# Statistical Analysis of Automobile Data

Group 5

OMIS 645: Applied Business Analytics SAS

OMIS Department, Northern Illinois University

Prof. Xinyan Cao

Lakshmi Suma Latha Gandham - Z1983150

Leela Sai Kiran Baru - Z1981392

Siruvuri Sai Ganesh Varma - Z2007457

Hasitha Sharon Sigatapu – Z1983992

## Introduction

This report presents an analysis of automobile data to investigate the relationships between various vehicle characteristics and key performance metrics. The analysis employs statistical techniques, including Analysis of Variance (ANOVA), regression modelling and Moderation Analysis. The goal is to gain insights into the factors influencing fuel efficiency and pricing, which can inform decision-making processes for automobile manufacturers, dealers, and consumers.

## About the dataset

The automobile dataset provides a comprehensive overview of various attributes of automobiles, which can be used for exploratory data analysis and modeling. The key attributes such as make, fuel type, dimensions, engine specifications, and drive configuration, each provides unique insights into automobile characteristics. Moreover, the dataset allows the investigation of relationships between different features and can be utilized to build predictive models for tasks such as price estimation, fuel efficiency analysis, or classification of vehicle types.

This dataset is taken from Kaggle, https://www.kaggle.com/datasets/toramky/automobile-dataset?resource=download .

To perform the analysis, several problem questions were to be analyzed and discussed:

**Question 1:** Is there evidence to suggest that the mean highway mpg varies significantly among various body styles of the car?

**Question 2:**Are there any notable relations or patterns between highway MPG and each of the predictor variables, including engine size, horsepower, curb weight, and body style?

**Question 3:**Does the multiple regression model demonstrate significant predictive capability in estimating the price of automobiles based on features such as engine size, horsepower, and fuel type?

**Question 4:**Is there evidence to suggest that the relationship between curb weight and price of the automobiles is influenced or moderated by horsepower?

## Insights, Interpretations and Key Discoveries

**Question 1:** *Is there evidence to suggest that the mean highway mpg varies significantly among various body styles of the car?*

For this question, we have utilized the ANOVA (Analysis of Variance) one-way test. This statistical method is used to compare the means of three or more groups to determine if there are significant differences among them.

In this scenario, the model tests whether there are significant differences in the mean highway mpg across different body styles. Let’s assume that the Null Hypothesis is , there is no statistically significant difference in the average highway mpg amongst the 5 different card body styles and the Alternative Hypothesis be there is a statistical difference in the average mileage highway between at least two of the five body styles.

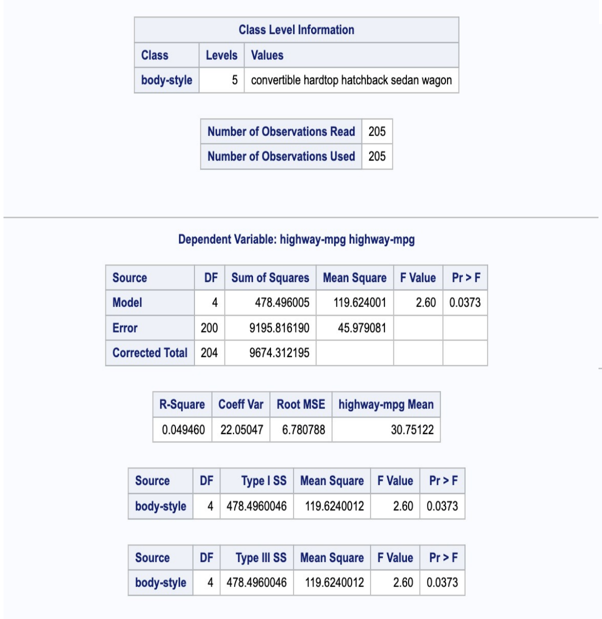
From the results of the test, it’s observed that the F-statistic value of 2.60 and the corresponding p-value of 0.03 at 10% significance level  indicate that the model is not significant, and we do not reject the null hypothesis. In other words, the evidence suggests that there is no statistically significant difference in the average highway mpg among the different body style categories.

However, it’s important to test for the Homogeneity of the Variance and here we performed the Levene’s test and here are the additional insights we have obtained. The p-value is 0.2402 which is greater than 0.01(significance). Considering this we can conclude that the variances are equal, and the homogeneity of variance is met. Additionally, the Levene's test for homogeneity of variance yielded a p-value of 0.2402, indicating that the assumption of equal variances across the different body styles is met. This further supports the validity of the ANOVA results.

Lastly, considering the ANOVA matrix it’s evident that all the p-values of the predictors are greater than the significance value i.e. 0.01 stating that the model is insignificant.

Considering all the above test results we conclude there is insufficient evidence to reject the null hypothesis. This means that there is no statistically significant difference in the average highway miles per gallon (mpg) among the different body styles of cars. Further investigation or larger sample sizes may be needed to draw more definitive conclusions.

