DAT603

ASSESSMENT 3

STATISTICAL CASE SCENARIO ANALYSIS

Table of Contents

Abstract

The study uses regression to investigate the association between car pricing and key vehicle attributes. 205 automobiles with 26 variables are examined using both simple and multiple linear regression models. Investigations are conducted into the confounding effects of curb weight, engine size, and horsepower, as well as the individual predictor of horsepower. The results show that the multiple regression model explains 81 percent of the pricing, while the basic model only explains 66 percent. Engine size was the most important variable (effect of $84.53 per unit), followed by horsepower (impact of 49.13), and curb weight (impact of 4.36).

1. Introduction

Automobile pricing research offers a lot of information as to what affects the cost and performance of vehicles. Knowledge of the connection between attributes and the price of cars in the market also helps the manufacturer, the consumer, and the policymaker in decision-making. Numerical attributes, such as horsepower, engine size, and curb weight, are analyzed quantitatively and demonstrated to be strong predictors of price changes. Statistical modeling and simple regression, and multiple regression allow testing these relationships and provide credibility by testing relationships such as engines, fuel price, and horsepower, which helps to maintain customer trust.

1.1 Objectives

The aim is to analyze the relationship between key vehicle attributes and market price, evaluate regression model performance, and identify significant predictors to support informed pricing and investment decisions in the automotive industry

- To examine the relationship between car horsepower and price using simple linear regression for predictive insights.

- To evaluate the combined effect of horsepower, engine size, and curb weight on price using multiple regression.
- To assess regression model performance, check assumptions, and determine the significance of predictors for informed decisions when determining factors that affect the price of an automobile.

2. Literature Review

Vehicle cost is influenced by multiple technical, demographic, and behavioral factors. Attributes such as engine size, horsepower, curb weight, and fuel efficiency significantly impact market value. Consumer decisions reflect both economic capacity and prior experience with vehicles, including conventional or electric models. Higher household income and familiarity with vehicles increase willingness to pay for advanced or higher-performance cars, whereas previous preferences for conventional vehicles may limit adoption of new technologies (Ling et al. 2021).

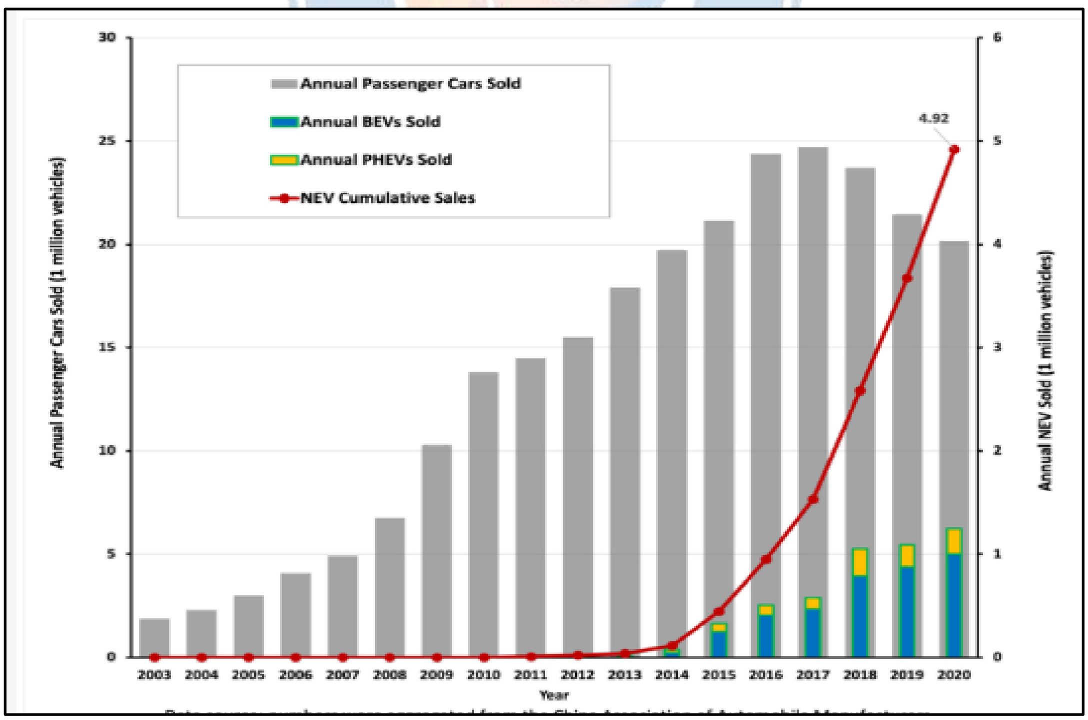

Figure 1: Conventional vehicles and NEVs selling chart

(Source: Ling et al. 2021)

