Motor trend: automatic or manual transmission?

José Antonio García Ramirez

January 10, 2017

Summary

With the data available from our Motor Trend magazine, the question of whether the performance of cars is improved by using manual or automatic transmission is assessed. First, a test of means is evaluated, indicating that the difference is notorious, then a multiple linear model is used, showing again the difference between the use of transmissions and the effect of the choice against the performance of miles per gallon

Data

Data from 32 cars of the models 1973 and 1974 have been collected and stored in a package of R 1, and the following variables are recorded:

- *mpg* Miles/(US) gallon
- *cyl* Number of cylinders
- *disp* Displacement
- *hp* Gross horsepower
- *drat* Rear axie ratio
- *wt* Weight (1000 lbs)
- *qsec* 1/4 mile time
- *vs* V/S
- *am* Transmission (0 = automatic, 1 = manual)
- *gear* Number of forward gears
- *carb* Number of carburetors

Table 1 shows the descriptive statistics of the variables, after coding some of the variables as categorical. In particular, we are interested in the relation of the performance by gasoline, since in fact this in Mexico has just had a considerable increase 2, which is reflected in the variable mpg with respect to the type of transmission of the cars.

```
cars <- mtcars
cars <- transform(cars, cyl = factor(cyl), am = factor(am)  )
```

Introduction

For this reason, we performed a Student t test (since we know neither the variance nor the population mean) with the following hypothesis:

- H_0 The performance of automobiles (mpg) with automatic transmission is on average equal to the average performance of cars with manual transmission

```
automatic <- subset(cars, am == 0, select = mpg )
manual <- subset(cars, am == 1, select = mpg )
student.test <- t.test(automatic, manual)
```

So, with a confidence of 95% we ** reject** the previous hypothesis, since the p- value of the test is less than 0.05, in fact it is 0.0013736 and the means are 17.1473684, 24.3923077 for the groups automatic and manual respectively, so we proceed to fit a multiple regression model which relates mpg to the other features.

In view of the fact that automobiles have a higher performance of miles per gallon of gasoline, *mpg* , when they have the manual transmission, we will try to predict performance depending on the other variables for which information is available.

As the variable *mpg* has a high linear correlation with the other variables, as can be seen in figure 2, a first proposed multiple linear model contemplates all variables, considering *cyl* and *am* as categorical, The confidence intervals for each of the parameters (in this saturated model) are as follows:

```
model.Saturate <- lm(mpg ~ ., data = cars)
kable(confint(model.Saturate))
```

	2.5 %	97.5 %
(Intercept)	-16.1939252	51.8336117
cyl6	-6.3793748	3.0587613
cyl8	-7.3650246	10.6399042
disp	-0.0223870	0.0502118
hp	-0.1027001	0.0104433
drat	-3.4707456	3.5234462
wt	-7.6582776	0.0457824
qsec	-0.8590047	2.1529189
vs	-2.9933241	6.4880979
am1	-1.5645684	6.7990993
gear	-2.2745524	3.8026107
carb	-1.4565480	2.4752503

Since no interval, with 95% confidence, does not go through zero, we proceed to discard from the model the variables that present less Akaike information, and the new model estimade has three parameters, α_{wt}, α_{qsec} and α_{am1} whose confidence intervals do not pass through zero, and their values are significant because the three values of the parameters pass the test T of significance ($p - value$<0.05) the constant term is decided to include to maintain a referent when the other variables are zero.:

```
model.Baseline <- step(model.Saturate, trace = FALSE)
library(pander)
panderOptions("digits", 2)
pander(summary(model.Baseline))
```

	Estimate	Std. Error	t value	Pr(>|t|)
wt	-3.9	0.71	-5.5	7e-06
qsec	1.2	0.29	4.2	0.00022
am1	2.9	1.4	2.1	0.047
(Intercept)	9.6	7	1.4	0.18

Table 3: Fitting linear model: mpg ~ wt + qsec + am

Observations	Residual Std. Error	R^2	Adjusted R^2
32	2.5	0.85	0.83

Conclusion

To diagnose the robustness of the model, the non-correlation of the residuals with our variables features and *mpg* was verified (both are small below to .4, see figure 3), also a Kolmogorov-Smirnov test was performed on the residues to contrast the hypothesis that the residues follow a normal distribution (see table 2).

