ST1131 Report

Chen Jiahao — A0287926J

1 Introduction

Main Goal of this Report: This report aims to explore the dataset about cars collected in 1983, to study which variable(s) may affect fuel consumption, measured in miles per gallon (mpg).
About the dataset: The dataset is about cars collected in 1983. Apart from the fuel consumption, other quantitative factors are as follows: cylinders, displacement, horsepower, weight, acceleration, model year. There are also three categorical variables: origin, car name, and year.
The steps to follow: First we would use plots and conduct hypothesis testing to each variable, to observe their significance. Then we would build the linear models, checking whether they satisfy the linearity, normality and constant variance. And we adjust the model at the same time, until we find the final good one. To be the final good model, it should satisfy all the assumptions as well as have a high R^2 value.

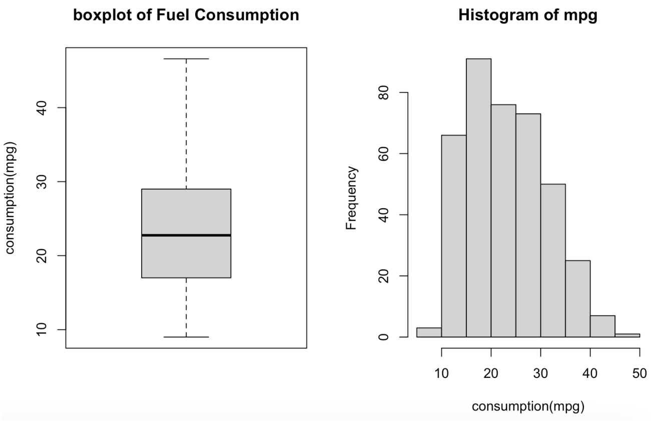
2 Explore the variables and association

In this report, the response variable is **Fuel Consumption**. We can run through the dataset to find out its statistical summary.

Table 1: Statistical Summary

Var/Stats	Consumption(mpg)
Min.	9.00
1st Qu.	17.00
Median	22.75
Mean	23.45
3rd Qu.	29.00
Max.	46.60

Figure 1: Boxplot and Histogram



For other variables, we divide them into quantitative variables and categorical variables.

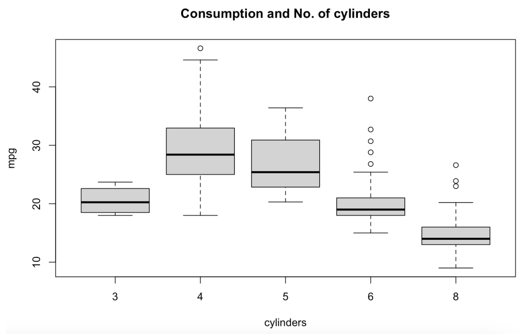
- **quantitative variables**

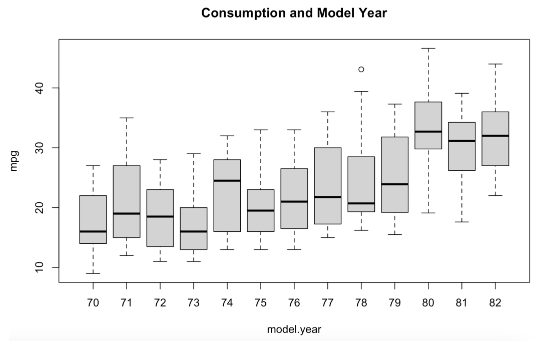
Var/Stats	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Understanding
Cylinders	3.000	4.000	4.000	5.472	8.000	8.000	Have a low outlier at 3, while most values are clustered around 4, 6, or 8.
Displacement	68.0	105.0	151.0	194.4	275.8	455.0	Have outliers on the higher end
Horsepower	46.0	75.0	93.5	104.5	126.0	230.0	Right-skewed, with outliers on the higher end
Weight	1613	2225	2804	2978	3615	5140	Mean close to median, a bit right skewed but not much
Acceleration	8.00	13.78	15.50	15.54	17.02	24.80	Mean almost equals to median, nearly symmetric
Model.Year	70.00	73.00	76.00	75.98	79.00	82.00	Mean almost equals to median, nearly symmetric