Attaching the results of the test’s performed for reference,



| Class Level Information | | |
|---|---|---|
| Class | Levels | Values |
| body-style | 5 | convertible hardtop hatchback sedan wagon |

| | |
|---|---|
| Number of Observations Read | 205 |
| Number of Observations Used | 205 |

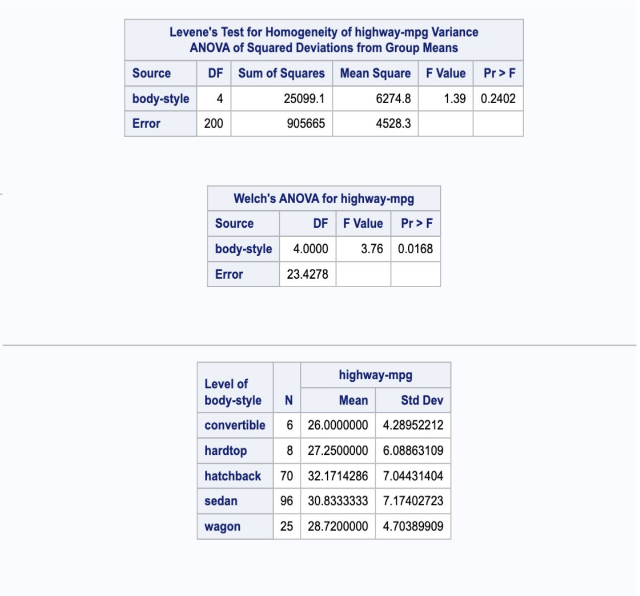
Dependent Variable: highway-mpg highway-mpg

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 478.496005 | 119.624001 | 2.60 | 0.0373 |
| Error | 200 | 9195.816190 | 45.979081 | | |
| Corrected Total | 204 | 9674.312195 | | | |

| R-Square | Coeff Var | Root MSE | highway-mpg Mean |
|---|---|---|---|
| 0.049460 | 22.05047 | 6.780788 | 30.75122 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| body-style | 4 | 478.4960046 | 119.6240012 | 2.60 | 0.0373 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| body-style | 4 | 478.4960046 | 119.6240012 | 2.60 | 0.0373 |



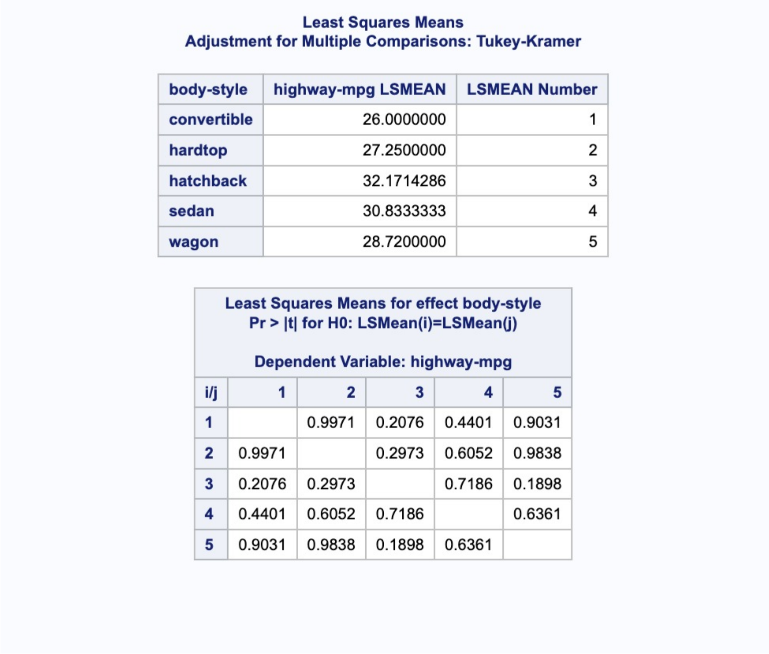
Levene's Test for Homogeneity of highway-mpg Variance
ANOVA of Squared Deviations from Group Means

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| body-style | 4 | 25099.1 | 6274.8 | 1.39 | 0.2402 |
| Error | 200 | 905665 | 4528.3 | | |

Welch's ANOVA for highway-mpg

| Source | DF | F Value | Pr > F |
|---|---|---|---|
| body-style | 4.0000 | 3.76 | 0.0168 |
| Error | 23.4278 | | |

| Level of body-style | N | highway-mpg | |
|---|---|---|---|
| | | Mean | Std Dev |
| convertible | 6 | 26.0000000 | 4.28952212 |
| hardtop | 8 | 27.2500000 | 6.08863109 |
| hatchback | 70 | 32.1714286 | 7.04431404 |
| sedan | 96 | 30.8333333 | 7.17402723 |
| wagon | 25 | 28.7200000 | 4.70389909 |



Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

| body-style | highway-mpg LSMEAN | LSMEAN Number |
|---|---|---|
| convertible | 26.0000000 | 1 |
| hardtop | 27.2500000 | 2 |
| hatchback | 32.1714286 | 3 |
| sedan | 30.8333333 | 4 |
| wagon | 28.7200000 | 5 |

Least Squares Means for effect body-style
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: highway-mpg

| i/j | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | | 0.9971 | 0.2076 | 0.4401 | 0.9031 |
| 2 | 0.9971 | | 0.2973 | 0.6052 | 0.9838 |
| 3 | 0.2076 | 0.2973 | | 0.7186 | 0.1898 |
| 4 | 0.4401 | 0.6052 | 0.7186 | | 0.6361 |
| 5 | 0.9031 | 0.9838 | 0.1898 | 0.6361 | |

**Question 2 :** *Are there any notable relations or patterns between highway MPG and each of the predictor variables, including engine size, horsepower, curb weight, height, wheel-base, stroke, compression ratio, peak rmp and body style?*

For this question, we have utilized the Multiple Regression test.