Statistical modeling, including simple and multiple regression, offers a reliable framework to analyze relationships between vehicle features and price while testing assumptions such as linearity, normality, and multicollinearity. Multilevel approaches account for both individual-level and regional-level factors, enhancing predictive accuracy. Through identifying significant determinants of price, regression analysis informs manufacturers, policymakers, and consumers, guiding evidence-based strategies for pricing, investment, and market positioning in the automotive industry.

Zhang (2025) has conducted an in-depth examination of the conditions that impact the prices of automobiles and consequently the market. The study uses quantitative techniques and regression models to provide significant determinants of economic cycles, supply and demand forces, government policies, and variations in prices of raw materials.

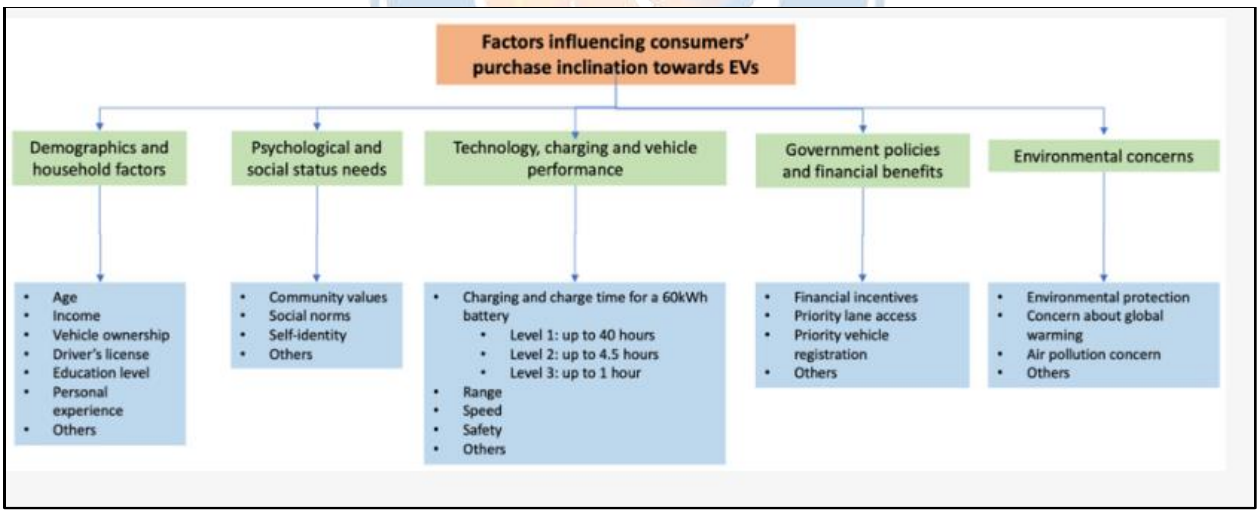

Figure 2: Factors Influencing Customer Behavior

(Source: Ling et al. 2021)

The journal also notes that in times of economic slowdown, consumers become particularly sensitive to price changes, which has an impact on market demand and stability in the supply chain.

The impact of government policies on the prices of automobiles has been determined to be relatively small. It can be concluded that the automakers and policymakers should aim to stabilize the price volatility by first dealing with the main sources of increasing and decreasing prices.

3. Methods

The study utilizes a secondary quantitative research design to determine the variables affecting car prices and how these affect the market (Zhang, 2025). Information gathered contains domestic and foreign automotive market reports, past price movements, and industry data. Economic cycles, supply and demand forces, cost of raw materials, and government policies are some of the key factors to be considered in the automotive industry. The primary analytical method used to examine the relationship between variables such as weight, height, prices, and horsepower. Simple and multiple regression models are used to ascertain the relative importance of each factor. Data preprocessing includes missing values, standardization of variables, and multicollinearity among predictors.

The fundamental assumptions of linear regression models, linearity, normality, and homoscedasticity, are systematically tested to generate robust results or assumptions used in models. The analysis also evaluates how price volatility affects market demand and the stability of market supply chains as input for automakers and policymakers to make evidence-based decisions.

