Automotive Data Analysis

By Tuomas Kuusela

Introduction

This project revolves around the exploration and analysis of automotive data, focusing on factors that influence fuel efficiency. Primary objective of this analysis is to identify and quantify the relationship between various variables, such as engine size, horsepower, weight, and their fuel consumption measured in miles per gallon (MPG). The problem addressed is determining which variables have the most significant impact on fuel efficiency.

The data for this analysis is sourced from the 'Auto' dataset included in the ISLR R-package. The dataset comprises 392 observations across 9 variables. Variables included in the dataset are:

- **MPG**: Fuel efficiency, miles per gallon.
- **Cylinders**: Number of cylinders in the vehicle's engine (3, 4, 5, 6, 8).
- **Displacement**: Engine volume, reflecting engine size (cu. inches).
- **Horsepower**: Engine power output.
- **Weight**: Total weight of the vehicle (lbs.).
- **Acceleration**: Seconds to accelerate 0 to 60 mph.
- **Year**: Model year of the vehicle.
- **Origin**: Region where the vehicle was manufactured (1: USA, 2: EU, 3: JAP).
- **Name**: Name of the car model.

During analysis, Generalized Additive Model (GAM) is used since it is suitable for exploring relationships between a response variable and predictor variables when those relationships may not be strictly linear. This flexibility allows for more accurate modelling compared to traditional linear models.

During model selection, preliminary analysis involved examining correlations and patterns within the data to identify which variables had a significant impact on MPG and to determine the nature of their relationships (linear or non-linear). The GAM was configured to use smooth functions for variables where non-linear relationships were observed, while retaining linear terms for other variables where appropriate.

Results

- **MPG** ranges from 9 to 46.6, where the average fuel efficiency is 23.45 MPG.
- **Cylinders** vary between 3 and 8, with 4 being the most common.
- **Displacement** spans 68 to 455 cu. inches, showing a wide variation in engine size.
- **Horsepower** has a range of 46 to 230, with a mean of 104.5.
- **Weight** of vehicles varies from 1613 to 5140 pounds, average weight being 2978 pounds.
- **Acceleration** times range from 8 to 24.8 seconds, suggesting varied performance levels.
- **Year** of the models covered is from 1970 to 1982.
- **Origin** has 63 cars from EU, 79 from Japan and 245 cars from USA.

	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Year	Origin
MPG	1.00	-0.78	-0.81	-0.78	-0.83	0.42	0.58	0.57
Cylinders	-0.78	1.00	0.95	0.84	0.90	-0.50	-0.35	-0.57
Displacement	-0.81	0.95	1.00	0.90	0.93	-0.54	-0.37	-0.61
Horsepower	-0.78	0.84	0.90	1.00	0.86	-0.69	-0.42	-0.46
Weight	-0.83	0.90	0.93	0.86	1.00	-0.42	-0.31	-0.59
Acceleration	0.42	-0.50	-0.54	-0.69	-0.42	1.00	0.29	0.21
Year	0.58	-0.35	-0.37	-0.42	-0.31	0.29	1.00	0.18
Origin	0.57	-0.57	-0.61	-0.46	-0.59	0.21	0.18	1.00

Figure 1. *Correlation matrix between variables.*

In Figure 1, the correlation matrix displays the relationships between variables. A strong negative correlation is observed between MPG and variables such as weight, displacement, and horsepower, indicating that heavier cars with larger engines and higher horsepower tend to have lower fuel efficiency. If we look at the relationship between MPG and year, we see a positive correlation suggesting improvements in fuel efficiency over time.

The matrix also reveals high positive correlation between displacement, cylinder, horsepower, and weight, which is obvious since more cylinders equal a bigger engine and thus more horsepower and weight.

