Statistical Analysis for Predicting Car Price

STAT 31631 – Statistical Modelling

Department of Statistics & Computer Science

University of Kelaniya

Academic Year 2022/2023

By

Group 09

PS/2018/129 – K.R.N.R. Kumara

PS/2020/019 – H.M.P.M. Herath

PS/2020/033 – R.D.R.P. Wickramasinghe

PS/2020/109 – R.S. Nayanathara

PS/2020/126 – G.L.P.C. Biseka

PS/2020/147 – W.H.H.H. Samarasinghe

PS/2020/193 – V.P. Vithanage

PS/2020/254 – L.D. Dilshan

Introduction

The automotive industry is a complex market where a wide range of factors influence the price of vehicles. Consumers, manufacturers, and dealers alike seek to understand how various features and specifications of a car contribute to its market value. With the vast amount of data available on car attributes, statistical models can be employed to predict car prices based on key characteristics such as body type, engine type, curb weight, horsepower, and fuel consumption. This study aims to develop and refine a predictive model for car prices using multiple linear regression, enabling stakeholders to make informed decisions and understand the pricing dynamics in the automotive market.

Problem Statement

Determining the fair price of a car is a multifaceted challenge influenced by numerous variables, including both quantitative and qualitative features. The primary problem is to identify which factors most significantly affect car prices and how these factors can be quantified to develop a reliable predictive model. With many potential predictors, it is also crucial to ensure that the model is both statistically significant and interpretable, avoiding the inclusion of irrelevant variables that could reduce the model's accuracy.

Methodology

To develop a multiple linear regression model to predict car prices based on the following predictor variables:

- Fuel type
- Number of doors
- Body type
- Curb weight
- Engine type
- Horsepower
- Average fuel consumption

Data Sources:

Obtain a comprehensive dataset from Kaggle. (www.kaggle.com)

(Source Link -: https://www.kaggle.com/datasets/hellbuoy/car-price-prediction)

➢ Focusing on above predictor variables (fuel type, number of doors, body type, curb weight, engine type, horsepower, and average fuel consumption).

➢ Preprocessing steps included handling missing values, removing outliers, normalizing variables, and encoding categorical data.

➢ Exploratory data analysis involved descriptive statistics, correlation analysis, and data visualization.

➢ A multiple linear regression model has been fitted and checked for assumptions such as linearity and normality.

➢ Model evaluation included assessing fit with R-squared, testing predictor significance, checking for multicollinearity, and performing residual analysis.

➢ Results were interpreted to understand the impact of each predictor on car price, identifying significant factors and their practical implications.

Results

➢ First of all, the full model, which includes all the variables the dataset, has been developed.

- **Intercept (-31566.469, p value < 0.05):** This is the baseline price when all predictors are set to their reference categories or zero. While it's statistically significant at 5% significance level.

- **Categorical variables:** fueltype-gas(629.07, p value>0.05) , doornumber□two(614.695, p value>0.05), carbody- hardtop(-2281.013 , p value>0.05), enginetype-dohcv(-2586.129, p value>0.05), enginetype-l(321.908, pvalue>0.05), enginetype-ohcv(1683.981, p value>0.05) are not statistically significant compared to there's other level. And the other categorical variables are statistically significant (p value < 0.05)

- **Numerical variables:** All the variables are statistically significant for the full model (p value < 0.05).

➢ For the full model, the Residual Standard error is 3392. That is the fitted value from the model can be deviated from the actual value by 3392 units on average.

➢ The multiple R squared value of full model is 0.833.

➢ Adjusted R-Squared is 0.8197 for full model. This indicates that 81.97% of the variability in car prices is explained by the full model. This suggests a strong fit, meaning the included variables are good predictors of car prices.

