other vehicles. This can be confirmed with the ***Pickup*** variable's estimate of -2.325e-01 and the t-value of -3.700 when incorporated into the model with weight and hp. The negative estimate indicates that being a pickup truck decreases the ***log(price)***, and the t-value of -3.700 shows this relationship is significant.

For our final steps, we incorporate ***AWD*** and ***rear*** into the model. ***AWD*** is for all consumers in the NorthEast, Mid-West, Mountain State, etc., as it snows a lot in such areas, and they require ***AWD*** for their winter tires. ***Rear***, as in rear wheel drives, are typically for high-end cars, which should correlate positively with the price of the vehicle. After observing the adjusted scatter plots, the results showed that both variables had meaningful relationships with the price. Their t-values were significant, 5.868 (***AWD***) and 3.626 (***rear***), and their coefficients made sense, 1.582e-01 (***AWD***) and 2.006e-01 (***rear***). Both positive coefficients indicate that these features increase car prices. As a result, we ended up with five significant variables and an adjusted R^2 of 0.8036, meaning these five variables explain approximately 80% of the variation in car prices. We also assessed the normality of residuals using a Q-Q plot. The points align nearly in a straight line, indicating that the residuals are approximately normally distributed.

Model 2

The analysis began by using logic and reasoning to determine which variables to analyze. One variable that stood out was ***Luxuary***, a binary variable indicating whether the car is a luxury model. It is fair to assume that any luxurious car would be more expensive than those that are not. Hence, the hypothesis was that the luxury variable would positively affect the price. A plot diagram was created to check the relationship between ***log(price)*** and ***Luxuray***, adjusting for ***Hp***, ***Curb Weight(lb)***, ***Pickup***, ***AWD***, and ***rear*** to ensure this was the case. The resulting scatter plot with a lowess line (smooth curve that helps show the trend) showed a clear positive trend (Figure 3.1), confirming that ***Luxuary*** explains price variation even after removing the effects of ***Hp***, ***Curb Weight(lb)***, ***Pickup***, ***AWD***, and ***rear***. A linear model was then created with variables from Model 1 and ***Luxuary***, and it was found that all variables had strong t-values, including ***Luxuray***, which had a t-value of 9.133 with an estimate of 2.406e-01, indicating that this relationship is significant.

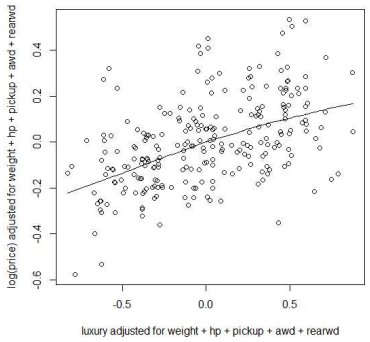

Figure 3.1: Adjusted plot diagram of luxury and log(price)

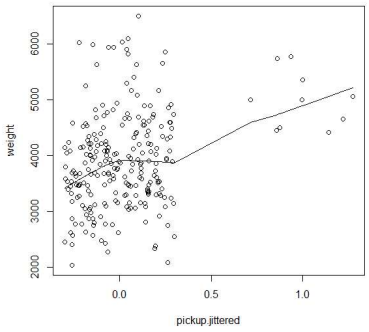

Figure 3.2: Plot diagram of Pickup vs. Weight

Hence, the decision was to use the ***Luxuary*** variable. As for which variable to substitute, ***Pickup*** was removed, as it had the lowest t-value in Model 1 and likely overlapped with weight. To confirm this relationship, a scatter plot of ***Pickup*** vs. ***Curb Weight(lb)*** was created (Figure 3.2), which showed a clear positive correlation, showing that pickup trucks tend to be heavier than other vehicles. This correlation justified removing ***Pickup*** in favor of ***Luxuary***, as the ***Luxuary*** variable had stronger predictive power and more apparent economic logic.