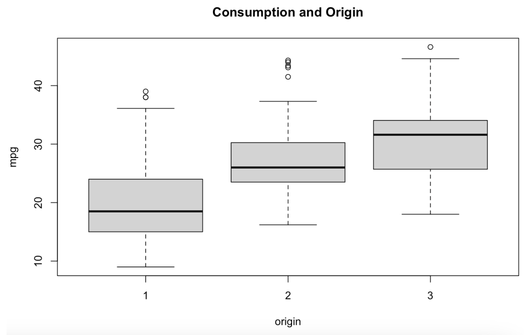
- **categorical variables**

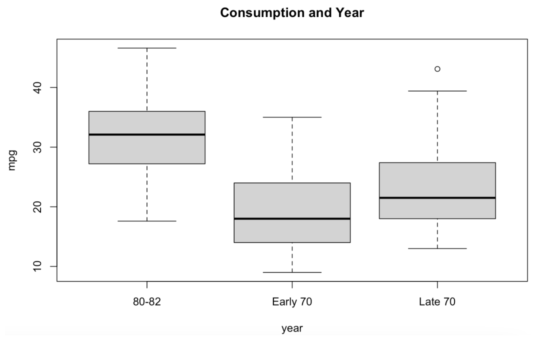
origin: More than half are 1 (from USA), few are 2 (Europe) or 3 (Japan)
car name: Name of the car brand
year: 150 ”early 70”, 157 ”late 70”, 85 ”80-82”

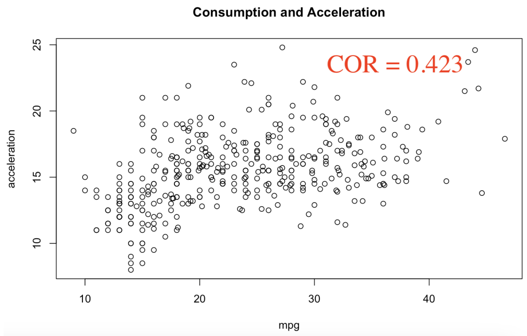
We can tell that the response variable is *quantitative*. And from the eight plots above, we can say that it also satisfies *linear relationship*. But from table 1 and Figure 1, we found that the distribution of mpg is *not that normal*. It is still **suitable** to fit a linear regression model for this response but we would like to make it symmetric by using log(y) or $\sqrt{y}$. As a result, we would **check the association** between the response and other variables below.

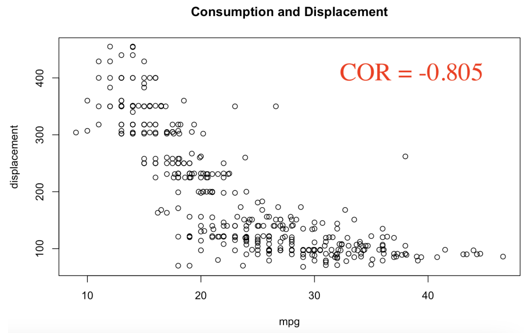

Consumption and No. of cylinders

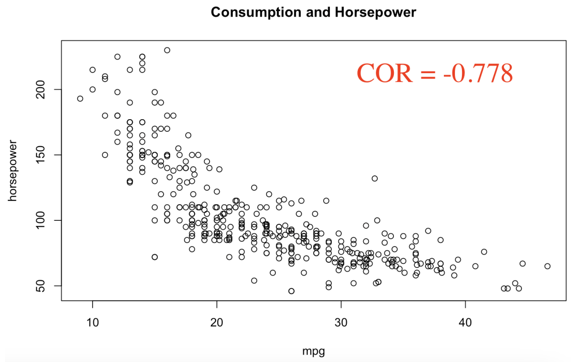

Consumption and Model Year

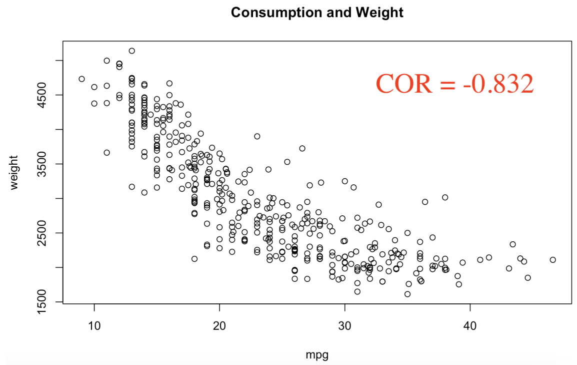

Consumption and Origin


Consumption and Year


Consumption and Acceleration


Consumption and Displacement


Consumption and Horsepower


Consumption and Weight

Result: From the four **histograms**, we can tell that No. of cylinders obviously is not a regressor. While *year (model year)* and *origin* may be regressors.