The Analysis of Variance table shows that the overall model is statistically significant (F-value = 111.04, p-value < 0.0001), indicating that at least one of the predictor variables is significantly associated with the highway MPG.

The R-Square value of 0.8192 suggests that approximately 81.92% of the variation in highway MPG can be explained by the combination of the predictor variables  included in the model.

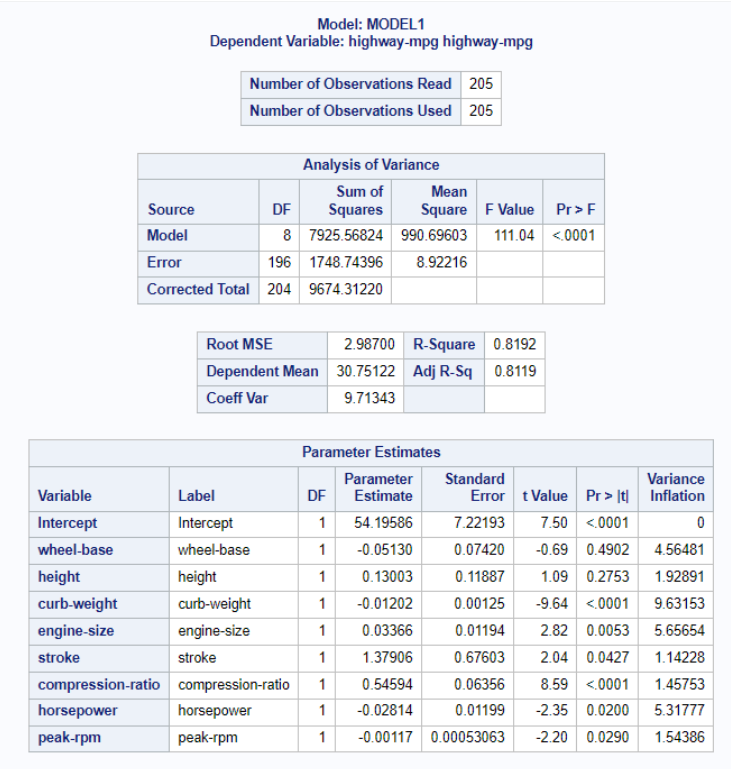
Looking at the Parameter Estimates from the test results obtained, we can interpret the coefficients for each predictor variable:

1. wheel-base: A coefficient of -0.0513 means that for every one-inch increase in wheelbase, highway-mpg is expected to decrease by 0.0513 mpg keeping the rest of the variables constant. This is an insignificant variable as the p-value >0.05.
2. height: A coefficient of 0.1300 means that for every one-inch increase in height, highway-mpg is expected to increase by 0.1300 mpg keeping the rest of the variables constant. This is an insignificant variable as the p-value >0.05.
3. curb-weight: A coefficient of -0.0120 means that for every one-pound increase in curb weight, highway-mpg is expected to decrease by 0.0120 mpg keeping the rest of the variables constant. This is an significant variable as the p-value <0.05.
4. engine-size: A coefficient of 0.0337 means that for every one-cubic inch increase in engine size, highway-mpg is expected to increase by 0.0337 mpg keeping the rest of the variables constant. This is an significant variable as the p-value <0.05
5. stroke: A coefficient of 1.3791 means that for every one-inch increase in stroke, highway-mpg is expected to increase by 1.3791 mpg keeping the rest of the variables constant. This is an significant variable as the p-value <0.05.
6. compression-ratio: A coefficient of 0.5459 means that for every one-unit increase in compression ratio, highway-mpg is expected to increase by 0.5459 mpg keeping the rest of the variables constant. This is an significant variable as the p-value <0.05.

7. horsepower: A coefficient of -0.0281 means that for every one-horsepower increase in horsepower, highway-mpg is expected to decrease by 0.0281 mpg keeping the rest of the variables constant. This is an significant variable as the p-value <0.05.
8. peak-rpm: A coefficient of -0.0012 means that for every one-rpm increase in peak rpm, highway-mpg is expected to decrease by 0.0012 mpg keeping the rest of the variables constant. This is an significant variable as the p-value <0.05.

In conclusion, this Regression Model states that the predictor variables like engine size, stroke, and compression ratio have a positive impact on fuel efficiency(highway-mpg), others like horsepower, curb-weight, and peak rpm have a negative influence. These findings can inform future car design and consumer choices when considering fuel efficiency alongside other car attributes.

Attaching the results of the test's performed for reference,



Model: MODEL1
Dependent Variable: highway-mpg highway-mpg

| | |
|---|---|
| Number of Observations Read | 205 |
| Number of Observations Used | 205 |