The model's adjusted R^2 improved from 0.7936 to 0.8501, indicating that the adjustment significantly improved the model's accuracy.

Before the second variable is selected, a clarification needs to be made: the second variable should not overlap with ***Curb Weight(lb)*** or ***Hp***, as these two variables are the foundational independent variables in this model. Weight and horsepower have demonstrated consistently strong statistical significance (with t-values well above the benchmark of 2) and logical economic relationships with car prices. These variables represent core characteristics that drive car pricing; engine performance and vehicle size. In other words, a variable that can meaningfully replace ***AWD*** or ***rear*** is needed.

Upon analyzing the matrix diagram (Figure 2.1), ***Hybrid*** variable was selected based on the logic that hybrid cars typically cost more upfront than conventional cars. To check if this variable adds any uniqueness to the model, ***log(price)*** and ***Hybrid*** were plotted, adjusting for ***Hp***, ***Curb Weight(lb)***, ***AWD***, ***rear***, and ***Luxuary***. The resulting scatter plot and lowess trend showed a weak to moderate positive relationship (Figure 3.3), indicating that hybrid status has some independent influence on price. The hybrid variable was then added to the linear model, and its t-values were checked. All variables showed statistical significance with t-values higher than 2, exceeding the standard rule of thumb for statistical significance.

However, ***AWD*** had the lowest t-value among the existing variables, making it the candidate for removal to accommodate the new ***Hybrid*** variable.

The decision to add ***Hybrid*** is justified from both statistical and business perspectives. Hybrid technology's impact on pricing is crucial for strategic decision-making as the car markets shift toward environmentally conscious consumers. Incorporating hybrid status indicates the pricing margins associated with hybrid technology, which is essential for product development and pricing strategies. While the model's adjusted R^2 declined slightly from 0.8501 to 0.8448, this tiny drop was deemed acceptable due to the strategic value of incorporating hybrid technology. It also allows for introducing a new type of variable, variety, and depth in our variables for Model 2. To ensure the substituting ***AWD*** was optimal, the adjusted R^2 was compared across different scenarios just for safety measures. This was done by analyzing the adjusted R^2 of when substituting ***AWD*** with when substituting ***rear*** for a ***Hybrid***. The latter had an adjusted R^2 value of 0.8378, showing that substituting ***AWD*** is, in fact, better than replacing ***rear***. Additional testing was done with ***Hp*** and ***Curb Weight(lb)***; the results were 0.8049 and 0.7668, further showing the importance of those two variables and that substituting ***AWD*** was the right move.

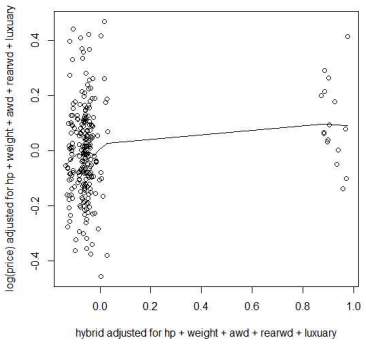

Figure 3.3: Adjusted plot diagram of hybrid and log(price)

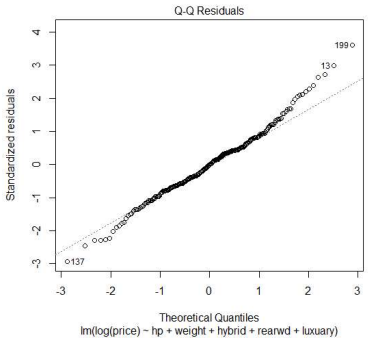

Figure 3.4: A Q-Q plot of Model 2

A Q-Q plot of the residuals from Model 2 (Figure 3.4) shows that they are approximately normally distributed but with some flagged outliers. After investigating what those are, three vehicles were identified: a Porsche Boxster that cost $74,610, an Audi A8 that cost $83,800, and a Dodge Challenger that cost $83,295. Further investigations show that these models were highly overpriced for their specs; The Porsche and Dodge, despite their high prices, had only 300 and 372 ***Hp*** and were not classified as ***Hybrid*** or ***Luxuary*** vehicles in the dataset. The Audi had only 230 ***Hp*** and was not a ***Hybrid*** car, nor did it have ***rear*** wheel drive.