From the four **scatter plots**, the absolute value of correlation of consumption and displacement, weight and horsepower are around 0.8, indicating that there may be an association between consumption and *displacement* / consumption and *weight*. So these three variables are also likely to be regressors.

I have also done **hypothesis testing** to each variable, making H0 to be regressor not significant, it turns out that for every variable, the P-value is super small(i.e. <2.2e-16), meaning that we can not directly reject any variable now.

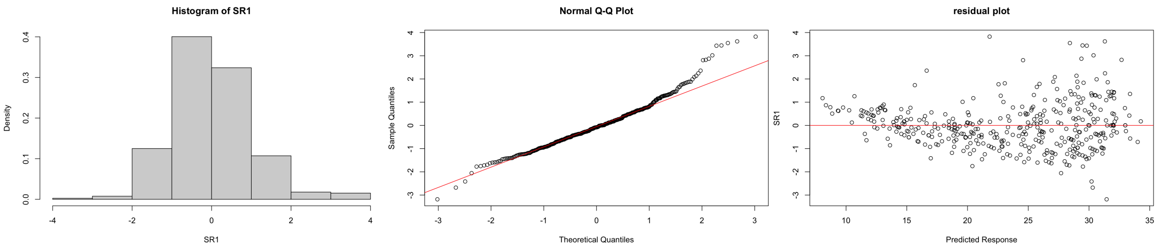
3 Build the model

3.1 Initial Model M1

For M1, we decided to use the three variable *displacement*, *weight* and *origin* as regressors. After fitting the model in R Studio, the result is

$$mpg = 42.258 - 0.014 \times displacement - 0.006 \times weight + 0.414 \times I(origin = Europe) + 1.985 \times I(origin = Japan)$$

Checking Residual plots:



The histogram and Q-Q plot indicate that the standardized residuals are not normally distributed. Additionally, the funnel shape in the third plot suggests a violation of the constant variance assumption.

Outliers and influential points:
There are several outliers (index = 111,243,321,324,325,382,389), but there is no influential point.

Checking each regressor:
The P-value of *displacement* is 0.02 ⟶ not very significant
The P-value of *weight* is 5.99e-14 ⟶ very significant
The P-value of *origin2* is 0.55 ⟶ not significant
The P-value of *origin3* is 0.004 ⟶ not very significant
R^2 value 0.706 and Adjusted R^2 0.703 ⟶ Okay

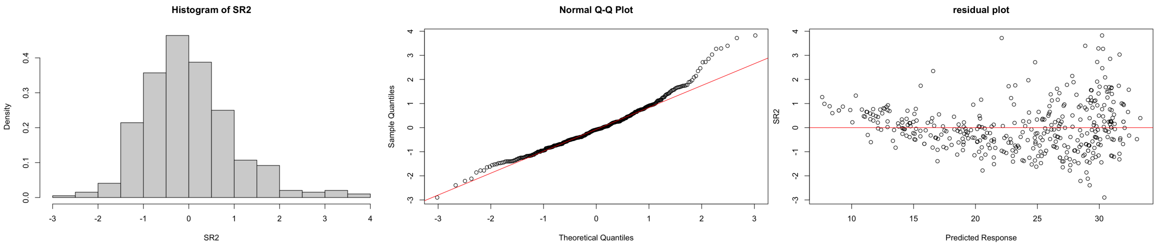
Interpretation:
As our initial model, it failed to satisfy the normality, constant variable assumption. Additionally the origin regressor is not significant. As a result, we would like to delete this variable from our model.

3.2 Second Model M2

For M2, we decided to delete the variable origin and use the two variables *displacement* and *weight* as regressors. After fitting the model in R Studio, the result is

$$mpg = 43.778 - 0.016 \times displacement - 0.006 \times weight$$

Checking Residual plots:



The histogram and qq plot illustrate that the standard residuals are not distributed normally. And in the third plot, there appears a funnel shape, which implies that the constant variance assumption is violated.

Outliers and influential points:
There are several outliers (index = 321,324,325,328,382,389), but there is no influential point.