4. Result analysis

4.1 Simple Linear Regression Results (Model 1)

```
> auto <- read.csv("D:/Automobile_data.csv", na.strings = c("?", ""))
> names(auto) <- make.names(names(auto))
> str(auto)
'data.frame':   205 obs. of  26 variables:
 $ symboling        : int  3 3 1 2 2 2 1 1 1 0 ...
 $ normalized.losses: int  NA NA NA 164 164 NA 158 NA 158 NA ...
 $ make             : chr  "alfa-romero" "alfa-romero" "alfa-romero" "audi" ...
 $ fuel.type        : chr  "gas" "gas" "gas" "gas" ...
 $ aspiration       : chr  "std" "std" "std" "std" ...
 $ num.of.doors     : chr  "two" "two" "two" "four" ...
 $ body.style       : chr  "convertible" "convertible" "hatchback" "sedan" ...
 $ drive.wheels     : chr  "rwd" "rwd" "rwd" "fwd" ...
 $ engine.location  : chr  "front" "front" "front" "front" ...
 $ wheel.base       : num  88.6 88.6 94.5 99.8 99.4 ...
 $ length           : num  169 169 171 177 177 ...
 $ width            : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
 $ height           : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
 $ curb.weight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
 $ engine.type      : chr  "dohc" "dohc" "ohcv" "ohc" ...
 $ num.of.cylinders : chr  "four" "four" "six" "four" ...
 $ engine.size      : int  130 130 152 109 136 136 136 136 131 131 ...
```

Figure 3: Importing and Viewing the Automobile dataset

The figure shows the importation and initial exploration of the Automobile dataset using R Studio. The 205 observations in the dataset have 26 of these variables, such as symbolizing, fuel type, engine size, horsepower, pricing, and other numeric and qualitative variables.

```
> model1 <- lm(price ~ horsepower, data = auto_m1)
> summary(model1)

Call:
lm(formula = price ~ horsepower, data = auto_m1)

Residuals:
     Min       1Q   Median       3Q      Max 
-10180.1  -2262.0   -471.1   1779.5  18276.2 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4562.175    974.995  -4.679 5.35e-06 ***
horsepower    172.206      8.866  19.424  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4685 on 197 degrees of freedom
Multiple R-squared:  0.657,	Adjusted R-squared:  0.6552 
F-statistic: 377.3 on 1 and 197 DF,  p-value: < 2.2e-16
```

Figure 4: Performance Metrics of Simple Linear Regression

The above image represents the performance metrics of a "Simple Linear Regression model" in which horsepower serves as the independent variable to determine automobile pricing. This outcome shows that the positive relationship between an increase in horsepower and a rise in car prices is statistically significant ($p < 0.001$). The adjusted R2 value of 0.6552 shows that the model explains this relationship appropriately.

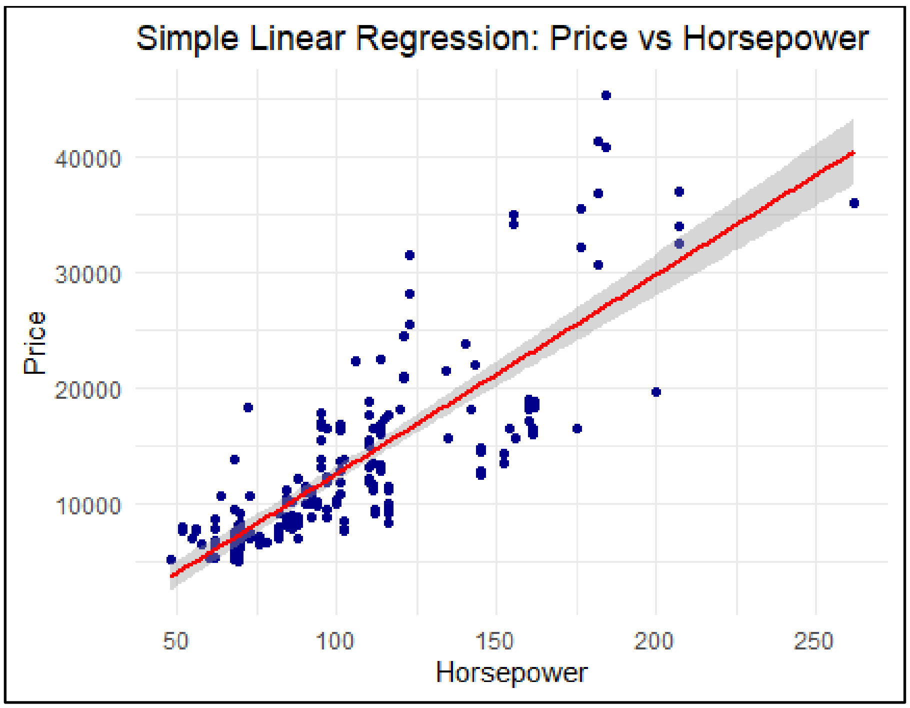

Figure 5: Simple Linear Regression depends on Price vs Horsepower

The figure presents a scatter plot with a regression line fitted on the scatter plot, depicting that there is a positive linear relationship between horsepower and price. This means that when the horsepower is high then the automobile price is also high.