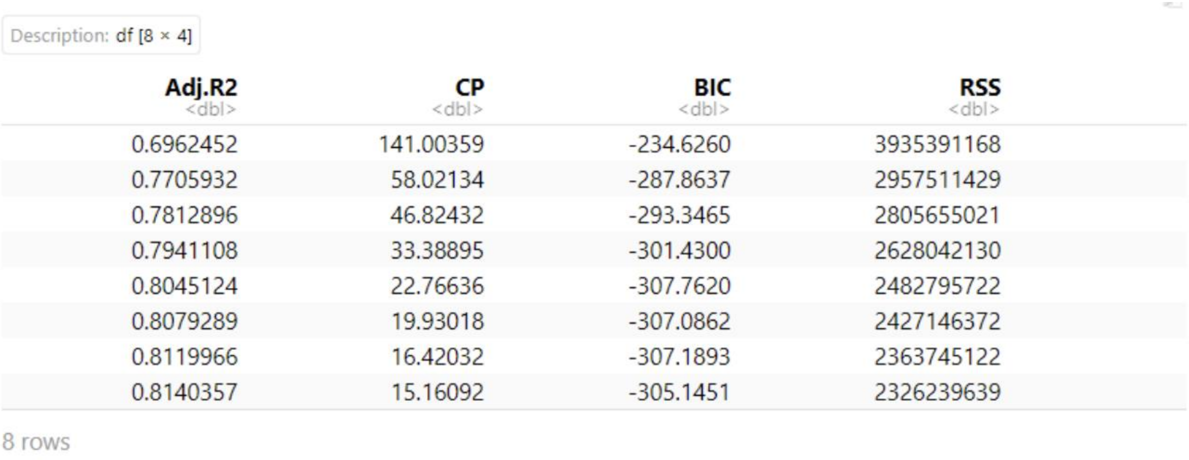
- Finally, p value for full model is 2.2e-16(p value < 2.2e-16). The full model is statistically significant overall, meaning that at least one of the predictors significantly relates to the car price.

The Procedure of choosing a better model.

- First and main, the full model has been created and analyzed for all factors that could indicate the model's suitability. The best model that is significant had to be chosen because some variables are not significant.

- The method of backward stepwise selection has been used to choose the best model.

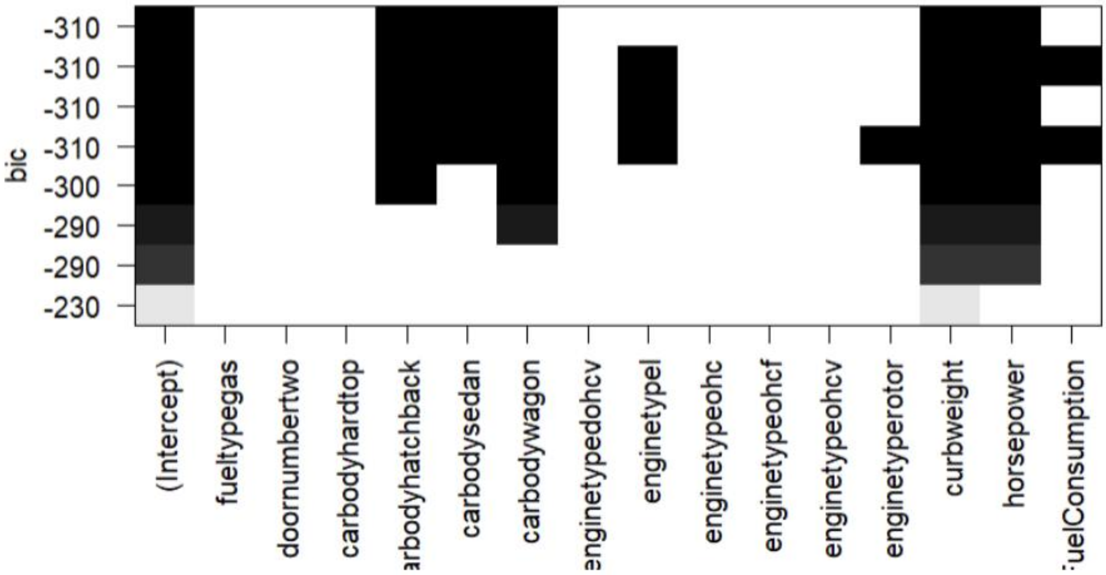
- From that method 8 models were obtained and selected the 8th model as the better model due to the highest Adjusted R squared value and lowest CP and RSS values. Results were obtained as follows.
-

