Effect of the car transmission type on the Miles Per Gallon efficiency.

Overview

Our goal is to analyse the links between the variables of the *mtcars* data set. More specifically, we will try to see if the miles per gallon (*mpg*) efficiency can be predicted by the transmission type (*am*), and if such a link exists, we will try to quantify it.

High level view of the data set

Before trying to get details about specific interaction from the transmission type over the miles per gallon efficiency, we will start by having a high level view of all the variables and their overall correlation.

In order to do so, we will use the very nice `ggpairs` method from the `GGally` package (https://github.com/ggobi/ggally).

The variable we are particulary interested in are *am* - the transmission type - and *mpg* - the miles per gallon measure - so we will lok at these two ones if one first exploratory question in mind: *what other variables have high correlation values, and how to they qualify against am and mgp?*. We can see in this correlation plot that indeed the transmission type and the mpg seems correclated. But we also see that lots of other variables have quite high correlation values:

- *disp*, the displacement of the car (in cu.in) has a correlation value of -0.85 with *mpg*.
- *wt*, the weight of the car has a correlation value of -0.87.
- *cyl*, the number of cylinders of the car also seems to have a very strong correlation (the value is not displayed on this plot). See *Fig 1* in the Appendix for the complete exploratory plot.

Comparison between transmission type over miles per gallon efficiency

Let's try to answer the "Is an automatic or manual transmission better for MPG?" question. A first boxplot to see how *mpg* and *am* are related confirms that the mpg values really differs from. See *Fig 2* in the appendix. Let's try to get a better idea of this relationship by fitting a simple model with *am* as the predictor, and *mpg* as the outcome.

```
fit1 <- lm(data = mtcars, mpg ~ I(factor(am)))
kable(summary(fit1)$coef, digits = c(2,2,2,15))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.15	1.12	15.25	1.000000e-15
I(factor(am))1	7.24	1.76	4.11	2.850207e-04

These P-Values are extremely small and at first glance we culd reject the null hypothesis and think that the transmission type, *am*, would be a good predictor for the miles per gallon efficiency. In the next section we will try to give more details and quantify this model.

This also give us a first answer. With a high level of confidence (because of the nice p-values), the variation of intercept for a predicted *mpg* is of 7.24 miles per gallon, between the two transmission types *am*.

Quantification of the transmission type effect on the miles per gallon outcome