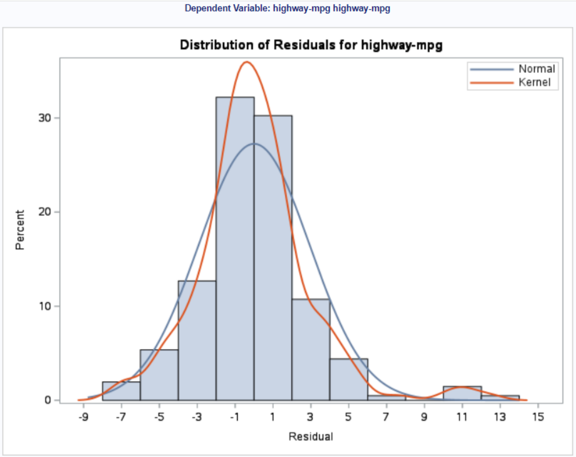
Analysis of Variance

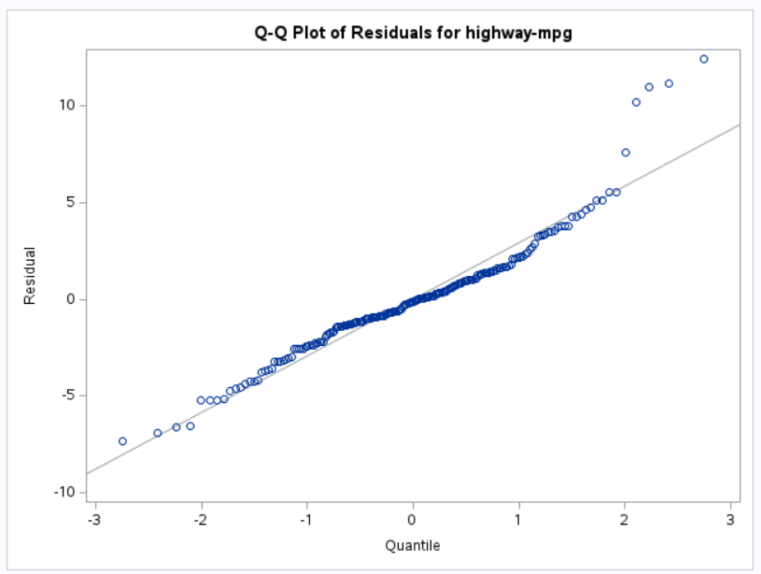
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 8 | 7925.56824 | 990.69603 | 111.04 | <.0001 |
| Error | 196 | 1748.74396 | 8.92216 | | |
| Corrected Total | 204 | 9674.31220 | | | |

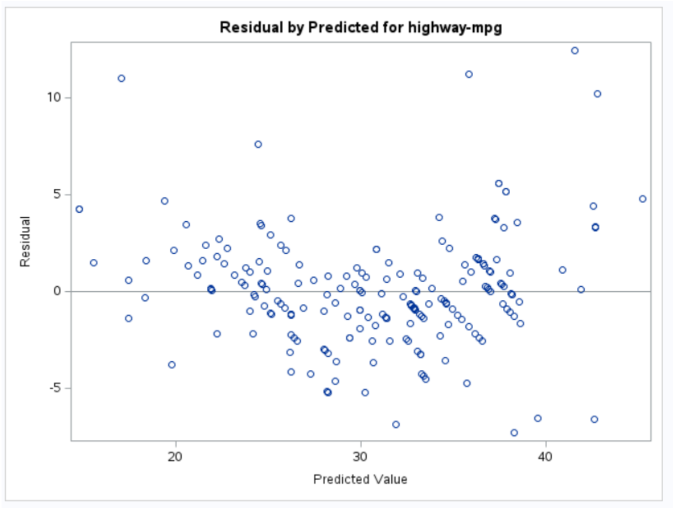
| | | | |
|---|---|---|---|
| Root MSE | 2.98700 | R-Square | 0.8192 |
| Dependent Mean | 30.75122 | Adj R-Sq | 0.8119 |
| Coeff Var | 9.71343 | | |

Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
|---|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 54.19586 | 7.22193 | 7.50 | <.0001 | 0 |
| wheel-base | wheel-base | 1 | -0.05130 | 0.07420 | -0.69 | 0.4902 | 4.56481 |
| height | height | 1 | 0.13003 | 0.11887 | 1.09 | 0.2753 | 1.92891 |
| curb-weight | curb-weight | 1 | -0.01202 | 0.00125 | -9.64 | <.0001 | 9.63153 |
| engine-size | engine-size | 1 | 0.03366 | 0.01194 | 2.82 | 0.0053 | 5.65654 |
| stroke | stroke | 1 | 1.37906 | 0.67603 | 2.04 | 0.0427 | 1.14228 |
| compression-ratio | compression-ratio | 1 | 0.54594 | 0.06356 | 8.59 | <.0001 | 1.45753 |
| horsepower | horsepower | 1 | -0.02814 | 0.01199 | -2.35 | 0.0200 | 5.31777 |
| peak-rpm | peak-rpm | 1 | -0.00117 | 0.00053063 | -2.20 | 0.0290 | 1.54386 |


Dependent Variable: highway-mpg highway-mpg
Distribution of Residuals for highway-mpg
Normal
Kernel
Percent
0
10
20
30
-9
-7
-5
-3
-1
1
3
5
7
9
11
13
15
Residual


Q-Q Plot of Residuals for highway-mpg
Residual
10
5
0
-5
-10
-3
-2
-1
0
1
2
3
Quantile


Residual by Predicted for highway-mpg
Residual
10
5
0
-5
20
30
40
Predicted Value

**Question 3:** *Does the multiple regression model demonstrate significant predictive capability in estimating the price of automobiles based on features such as engine size, horsepower, and fuel type?*

For this question, we have utilized the Multiple Regression technique with Log Transformation. This is a statistical technique used to understand the relationship between one dependent variable and two or more independent variables.

The output of the analysis shows the results of a multiple linear regression analysis performed to predict the price of automobiles based on their characteristics such as engine size, horsepower, and fuel type.