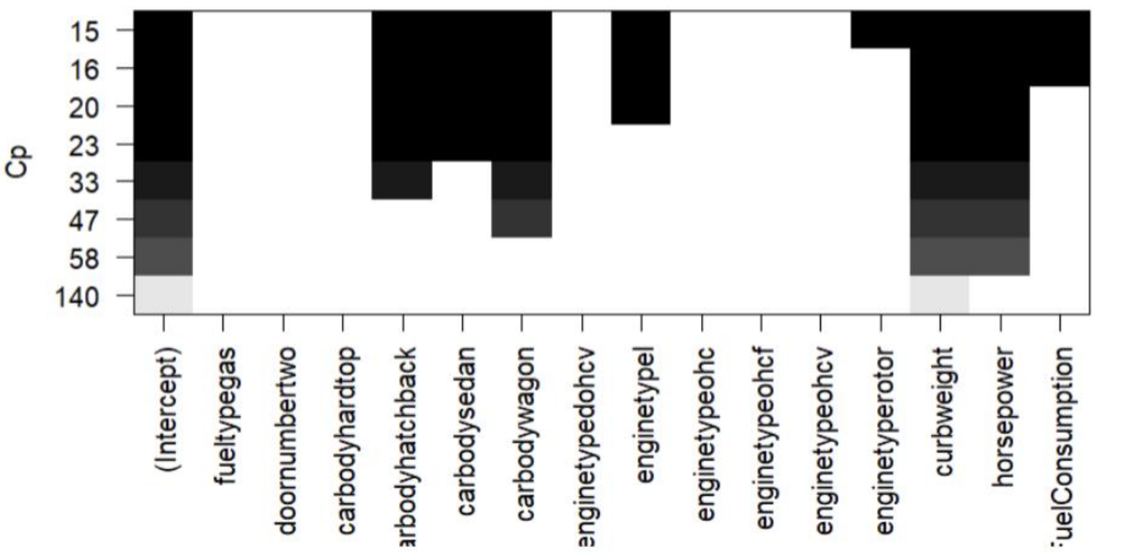


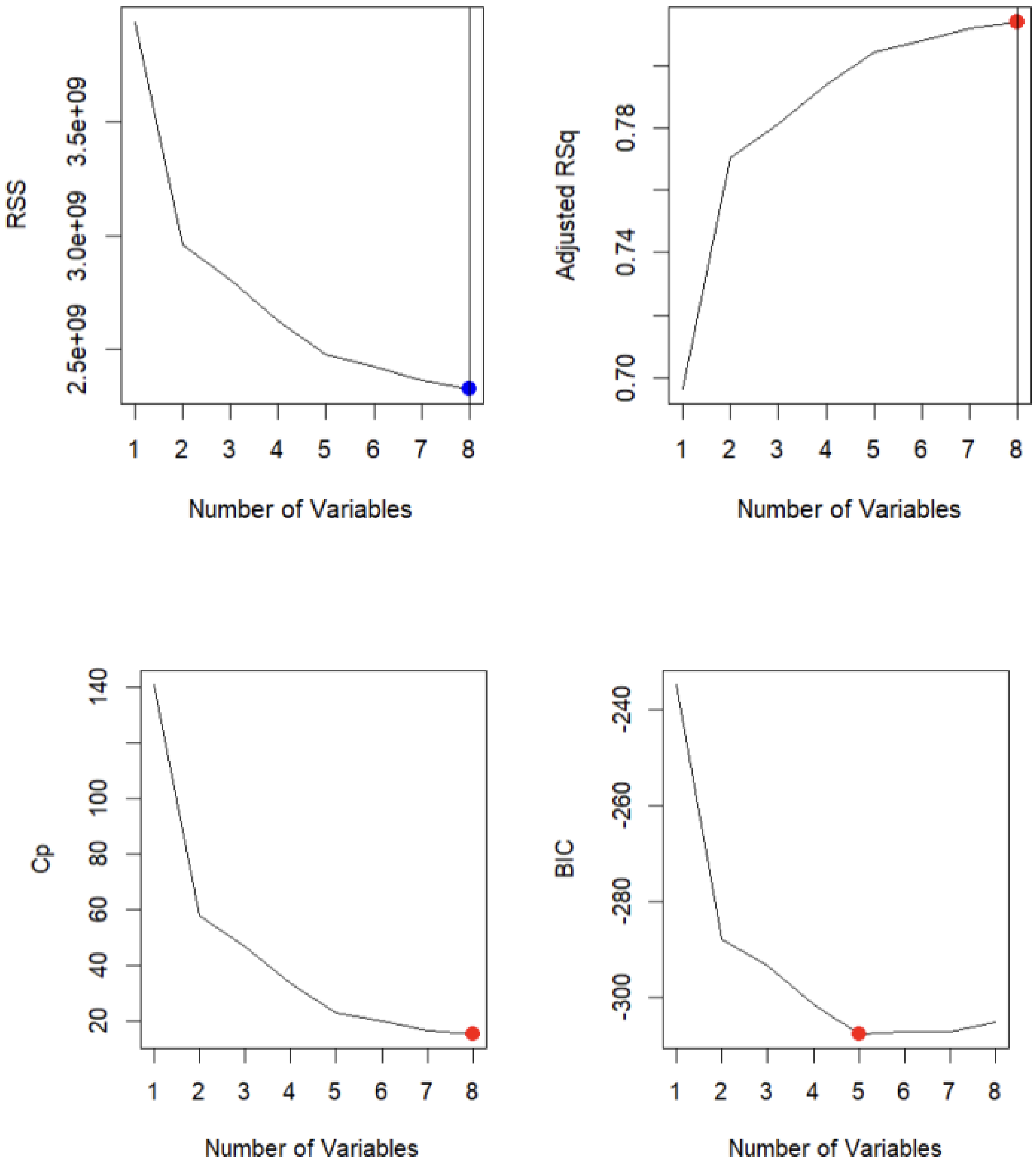
Description: df [8 × 4]