In conclusion with the adjusted model, because the coefficient with the variable am1 *which refers to the manual automobile transimision (due to encoding of factor type used) has a value of 9.6 we can say that if the other variables are left fixed An automobile will have an average of 9.6 miles per gallon higher than cars that do not have manual transmission, this estimate in the difference between the original means is due in part to the information provided by the variables* wt *and* qsec *to the model and both are Part of two different components as mentioned in the legend of figure 2.*

Adjuncts

```
kable(summary(cars))
```

mpg	cyl	disp	hp	drat	wt	qsec	vs
Min. :10.40	4:11	Min. : 71.1	Min. : 52.0	Min. :2.760	Min. :1.513	Min. :14.50	Min. :0.0000
1st Qu.:15.43	6: 7	1st Qu.:120.8	1st Qu.: 96.5	1st Qu.:3.080	1st Qu.:2.581	1st Qu.:16.89	1st Qu.:0.000
Median :19.20	8:14	Median :196.3	Median :123.0	Median :3.695	Median :3.325	Median :17.71	Median :0.000
Mean :20.09	NA	Mean :230.7	Mean :146.7	Mean :3.597	Mean :3.217	Mean :17.85	Mean :0.4375
3rd Qu.:22.80	NA	3rd Qu.:326.0	3rd Qu.:180.0	3rd Qu.:3.920	3rd Qu.:3.610	3rd Qu.:18.90	3rd Qu.:1.000
Max. :33.90	NA	Max. :472.0	Max. :335.0	Max. :4.930	Max. :5.424	Max. :22.90	Max. :1.0000

Table 1: Descriptive statistics of the variables. Note

```
cars2 <- mtcars
library(ggcorrplot)
```

```
## Loading required package: ggplot2
```

```
r <- cor(mtcars)
ggcorrplot(cor(mtcars),, title = "Correlations between 11 initial variables",
           colors = c("deeppink", "white", "deepskyblue"),  type = "lower",
           outline.col = "white")
```

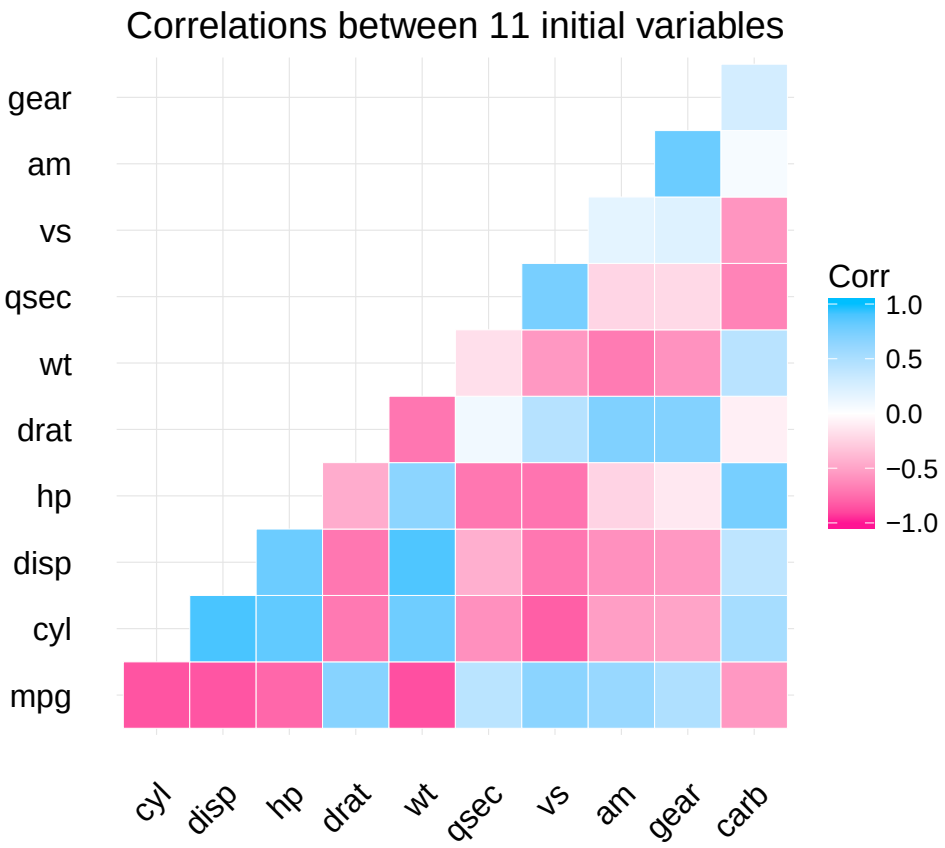

Figure 2: Correlation between the variables of the data set, note that the variable *wt* is positively correlated with *hp*, *disp*, *cyl* and *carb* which reflect directly the weight of the car, this same set of variables $ {wt, hp, Disp. , Cyl, carb}$ negatively correlate with the *mpg* variable while the other variables correlate positively with respect to *mpg*.

```
library(RColorBrewer)
r <- data.frame(res = model.Baseline$residuals, mpg = cars$mpg, qsec= cars$qsec)
r1 <- round(cor(r$res, r$mpg), 2)
r2 <- round(cor(r$res, r$qsec), 2)
g1 <- ggplot(r, aes(res, mpg)) + geom_point(aes(
    colour = rep(brewer.pal(8, "Set2"),4))) + xlab('residuals') + ylab('mpg') +
    theme_bw() +  ggtitle('Residuals vs mpg') + #xlim(c(0,20))+
    theme(legend.position = "none")
g2 <- ggplot(r, aes(res, qsec)) + geom_point(aes(
    colour = rep(brewer.pal(8, "Set2"),4))) + xlab('residuals') + ylab('qsec') +
    theme_bw() +  ggtitle('Residuals vs qsec') + #xlim(c(0,20))+
    theme(legend.position = "none")
multiplot(g1, g2, cols = 2)
```

