Regression models course project

Kevin Payet

Thursday, September 18, 2014

Executive summary

This report explores the relationship between a car transmission type (automatic and manual transmission) and its average fuel consumption. We used the mtcars dataset that comprises the fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973 74 models).

The study showed that manual cars seem to consume less fuels. However, the effect is hard to quantify because of a high correlation between the different features.

Exploratory data analysis

The dataset gives several variables that are linked to the fuel consumption (mpg for miles per gallon). However, there is a pretty high correlation between most of these variables (see figure 1). Hence, as we will see later, it is hard to separate the effect of a single variable from the rest of the features.

Let's now try to evaluate the impact of the transmission type (am variable) on the mpg. Figure 2 shows a boxplot of mpg vs am. It is clear to the eye that the transmission type seems to have a non negligible influence on the fuel consumption. We need to assess this effect more precisely, which leads us to our first question.

Is an automatic or manual transmission better for MPG ?

I extract the mpg values into two vectors, for both transmission types. I then perform a t-test on these two samples to answer the question.

```
## [1] "P-value for the t-test 0.000686819166535514"
```

We tested the hypothesis H_α : "The mean of mpg for manual transmission is greater than for automatic transmission". We see that we can reject the null hypothesis (manual is not greater than automatic) with 99% confidence.

The answer to the first question is that **a manual transmission seems to be better for MPG**. In average, a manual car can drive **7.24494** more miles per gallon of fuel.

However, we have seen that the features in the dataset are quite correlated. To obtain a better estimate of the impact of the transmission type, we need to run a multivariate regression.

Quantify the MPG difference between automatic and manual transmissions

First, let us confirm that the am variable alone is not enough to explain the difference in mpg that we observed above.

```
summ <- summary(lm(mpg ~ am, data = mtcars))
summ$r.squared
```

```
## [1] 0.3598
```

The linear regression coefficients give us the same estimate than we found above; however, the value of $R^2 = 0.3598$ tells us that using the am variable alone can only explain around 35% of the variance. This means that we have to add other variables.

We can try simply using all the features at once `lm(mpg ~ .)`; however, the t-values and p-values obtained are clearly not significant, which means that we have to find another way to combine the variables.

This is a pretty long process, and given the lack of space for this report, I will directly give the model I finally used.

```
summ2 <- summary(lm(mpg ~ am*wt + wt:qsec, data = mtcars))
summ2$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  28.2691     2.6885  10.515 4.786e-11
## amManual     14.1990     3.5892   3.956 4.972e-04
## wt           -8.3286     1.4387  -5.789 3.696e-06
## amManual:wt  -4.2533     1.2494  -3.404 2.088e-03
## wt:qsec       0.2978     0.0838   3.554 1.421e-03
```

This model explains ~88% of the variance. We see that we had to introduce interactions between features, in particular between am and wt, which will make it harder to interpret the effect of the transmission alone. It would be possible to obtain a higher R^2, but for that we need to add other variables, and other interaction terms, that would make the interpretation in terms of am much harder. Since the goal of this section is to be able to evaluate quantitatively the overall effect of a change in transmission, I will keep this simpler model.

Figure 3 shows the mpg value as a function of the weight wt, for both type of transmissions. qsec is fixed to its average value in the dataset.

The red and black lines are automatic and manual transmissions respectively. We see that for low weights, the effect is what we expect, based on what we saw earlier: manual transmission implies higher mpg than automatic one. However, for higher weights, this order is reversed. It is hard to tell if this effect is real, and if it is, if the difference is that big. As said earlier, the model used is most probably incomplete. To study this effect in more details would require to push the study further, which goes beyond the goal of this project.

Figure 4 shows some diagnostic plots for the model used. We see that there is no

Conclusions

We have seen that the type of transmission can influence the average fuel consumption of a car. It seems that manual cars tend to use less fuel than automatic ones. However, the effect is hard to quantify, due to the high correlation between most variables, and the fact that the real model seems to include interactions betweens the various features.