Adj.R2	CP	BIC	RSS
<dbl>	<dbl>	<dbl>	<dbl>
0.6962452	141.00359	-234.6260	3935391168
0.7705932	58.02134	-287.8637	2957511429
0.7812896	46.82432	-293.3465	2805655021
0.7941108	33.38895	-301.4300	2628042130
0.8045124	22.76636	-307.7620	2482795722
0.8079289	19.93018	-307.0862	2427146372
0.8119966	16.42032	-307.1893	2363745122
0.8140357	15.16092	-305.1451	2326239639

8 rows



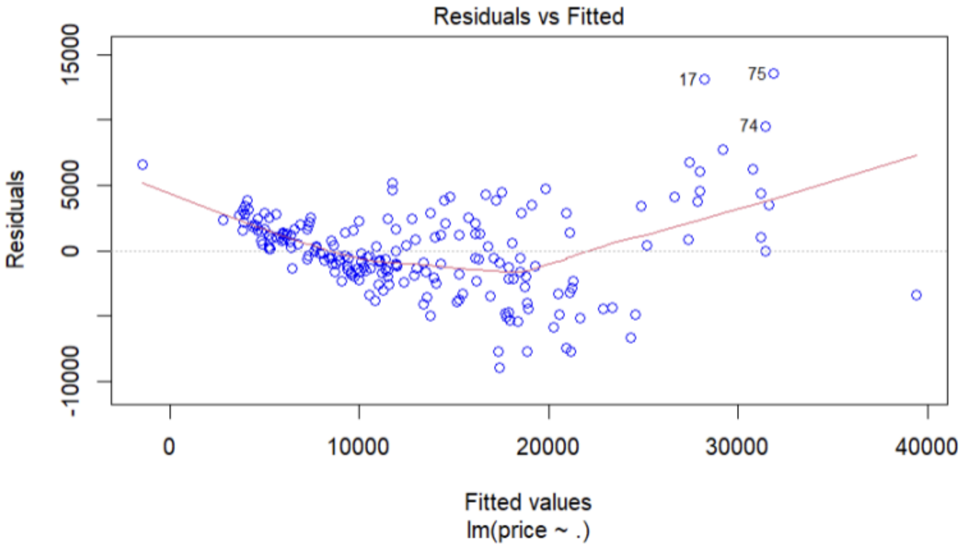

Cp
15
16
20
23
33
47
58
140
(Intercept)
fueltypegas
doornumbertwo
carbodyhardtop
arbodyhatchback
carbodysedan
carbodywagon
enginetypedohcv
enginetypel
enginetypeohc
enginetypeohcf
enginetypeohcv
enginetyperotor
curbweight
horsepower
uelConsumption


RSS
2.5e+09
3.0e+09
3.5e+09
1
2
3
4
5
6
7
8
Number of Variables
Adjusted RSq
0.70
0.74
0.78
1
2
3
4
5
6
7
8
Number of Variables
Cp
20
40
60
80
100
140
1
2
3
4
5
6
7
8
Number of Variables
BIC
-240
-260
-280
-300
1
2
3
4
5
6
7
8
Number of Variables