Horsepower is selected after correlation analysis as it is one of the strongest predictors of price among numerical variables. This choice is accurate because of its theoretical importance to the performance of the vehicles (Muti, & Yıldız, 2023). The regression results, with a large coefficient and high fit values, justified this decision and enabled the use of horsepower as an appropriate variable in a simple linear regression analysis.

4.2 Model 1 Assumptions Check

```
> par(mfrow=c(2,2))
> plot(model1)
> par(mfrow=c(1,1))
> shapiro.test(residuals(model1))    # Normality

        Shapiro-Wilk normality test

data:  residuals(model1)
W = 0.93443, p-value = 8.206e-08

> bptest(model1)                     # Homoscedasticity

        studentized Breusch-Pagan test

data:  model1
BP = 52.04, df = 1, p-value = 5.437e-13

> dwtest(model1)                     # Independence

        Durbin-Watson test

data:  model1
DW = 0.74858, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

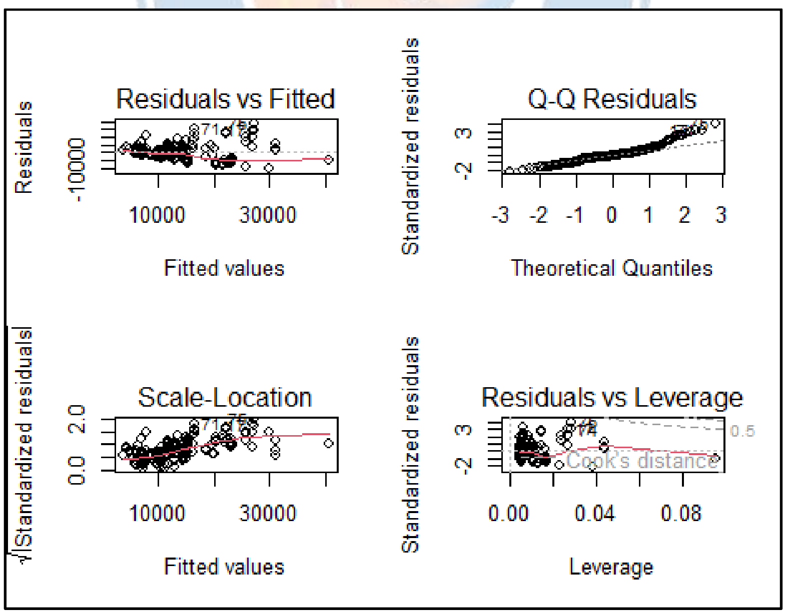

Figure 6: Assumptions of the Simple Linear Regression Model

The model assumptions are combined within diagnostic plots and statistical tests. Shapiro-Wilk test (p less than 0.05), W = 0.93443, and p = 8.206e-08, showed non-normal residuals, and

Breusch-Pagan test BP = 52.04, df = 1, p = 5.437e-13, showed the presence of heteroscedasticity. The Durbin-Watson test used confirmed autocorrelation, and the Deviations of linearity, constant variance, and normality are validated by residual plots.

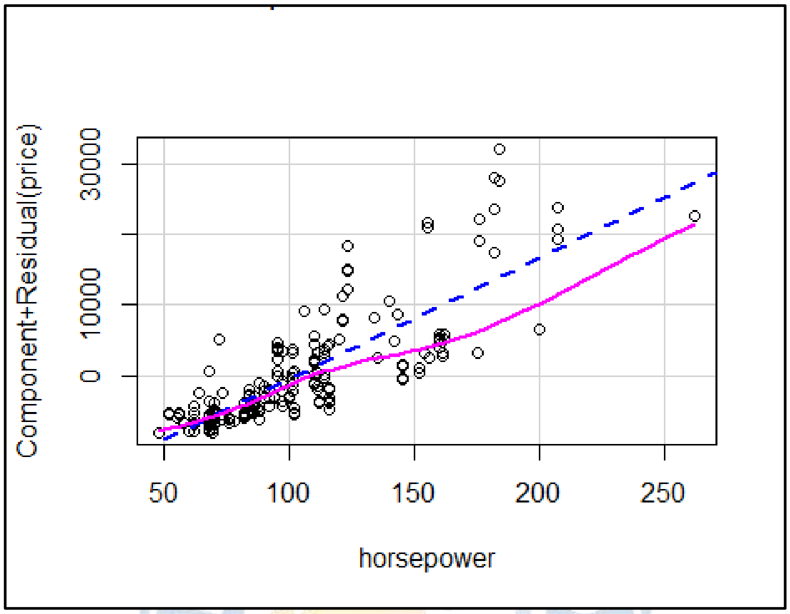

Figure 7: Linearity Plot by Model 1

The linearity plot demonstrates the connection between the price and horsepower. The blue dashed line reflects the linear regression fit, whereas the curved magenta indicates deviations, meaning that such a relationship is not linear and a transformation can be needed.

4.3 Multiple Linear Regression Results

