FACTORS AFFECTING ACCELERATION

OF A VEHILE

SUBMITTED TO: PROF. ELIZABETH SUBMITTED BY: AYUSHI SINGH

AVERY GOMEZ

A car parked in front of a body of water

Description automatically generated

**TABLE OF CONTENTS**

[1) Introduction: 2](#__RefHeading___Toc4481_1188962741)

[2) Data preparation and description: 2](#__RefHeading___Toc4483_1188962741)

[2.1) Data preparation: 2](#__RefHeading___Toc4485_1188962741)

[2.2) Data description: 3](#__RefHeading___Toc4487_1188962741)

[i) Correlation between variables: 5](#__RefHeading___Toc4489_1188962741)

[ii) Graphics of the dependent variable VS independent variables: 6](#__RefHeading___Toc4491_1188962741)

[Plot 1: 6](#__RefHeading___Toc4493_1188962741)

[Plot 2: 7](#__RefHeading___Toc4495_1188962741)

[Plot 3: 7](#__RefHeading___Toc4497_1188962741)

[Plot 4: 8](#__RefHeading___Toc4499_1188962741)

[3) Model Planning and building: 9](#__RefHeading___Toc4501_1188962741)

[Feature selection: 10](#__RefHeading___Toc4503_1188962741)

[Verifying the best model: 14](#__RefHeading___Toc4505_1188962741)

[Hypothesis of the linear model : 15](#__RefHeading___Toc4507_1188962741)

[i) H1:Normality of residuals 15](#__RefHeading___Toc4509_1188962741)

[ii) H2: Homoscedasticity 16](#__RefHeading___Toc4511_1188962741)

[iii) H3: Multicolinearity of variable 16](#__RefHeading___Toc4513_1188962741)

[iv) H4: Exogeneity of variables 17](#__RefHeading___Toc4515_1188962741)

[Plot of the prediction : 17](#__RefHeading___Toc4517_1188962741)

[4) Conclusion: 18](#__RefHeading___Toc4519_1188962741)

# 1) Introduction:

In this report , we will try to detect variables that can determine the acceleration of a vehicle, in our life we notice that many factors can influence the acceleration of a vehicle, for example the consumption per mile of diesel or gasoline of the vehicle, number of horsepower inside the engine of the vehicle, brand of the vehicle and number of years that the vehicle had being used.

To know more about determinants of the acceleration, we use a data that gather information about some 392 vehicles from different brand types.

In a first part, we will discover the variables (we will do both statistical and graphical exploration).

In a second part we will start building our model by including all the variables , and then making a selection of the best that explain the best acceleration, the selection of variables will be based on testing hypothesis and on comparing the full and the reduced model using ANOVA method.

after model selection we will discuss the effect of those selected variables on the acceleration and finally we will verify hypothesis related to the model we build, to verify if the model is strong enough.

# 2) Data preparation and description:

## 2.1) Data preparation:

bd = Auto  
dim(bd) ## dimension of the data set

## [1] 392 9

The data set is composed of 392 rows and 9 variables, those variables are:

names(bd) ## names of attributes

## [1] "mpg" "cylinders" "displacement" "horsepower"   
## [5] "weight" "acceleration" "year" "origin"   
## [9] "name"

*mpg* : is Miles per gallon  
*cylinders* : Number of cylinders inside the vehicles   
*displacement* : Engine displacement  
*horsepower* : Engine horsepower  *weight* : Weight of the vehicle   
*acceleration* : Time to accelerate from 0 to 60 mile per hour   
*year* : Model year   
*origin* : Vehicle origin   
*name* : The name of the vehicle

We should notice that the variable **name** should be removed, and the variable origin should be transforming to factor type.

## 2.2) Data description:

By doing those transformations, we give a statistical description about our variables.

We can notice that there is no missing value in all continuous variables, the table above give a summary about many statistics like the mean and median, the max and the mean, the variation and coefficient of variation for each continuous variable.

regarding the categorical variable, we present the frequency of each modality.