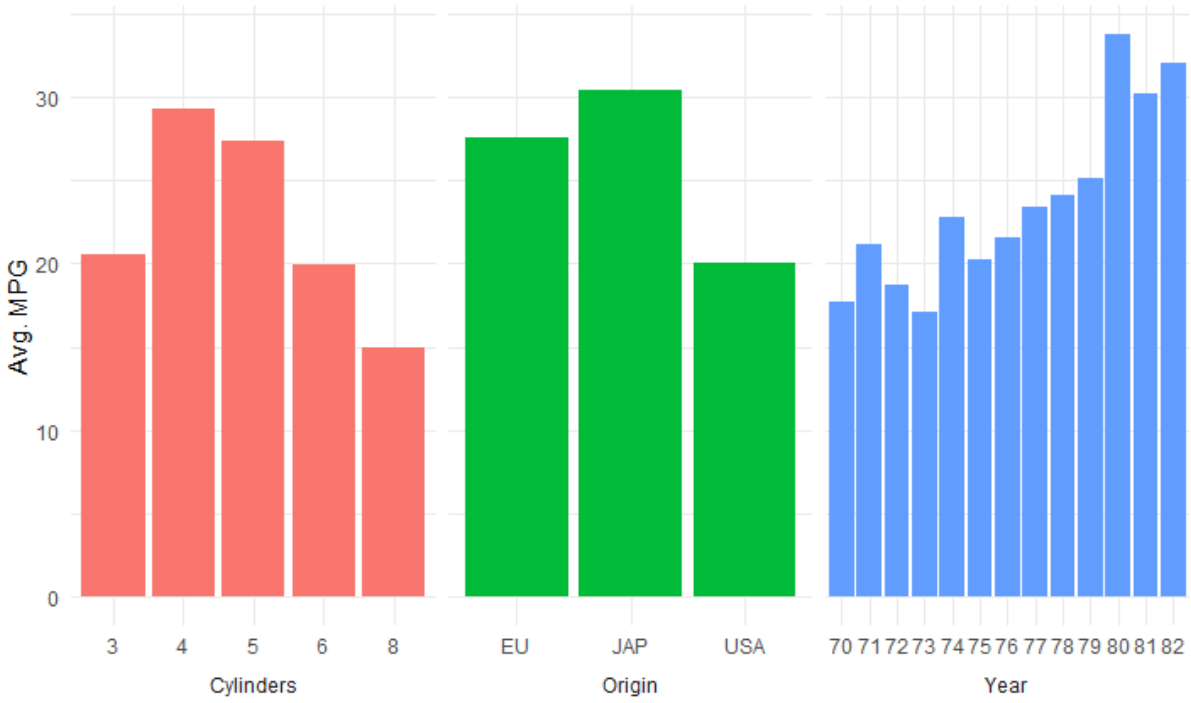

Figure 2. *Graph showing the average MPG between selected variables (Cylinders, Origin, Year).*

The average MPG is highlighted in Figure 2, where 4-cylinder vehicles being on average the most efficient, and cars made in the USA being the least efficient. We can also see the yearly increase in fuel efficiency over time from 1970 to 1982.

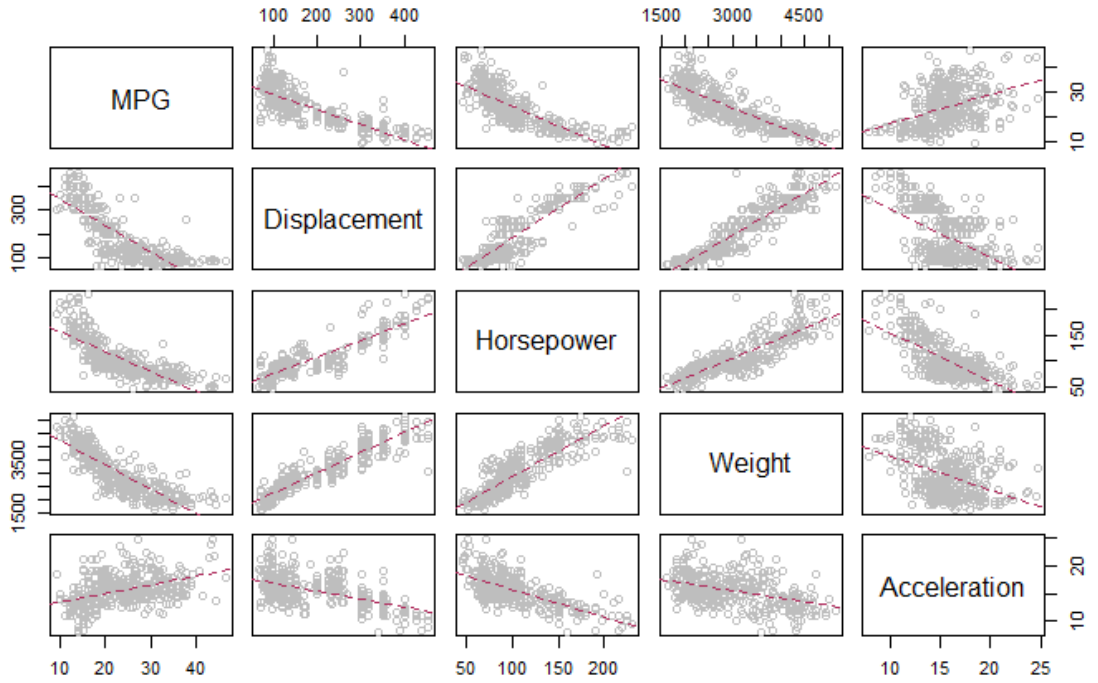

Figure 3. *Pairwise scatter plot between selected variables.*

When selecting variables for our model using the GAM approach, we target those exhibiting non-linear relationships with MPG, as observed with displacement, horsepower, and weight in Figure 3. Figure 4 show how GAM adjusts the regression line to accommodate these relationships.