```
> model2 <- lm(price ~ horsepower + engine.size + curb.weight, data=auto_m2)
> summary(model2)

Call:
lm(formula = price ~ horsepower + engine.size + curb.weight,
    data = auto_m2)

Residuals:
    Min      1Q  Median      3Q     Max
-9061.9 -1673.2   -37.6  1292.7 13670.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.370e+04  1.360e+03 -10.076  < 2e-16 ***
horsepower   4.913e+01  1.186e+01   4.141 5.14e-05 ***
engine.size  8.453e+01  1.318e+01   6.413 1.05e-09 ***
curb.weight  4.361e+00  9.227e-01   4.726 4.38e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3493 on 195 degrees of freedom
Multiple R-squared:  0.8112,    Adjusted R-squared:  0.8083
F-statistic: 279.4 on 3 and 195 DF,  p-value: < 2.2e-16
```

Figure 8: Evaluation Output of Multiple Linear Regression

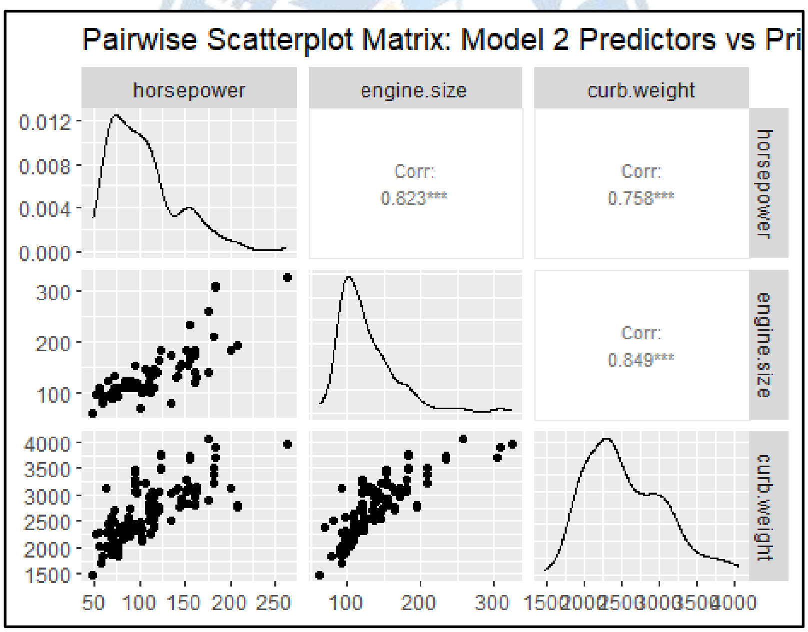

Figure 9: Pairwise scatterplot matrix for Multiple Linear Regression

The Engine size, horsepower, and curb weight are used as predictors of car pricing in a multiple linear regression equation. According to these findings, the model explains about 81% of price movements, with R 2 = 0.8112 and adjusted R 2 = 0.8083. Additionally, all of the predictors are statistically significant ($p < 0.001$), indicating that they have a considerable impact on the price. The residual's standard error (3493) is lower than that of the simple regression, indicating higher accuracy.

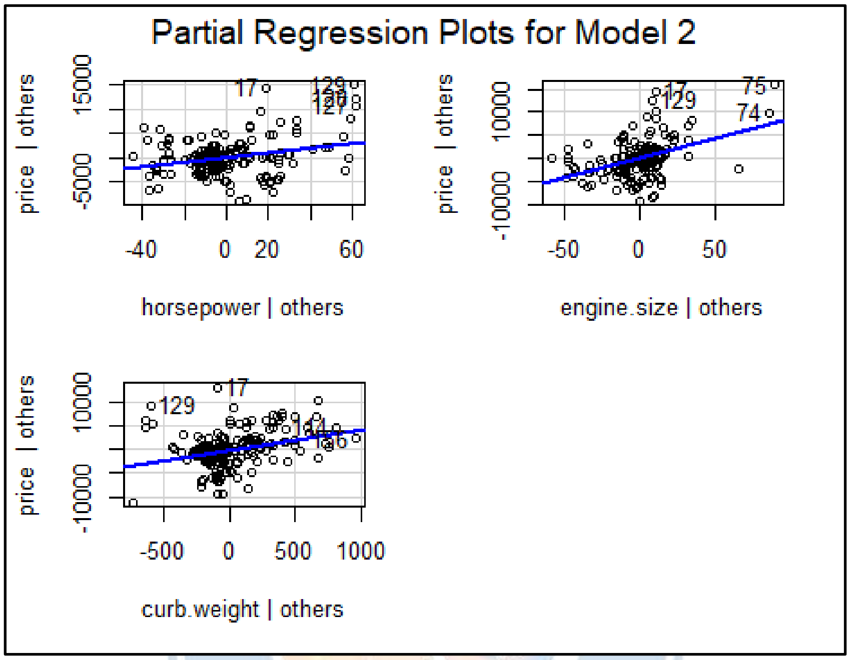

Figure 10: Partial Regression Plots for Multiple Regression Model

The predictors (engine size-curb weight, horsepower-engine size) are selected based on their theoretical significance to car pricing and supported by correlation analysis. The explanatory power is significantly increased by these variables in comparison to the simple model (R 2 = 0.657).

4.4 Model 2 Assumptions Check