Residuals Analysis

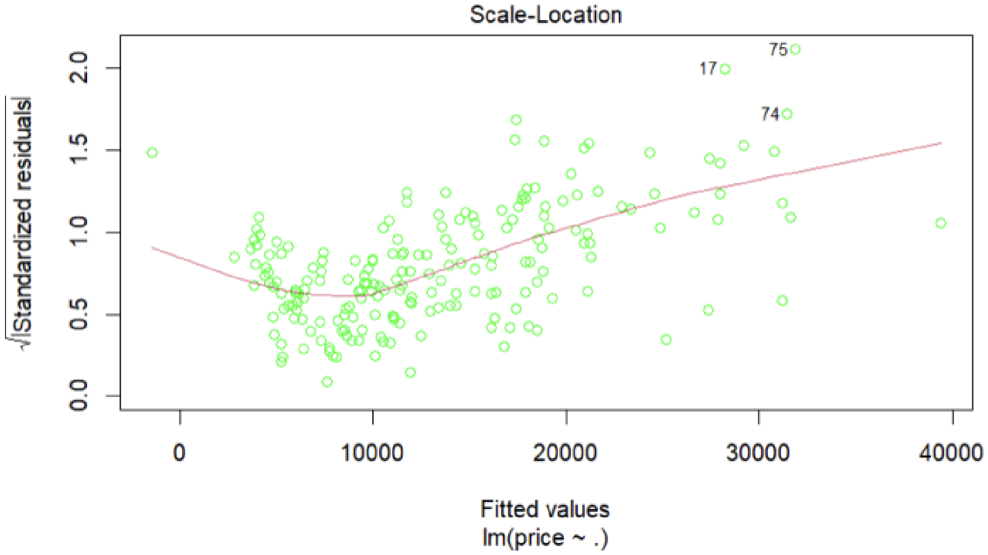
These would be the conclusions from our residual analysis, which we used to check the accuracy the of the model 9's assumptions.

01) Residuals vs Fitted



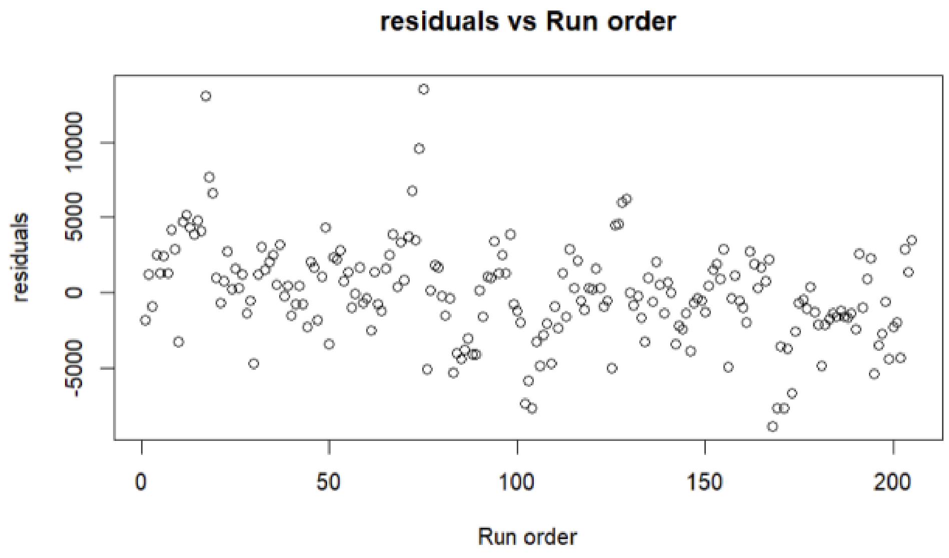
Here, the plot sees that linearity is violated: there seems to be a quadratic relationship. Whether there is homoscedastic or not is less obvious: So we will need to investigate more plots. There are several outliers(17,74,75) with residuals close to 10000.

