Assignment 3

Martin Lindqvist

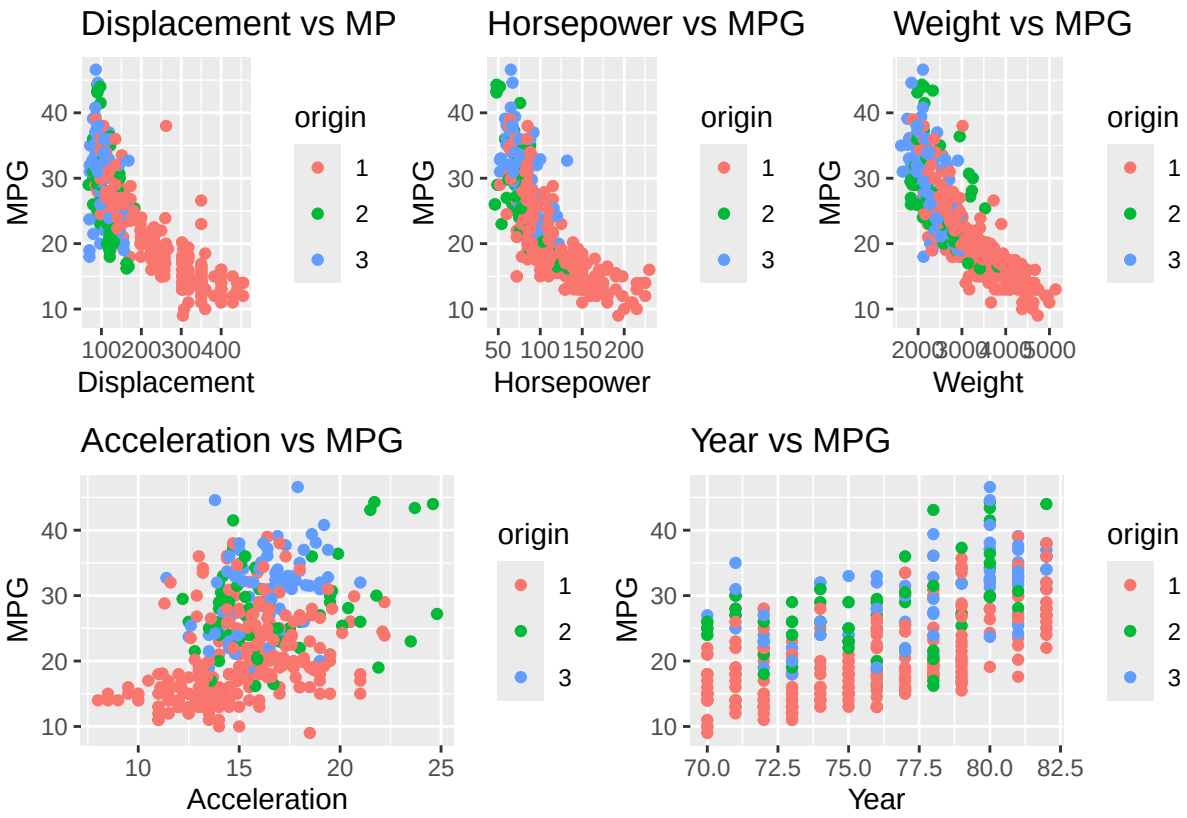
2024-09-09

GAM modelling A Generalised Additive Model (GAM) has been fit to the Auto data to predict mpg (fuel consumption per mile). The is split in to training data (80%) and training data (20%). For this the package `caret` is used to make the the response variable be evenly distributed between the two groups.

