Motor Trend Magazine

Cristóbal Alcázar

6 de agosto de 2016

Executive Summary

This report has the goal of build a linear regression model to study the miles per gallon on automobiles, the variable is typified as **mpg** in the **mtcars** dataset. Of particular interest is the relationship between **mpg** and the binary variable **am**, that indicate with a 0 if the car has automatic transmission and with a 1 the case of manual transmission. The model will help to answer the following two questions.

1. *"Is an automatic or manual transmission better for MPG"*
2. *"Quantify the MPG difference between automatic and manual transmissions"*

Building the model

So the first task to do is explore which variables can be useful to build a model, and a good starting point is to look the correlation of all pairs of variables that can be made with the response variable.

Table 1: mpg correlation pairs

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.6	0.48	-0.55

In the **table 1**, we can see that the 3 variables with the highest correlation with **mpg** are **wt** (*Weight*), **cyl** (*Number of cylinders*) and **disp** (*Displacement*). So, one important aspect when we build models with more than one variable, is the potential problem of multicollinearity, this it could be presented when the correlation between two or more independent variables is high.

Table 2: Independent variables correlation pairs

	am	wt	cyl	disp
am	1.00	-0.69	-0.52	-0.59
wt	-0.69	1.00	0.78	0.89
cyl	-0.52	0.78	1.00	0.90
disp	-0.59	0.89	0.90	1.00

As shown in the **table 2**, the correlation between the pairs (disp, wt) and $(\text{disp}, \text{cyl})$ are 0.89 and 0.90 respectively. So the **disp** variable has a warning flag, because if we incorporate it with the other two variables into a model, the information that **disp** add to the model is redundant. This information is useful to take in consideration later.

Now we proceed to fit models in an additive approach, this means formulate a simple model first, and then incorporate an extra variable until formulate a model with all the pre-selected variables. Then make an analysis of variance of all the models.

Table 3: Analysis of Variace between models

model	res.df	rss	df	sumsq	statistic	p.value
mpg ~ am	30	720.8966	NA	NA	NA	NA
mpg ~ am + wt	29	278.3197	1	442.5769020	62.9247094	0.0000000
mpg ~ am + wt + cyl	27	182.9683	2	95.3513637	6.7784342	0.0042734
mpg ~ am + wt + cyl + disp	26	182.8693	1	0.0990033	0.0140761	0.9064703

From the above result, **the most important to highlight is the high significance when we added the variables wt and cyl into the model. Therefore, the lowest p-value of the F-statistic in the model 2 and 3, translate into a significant reduction of the residual sum of squares (RSS)**. For the other hand, we confirm the unnecesary participation of the variable **displ** into the model because doesn't add relevant information and not contribute significantly in reduce the RSS. Finally the following is the fitted model:

$$\text{mpg} = \beta_0 + \beta_1 \text{am} + \beta_2 \text{wt} + \beta_3 \text{cyl}_6 + \beta_4 \text{cyl}_8 + \varepsilon$$

Residual Analysis

$$y = f(x) + \varepsilon$$

Once the model is adjusted is important to analyze the residuals as a kind of validation test of the adjustment. Before we fitted a function to explain the response variable (y), specifically, a linear function of a set of independent variables (x), and is not realistic to say that this function capture all the behavior of the response variable. Thus the unexplained portion of y is confined into an error term (ε). So the residuals of our model can be view as a representation of this term and is necessary to check some assumptions.

$$\varepsilon \approx N(0, \sigma^2)$$

The plot of the left in **figure 1** is a *quantile versus quantile plot* and is useful to verify if the residuals are normally distributed. Briefly speaking, if the residuals are normal, both quantiles (*sample and theorical*) come from the same distributions and the points form a straight line. As you can see, the plot is not a perfect straight line but it still conserving the shape, so the assumption is plausible.

Right plot is used to detect patterns, the residuals represent the unexplained portion of the response variables, this term should to be a random error term. The presence of some pattern in this plot means that the residuals contain important information about the response variable. As you can see in the chart, it is not possible to identify a pattern. So we can conclude that the model works well.

Interpreting the model

In this last part we focus on answering the questions that are formulate in the executive summary. The table 4 contain the most important information of the outcomes of the model, the corresponding estimate, standard error, statistic and p-value of each parameter.

$$E[Y = \text{mpg} | X = \text{am, wt, cyl}] = 33.75 + 0.15\text{am} - 3.15\text{wt} - 4.26\text{cyl}_6 - 6.08\text{cyl}_8$$

Is an automatic or manual transmission better for MPG?

The variable **am** is a binary variable that control the effect of the transmission class on the response variable **mpg**. So the sense of the word "better" on *miles per galon* is refered on efficiency, **what type of**