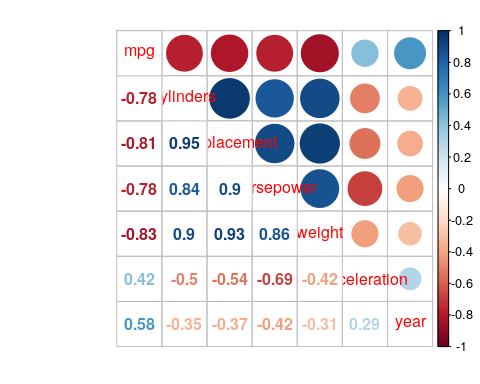
1 : American

2 : European

3 : Japanese

We notice that 62.5% of vehicles are from USA, 17.3% are from UE and the rest of vehicles are from Japan.

### i) correlation between variables:



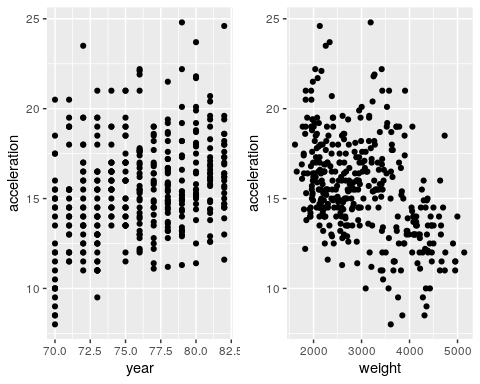
We are interested in a first stage about correlation between acceleration and other variables, we notice that there is a negative correlation between acceleration and horsepower, displacement, cylinders that exceed 0.5, the correlation between acceleration and other variables does not exceed 0.5.

### ii) Graphics of the dependent variable VS independent variables:

We will now present some graphs where Y axis represent the acceleration variable and the X axis will vary for each of the independent variable.

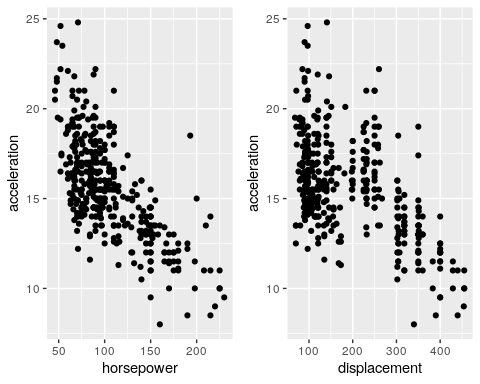
#### 

#### Plot 1:



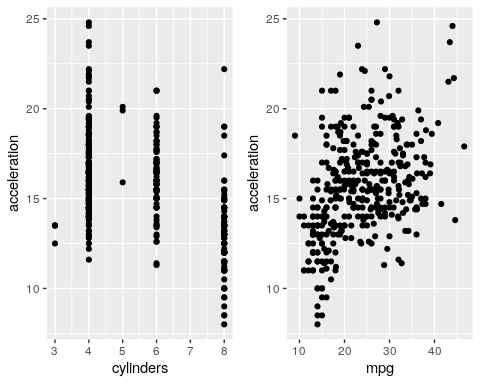
We can notice that acceleration does not vary much within year, but considering weight, we can see that the more vehicle is heavy the more acceleration is low.

#### Plot 2:



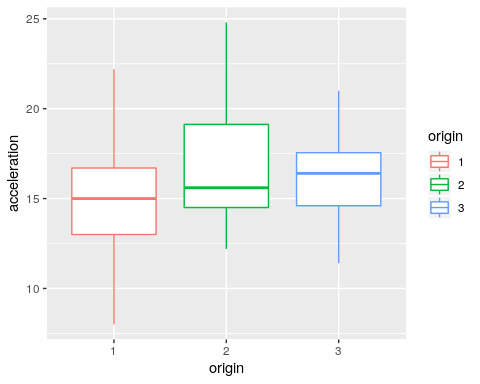
From those two plots, we notice that by increasing displacement or horsepower acceleration become low.