02) Scale-location



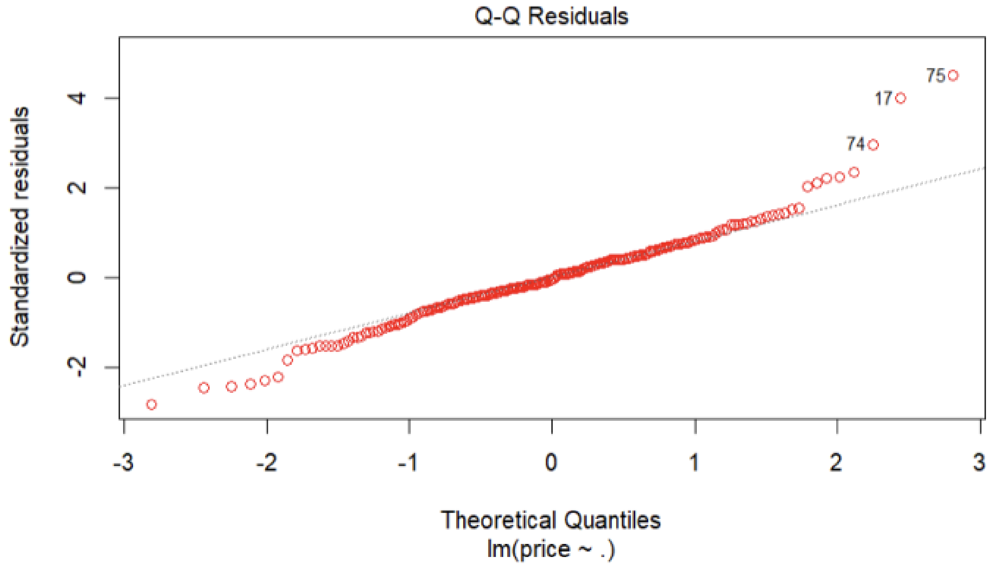
Here the scale location plot suggests some non-linearity here, but what we can also see is that the spread of magnitudes seems to be lowest in the fitted values close to 0, highest in the fitted values around 30000. This suggests heteroscedasticity.

03) Residuals vs Run order



It seems like residuals are randomly distributed along the time. Therefore residuals could assumed to be independent of each other.

04) Q-Q plot



It seems to be most of the residuals fall approximately along the reference line. Therefore, residuals assumed to be normally distributed. Here few leverage points also can be observed.

The categorical variables of the full model are fuel type and door number removed using backward model selection method.(These variables are non-significant)Then this is the final summary result of the model and the best model of the data set.

Residuals:

Min, 1Q, Median, 3Q, Max: The residuals range from -8924.0 to 13563.5, with a median close to zero (-61.1), indicating the model's predictions are generally centered around the actual values. The spread suggests some variability in the predictions.

Coefficients:

Intercept: The negative estimate (-28492.614) indicates the baseline price when all other predictors are zero. This number is more of a reference point rather than a meaningful value in isolation.

Car Body Type:

- Hardtop: The estimate (-2335.518) is negative and not significant (p = 0.217573), suggesting that hardtop cars are not significantly different in price compared to the reference category.

- Hatchback: The estimate (-6271.020) is negative and significant (p = 4.08e-05), indicating that hatchback cars are significantly cheaper than the reference category.

- Sedan: The estimate (-4730.379) is negative and significant (p = 0.001516), indicating that sedans are significantly cheaper.

- Wagon: The estimate (-7520.528) is negative and significant (p = 5.07e-06), indicating that wagons are also significantly cheaper.

- Curb Weight: The estimate (10.650) is positive and highly significant (p < 2e-16), suggesting that cars with higher curb weight are more expensive.

Engine Type:

- DOHC: The estimate (-2872.702) is negative and not significant (p = 0.473481), indicating no significant effect on price.

- L: The estimate (308.370) is positive and not significant (p = 0.841729), indicating no significant effect on price.

- OHC: The estimate (3271.840) is positive and significant (p = 0.004020), suggesting that cars with OHC engines are more expensive.

- OHCF: The estimate (4391.843) is positive and significant (p = 0.001961), suggesting that cars with OHCF engines are more expensive.

- OHCV: The estimate (1738.032) is positive and not significant (p = 0.209345), indicating no significant effect on price.

- Rotor: The estimate (6638.398) is positive and significant (p = 0.002084), suggesting that cars with rotor engines are more expensive.

Horsepower:

The estimate (95.592) is positive and highly significant (p = 2.18e-10), indicating that higher horsepower is strongly associated with higher car prices.

Average Fuel Consumption:

The estimate (251.430) is positive and significant (p = 0.000928), suggesting that cars with higher average fuel consumption are more expensive.

Model Fit:

Residual Standard Error:

3382 indicates the typical error in predicting car prices is about $3,382.

Multiple R-squared (0.8322):

About 83.22% of the variance in car prices is explained by the model, indicating a strong fit.