Appendix

Figure 1: Correlation matrix of the data

```r
library(corrplot)

corrplot(abs(cor(mtcars2)))
```

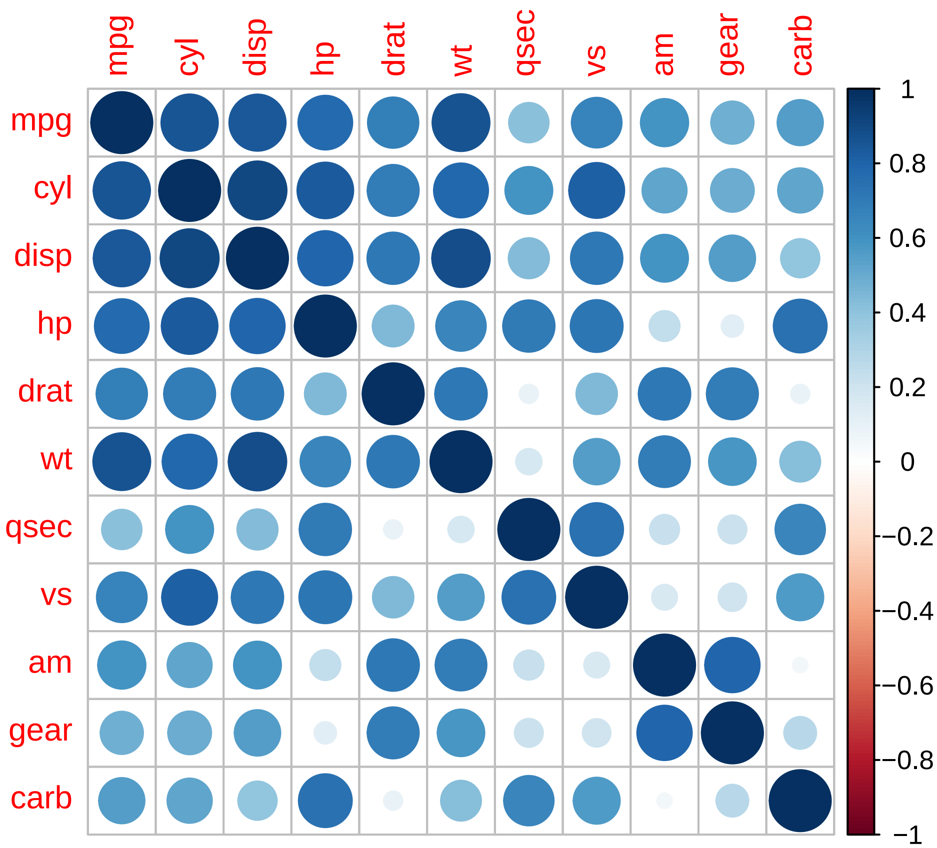

Figure 2: Difference in mpg between automatic and manual transmissions

```r
library(ggplot2)

g <- ggplot(data = mtcars, aes(x = am, y = mpg))
g <- g + theme_bw() + xlab("Transmission type") + ylab("# miles per gallon of fuel")
g <- g + geom_boxplot()
g
```