Auto data

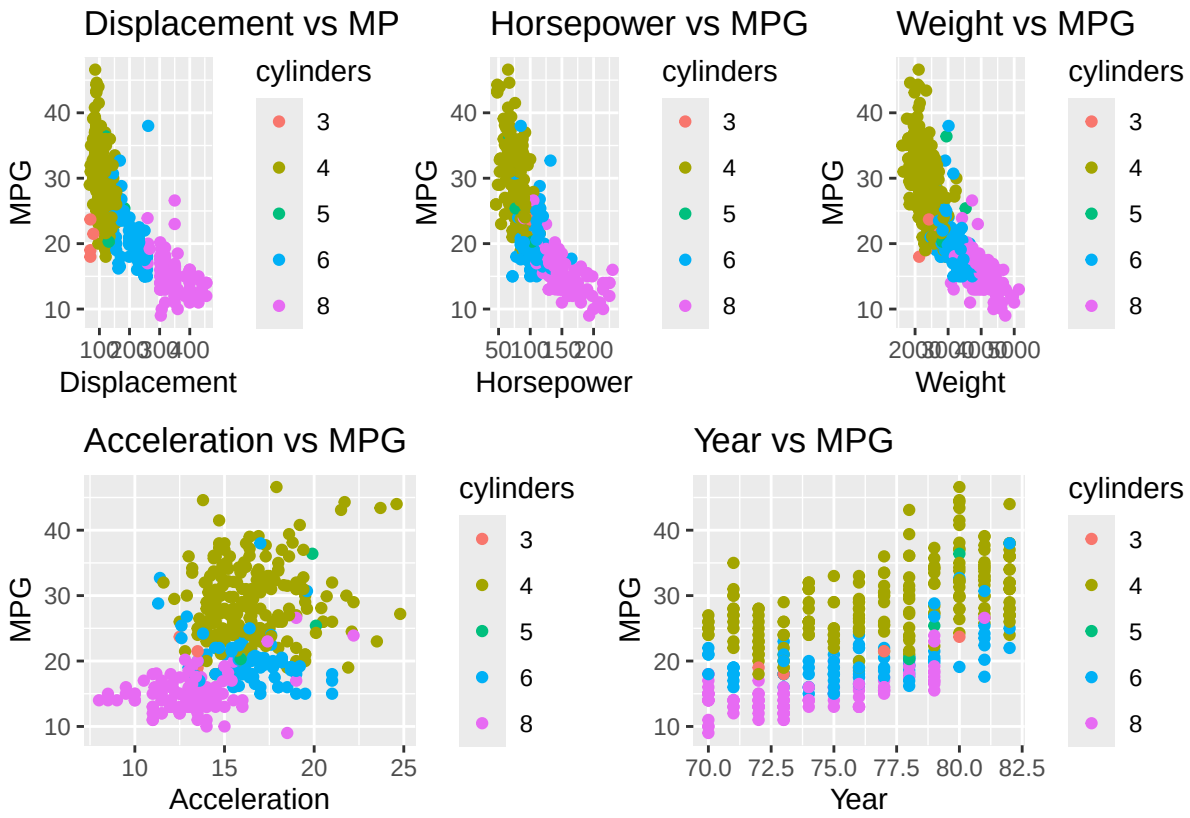
- **Mpg:** Miles per gallon. A numerical variable indicating the distance travelled per gallon of fuel burned.
- **Cylinders:** A categorical factor indicating the number of cylinders (3, 4, 5, 6, 8).
- **Displacement:** A numerical measure of the cylinder volume.
- **Horsepower:** A numerical measure of engine power.
- **Weight:** A numerical measure indicating the car weight in pounds.
- **Acceleration:** A numerical measure indicating the time in seconds to reach 60 mph from standstill.
- **Year:** A numerical variable indicating the model year of the car.
- **Origin:** A factor indicating where the car was manufactured (1 = America, 2 = Europe, 3 = Japan)
- **Name:** A categorical factor representing the model name of the car (will not be used in model training or testing).

In **plot 1** we can see explanatory variables being plotted against the response (`mpg`), grouped by `origin`. We can see that American cars seem to have higher `displacement`, more `horsepower`, higher `weight` and faster `acceleration` compared to European and Japanese cars, possibly resulting in higher `mpg`. `Displacement`, `horsepower` and `weight` shows non-linear effects. The shape of `acceleration` and `year` is more difficult to determine just from looking at the graphs.



Plot 1

Plot 2 shows the same explanatory variables being plotted against `mpg`, but now grouped by number of `cylinders`. We can see clear effects of `cylinders` on `displacement`, `horsepower`, `weight` and `acceleration`. It seems like engines with more `cylinders` are more powerful and thus consuming more fuel.



Plot 2

Model Specifications

- **Response Variable:** Mpg (miles per gallon).
- **Explanatory variables:**
  - **Smooth Terms:** These variables model non-linear effects using splines.
    - **s(displacement)**, **s(horsepower)**, **s(weight)**, **s(acceleration)**, **s(year)**
  - **Linear Variables:** These variables model categorical effects.
    - **Cylinders**,**Origin**

Model fitting All continuous variables has been set to smooth. If the relationship is linear the Thin Plate Regression splines will automatically be adjusted to be linear. It is possible to decide to fit smooth or linear terms from looking at **plot 1** and **plot 2** or by having prior knowledge about the underlying relationships. But since this increases the subjectivity of the model and I lack prior knowledge about the relationships between these variables, I have decided to fit the variables automatically.

Factor variables are treated as linear terms.

Since all the variables were significant at a 0.05 significance level, non were removed.

When running `gam.check` the diagnostic test suggested that the basis dimensions for `year` were too low, indicated by the low p-value and effective degrees of freedom being close to the maximum basis dimensions. Therefore the maximum basis dimensions for `year` was increased to 13.

Thin Plate Regression Splines: Thin Plate Regression Splines, the default spline in the `mgcv` package are used for all the smooths. They fit smooth curves to data without requiring manual knot specification. They adjust the curve's smoothness automatically and apply a penalty to prevent overfitting, ensuring a model that captures the true underlying trends.