```
> par(mfrow=c(2,2))
> plot(model2)
> par(mfrow=c(1,1))
> shapiro.test(residuals(model2))    # Normality

        Shapiro-Wilk normality test

data:  residuals(model2)
W = 0.9599, p-value = 2.025e-05

> bptest(model2)                     # Homoscedasticity

        studentized Breusch-Pagan test

data:  model2
BP = 75.14, df = 3, p-value = 3.381e-16

> vif(model2)                        # Multicollinearity
 horsepower  engine.size curb.weight
   3.220484     4.915198    3.734217
> crPlots(model2)                     # Linearity
```

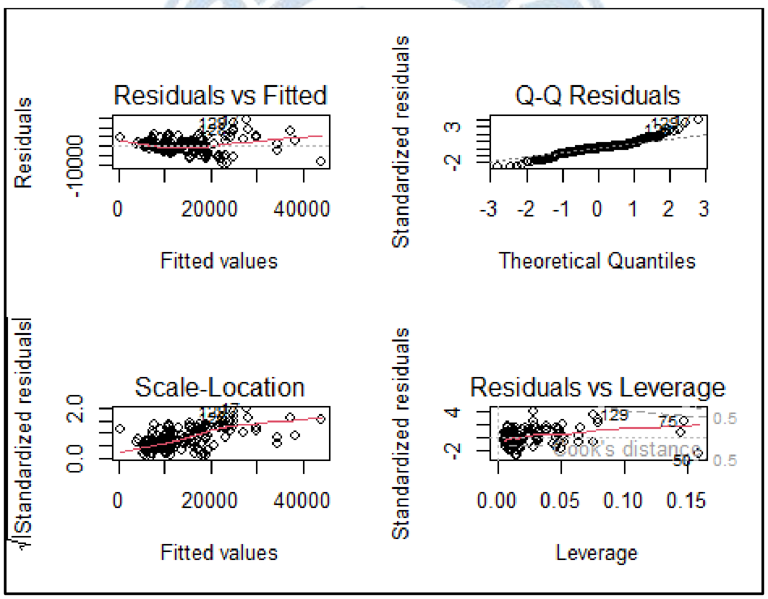

Figure 11: Assumptions of the Model 2

The multiple regression model has performed according to the assumptions check. The Breusch-Pagan test (p = 3.381e-16) verifies heteroscedasticity, whereas the Shapiro-Wilk test (p = 2.025e-05) shows non-normal residuals. Moderate multicollinearity is seen in the VIF data (horsepower: 3.22, engine.size: 4.91, curb.weight: 3.73). All predictors show positive linear connections in component-residual plots, but engine size and curb weight have better linearity than horsepower, indicating that they make a significant contribution even in assumption violations.

4.5 Comparison of Models