In order to evaluate our first model, we can start by checking the standardinzed residuals and display them in a QQ plot. *Fig 3* (see Appendix) is this Q-Q plot. So far, our model looks good enough.

(Intercept)	I(factor(am))1
Min. :-0.3392202	Min. :-0.4682836
1st Qu.:-0.0472788	1st Qu.:-0.1374647
Median : 0.0000000	Median :-0.0124676
Mean : 0.0000163	Mean : 0.0001677
3rd Qu.: 0.0353561	3rd Qu.: 0.0966102
Max. : 0.3666622	Max. : 0.4749345

As we can see in this *df betas* summary, all values are quite low. We could be quite confident that this model is not very influenced by some outlier.

Our first exploration shown us that other variables might be as good - if not better - predictors for the mpg outcome. We will fit this model and compare them against the previous single am model. For each model we will use a single predictor, and measure the confidence level, the p-value, and the residuals via the adjusted R^2. The closest the value of R^2 is to 1, the better the model fits the variance.

Coefficients	*Response*											
	am			cyl			gear			disp		
	std. Beta	*Conf. Int.*	*p-value*	*std. Beta*	*Conf. Int.*	*p-value*	*std. Beta*	*Conf. Int.*	*p-value*	*std. Beta*	*Conf. Int.*	*p-value*
(Intercept)			**<.001**									
am1	2.02	1.80 - 2.24	**<.001**									
cyl4						**<.001**						
cyl6				2.93	2.57 - 3.28	**<.001**						
cyl8				2.24	1.99 - 2.49	**<.001**						
gear3									**<.001**			
gear4							3.00	2.68 - 3.33	**<.001**			
gear5							2.62	2.11 - 3.12	**<.001**			
disp												**<.001**
Observations	32			32			32			32		
R^2 / adj. R^2	.949 / .945			.979 / .976			.954 / .949			.541 / .526		
F-statistics	277.184***			440.911***			201.506***			36.570***		

We can see that, although the *am* fitted model does a good job (nice p-value, 0.945 adjuted R^2), other model could perform as good, and even better. The cylinder fitted model in particular looks better.

Conclusion

We have seen that the transmssion type is a nice - but not the best - predictor for the mpg outcome. We also manage to quantify the mpg change when the transmission type changes, and we did some residuals measurements to get some confidence with our fitted models. That said, transmission type seems to be only one of the various predictors of the mpg outcome, and a better model could probably by found in another (longer than this 2 pages) analysis to compare the predictor combinations and end with better results.

Appendix

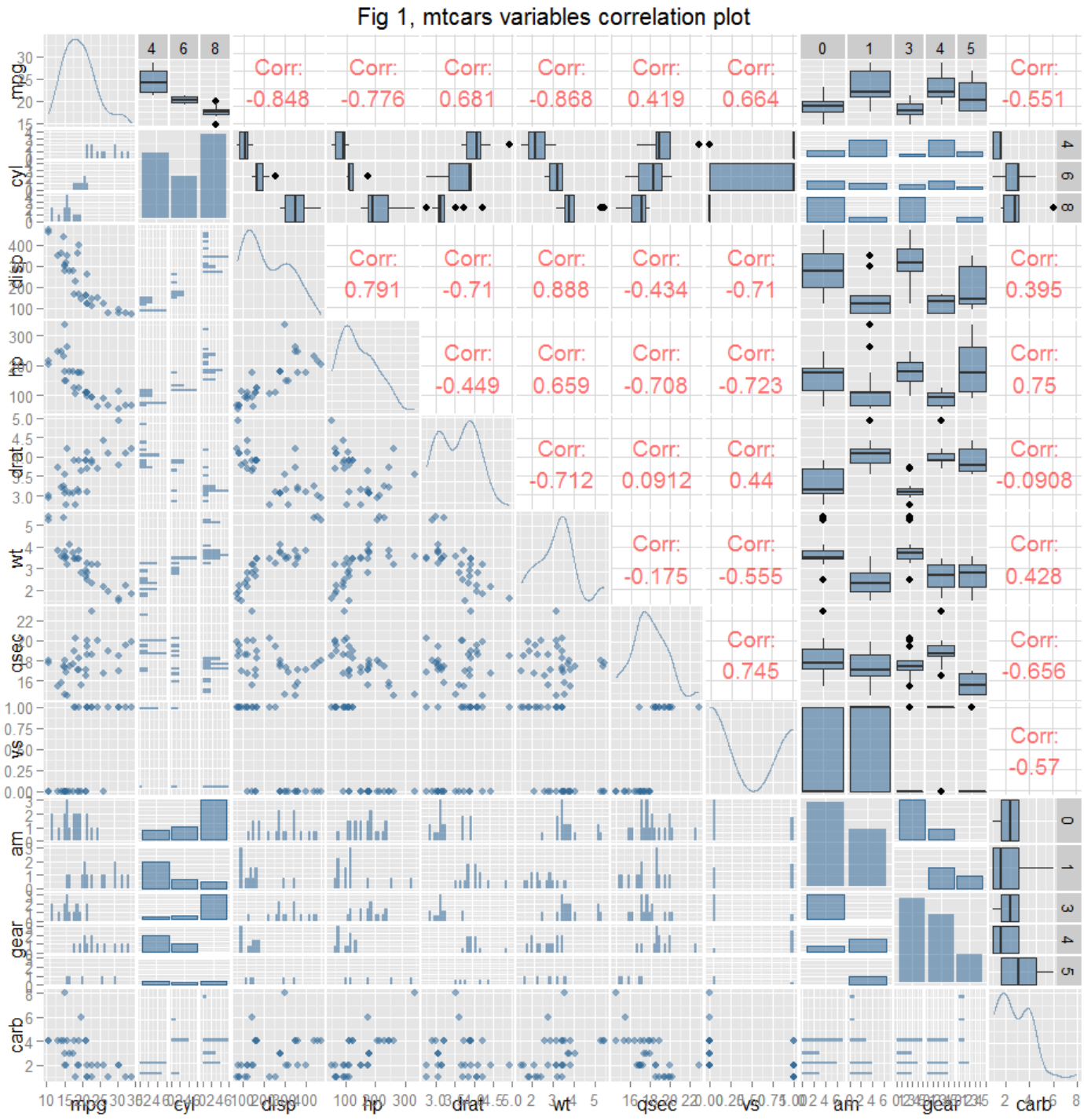

Fig 1, mtcars variables correlation plot

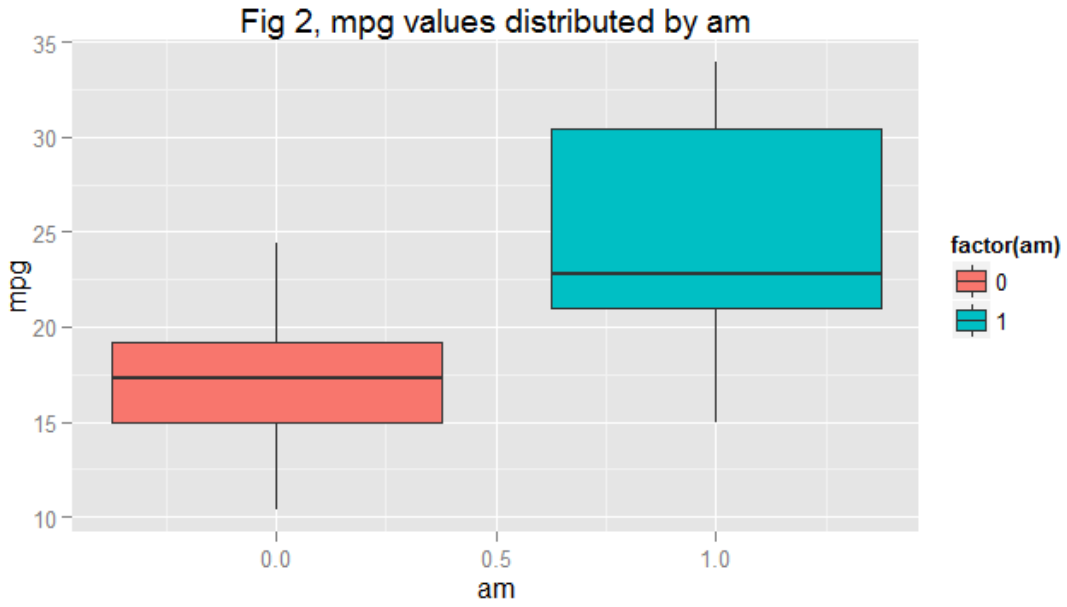

Fig 2, mpg distributed by am

Fig 3, QQ plot of the residuals for the linear regression mpg~am-1

R markdown exported to html to keep the `sjt.lm` model comparison table, then saved to HTML. All the sources of the R Markdown file are available on the dedicated Github repo (https://github.com/sportebois/coursera-reggressionModels-project/blob/master/mpg-analysis.Rmd)