#### Plot 3:



From this plot we notice that vehicle with 4 cylinders have high acceleration than vehicles with 8 cylinders. and regarding mpg, we notice that vehicles with more mpg have high acceleration and those with low mpg have low acceleration.

#### Plot 4:



Finally, by considering the origin of the vehicle, we notice that vehicle from US have a mean acceleration low than vehicle from UE or Japan. it’s maybe related to some norm applied in US but not in other countries.

# 3) Model Planning and building:

The model we will use to figure out variables that influence acceleration is the linear model. we will first run a model with all variables inside and interpret the results.

From the output of our regression, we notice that significant variables are:

displacement, horsepower, weight and year.

*displacement* is significant at a 5%, since it’s p-value is less than 0.05, an increase (decrease) of displacement of on unit will decrease (increase) acceleration by 0.0081.

*horsepower* is significant at a 5% since it’s p-value is less than 0.05, an increase (decrease) of horsepower of on unit will decrease (increase) acceleration by 0.085.

*weight* is significant at a 5% since it’s p-value is less than 0.05, an increase (decrease) of weight of on unit will increase (decrease) acceleration by 0.003.

*year* is significant at a 10% since it’s p-value is less than 0.1, an increase (decrease) of year of on unit will decrease (increase) acceleration by 0.096.

The R squared is equal to 0.61, it means that 61% of variability of acceleration is explained by the independent variables.

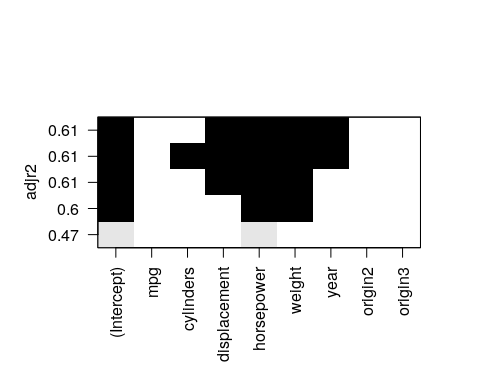
Since not all variables are significant, we will make a sub selection of the best variables using reg-subsets function.

## Feature selection:

By the adjusted R squared and CP(Mallows Cp) criteria, 4 variables were retained, and by the BIC criteria 3 variables were retained.

the variable retained taking R squared as criteria for selecting the best model among all models are:

displacement, horsepower, weight and year

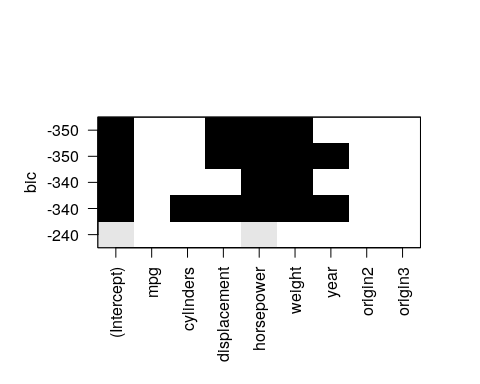


the variable retained taking Cp as criteria for selecting the best model among all models are: displacement, horsepower, weight and year



the variable retained taking BIC as criteria for selecting the best model among all models are:

displacement, horsepower and weight



We will retain only variables selected by the BIC criteria, and we will fit a regression model and print the results.

## Verifying the best model:

The r-squared is the same as the r-squared of the full model (model with all variables).

Let’s make some tests to see if those selected are good or not. We will test in a first stage the null hypothesis of the null coefficient of mpg, cylinders, year and origin variables.

Since the p-value is greater than 5%, it means that the restricted model is better than the full model.

a second test that gives the same results is the anova test.

