Regressional Analysis of Car's Gas Mileage

Ngoc Lan Nguyen

2015-04-13

Executive Summary

In this study, we want to analyze the gas efficiency as related to some aspect of automobile design and performance. Specifically, based on the dataset *mtcars* that comes from the 1974 issue of the magazine Motor Trend, we want to answer the questions "Is an automatic or manual transmission better for the gas efficiency (mpg)?" and in either case, "quantify the mpg difference between these two modes". At first glance, these questions look trivial; however, some other factors should be taken into account to have a reliable response. Through this report, we did a regressional analysis of the gas efficiency versus its influential factors as the weight, the acceleration, the transmission type, etc. and found that the manual transmission has a better gas consumption than the automatic one.

1. Exploratory Analysis

Table 1 shows that *mtcars* is a data frame that comprises of 11 variables describing the fuel comsumption (mpg) and other physical attribute aspects for 32 models of vehicle from 1973-1974.

The table 2 gives the correlations between the different variables. From the common sense, one can think that the weight should greatly influence the fuel consumption. This suspection is somewhat confirmed by the high correlation between this variable *wt* and the *mpg* as seen in table 2, and thus the weight ought to be one of the keys for explaning the mileage performance. Some other variables as *cyl, disp, hp* and *drat* have also a high correlation with the outcome *mpg*, but they are also highly correlated with the weight and therefore, we suspect the existence of multicollinearities between those.

The variable *am* has a more moderate correlation with the *mpg* but, since our interest is related to the type of transmission, we have also to consider it as a key variable as well. A first plot for the *mpg* against this variable *am* can provide a prior visual cue about the response. The box plot from figure1 suggests that the automatic transmission could have a negative influence on the gas mileage when compared to the manual one.

2. Regression Model Analysis

2.1 Variables selection

We want to derive a parcimonious linear regression model (LRM) but still capable to explain the cause-effect relationship. The goal is therefore to identify those key variables to fit our LRM, and which can minimize the RSS (or equivalently maximize the R^2). Our approach consist to identify the unnecessary variables to drop, the base ones we have to include and those which need further tests due to uncertainty. For each step, we test the null hypothesis ($\beta_i = 0$) vs an alternative ($\beta_i \neq 0$) about the parameter's coefficients.

1. Ignored variables
   For this step, we want to develop a first intuition about the relationships between the outcome *mpg* and the other variables. To develop this intuition, we first fit the full model, ie. all the regressors are considered. Table 3 resumes the estimates for each variables in the dataset. Variables *cyl, drat, vs, gear* and *carb* have their estimate's p-values clearly too high (>0.5) to be significant. They can be ignored from our selection in the upcoming steps.

2. Base variables
   From our previous exploration, we choose to force the inclusion of the predictors weight *wt* and gear type *am*. This is our base model, fitbase. Note that the choice of *wt* is fully justified when we observed table 3 where the LRM only includes these base predictors: the corresponding p-values of the weight coefficient is significant and therefore, its inclusion in the model is necessary. Though we cannot say much about the transmission type at this time.

3. Tested variables
   The uncertain variables, ie. those that remain after removing the ignored and key variables, are *disp, hp* and *qsec*. We fit these later in addition to our base variables *wt* and *am*, to build the fittest model. The statistic tests performed on this model (see table 3) suggest to omit the variables *disp* and *hp* since their p-values are not significant. Only the estimate of *qsec* has a significant p-value.

4. Retained variables
   The previous elimination steps result in our final subset which is composed of the variables *wt, am* and *qsec*. The model explained by these regressors (fitbest) shows that they are all significant, as seen from table 3. We could go further to include the interaction terms between the selected variables but for this causal effects study, we prefer to keep the model simple for interpretability. A prediction model could need to include them thought.

5. Likelyhood ratio tests
   The summary for the likelyhood ratio test for the fitbase, fitbest, fittest and fitfull models, is provided in table 4. The p-value for the likelyhood ratio test comparing those models suggests that only the fitbest model is significant.

2.2 Model Analysis

1. residual plot diagnostics
   The residual plot for the fitbest model in figure 2 has not any particular pattern and the residuals seem to be distributed evenly around the mean of 0. Also, the standardized residuals of our model approximate the normality pretty good, as seen from the Q-Q plot.

2. parameter's confidence interval
   Table 5 presents the 95% CI for our parameter's estimates. None of them include the null value, which mean that we are 95% confident about their estimates.

3. Interpretations

With the selected model fitbest, we have: $mpg = \beta_{au} + \beta_{mn} am_{mn} + \beta_{wt} wt + \beta_{ac} qsec$

where:

- $\beta_{au} = 9.62$ is the average mpg for automatic transmissions when the weight and the acceleration are *"null"* (please don't jump out of your chair yet), which is just a hypothetic case. In reality, we should interpret this as a *reference* mpg which adds to the quantities generated from the weight and the acceleration.

- $\beta_{mn} = 2.94$ is the increase in mpg when the manual transmission is considered.