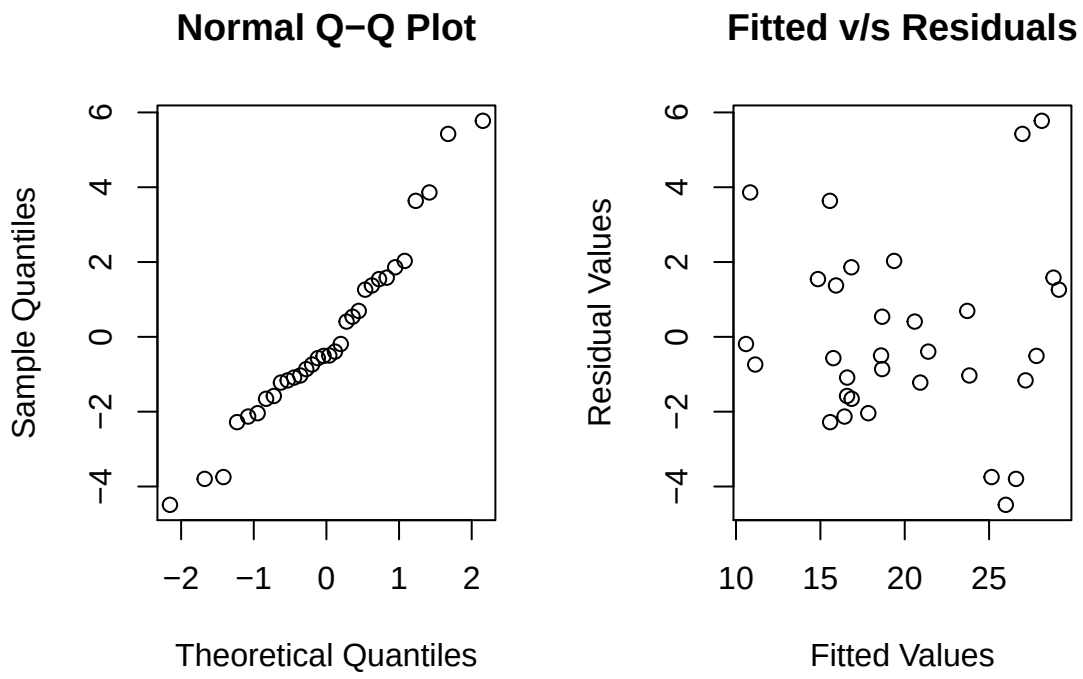

Figure 1: Residual Diagnosis

transmission gives more mpg? Linear regression model suggests that **an automobile with manual transmission ($am = 1$) is more efficiently on fuel consumption**, gives on average 0.1501 ($+/-1.3002$) more than an automobile with automatic transmission **ceteris paribus**.

Table 4: Final Model

term	estimate	std.error	statistic	p.value
(Intercept)	33.7536	2.8135	11.9971	0.0000
am	0.1501	1.3002	0.1154	0.9089
wt	-3.1496	0.9080	-3.4685	0.0018
factor(cyl)6	-4.2573	1.4112	-3.0167	0.0055
factor(cyl)8	-6.0791	1.6837	-3.6105	0.0012

Quantify the MPG difference between automatic and manual transmissions

An interesting observation is that the standard error of the coefficient has a higher absolute value than the estimate coefficient (0.1501 *v/s* 1.3002). Constructing a 95% confidence interval gives us a useful measure of a uncertainty range . Therefore with a 95% confidence interval we estimate that an automobile with manual transmission compared with an automobile with automatic transmission has a change on miles per galon of -2.518 to 2.818. **Considering the confidence interval, the linear regression model not assure a better class on fuel consumption efficiency given the transmission variable because the range contain the zero, that means that the coefficient estimate can take negative values and sometimes the manual transmission is worst than the automatic transmission. In conclusion, the manual transmission could be better or worst in terms of efficiency. It is necessary collect more data in order to reduce the standard error and obtain a more robust conclusion.**

Appendix

This section contain all the code used in the report.

Table 1

```
library(knitr); library(broom);
data(mtcars)
tabla <- cor(mtcars[c("mpg", "cyl", "disp", "hp", "drat", "wt", "qsec", "vs",
                      "am", "gear", "carb")])
tabla <- tabla[1,]
attr(tabla, "names") <- NULL
attr(tabla, "dim") <- c(1, 11)
colnames(tabla) <- c("mpg", "cyl", "disp", "hp", "drat", "wt", "qsec", "vs",
                     "am", "gear", "carb")
rownames(tabla) <- c("mpg")
kable(tabla, align = 'c', digits = round(2), caption = "mpg correlation pairs")
```

Table 2

```
tabla2 <- cor(mtcars[c("am", "wt", "cyl", "disp")])
kable(tabla2, align = 'c', digits = round(2), caption = "Independent variables
      correlation pairs")
```

Table 3

```
base <- lm(mpg ~ am, data = mtcars)
fit1 <- update(base, mpg ~ am + wt)
fit2 <- update(base, mpg ~ am + wt + factor(cyl))
fit3 <- update(base, mpg ~ am + wt + factor(cyl) + disp)
models <- anova(base, fit1, fit2, fit3)
models <-tidy(models)
models$model <- c("mpg ~ am", "mpg ~ am + wt",
                  "mpg ~ am + wt + cyl", "mpg ~ am + wt + cyl + disp")
models <- models[, c(7, 1:6)]
kable(models, align = 'c', caption = "Analysis of Variace between models")
```

Figure 1

```
model <- lm(mpg ~ am + wt + factor(cyl), data = mtcars)
par(mfrow = c(1,2))
qqnorm(resid(model))
plot(predict(model), resid(model), xlab = "Fitted Values",
     ylab = "Residual Values", main = "Fitted v/s Residuals")
```

Table 4 (*plus confidence interval calculation*)

```
coeff <- summary(model)$coefficients
# construct a 95% confidence interval
conf.interval <- coeff[2,1] + c(-1,1) * qt(0.975, df = model$df) * coeff[2,2]
fit <- tidy(model)
kable(fit, align = 'c', digits = round(4), caption = "Final Model")
```