REML (Restricted Maximum Likelihood): REML is used and complements the splines by estimating the optimal smoothness level of these spline functions. It maximizes a likelihood function that considers the residuals, thus refining the model's accuracy and preventing overcomplexity.

Model Interpretation: The factor variables can be interpreted in the same way as in linear models. For example vehicles with 4 cylinder are expected to have an `mpg` that is 9.2874 higher than the baseline category (vehicles with 3 cylinders), given that the other variables are included in the model and held constant.

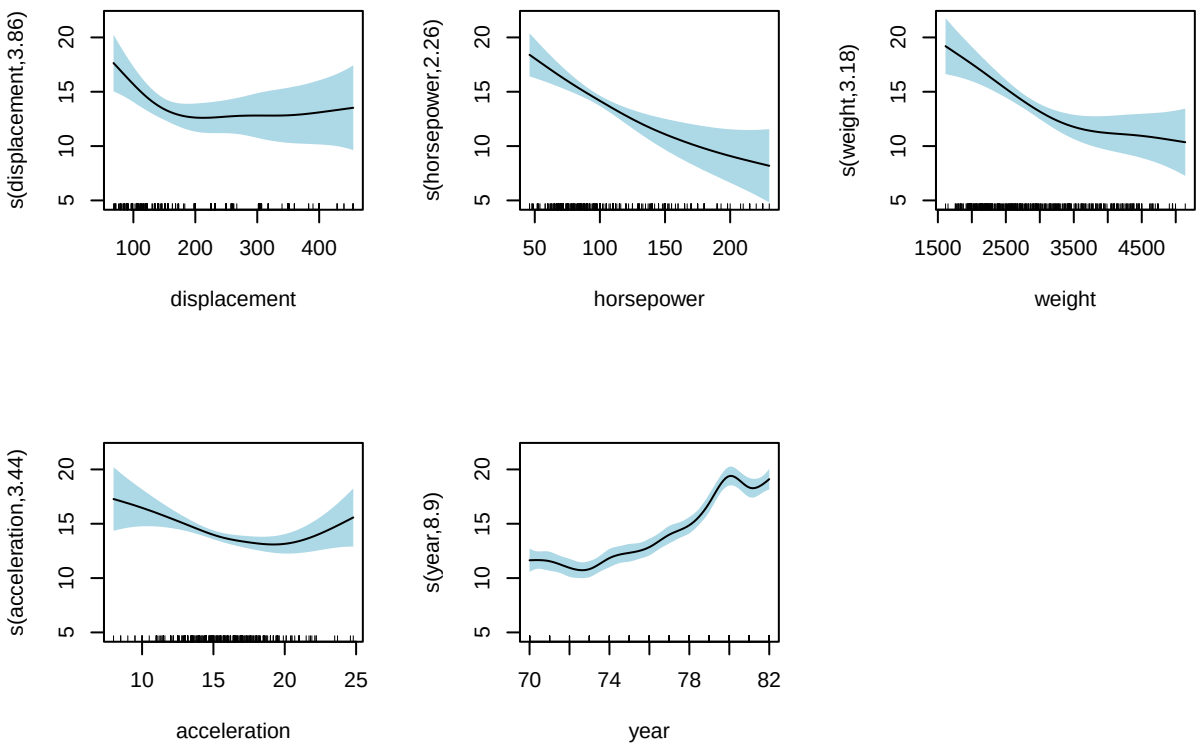
To summarise, both `cylinders` and `origin` are significant at significance level of 0.05, although `cylinders` much more so. All the `cylinder`-groups has a much lower expected fuel consumption (approximately 9 to 14 units higher `mpg`) than the baseline group of 3 `cylinders`. `Origin` shows an affect of American cars having a lower higher fuels consumption compared to European cars (approximately 0.3 units of `mpg`) and Japanese (approximately 1 unit of `mpg`) cars, although this effect is small.

For the smooths each terms has several coefficients. It is therefore more interpretable to look at the overall shape and direction of the curves than each coefficient separately.

Effective degrees of freedom (`edf`): `s(displacement)`: 3.856 `s(horsepower)`: 2.263 `s(weight)`: 3.183 `s(acceleration)`: 3.445 `s(year)`: 8.899

The `edf` indicates how complex the smooth is. An `edf` of one means that it is a straight line, an `edf` of two means that it is equivalent to a quadratic curve. Higher `edf`s describes more wiggly curves. This means that non of the smooths shows a linear effect, with `year` being the most wiggly. By looking at **plot 3** we can see both the willingness and the directions of the smooths. Increases in `displacement` up to 200, decreases `mpg` before the effect tampers off (given that all the other variables are included, held constant and are at their average values). As `horsepower` increases `mpg` decreases. As `weight` increases `mpg` decreases, with the effect slowing down after a `weight` of 3500 lbs. Initially as `acceleration` increases `mpg` decreases, but after approximately 20 this trend reverses. `Year` shows an increasing trend of cars becoming more fuel efficient with time. This smooth is more wiggly than the others with an specially large effect at `year` 1980.