Further investigation shows that cars with similar model characteristics to the Audi (***Hp*** of 200-250, not ***Hybrid***, not ***rear*** wheel drive, and ***Luxuary***) have an average price of $45,379, much less than the Audi A8's $83,800. The same applies to the Porsche Boxster and Dodge Challenger; cars with similar characteristics (***Hp*** of 300-350, not ***Hybrid***, mixed drivetrain types, and not ***Luxuary***) have an average price of $40,460, much less than the Porsche Boxster ($74,610) or Dodge Challenger ($83,295). Considering these factors, it would be fair to remove these models, as these price fluctuations are most likely caused by their "branding" compared to their model characteristics.

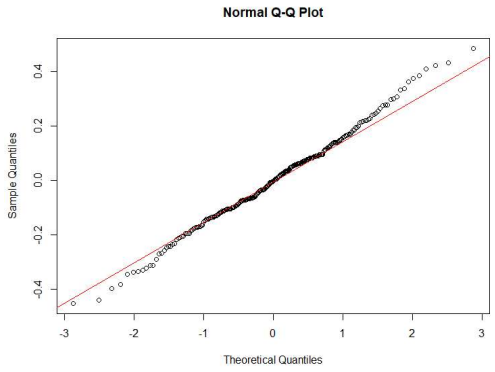

Figure 3.5: A Q–Q plot of Model 2 without the outliers

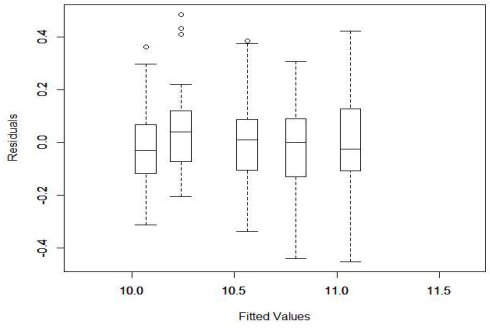

Figure 3.6: Boxplot of Residuals vs. Fitted Values for Model 2 (Outliers Removed)

Now that those outliers are removed, Figure 3.5 shows a Q-Q plot of the residuals from Model 2 that follow the straight line (theoretical reference line), indicating that they are approximately normally distributed, with only minor deviations at the tails. This plot supports the validity of the regression model and indicates that the assumptions of normality of residuals are reasonably satisfied.

Finally, Figure 3.6 illustrates the residuals split into five segments of fitted ***log(price)*** and shows a small boxplot for each. Each box summarizes the spread of residuals in one segment of the predicted ***log(price)***. All boxes are centered around zero, meaning the model does not over or under-predict in any range. The box heights are similar across bins, indicating roughly constant variance. A few new isolated points lie outside the whiskers, but these are different from the original brand-power outliers and have been retained as valid data. Overall, these patterns confirm that the error variance is roughly constant and that no apparent non-linearity or model misspecification remains.

Model 3

For our final model, these two variables were considered: ***MPG*** and ***7over***. MPG is a crucial factor when it comes to buying a car; consumers in North America typically prioritize fuel efficiency. Similarly,

vehicles with more than seven gears typically offer better performance and efficiency. Hence, these are other variables that consumers look into. The adjusted scatter plot shows that ***7over*** and ***MPG*** positively correlate with the ***log(price)***.