Checking each regressor:
The P-value of *displacement* is 0.004 ⟶ not very significant
The P-value of *weight* is 7.31e-15 ⟶ very significant
R^2 value 0.699 and Adjusted R^2 0.6974 ⟶ Too low

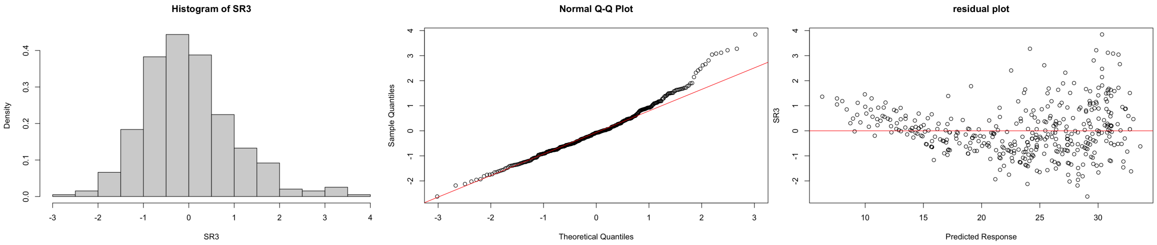
Interpretation:
Given that the SR is not distributed normally as well as the constant variance assumption is violated, This model is not a strong model. As a result, we would like to remove the variable *displacement*, which is not that significant. Consequently we would also introduce in a new variable *horsepower*, which is also likely to be a regressor shown in Part 2.

3.3 Third Model M3

For M3, we decided to delete the variable origin and use the two variables *horsepower* and *weight* as regressors. After fitting the model in R Studio, the result is

$$mpg = 45.640 - 0.047 \times horsepower - 0.006 \times weight$$

Checking Residual plots:



Similar to M2, the histogram and qq plot illustrate that the standard residuals are stillnot distributed normally. And in the third plot, the funnel shape is still there, meaning that the constant variance assumption is violated.

Outliers and influential points:
There are several outliers (index = 321,324,325,328,382,389), but there is no influential point.
P.S. The indexes of outliers are exactly the same with M2!

Checking each regressor:
The P-value of *horsepower* is 2.49e-05 ⟶ very significant
The P-value of *weight* is <2e-16 ⟶ very significant
R^2 value 0.7064 and Adjusted R^2 0.7049 ⟶ Better! But still needs improving

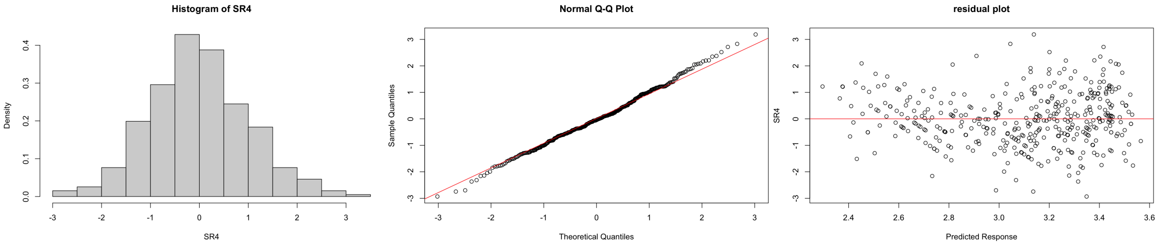
Interpretation:
The good thing is that now both regressors are significant now. However similar to M2, the SR is not distributed normally as well as the constant variance assumption is violated. This model is not a strong model.
Specifically to address the constant variance issue, we would like to make mpg to be log(mpg), trying to achieve constant variance.

3.4 Fourth Model M4 (Final Model)

For M4, we decided to use the same two variables *horsepower* and *weight* as regressors, but make mpg to be log(mpg). After fitting the model in R Studio, the result is

$$log(mpg) = 4.111 - 0.0003 \times horsepower - 0.002 \times weight$$

Checking Residual plots:



This time, the histogram is quite symmetric, and in the qq plot the distribution aligns quite well with the line. In the third plot, we can tell that the constant variance assumption is satisfied.

Outliers and influential points:
There is only one outlier (index = 382), and there is no influential point.

Checking each regressor:
The P-value of *horsepower* is 1.2e-09 ⟶ even more significant than that in M3!
The P-value of *weight* is <2e-16 ⟶ still very significant
R^2 value 0.7879 and Adjusted R^2 0.7869 ⟶ Good job!

Interpretation:
This time, we can say that it is a good model. It not only satisfies the normality and constant variance assumptions which the previous models failed to, but also increased the R^2 value, making the model better.
As a result, we would declare it to be our final model.

4 Final model

Our final model is M4, where

$$log(mpg) = 4.111 - 0.0003 \times horsepower - 0.002 \times weight$$

Effect of each variable:
For a unit increase in horsepower, log(mpg) would decrease by 0.0003. For a unit increase in weight, log(mpg) would decrease by 0.002.