After utilizing the Multiple Regression technique, we can observe that the points in the Q-Q Plot and Residual Plots the variables are all together so it's not random hence we did the log transformation for the variables.

The Analysis of Variance table shows that the overall model is statistically significant as p-value < 0.0001, indicating that at least one of the predictor variables is significantly associated with the price of automobiles. We have performed the Log Transformation on the variables to improve normality and stabilize the variance and here are the final test statistics we obtained where the F-value is 87.99 and the R-Square value being 0.8193.

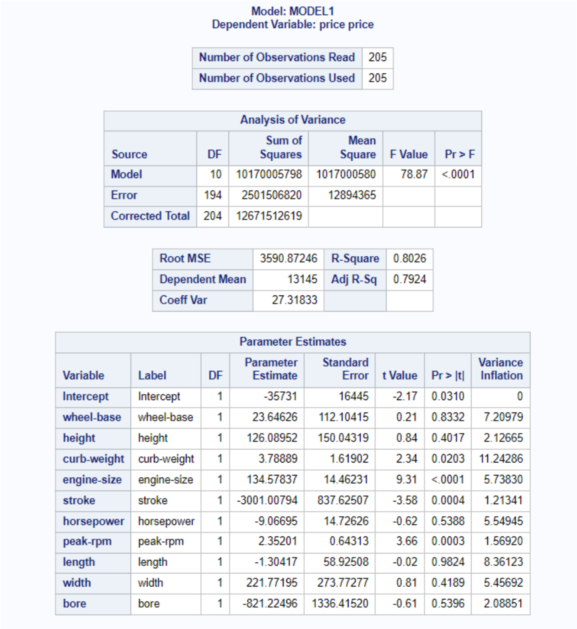
Looking at the Parameter Estimates from the test results obtained, we can interpret the coefficients for each predictor variables significance considering the p-value<0.05 and Variation Inflation <10. From the test results we can understand that the highest predicted variables are engine-size, stroke and peak-rpm.

1. **Engine Size:** A coefficient of 0.00348 means that for every one unit increase in engine size, highway-mpg is expected to increase by 0.00348 mpg, keeping all other variables constant. This indicates a positive relationship between engine size and fuel efficiency.
2. **Stroke:** A coefficient of 0.11236 means that for every one unit increase in stroke, highway-mpg is expected to decrease by 0.11236 mpg, keeping all other variables constant. Stroke is the distance a piston travels in an engine in one complete cycle. A longer stroke may allow for less fuel burning.
3. **Peak-rpm:** A coefficient of 0.000124 means that for every one unit increase in stroke, highway-mpg is expected to increase by 0.000124 mpg, keeping all other variables

constant. Cars that can reach higher maximum engine speeds (rpm) tend to have slightly higher fuel efficiency.

Finally, this analysis shows that engine design characteristics such as the size and peak-rpm and stroke significantly influence highway miles per gallon. Interestingly, the model suggests a negative association between stroke and fuel efficiency. Here, cars with a longer stroke tend to have slightly lower fuel economy. While a substantial portion (81.92%) of the variation in mpg is explained by the model.

Attaching the results of the test’s performed for reference,



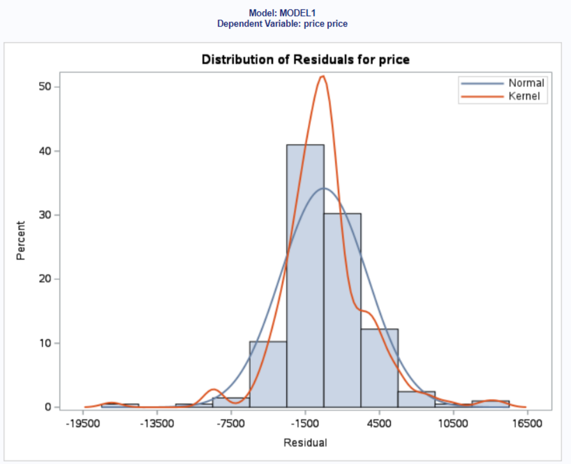
Model: MODEL1
Dependent Variable: price price

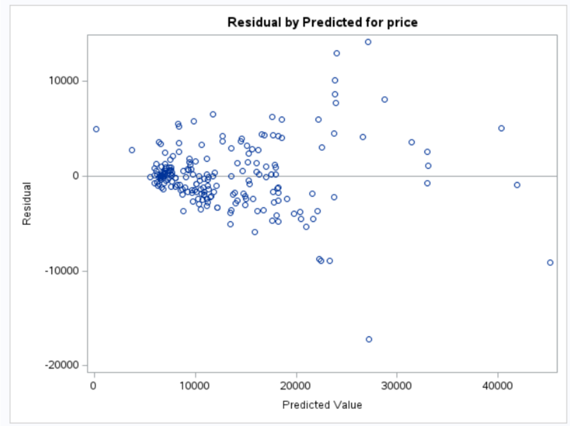
| | |
|---|---|
| Number of Observations Read | 205 |
| Number of Observations Used | 205 |