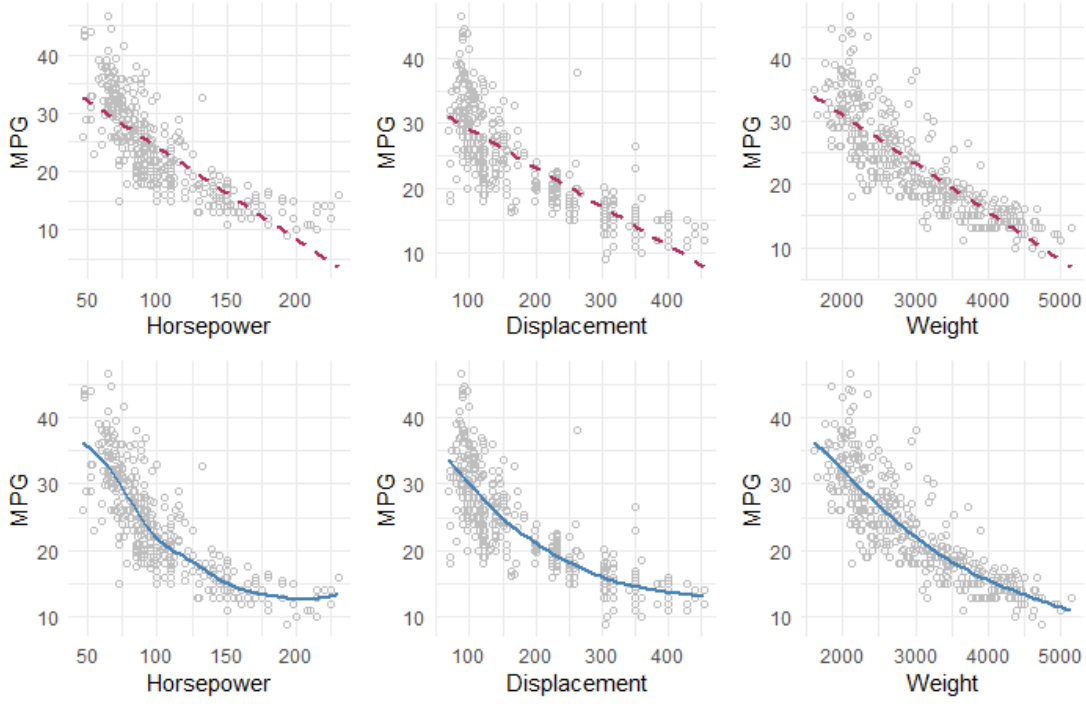

Figure 4. *Comparison of linear and GAM fits. Traditional linear model above and the smoothened GAM below.*

During the refinement of the GAM, it was observed that cylinders and displacement exhibit a redundancy in their contribution to the model's explanatory power regarding fuel efficiency. This redundancy is theoretically anticipated, as mentioned before about the number of cylinders is typically correlated with engine displacement.

The analysis confirmed that excluding either cylinders or displacement from the model resulted in the remaining variable becoming statistically non-significant. However, inclusion of displacement rather than cylinders was substantiated by improved model performance: lower GCV score, higher adjusted R^2, and greater explained deviance were observed.

The final model is predicting fuel efficiency as a function of engine horsepower, engine displacement, vehicle weight, model year and origin. Smooth terms were applied to horsepower, displacement, and weight to account for non-linear relationships, while year and origin are included as linear predictors. The models adjusted R^2 came out as 0.863.

It was also compared against a traditional linear regression model, which utilized the same predictors but treated all as linear terms without incorporating smooth functions for any of the variables. The models adjusted R^2 came out as 0.818.

While the adjusted R^2 values indicate a superior fit for the GAM compared to the traditional linear regression model, it is crucial to recognize that R^2 alone does not encapsulate the full picture of model performance. Higher R^2 might suggest a better fit to the training data, but it does not inherently guarantee that the model will perform better on unseen data. This can also introduce overfitting, where a model might capture the noise within the training data, leading to poor generalization on new data.

The finalized model is represented as follows:

$$\boldsymbol{MPG = \beta 0 + f1(Horsepower) + f2(Displacement) + f3(Weight) + \beta 1(Year) + \beta 2(Origin) + \epsilon}$$

- *MPG* being the fuel efficiency we're trying to predict.
- *β0* is the intercept of the model.
- *f1(Horsepower), f2(Displacement), f3(Weight)* are smooth functions of the respective variables to capture the non-linear relationships with MPG.
- *β1(Year)* and *β2(Origin)* are the coefficients for the linear terms of year and origin.
- ϵ is the error term.

The model's adjusted R^2 of 0.863 suggests a good fit, explaining substantial portion of the variability in MPG. Despite displacement not showing a statistically significant effect alone, its inclusion as a non-linear term helps enhance the model's overall accuracy. Year and origin emerge as significant linear predictors, reflecting not only the impact of technological progress on fuel efficiency but also the distinct influence of manufacturing origin. Overall, the analysis underscores the GAM's capability to offer a more nuanced view of the variables influencing fuel efficiency, highlighting its strength in capturing the complect interplay of factors that traditional linear models might overlook.