Study of wich variables affect miles per galon

Cristobal Morgado

2022-10-01

1 Introduction

The overall aim of this proyect is try to explain the variables that effect the most the miles per galon of a car. For this purpose, we use the mtcars data set. Also the packages dplyr, ggplot2, car and lmtest are used. The methodology is to use an ANOVA analysis to find relationships between mpg and other variables, with them construct a model and test some transformations of it. Finally we test the models to check if a variable is inflating the variance (VIF test) and if there is heterocedasticity (bptest). With this results we will select a model and make inferences on the resultant coefficients.

Code available at: https://github.com/CMorgadoM/Proyecto_regresiones_mtcars

2 Summary and transformations

The first step is to get some descriptions of the data and transform the variables "am", "vs", "cyl" to class factor.

3 Modeling

Once we understand the data set, the next step is to create a model with all original variables and test the model with and anova analysis to find the variables which impact in mpg is significant. Here we found that variables "am", "wt" and "hp" are significant, so that variables are selected to construct the initial models.

Table 1: ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
am	1	405.151	405.151	57.685	0.00000
wt	1	442.577	442.577	63.013	0.00000
hp	1	98.029	98.029	13.957	0.001
cyl	1	10.293	10.293	1.466	0.239
qsec	1	10.180	10.180	1.449	0.242
disp	1	8.826	8.826	1.257	0.275
vs	1	0.233	0.233	0.033	0.857
carb	1	0.044	0.044	0.006	0.938
gear	1	1.593	1.593	0.227	0.639
drat	1	1.627	1.627	0.232	0.635
Residuals	21	147.494	7.024		

we build 3 models, one with am, cyl, wt and hp as independent variables; a second one without cyl (as it wasn't significant in the model); and a third one with an interaction between am and wt and hp. All three models ended up having a large R-squared and a significant F test.

The test VIF shows that all models have high variance inflation factors, the model3 is the one with the lesser values so we test taking out the "wt" variable and the result is that the Adjusted R-squared pass from 0.985 to 0.767, since the difference is huge and the variable is significant we keep working with the model3.

Also, we construct model_5 and model_6, using model4 formula and taking out one interaction; model_5 maintain "am * wt" and model_6 maintain "am * hp". From this two models the model_5 has the best VIF results of all models tested.

The test Brauch-Pagan to test heterocedasticity. With the 4 out of 5 models we reject the null Hypothesis so we assume homocedasticity in every model except in model 6.

Table 2: BP test results

	model	BP test P-value
1	model 2	0.24
2	model 3	0.14
3	model 4	0.56
4	model 5	0.69
5	model 6	0.02

4 conclusions

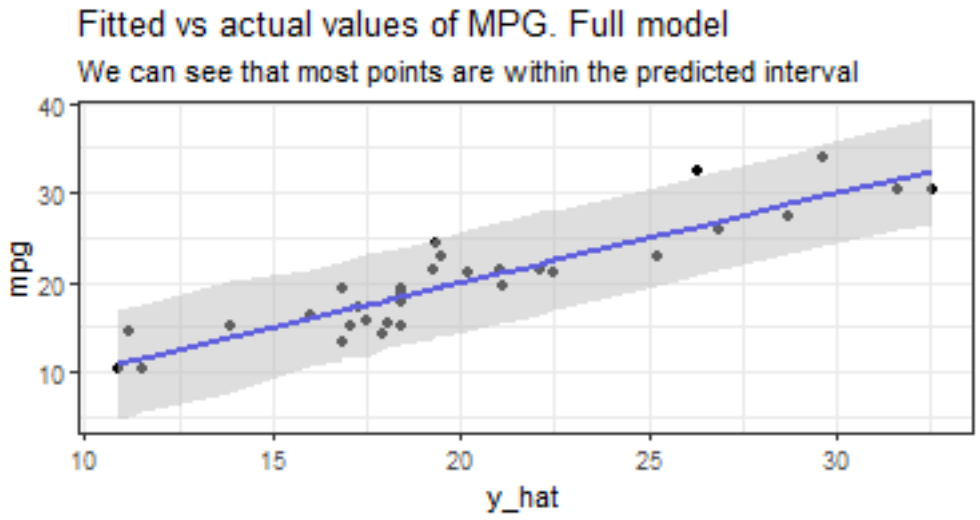
After all these analysis we decide to use the model_5 to make inference of the population. This is because, in this model, the interaction term allows to understand better the impact of the variables in the miles per galon consumption for each type of transmission.

With this model we can assume that the mean of the impact for cars with automatic transmission is 17.15 MPG per 1 lb more of weight and for cars with manual transmission, the mean of the effects is 24.39 MPG per 1 lb more of weight.

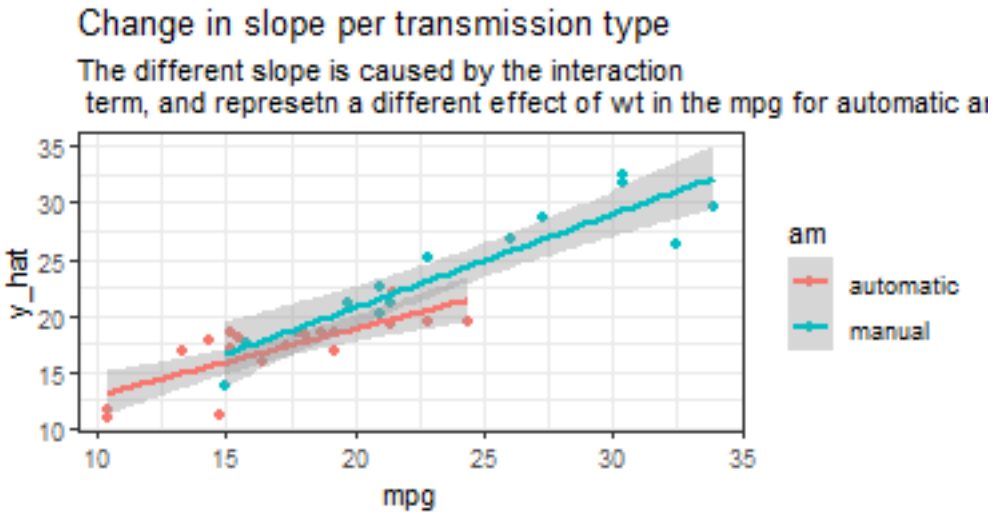
Finally, by performing a median test, we try to find out if the difference of the means is significant. The result of the test is a confidence interval between 5.07 and 9.42. Since the interval doesn't contain zero we can assume that the difference predicted by the model between manual and automatic transmission is significant.