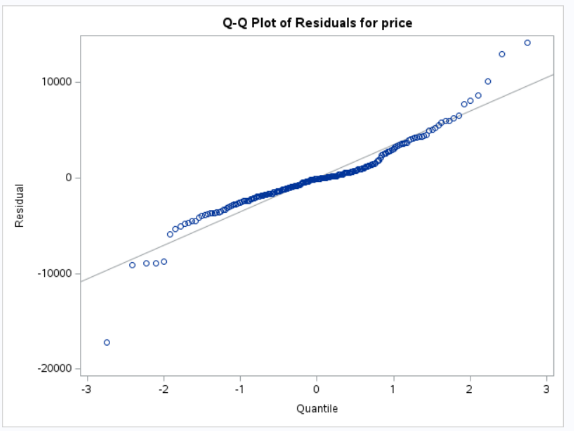
| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 10 | 10170005798 | 1017000580 | 78.87 | <.0001 |
| Error | 194 | 2501506820 | 12894365 | | |
| Corrected Total | 204 | 12671512619 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 3590.87246 | R-Square | 0.8026 |
| Dependent Mean | 13145 | Adj R-Sq | 0.7924 |
| Coeff Var | 27.31833 | | |

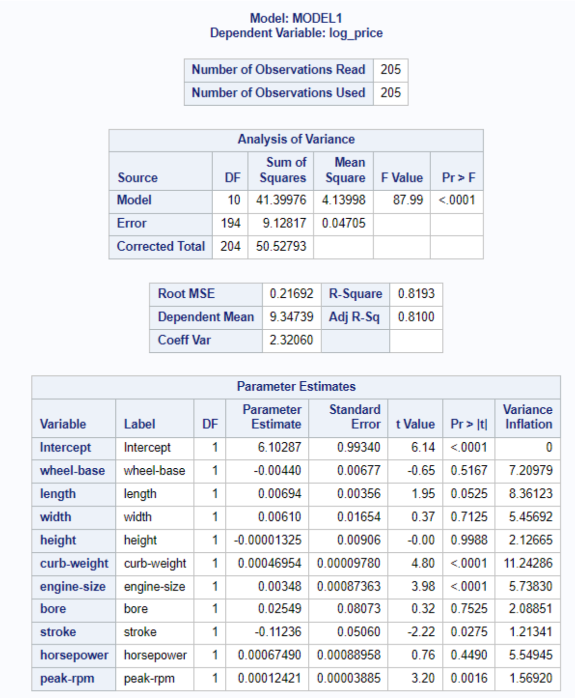
| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
| Intercept | Intercept | 1 | -35731 | 16445 | -2.17 | 0.0310 | 0 |
| wheel-base | wheel-base | 1 | 23.64626 | 112.10415 | 0.21 | 0.8332 | 7.20979 |
| height | height | 1 | 126.08952 | 150.04319 | 0.84 | 0.4017 | 2.12665 |
| curb-weight | curb-weight | 1 | 3.78889 | 1.61902 | 2.34 | 0.0203 | 11.24286 |
| engine-size | engine-size | 1 | 134.57837 | 14.46231 | 9.31 | <.0001 | 5.73830 |
| stroke | stroke | 1 | -3001.00794 | 837.62507 | -3.58 | 0.0004 | 1.21341 |
| horsepower | horsepower | 1 | -9.06695 | 14.72626 | -0.62 | 0.5388 | 5.54945 |
| peak-rpm | peak-rpm | 1 | 2.35201 | 0.64313 | 3.66 | 0.0003 | 1.56920 |
| length | length | 1 | -1.30417 | 58.92508 | -0.02 | 0.9824 | 8.36123 |
| width | width | 1 | 221.77195 | 273.77277 | 0.81 | 0.4189 | 5.45692 |
| bore | bore | 1 | -821.22496 | 1336.41520 | -0.61 | 0.5396 | 2.08851 |







After Performing the Log Transformation on the variables, the test results are as follows,



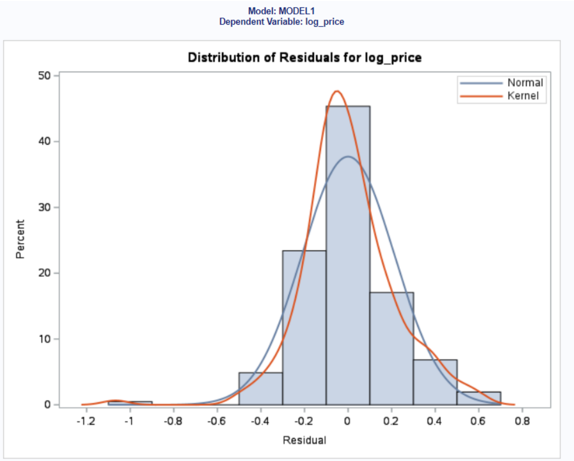
Model: MODEL1
Dependent Variable: log_price

| | |
|---|---|
| Number of Observations Read | 205 |
| Number of Observations Used | 205 |

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 10 | 41.39976 | 4.13998 | 87.99 | <.0001 |
| Error | 194 | 9.12817 | 0.04705 | | |
| Corrected Total | 204 | 50.52793 | | | |

