Automatic or manual transmission? A statistical approach

Omar Alfaro-Rivera

Executive Summary

The following analysis emphasizes the relationship between automobile performance (measured in Miles Per Gallon) and the type of transmission (automatic or manual). Research shows that **performance is favored by manual transmission**, a relationship that is maintained by including other variables to the model, such as *Gross horsepower* and *Number of cylinders*. It should be noted that the conclusions drawn here are completely partial, since the large number of variables that can be decisive in the performance of variables. In the next issue of the magazine this topic is completed by conducting a Principal Component Analysis.

Introduction

Based on the data obtained in the Motor Trend Car Road Tests (the following table shows the structure of the data), it is proposed to use regression models to check the impact of the type of transmission on the performance of automobiles (measured in Miles Per Gallon).

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

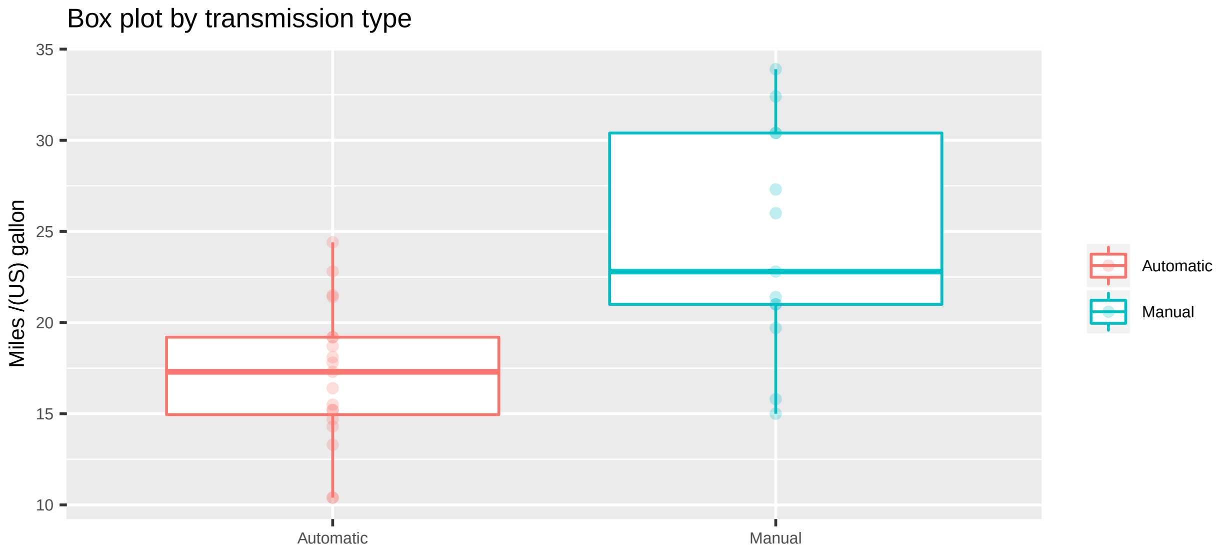
The following variables will be used for this analysis:

- mpg Miles/(US) gallon
- hp Gross horsepower
- cyl Number of cylinders

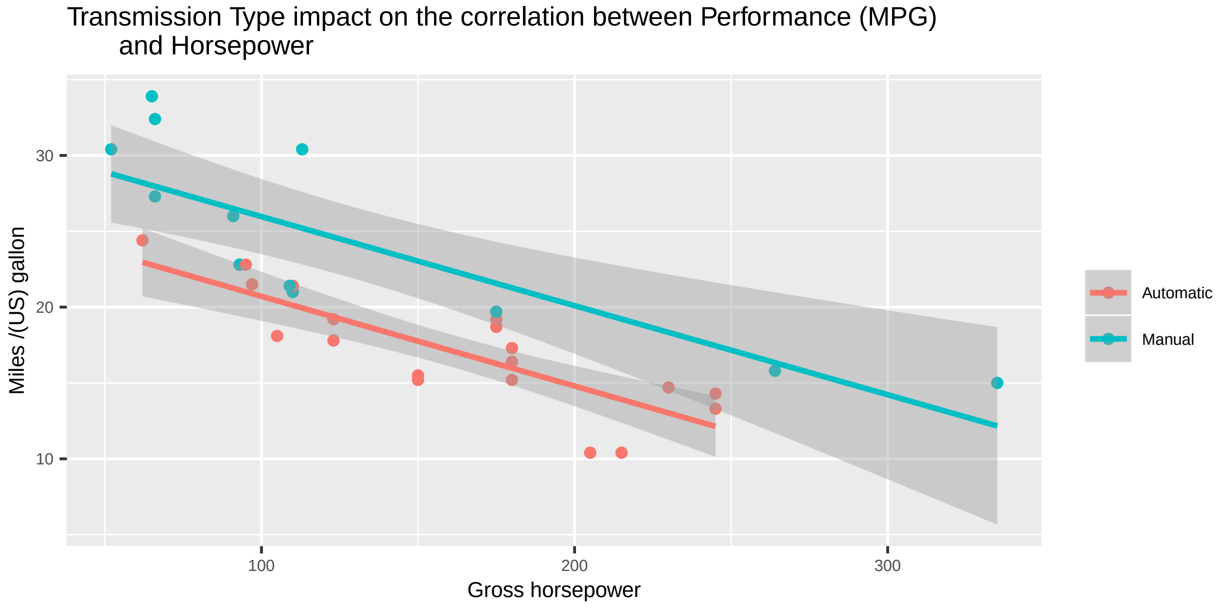
For readers looking to reproduce this analysis, visit the following site:

Exploratory Analysis

In this section we will review two graphs that shed light on the best way to estimate the model to measure the relationship between the type of transmission and the fuel efficiency of automobiles. First off, the box diagrams below show that, on average, cars with a manual transmission perform better and the maximum values are also favorable for this transmission.The magnitude of this difference and its statistical significance will be reviewed in the next section.



One of the variables that is given great importance in this analysis is the horsepower of each car. This variable is assumed to be beneficial in modeling the desired relationship by adding more variables and keeping the principle of parsimony, since this variable is a good proxy for many other variables, we delve further into this in the next section. For now, the negative relationship between this variable and performance is clear, while the type of transmission also shows a difference in the ordinate to the origin.



Estimated models

The following regression models are proposed to capture the desired relationship, as can be seen from the simpler relationship, and more variables are gradually added to the model.

$$MPG = \beta_0 + \beta_1 Transmission + \epsilon$$

$$MPG = \beta_0 + \beta_1 Transmission + \beta_2 horsepower \epsilon$$

$$MPG = \beta_0 + \beta_1 Transmission + \beta_2 horsepower + \beta_3 Cylinders + \epsilon$$

The estimates shown below show the p value associated with the null hypothesis that the estimator is equal to zero, that is, that said variable has no effect on the dependent variable, in other words:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

Thus, being a two-tailed test, the p value should be less than 0.025 to fail to reject the null hypothesis. In this order of ideas, the following table shows the model made in the columns, and the independent variables and the statistics associated with that model in the rows. Los espacios en blanco quieren decir que para esa estimación no se utilizo la variable en cuestión, a la vez que estaríamos buscando tres asteriscos en nuestros regresores (significancia al 95%).

Table 2: Regression results

	Miles /(US) gallon (MPG)		
	(1)	(2)	(3)
Transmission	7.245***	5.277***	3.796**
	(1.764)	(1.080)	(1.424)
Gross horsepower		−0.059***	−0.041***
		(0.008)	(0.014)
Number of cylinders			−0.014
			(0.009)
Constant	17.147***	26.585***	27.866***
	(1.125)	(1.425)	(1.620)
Observations	32	32	32
R^2	0.360	0.782	0.799
Adjusted R^2	0.338	0.767	0.778
Residual Std. Error	4.902 (df = 30)	2.909 (df = 29)	2.842 (df = 28)
F Statistic	16.860*** (df = 1; 30)	52.024*** (df = 2; 29)	37.149*** (df = 3; 28)
Notes:	***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.		

As the table allows observing, making value analyzes of the value associated with each estimator, it can be seen that the first models are specific in all their parameters. On the other hand, a weight of which the Number of Cylinders is not significant, this specification contains the highest correlation. By using the Analysis of Variances, you can tell that there is a better way to choose the best specification for the model.

Incidentally, the selection of the variable 'Horse grosspower' was strategically chosen since it retains a high correlation with other variables. Cars with more horsepower are generally associated with cars with higher cylinder numbers, and these in turn are heavier cars. In addition to the fact that the chosen variable, unlike others, does not maintain correlation with the type of transmission and this rules out a multicollinearity problem from the model.