5 Appendix

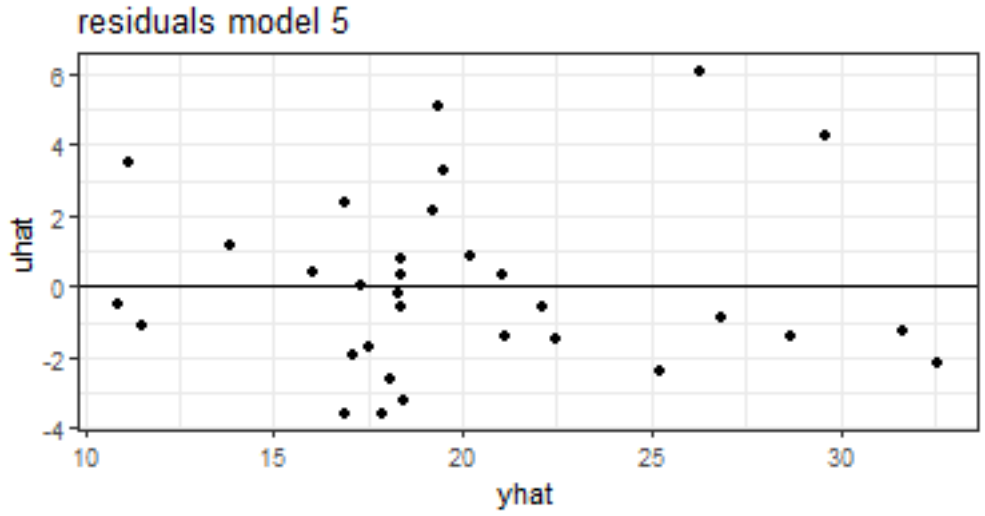
5.1 Predicted vs actual values for mpg with model_5



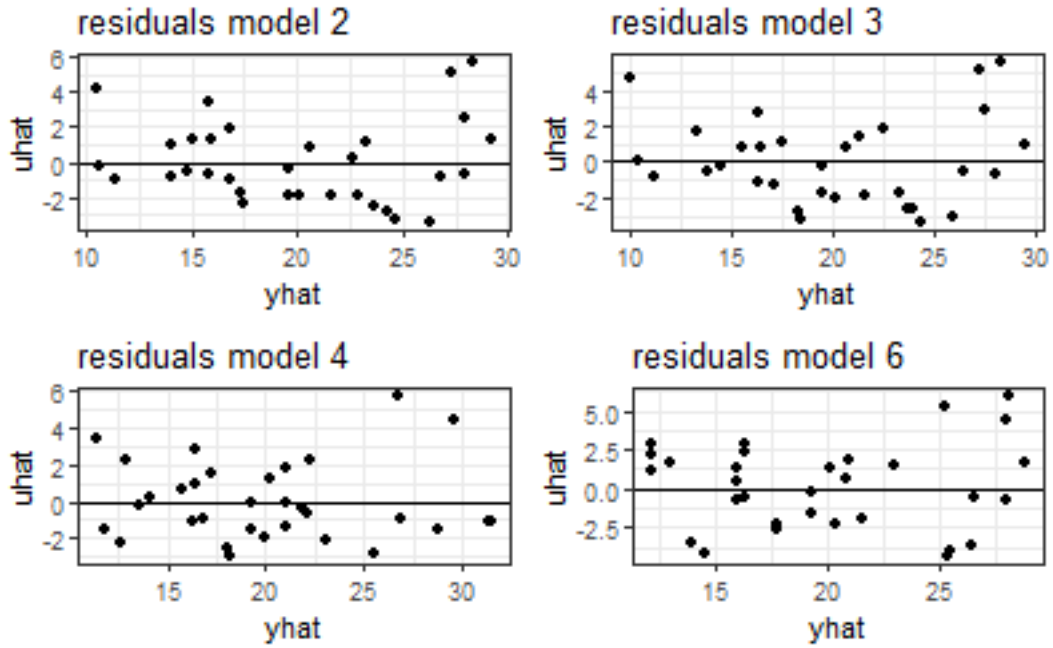
5.2 Change in slope per transmission type



5.3 residuals vs fitted values for model_5



5.4 Other models residuals plots



5.5 Model 5 coefficients and summary

Table 3: Model 5 Summary

	Dependent variable:
	mpg
amautomatic	31.416^{***}
	(3.020)
ammanual	46.294^{***}
	(3.010)
wt	-3.786^{***}
	(0.786)
ammanual:wt	-5.298^{***}
	(1.445)
Observations	32
R^2	0.987
Adjusted R^2	0.985
Residual Std. Error	2.591 (df = 28)
F Statistic	515.831^{***} (df = 4; 28)
Note:	*p<0.1; **p<0.05; ***p<0.01