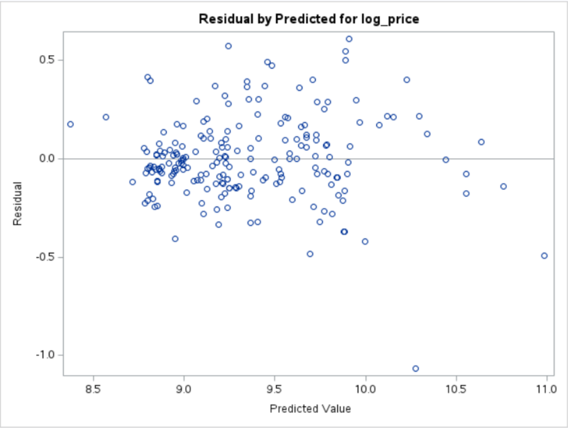
| | | | |
|---|---|---|---|
| Root MSE | 0.21692 | R-Square | 0.8193 |
| Dependent Mean | 9.34739 | Adj R-Sq | 0.8100 |
| Coeff Var | 2.32060 | | |

| Parameter Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
| Intercept | Intercept | 1 | 6.10287 | 0.99340 | 6.14 | <.0001 | 0 |
| wheel-base | wheel-base | 1 | -0.00440 | 0.00677 | -0.65 | 0.5167 | 7.20979 |
| length | length | 1 | 0.00694 | 0.00356 | 1.95 | 0.0525 | 8.36123 |
| width | width | 1 | 0.00610 | 0.01654 | 0.37 | 0.7125 | 5.45692 |
| height | height | 1 | -0.00001325 | 0.00906 | -0.00 | 0.9988 | 2.12665 |
| curb-weight | curb-weight | 1 | 0.00046954 | 0.00009780 | 4.80 | <.0001 | 11.24286 |
| engine-size | engine-size | 1 | 0.00348 | 0.00087363 | 3.98 | <.0001 | 5.73830 |
| bore | bore | 1 | 0.02549 | 0.08073 | 0.32 | 0.7525 | 2.08851 |
| stroke | stroke | 1 | -0.11236 | 0.05060 | -2.22 | 0.0275 | 1.21341 |
| horsepower | horsepower | 1 | 0.00067490 | 0.00088958 | 0.76 | 0.4490 | 5.54945 |
| peak-rpm | peak-rpm | 1 | 0.00012421 | 0.00003885 | 3.20 | 0.0016 | 1.56920 |

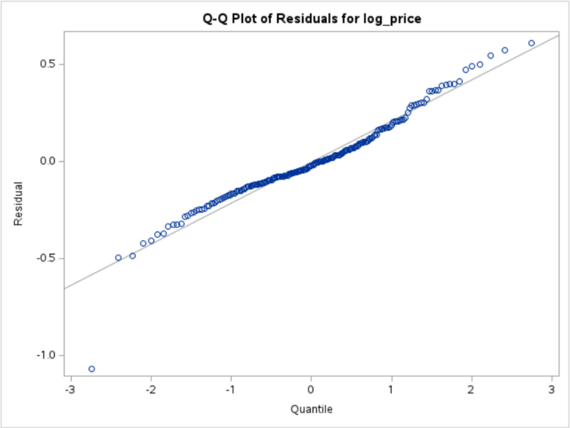


Model: MODEL1
Dependent Variable: log_price

Distribution of Residuals for log_price



Residual by Predicted for log_price



Q-Q Plot of Residuals for log_price

**Question 4:** *Is there evidence to suggest that the relationship between curb weight and price of the automobiles is influenced or moderated by horsepower?*

For this question, we have utilized the Moderation test. In statistics, a moderation test (analysis) investigates how the relationship between two variables (independent and dependent) is influenced by a third variable (moderator).

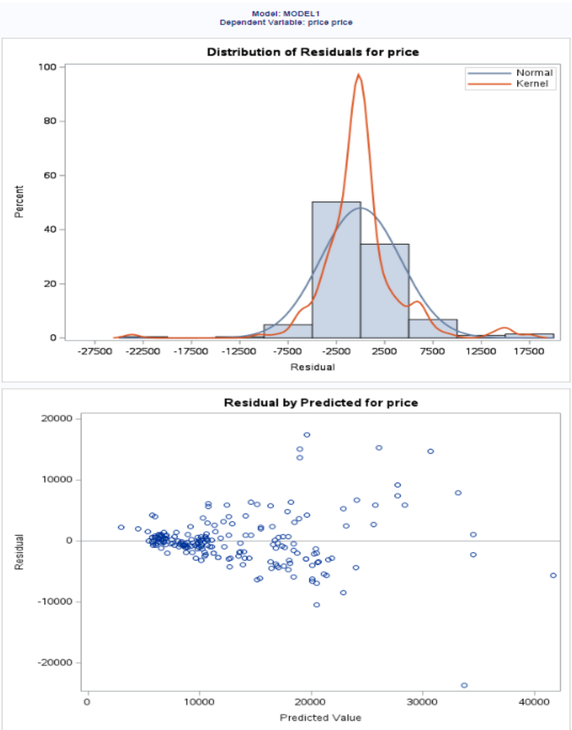
The output of the analysis shows the results of there is a evidence to suggest that the relationship between curb weight and price of the automobiles is influenced or moderated by horsepower.