```
> vif(model2)                          # Multicollinearity
 horsepower engine.size curb.weight
   3.220484    4.915198    3.734217
> crPlots(model2)                       # Linearity
> adj_r2_model1 <- summary(model1)$adj.r.squared
> adj_r2_model2 <- summary(model2)$adj.r.squared
> cat("Adjusted R² - Model 1:", adj_r2_model1, " Model 2:", adj_r2_model2, "\n")
Adjusted R² - Model 1: 0.6552226  Model 2: 0.808342
> AIC(model1, model2)
       df      AIC
model1  3 3932.665
model2  5 3817.784
> BIC(model1, model2)
       df      BIC
model1  3 3942.545
model2  5 3834.250
> # RMSE comparison
> set.seed(123)
> idx <- createDataPartition(auto_m2$price, p=0.8, list=FALSE)
> train <- auto_m2[idx,]; test <- auto_m2[-idx,]
> rmse <- function(model, data) sqrt(mean((data$price - predict(model, newdata=data))^2))
> cat("RMSE M1:", rmse(model1,test), " RMSE M2:", rmse(model2,test), "\n")
RMSE M1: 4980.016  RMSE M2: 4059.804
>
```

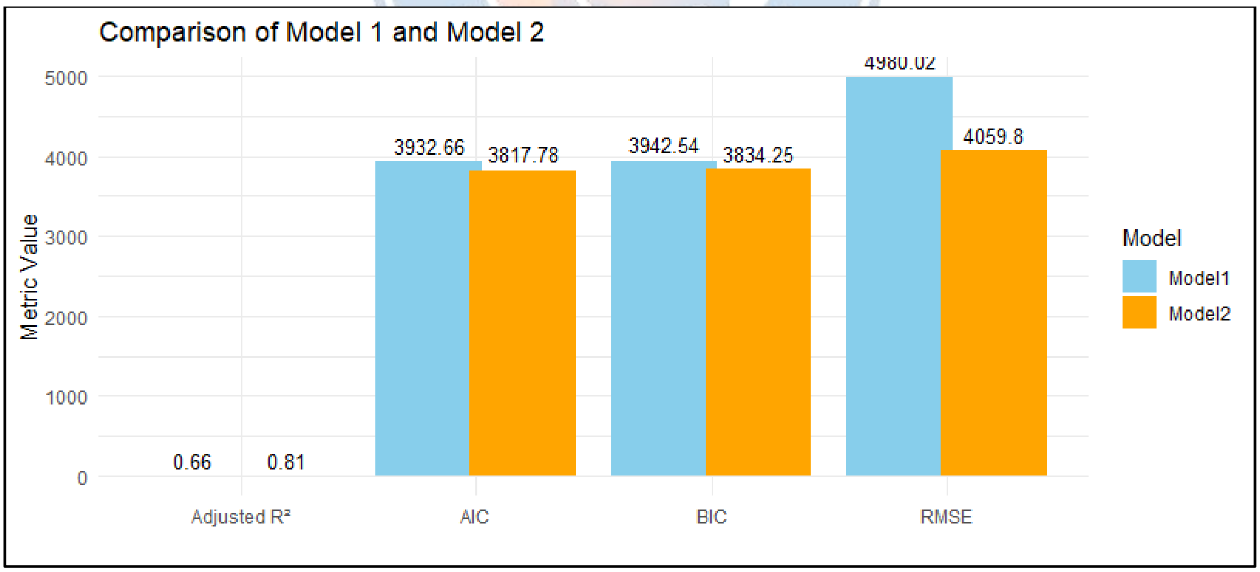

Figure 12: Comparing Model 1 vs Model 2

Model 2 "(multiple linear regression)" performs better than Model 1 (explaining 66% of price variance), with a higher adjusted R2 (0.81 vs. 0.66). Also, Model 2 exhibits lower RMSE

(4059.8 vs. 4980.02), AIC (3817.78 vs. 3932.66), and BIC (3834.25 vs. 3942.54), suggesting improved model fit and prediction accuracy even with assumption violations (Mammadov, 2021).

4.6 Significance of Independent Variable in Simple Linear Regression (Model 1)

