Linear models R practical #2

Aurélien Bloch & Ogier Bouvier

Summary

Contents

1 Introduction

In our previous report we fitted a Gaussian liner model to our cars data. This has given an insight on how the variables we used in this analysis influenced the fuel efficiency of each car. But how do we know the variables we ignored in our previous model have no influence on this fuel efficiency?

Throughout this report we are going to determine which of these variables do have an effect on the fuel consumption of cars. This will allow us to use the simplest possible model for the interaction between each of our 26 variables and the fuel efficiency.

This will then try to verify the assumptions we made when starting our analysis seem to hold. This ensures that our model is indeed valid.

2 The Data

We'll use data from the April 1993 issue of U.S. consumer report which covers 82 different models of cars. We'll focus our analysis on three of these variables, namely the Gallon per mile in city, the weight and the horsepower of the cars.

In our analysis we'll consider the fuel efficiency of the cars as $Y = \frac{100}{CityMPG}$ which will represent how many gallons of fuel each car consumes to travel 100 miles.

We are also going to consider the power of the car but since this is also influenced by the weight we'll use $X1 = \frac{Horsepower}{Weight}$

3 Methodology

To assess the effect of each variable on the response variable we are going to use a process named ANOVA (ANalysis Of VAriance) [**?**]

This process allows use to assess the effect of each covariate on the response variable. This way we can filter out covariate which seem to have no or little effect on the response variable. Doing this we can select the simplest model while still retaining every useful covariate.

ANOVA consists of analyzing by how much the inclusion of additional columns to our design matrix reduces the residual sum of squares. Then we will need to consider whether or not this reduction is sufficient to consider the corresponding columns significant in our model. To this end we will use a F-statistic, under the hypothesis that the parameter β_i that correspond to the column is equal to 0 we know that $COMPLETEHEREFORTHEFORMULA$ follows a $F_{2,3}$ distribution. We can then compute the value for this particular column and reject the hypothesis that $\beta_i = 0$ if the value is too far away from plausible values under the F distribution.

4 Analysis

5 Discussion and Conclusions

A R source code

% TODO