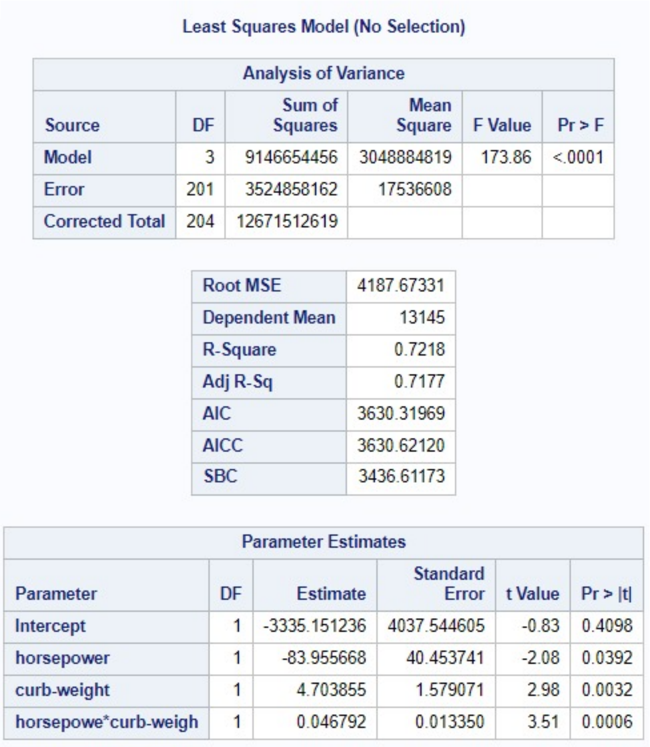
The model is considered significant as the p-value(i.e <.001) is less than the significance level. The horsepower*curb-weight have p-value < 0.05 so these both variables are significant.

$$\hat{Y}_i = -3335.151236 - 83.955668\, horsepower + 4.703855\, curbweight + 0.046792\, (horsepower\ X\ curbweight)$$

Horsepower*curbweight: The coefficient for horsepower*curbweight is 0.046792 which means that One unit increase in horsepower*curbweight assuming all other variables constant, the price is expected to increase by 0.046792. The p-value is less that alpha(0.05). By this we can say that horsepower moderates the relationship between curb weight and price.

Attaching the results of the test’s performed for reference,





Least Squares Model (No Selection)

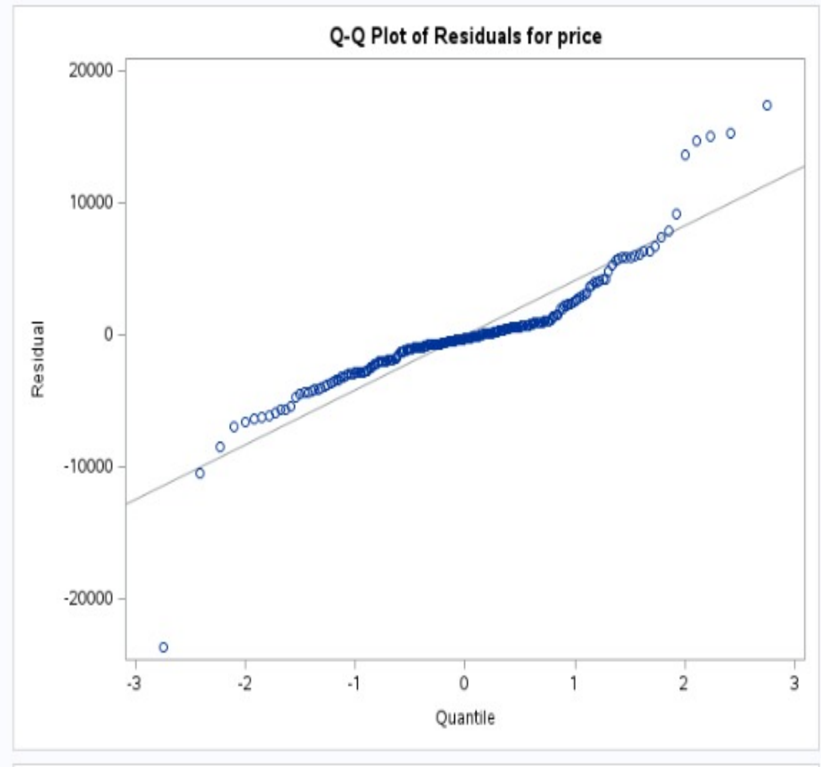
Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 9146654456 | 3048884819 | 173.86 | <.0001 |
| Error | 201 | 3524858162 | 17536608 | | |
| Corrected Total | 204 | 12671512619 | | | |

| | |
|---|---|
| Root MSE | 4187.67331 |
| Dependent Mean | 13145 |
| R-Square | 0.7218 |
| Adj R-Sq | 0.7177 |
| AIC | 3630.31969 |
| AICC | 3630.62120 |
| SBC | 3436.61173 |

Parameter Estimates

| Parameter | DF | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -3335.151236 | 4037.544605 | -0.83 | 0.4098 |
| horsepower | 1 | -83.955668 | 40.453741 | -2.08 | 0.0392 |
| curb-weight | 1 | 4.703855 | 1.579071 | 2.98 | 0.0032 |
| horsepowe*curb-weigh | 1 | 0.046792 | 0.013350 | 3.51 | 0.0006 |



## Conclusion of the Analysis

In conclusion, this analysis demonstrates the significant impact of factors such as engine size, horsepower, curb weight, rpm, and fuel type on automobile performance and pricing. The findings provide quantitative insights to guide the design, pricing, and selection of fuel-efficient and cost-effective vehicles, benefiting both manufacturers and consumers alike. Moreover, these findings have broad implications for automakers, allowing them to fine-tune design and pricing strategies to match fuel efficiency and affordability consumer preferences. Also, the consumers can utilize these insights to make informed decisions when choosing vehicles that best match their individual needs and preferences.