Before checking for t-values, it is important to investigate if any of these variables correlate with the others (multicollinearity). The first to investigate was the ***Curb Weight(lb)*** and ***MPG***, as it is fair to assume a relationship that MPG will worsen as the cars get heavier. As expected, they had a pretty strong negative relationship (you can see this in Figure 2.1). T-values are examined as ***Curb Weight(lb)*** is replaced by ***MPG***, and we find that it is significant (-3.487); we also notice that ***rear*** has a t-value of 0.426, so it gets swapped out for ***7over*** , they are checked again, and the results indicate that all t-values are significant (***MPG***: -3.673 and ***7over***: 4.260) with an adjusted R^2 of 0.8235.

However, the coefficient signs require explanation. ***7over*** has a positive coefficient (0.11768), which makes sense since more gears typically indicate higher-end vehicles. However, ***MPG*** has a negative coefficient (-0.01032), which requires some explanation. It is fair to say that the higher the ***MPG***, the higher the price; hence, a positive coefficient is expected. However, in our dataset, most high-MPG vehicles (27/36 cars above 30 MPG) are from Asian car companies such as Toyota, Nissan, Honda, Kia, etc; these cars are cheaper than the other cars on the market/data, as they position themselves as brands with lower prices despite superior fuel efficiency. Hence, the negative coefficient.

Final Model and Cp value

The Cp measure was used to compare the three models to finalize our model selection process. Cp measures how well each model balances accuracy with simplicity, helping identify the most efficient combination of variables. The full model was used as the benchmark (ideal model) for calculating the Cp.

The results show:

Models	Variables	Adjusted R^2	Cp Value
Model 1	Hp, Curb Weight(lb), Pickup, AWD, rear	0.8036	120.71
Model 2	Hp, Curb Weight(lb), rear, Luxuary, Hybrid	0.8448	45.71
Mode 3	Hp, Pickup, AWD, MPG, 7over	0.8235	84.59
Full Model	Hp, Curb Weight(lb), Pickup, AWD, rear, Luxuary, Hyrbid, MPG, 7over	0.8666	10

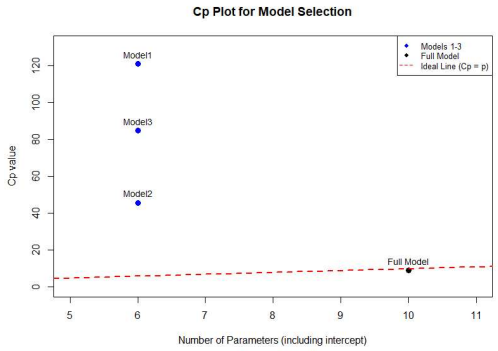

Figure 4.1: Cp Plot

Model 2 demonstrates the best accuracy with the lowest Cp value of 45.71. Model 1's high Cp value (120.71) suggests that it is less effective for price prediction. Model 3's high Cp (84.59) indicates that substituting MPG for Curb Weight(lb) and 7over for rear lowered the predictive accuracy compared to Model 2. The relatively high Cp values across all three models are mainly due to the full model's nine variables; the exploration of different variable combinations (replacing two different variables in each model) demonstrates the challenge of selecting the optimal subset from the available options while attempting to incorporate some diversity in the models. Figure 4.1 visually illustrates these Cp comparisons, clearly showing Model 2's optimal position closest to the ideal Cp = p (variables) line. This demonstrates that while our simplified models are more practical and interpretable, they sacrifice some predictive power by excluding four variables from the full model.

Model 2 delivers the optimal framework for price prediction and strategic decision-making. Its combination of horsepower, weight, pickup status, luxury features, and hybrid technology captures the most important cost drivers while maintaining statistical efficiency.

Conclusion

In conclusion, this analysis determined the most optimal model for automotive price prediction. Model 2, which incorporates ***Hp, Curb Weight(lb), rear, Luxuary,*** and ***Hybrid***, provides the most right balance of statistical accuracy, 84.48% variance explained, and transparency for engineering purposes. The systematic and logical method of variable selection, multicollinearity review, and iterate validation makes it safe to say that predictions for strategic decision-making in auto product development are reliable.