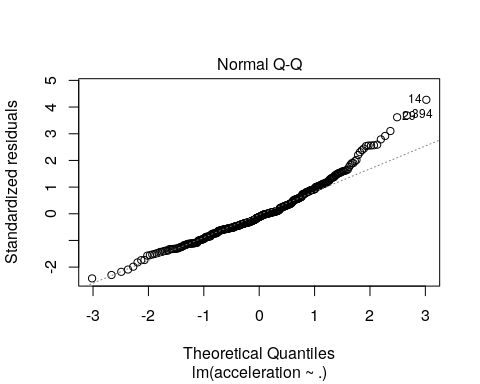
We get the same conclusion as the previous one, the restricted model is better than the full model

Now that we have selected the best variables that explain acceleration, let’s verify the strangeness of our model by verifying the hypothesis related to the linear model.

## Hypothesis of the linear model:

### i) H1:Normality of residuals

We should verify either the residues of the regression follow a normal distribution or not.



From this plot, we notice that not all observations are aligned to the dashed line, so we can suspect the validity of this hypothesis. To be sure of that we fit a shapiro test:

When the p-value is less that 5% we reject the hypothesis of normality of residues. the violation of this hypothesis is due to the restricted observations we have, and so this hypothesis will be respected if we add more and more variables to our data.

### ii) H2: Homoscedasticity

Homoscedasticity of errors is a fundamental assumption of the regression model, when this assumption is not verified the estimator of our regression is unbiased but is no longer minimum variance. To check if the residuals are of the same variance, we use the test of Breucsh-Pagan defines as follows:

H0: Residuals don’t have the same variance

H1 : Residuals have the same variance

for a p-value greater than 5% we accept H0, results of the test are:

We notice from the result that the p-value of the test is greater than 5%, we deduce that the hypothesis of homoscedasticity of the variance is not respected.

### iii) H3: Multi collinearity of variable

The multi-collinearity is a hypothesis to verify for a model of linear regression, a multi col-linearity means that among our exogenous, there is a linear relationship between one or more variables with the rest of the variables. To check for the existence or absence of multi-col-linearity, we calculate the VIF of each exogenous variable.

For the case of a linear regression, we suspect the existence of multiple collinearity when the VIF of the variable is greater than 10. The VIF variable table Exogenous is as follows:

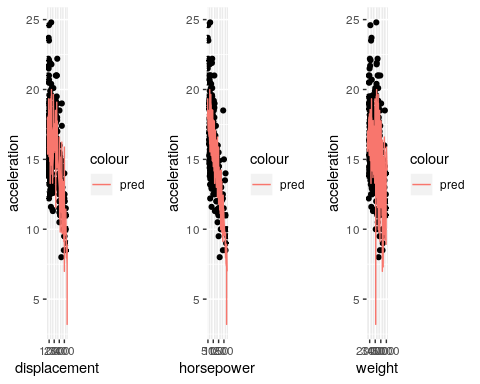
Since there is a high correlation between displacement and horsepower (0.9), vif of horsepower is greater than 10 and hence we can suspect existence of co-linearity.

### iv) H4: Exogeneity of variables

We speak of exogeneity if the explanatory variables are not correlated with the residues, for this reason let us take the residues of our regression and calculate its correlation with Exogenous variables.

We find that the correlation between residues and exogenous variables is almost zero, which validates the hypothesis of exogeneity of our model.

## Plot of the Prediction:



# 4) Conclusion:

As a conclusion we can say that among all the characteristics related to vehicles present in the data set, only three characteristics were retained to explain acceleration, those three characteristics are displacement, horsepower and weight.

Those three characteristics were retained using two different methods, the first method was the reg-subsets which give us the best model within a criterion, and the second method was using hypothesis and testing whether coefficient of another variable could be equal to zero or not. Both methods agreed that those three variables are the best to explain acceleration.

The regression model we fitted can explain 61% of variability of acceleration which is something good, but the model sick to respect hypothesis related to regression model since 3 hypotheses are not respected among 4.

For the first hypothesis *Normality of residuals* the solution is to increase number of variables For the second hypothesis *Homoscedastivity* , we can propose three solutions , the first one is scaling our data and then fitting the model, the second solution is instead of using acceleration we will use the logarithm of acceleration, the third solution is to combine the two previous solutions. For the third hypothesis *Multicolinearity* the solution is to delete displacement from the model since there is strong correlation between this variable and horsepower.