Adjusted R-squared (0.8208):

After adjusting for the number of predictors, about 82.08% of the variance is explained, suggesting that the model performs well in explaining car prices.

F-statistics (72.86, p < 2.2e-16):

The model is statistically significant, meaning that at least one of the predictors significantly relates to the car price.

Discussion

The model indicates that car prices are significantly influenced by car body type, curb weight, engine type, horsepower, and average fuel consumption. The high R-squared values suggest that the model fits the data well. The F-statistics confirms the overall significance of the model.

Fitted Model

Price = -28492.614 -2335.518(carbodyhardtop) – 6271.02(carbodyhatchback) - 4730.379(carbodysedan) -7520.528(carbodywagon) +10.650(curbweight) - 2872.702(enginetypegohcv) +308.370(enginetypel) +3271.840(enginetypeohc) +4391.843(enginetypeohcf) +1738.032(enginetypeohcv) +6638.398(enginetyperotor) +95.592(horsepower) +251.43(AvgFuelConsumption)

Conclusion

The analysis aimed to develop and refine a statistical model to predict car prices using various features such as fuel type, number of doors, car body type, curb weight, engine type, horsepower, and average fuel consumption. Here's a summary of the findings and the final model's performance:

1. **Model Selection:**
   - The full model, which included all variables, was initially considered but included some predictors that were not statistically significant (fuel type, number of doors). These non-significant variables were identified through their p-values and removed using a backward stepwise selection method.
   - The final model retained the variables that significantly contribute to predicting car prices.
2. **Key Predictors:**
   - **Car Body Type:** Certain car body types, like hatchback, sedan, and wagon, were found to significantly decrease the car price compared to the reference category. For example, wagons reduce the price by around $7,521, indicating that these body types generally lead to lower car prices.
   - **Curb Weight:** Heavier cars are generally more expensive, with each unit increasing in curb weight adding about $10.65 to the price.
   - **Engine Type:** Some engine types, like OHC, OHCF, and rotor engines, significantly increase the car price, suggesting that these engines are associated with more expensive vehicles.
   - **Horsepower:** Higher horsepower is strongly associated with higher car prices, reflecting the market trend where more powerful cars typically command higher prices.
   - **Average Fuel Consumption:** Interestingly, higher fuel consumption (indicating lower fuel efficiency) is associated with higher car prices, because it correlates with more powerful vehicles.

3. **Model Performance:**
   - The final model has a **Residual Standard Error** of approximately $3,382, indicating that the typical error in predicting car prices is about $3,382.
   - The **Multiple R-squared** value of 0.8322 suggests that the model explains about 83.22% of the variability in car prices, indicating a strong fit.
   - The **Adjusted R-squared** value of 0.8208 further confirms the model's robustness, showing that it explains 82.08% of the variance when accounting for the number of predictors.
   - The **F-statistics** (72.86, p < 2.2e-16) indicates that the model is statistically significant overall, meaning that at least one of the predictors significantly relates to car prices.
4. **Residuals Analysis:**
   - The residuals analysis showed some issues with linearity and potential heteroscedasticity, suggesting that the relationship between the predictors and car prices might not be perfectly linear, and the variance of residuals may not be constant.
   - However, residuals were generally centered around zero and normally distributed, indicating that the model's predictions are reasonably accurate.

The final model is a strong predictor of car prices, explaining over 80% of the variability in the data. It successfully identifies key factors like car body type, curb weight, engine type, horsepower, and fuel consumption that significantly influence car prices. However, some non-linear relationships and potential heteroscedasticity suggest that there might be room for further refinement or exploration of alternative modeling approaches.

Overall, this model provides valuable insights for understanding the factors that drive car prices and can be a useful tool for predicting prices based on a car's characteristics.

Individual Contribution

Task	Student No:
Finding and reading research papers	PS/2020/147 , PS/2020/193
Finding the research gap and problem	PS/2020/193
Finding objective for the problem	PS/2018/129
Creating Project Proposal	PS/2020/033 , PS/2020/109
Data collection	PS/2020/019 , PS/2020/254
Data cleaning and preparation	PS/2020/126 , PS/2020/147
Descriptive analysis	PS/2020/254 , PS/2018/129
Model Evaluation and refinement	PS/2020/109 , PS/2020/033
Report writing and presentation preparation	PS/2020/019 , PS/2020/109 , PS/2020/126