ANOVA analysis of the models

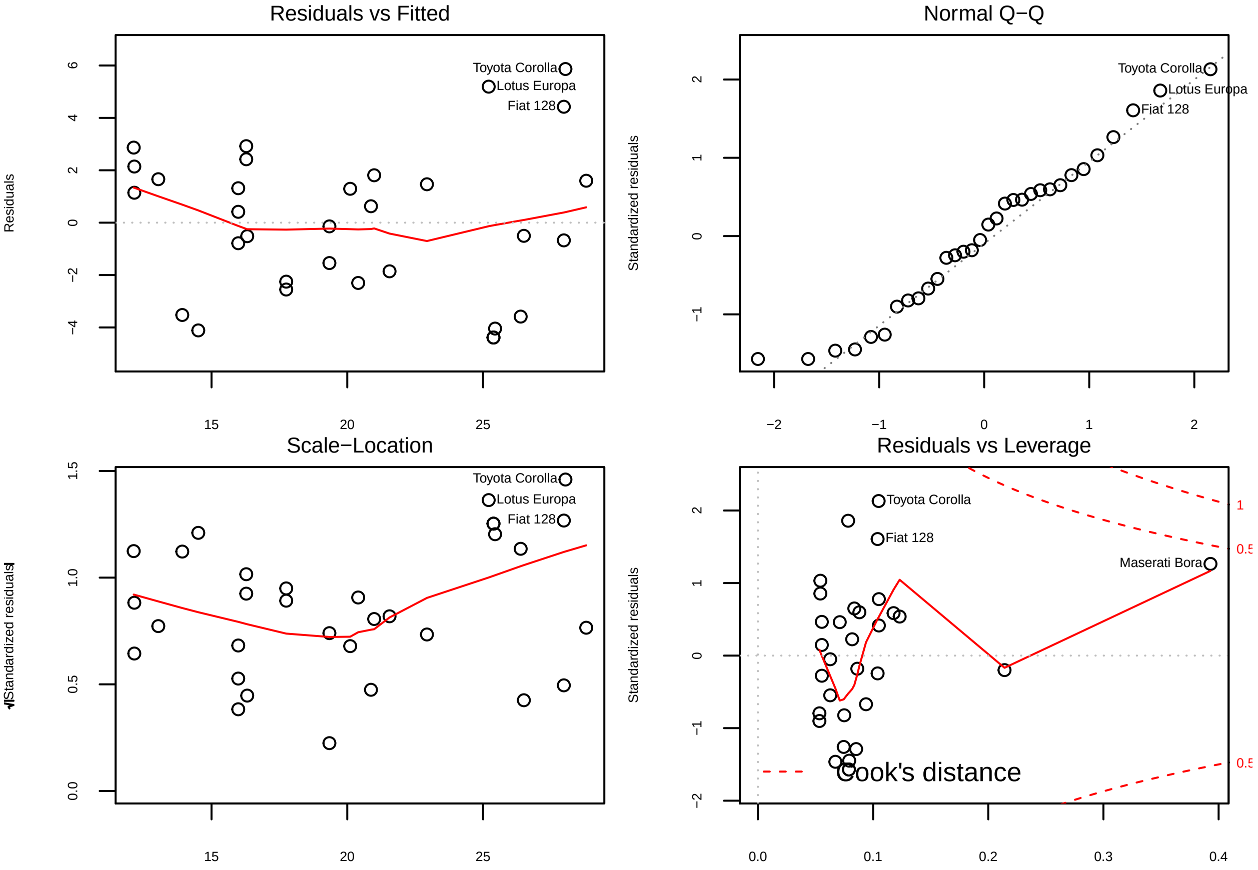
In the Analysis of Variances (ANOVA) an F test is performed, which is one-tailed. This means that to obtain statistical significance at 95%, it is necessary to obtain a p-value less than or equal to 0.05. If so, it is concluded that adding an extra variable to the model and inflating the variances of the regressors is justifiable.

The following table shows that of the elected models, the second is the one that best keeps the principle of parsimony. Since it does not really make it necessary to add more variables to the model and with only two variables we have an autocorrelation coefficient drawn.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.90				
29	245.44	1	475.46	58.88	0.00
28	226.10	1	19.34	2.39	0.13

Analysis of errors

The analysis of the errors shows that the errors in our electro model are randomly distributed and do not follow any pattern. The QQ plot shows the good fit that the model achieves. In addition to that the coefficient of Leverage does not show values that generate great concern, that is, there are no outliers that concentrate a great weight in the form of our regression.



Conclusions

It is possible to assume that the type of transmission has an impact on the fuel efficiency of a vehicle. This analysis generates a model with a really high level of fit and keeping the principle of parsimony in mind. Other models, taking into account other independent variables, may show a different effect from the type of transmission, but in all likelihood they will converge in that you opt for a manual car, allowing an improvement in the sense of efficiency.