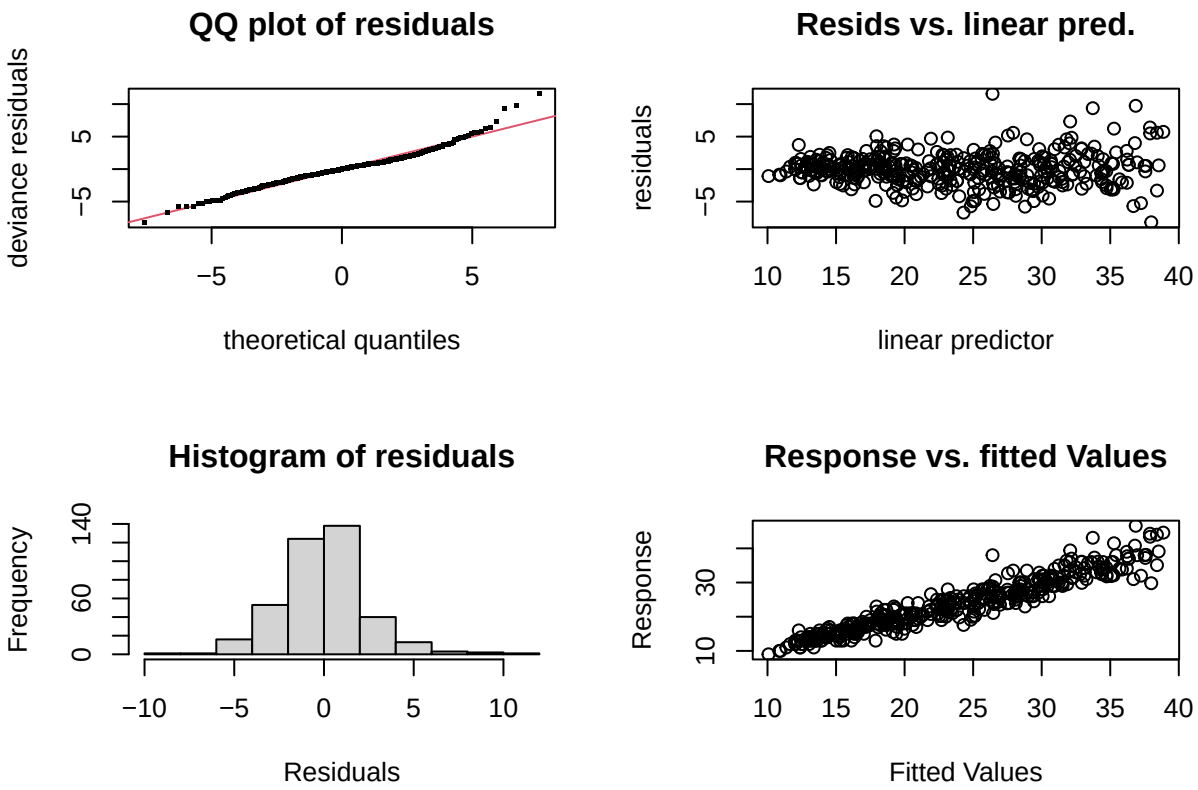
The adjusted R-squared is 0.897. This means that 89.7% of the variance in the training data can be explained by the model. This value has been adjusted for the number of predictors in the model.



Plot 3

Model checking By using `mgcv`s concurvity function we can see that we have a large problem with concurvity. `s(displacement)`, `s(horsepower)`, `s(weight)` and `s(acceleration)` has worst concurvity statistics higher than the commonly used threshold of 0.8. This means that the variables could have similar shapes or a concave relationship to each other. This can lead to unreliable estimations of the smooth terms, undermines interpretability and may affect the accuracy of out predictions. I would recommend removing or transforming variables until this problem is reduced.

Plot 4 shows four different plots to check model assumptions. In the QQ plot of residuals we can see that the residual does not follow the line, indicating that they are not normally distributed. In the histogram of residuals we can see can see that the residuals are left skewed. In the residuals vs linear predictions plot we can see non-constant variance. The response vs fitted values is the only one indicating a decently good fit with residuals being randomly scattered around the around a straight line.


QQ plot of residuals
deviance residuals
−5
5
−5
0
5
theoretical quantiles
Resids vs. linear pred.
residuals
−5
5
10
15
20
25
30
35
40
linear predictor
Histogram of residuals
Frequency
0
60
140
−10
−5
0
5
10
Residuals
Response vs. fitted Values
Response
10
30
10
15
20
25
30
35
40
Fitted Values

Plot 4