- $\beta_{wt} = -3.92$ is the decrease in mpg for each unit of weight ($\frac{1}{1000}$ lbs) increased.

- $\beta_{ac} = 1.23$ is the increase in mpg for each unit of acceleration ($\frac{1}{4}$ mile per sec) increased.

Thus, given the proposed model, we can state with 95% confidence that: for comparable weight and acceleration, automatic cars have less energetic efficiency over manual ones and their average mpg difference is about 2.94.

Appendices

1. Figures

figure 1: transmission performance

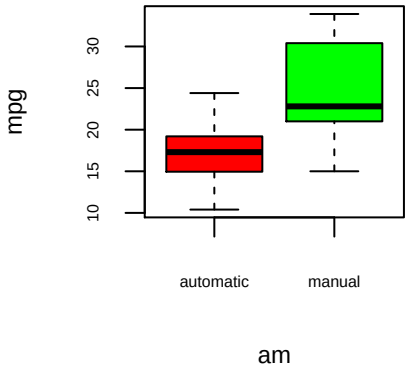
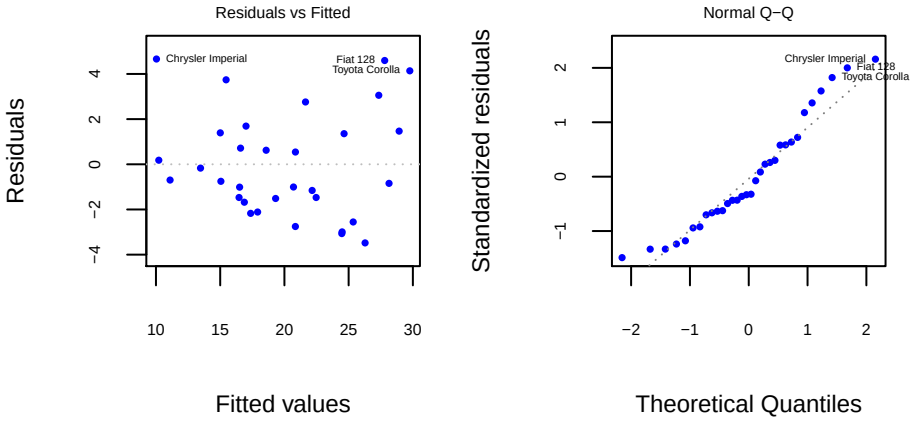


figure 2: residual diagnostics



2. Tables

Table 1: the mtcars dataset

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.900	2.620	16.460	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.900	2.875	17.020	0	1	4	4
Datsun 710	22.800	4	108	93	3.850	2.320	18.610	1	1	4	1

Table 2: correlations between variables

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.60	0.48	-0.55
cyl	-0.85	1.00	0.90	0.83	-0.70	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.90	1.00	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1.00	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.70	-0.71	-0.45	1.00	-0.71	0.09	0.44	0.71	0.70	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1.00	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1.00	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1.00	0.17	0.21	-0.57
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1.00	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1.00	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1.00

Table 3: regression results

	Dependent variable:			
	mpg			
	(1)	(2)	(3)	(4)
	fitfull	fitbase	fittest	fitbest
Constant	12.30	37.32	14.36	9.62
	p = 0.52	p = 0.00***	p = 0.16	p = 0.18
cyl	−0.11			
	p = 0.92			
disp	0.01		0.01	
	p = 0.47		p = 0.30	
hp	−0.02		−0.02	
	p = 0.34		p = 0.16	
drat	0.79			
	p = 0.64			
wt	−3.72	−5.35	−4.08	−3.92
	p = 0.07*	p = 0.0000***	p = 0.003***	p = 0.0000***
qsec	0.82		1.01	1.23
	p = 0.28		p = 0.05**	p = 0.0003***
vs	0.32			
	p = 0.89			
am	2.52	−0.02	3.47	2.94
	p = 0.24	p = 0.99	p = 0.03**	p = 0.05**
gear	0.66			
	p = 0.67			
carb	−0.20			
	p = 0.82			
Observations	32	32	32	32
R^2	0.87	0.75	0.86	0.85
Adjusted R^2	0.81	0.74	0.84	0.83
Residual Std. Error	2.65 (df = 21)	3.10 (df = 29)	2.43 (df = 26)	2.46 (df = 28)
F Statistic	13.93*** (df = 10; 21)	44.17*** (df = 2; 29)	32.96*** (df = 5; 26)	52.75*** (df = 3; 28)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: ANOVA for the (1)base, (2)best, (3)test and (4)full models

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
29	278.3197	NA	NA	NA	NA
28	169.2859	1	109.0338	15.5240	0.0007
26	153.4378	2	15.8481	1.1282	0.3425
21	147.4944	5	5.9434	0.1692	0.9711

Table 5: coefficient's 95% CI

	2.5 %	97.5 %
(Intercept)	-4.64	23.87
wt	-5.37	-2.46
am	0.05	5.83
qsec	0.63	1.82