

Regression Models: Motor Trends

noootsab

2014-08-16

Summary

Analyse cars data to detect if some difference exists between gaz consumption of manual cars, or automatic ones.

For that, we'll use the data from mtcars which containing 11 measures for 32 cars ~70.

In the following sections we'll explain why we think that the difference between the manual and automatic transmission is statically significant. But also, we'll see that the weight has a dramatic importance when choosing a car's transmission regarding mpg.

In fact, when the car is light enough (~1000lb), manual transmission is initially at 16.8 mpg which is around 10 more than for automatic ones. However, when the weight increases, there is a decrease of the mpg of (around) -7 for the manual cars, and only -3 for the automatic ones. Thus, starting, with cars more than 3500lb, the automatic cars should be the choice of preference.

Analysis

a look at data > head

give the mean of both population > mean manual > mean automatic

See more in the boxplot in the appendix > boxplot, showing the clear difference

To ensure that the difference is significant, we can test both pops using t test, the indenpendence between population is clear (but assuming normality). > t.test

Maybe the am is sufficient to model the mpg > fit mpg am > low R² => need more

load leaps for stepwise model selection (since n <<) > best + forward + backward

The forward and backward are diverging (qsec, hp) at the third model, however they both select wt at first. However, best model selection cuts the choice by selecting am with wt and qsec.

So let's fit these variables in a linear regression. > lm without interaction

Quite good, but maybe could we improve it a bit by looking if am is or not interacting with other variables. Like with wt, it's a good choice since it's so importance in explaining mpeg. See in the appendix how wt and mpg are related in presence of am. > plot wt mpg with am

If we draw a regression line per group, it gives us the impression that wt and am are interacting, and actually could inverse the relation with mpg at some point. Let's add this interaction to the model > lm with interaction

Now we can test how this new variable in the model is significant to explain the variance, for this we can run an anova > anova

Including interaction with qsec? Will considerably change the
model <<<<<<

We can have a quick look at the VIF of the models. > vif > vif

Okay, looks like the model including the interaction inflates the variance due to colinearity, but we could have foresee it regarding the plots we've seen. However, still it's inclusion allow better explanation of the difference between the groups.