```
> rmse <- function(model, data) sqrt(mean((data$price - predict(model, newdata=data))^2))
> cat("RMSE M1:", rmse(model1,test), " RMSE M2:", rmse(model2,test), "\n")
RMSE M1: 4980.016  RMSE M2: 4059.804
> p_value_hp <- summary(model1)$coefficients["horsepower", "Pr(>|t|)"]
> alpha <- 0.05
> cat("P-value for horsepower in Model 1:", p_value_hp, "\n")
P-value for horsepower in Model 1: 1.189128e-47
> if (p_value_hp < alpha) {
+   cat("Conclusion: At 0.05 significance, horsepower significantly predicts price.\n")
+ } else {
+   cat("Conclusion: At 0.05 significance, horsepower does NOT significantly predict price.\n")
+ }
Conclusion: At 0.05 significance, horsepower significantly predicts price.
```

Figure 13: Significance of the Independent Variable in Model 1

At the 0.05 significance level, horsepower makes a highly significant contribution to Model 1. With a p-value of 1.189128e-47 (much less than $\alpha = 0.05$), this rejects the null hypothesis and concludes that horsepower significantly predicts automobile price in the simple linear regression model, demonstrating strong statistical evidence for this relationship (Chaubey *et al.* 2023).

4.7 Significance of Independent Variables in Multiple Linear Regression (Model 2)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.370e+04  1.360e+03 -10.076  < 2e-16 ***
horsepower   4.913e+01  1.186e+01   4.141 5.14e-05 ***
engine.size  8.453e+01  1.318e+01   6.413 1.05e-09 ***
curb.weight  4.361e+00  9.227e-01   4.726 4.38e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3493 on 195 degrees of freedom
Multiple R-squared:  0.8112,    Adjusted R-squared:  0.8083
F-statistic: 279.4 on 3 and 195 DF,  p-value: < 2.2e-16

> fstat <- summary(model2)$fstatistic
> f_value <- fstat[1]
> df1 <- fstat[2]
> df2 <- fstat[3]
> f_p_value_model2 <- pf(f_value, df1, df2, lower.tail = FALSE)
> cat("Overall F-test p-value for Model 2:", f_p_value_model2, "\n")
Overall F-test p-value for Model 2: 2.532819e-70
> if (f_p_value_model2 < alpha) {
+   cat("Overall Conclusion: At 0.05 significance, Model 2 is statistically significant.\n")
+ } else {
+   cat("Overall Conclusion: At 0.05 significance, Model 2 is NOT statistically significant.\n")
+ }
Overall Conclusion: At 0.05 significance, Model 2 is statistically significant.
```

Figure 14: Significance of Independent Variables in Model 2

At the 0.05 significance level, all independent variables in Model 2 make significant contributions to predicting automobile price. Each unit increase in horsepower adds $49.13 to the price (p = 5.14e-05), engine size increases the price by $84.53 per unit (p = 1.05e-09), and curb weight contributes $4.36 per unit (p = 4.38e-06). The overall F-test confirms that these predictors collectively explain price variation significantly.

6. Conclusion

This statistical investigation used regression modeling to successfully identify important factors that influence car prices. The multiple regression model outperformed the basic model, accounting for 81% of the price variance as opposed to 66%. Curb weight, horsepower, and engine size were found to be important predictors, and each had a considerable impact on price.

References

Chaubey, V. P., Sareen, D., Harshita, C., Sharma, S., Rana, R. P. S., & Batra, I. (2023, December). Enhancing Predictive Accuracy in Used Car Pricing Models: A Comparative Analysis of Regression Techniques. In *International Conference on Advanced Informatics for Computing Research* (pp. 24-35). Cham: Springer Nature Switzerland. https://link.springer.com/chapter/10.1007/978-3-031-84059-33

Ling, Z., Cherry, C. R., & Wen, Y. (2021). Determining the factors that influence electric vehicle adoption: A stated preference survey study in Beijing, China. *Sustainability*, *13*(21), 11719. DOI: https://doi.org/10.3390/su132111719

Mammadov, H. (2021). Car price prediction in the USA by using liner regression. *International Journal of Economic Behavior (IJEB)*, *11*(1), 99-108. DOI: https://doi.org/10.14276/2285-0430.3049

Muti, S., & Yıldız, K. (2023). Using linear regression for used car price prediction. *International Journal of Computational and Experimental Science and Engineering*, *9*(1), 11-16. DOI: https://doi.org/10.22399/ijcesen.1070505

Zhang, J. (2025). Analysis of Factors Affecting Automobile Prices and Their Impact on the Market. *Science and Technology of Engineering, Chemistry and Environmental Protection*, *1*(4). DOI: https://doi.org/10.61173/v0nnx252

Dataset Link: https://www.kaggle.com/code/mobinapoulaei/automobile-prediction/input