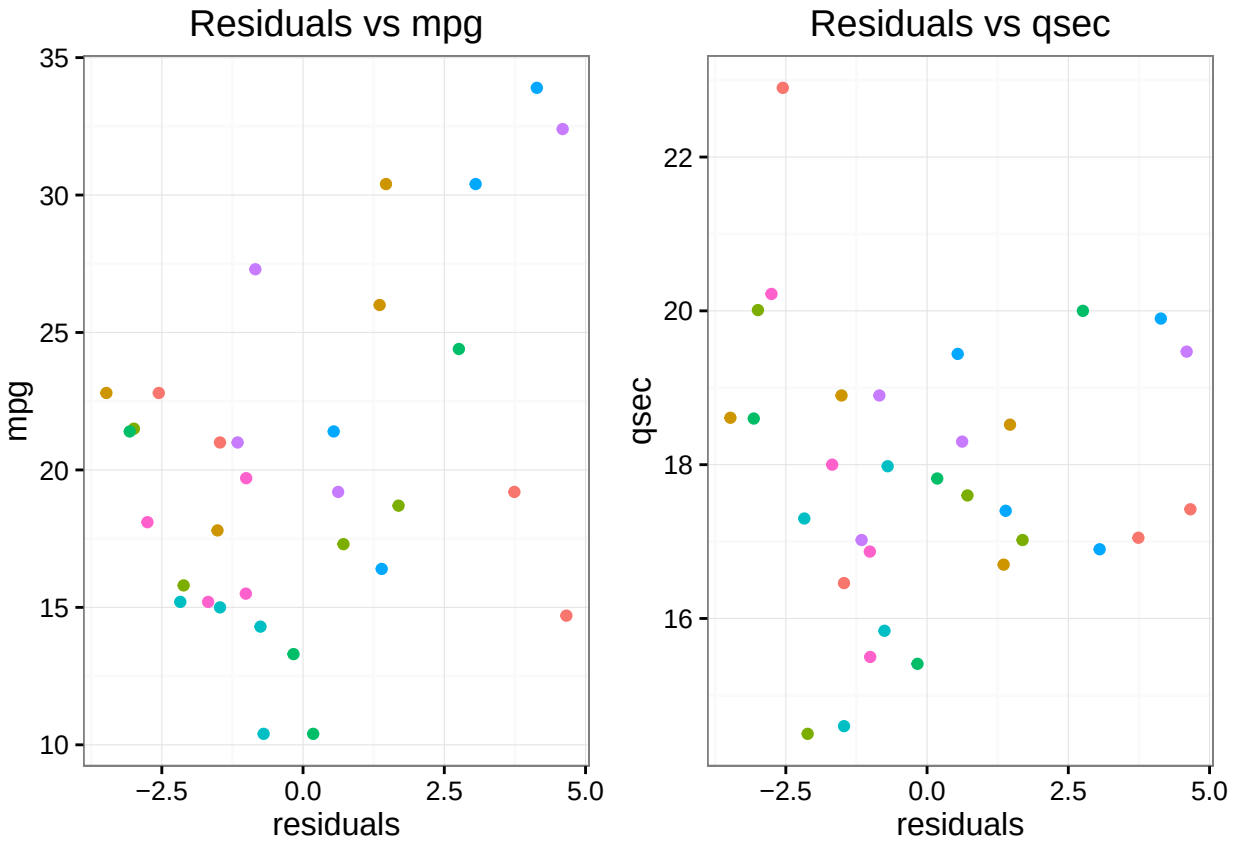

Figure 3: Correlation of the residuals with the variable to be predicted *mpg* (right: correlation = 0.39). Correlation of residues with the variable *qsec* (left: correlation = 0)

```
meanks <-  round(mean(model.Baseline$residuals), 2)
sdks <- round(sd(model.Baseline$residuals), 2)
pander(ks.test(model.Baseline$residuals, "pnorm", meanks, sdks))
```

Table 5: One-sample Kolmogorov-Smirnov test: `model.Baseline$residuals`

Test statistic	P value	Alternative hypothesis
0.15	0.44	two-sided

Table 2: Kolmogorov-Smirnov test on the residuals of the model with three parameters. The proof that residuals is distributed as a normal with average 0 (another important assuption) and 2.34 sd is accepted.