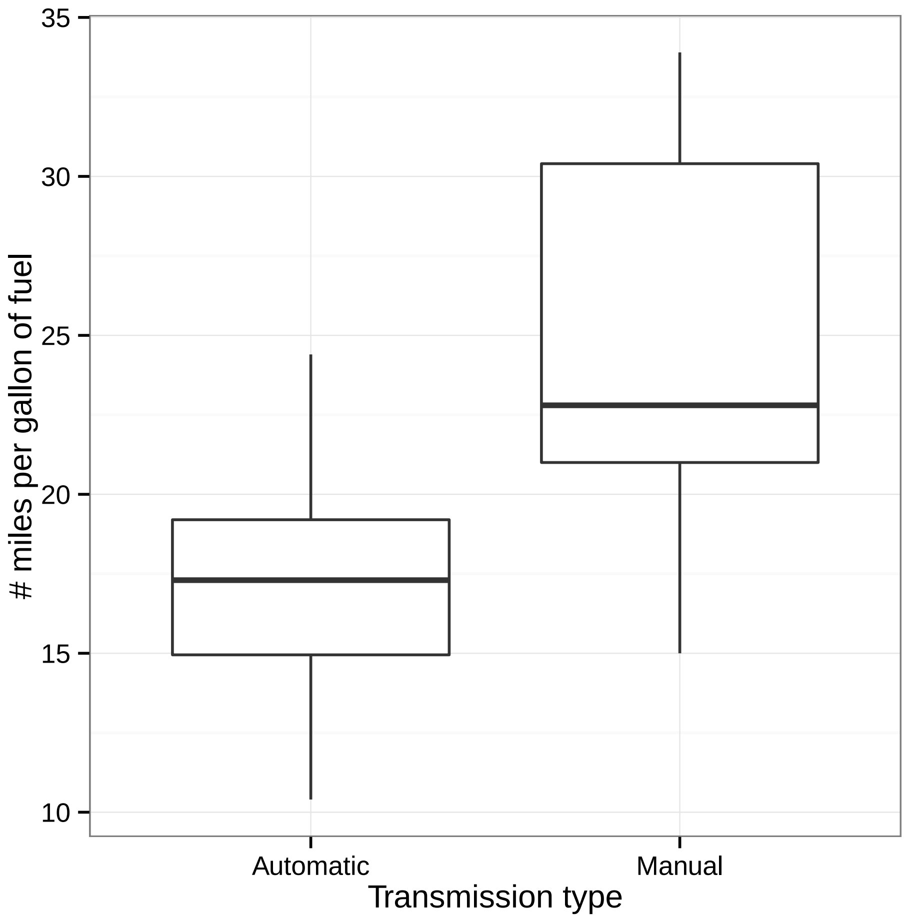

Figure 3: MPG vs Weight, for both transmissions

```
ggplot(data = mtcars, aes(x = wt, y = mpg)) + geom_point() +  geom_abline(intercept = 28.2691, slope =
```

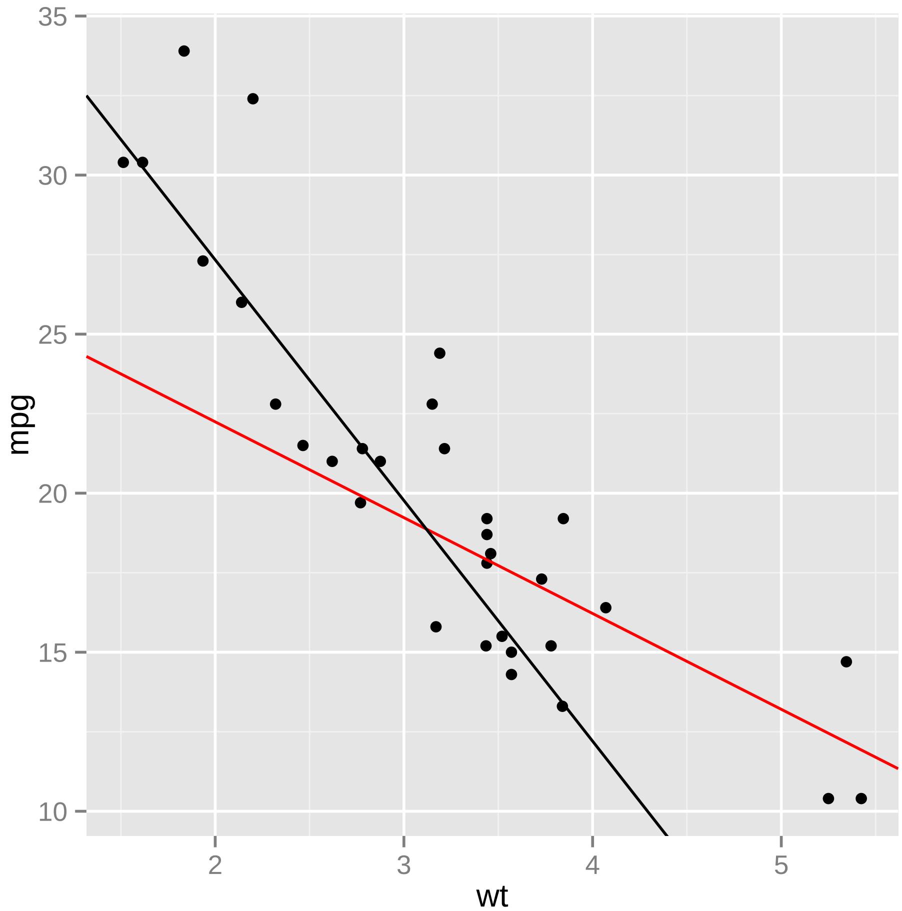

Figure 4: Diagnostic plots for the final model

```
fit <- lm(mpg ~ am*wt + wt:qsec, data = mtcars)
par(mfrow = c(2,2))
plot(fit)
```