Motor Trend

MATIAS, V. A.

30/03/2021

Summary

This relatory is the final project of the Regression Models Course offered by Johns Hopkings University in the Coursera Plataform. The objective of the paper is review a collection of cars and explore the relationship that explains in a quantitative way the miles per gallon spent. To do this, will be used strategies of Exploratory Data Analysis and Regression Models.

Trends: MPG and Transmission

The data set used can be retrieved in the base R software by the command `data("mtcars")`. It have 32 observations with 11 numeric attributes for each (more information about the dataset and the variables can be found with the command `?mtcars`). This paper analyzes the available data to understand if automatic transmissions are better than manual for mpg and quantify the difference. The answers are retrieved in the Conclusions section.

Simple Linear Regression

Some exploratory analysis about the transmissions shows that automatic transmissions has lower values (apendix, figure 1). To see if this relation follows a pattern, we'll do a simple linear regression with mpg (Miles/US gallon) as outcome and am (transmission) as explanatory variable.

```
fit <- lm(mpg ~ factor(am), data = mtcars)
summary(fit)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

```
message("R Squared:", summary(fit)$r.squared, "\nAdjusted R Squared:",
        summary(fit)$adj.r.squared)
```

```
## R Squared:0.359798943425465
## Adjusted R Squared:0.338458908206314
```

Is expected that the values of mpg for automatic transmission be around 17.147 ± 1.125 and the cost to change the transmission for manual is equal to 7.245, getting around 24.392 ± 1.764 for manual transmission. The result of t-test results in a P-value very low for transmissions, what is good. The results of Multiple R^2 and Adjusted R^2, says that the linear model can explain round of one third of the variation in the data, so *am* alone brings some information about the mpg, but is better add more variables.

Multivariable Linear Regression

First, we need to find the most important variables to put on a new model. This can be done using variables with significant correlation with mpg, but well uncorrelated between itselfs. The correlation table can draw some insights to understand the pattern (apendix, figure 2), else can be done a model with all variables and, with him, extract that with lower p-values on t-test.

```
coefs <- summary(lm(mpg ~., mtcars))$coefficients
coefs[coefs[,4] < 0.35,4]
```

```
##         hp         wt       qsec         am
## 0.33495531 0.06325215 0.27394127 0.23398971
```

As we can see, the most influencible variables (lower p-values) are wt (Weigth in 1000 lbs), qsec (1/4 mile time) and hp (Gross horsepower), plus am. Now we'll build one model with these 4 variables and another without hp. Then is used a ANOVA to find the model with lower p-value for F-test and get his results.

```
mult_fit <- lm(mpg ~ factor(am) + wt + qsec, data = mtcars)
mult_fit_hp <- lm(mpg ~ factor(am) + wt + qsec + hp, data = mtcars)

p_values <- anova(fit, mult_fit, mult_fit_hp)$`Pr(>F)`[2:3]
message("P-Value without hp(", p_values[1], ")",
        "\nis lower than the P-Value with hp(", p_values[2], ")")
```

```
## P-Value without hp(1.78745027518802e-09)
## is lower than the P-Value with hp(0.223087931975388)
```

```
summary(mult_fit)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## factor(am)1  2.935837  1.4109045  2.080819 4.671551e-02
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
```

```
message("R Squared:", summary(mult_fit)$r.squared, "\nAdjusted R Squared:",
        summary(mult_fit)$adj.r.squared)
```

```
## R Squared:0.849663556361707
## Adjusted R Squared:0.833556080257604
```

The results say that qsec and wt were added well in the model, with low p-values. The R-squared and adjusted R-square grows to aprox. 0.85 and 0.83, respectively.

About the residuals (apendix, figure 3), they showed no relevant problems. Compared with the fitted values, they had well distribution whitout relevant pattern. They yet tended to a linear tendence in normal QQ plot (a slow senoidal pattern can be detected, but is not something so anormal). The other residual plots show not a tendence too, being spaced in the plots.

Conclusions

As can be derived by the analysis of the boxplot of the data and by the pattern **with other variables** showed before, automatic transmission tends to be better than manual transmissions for mpg. Analyzing the trend of mpg just with the fact to be manual or automatic don't shows an information with good confidence.

So, mpg can be calculated as: mpg = 9.6175 + 2.9358∗am - 3.9165∗wt + 1.2259∗qsec, where am = 0 for automatic transmission and 1 for manual transmission. The trend of the chosen model is that automatic tranmission starts with 9.6178 of mpg and increases/decreases with the values of wt and qsec. Manual transmissions, in other side, have a start increased in 2.9358 mpg, so, starts with 12.5536 and change the value by the values of qsec and wt.

Apendix

```r
library(ggplot2)
library(dplyr)

mtcars$am <- if_else(mtcars$am == 0, true = "auto", false = "manual")

ggplot(mtcars) +
  geom_boxplot(aes(x = am, y = mpg, group = am)) +
  labs(x = "Transmission",
       y = "Milles per gallon (mpg)")
```

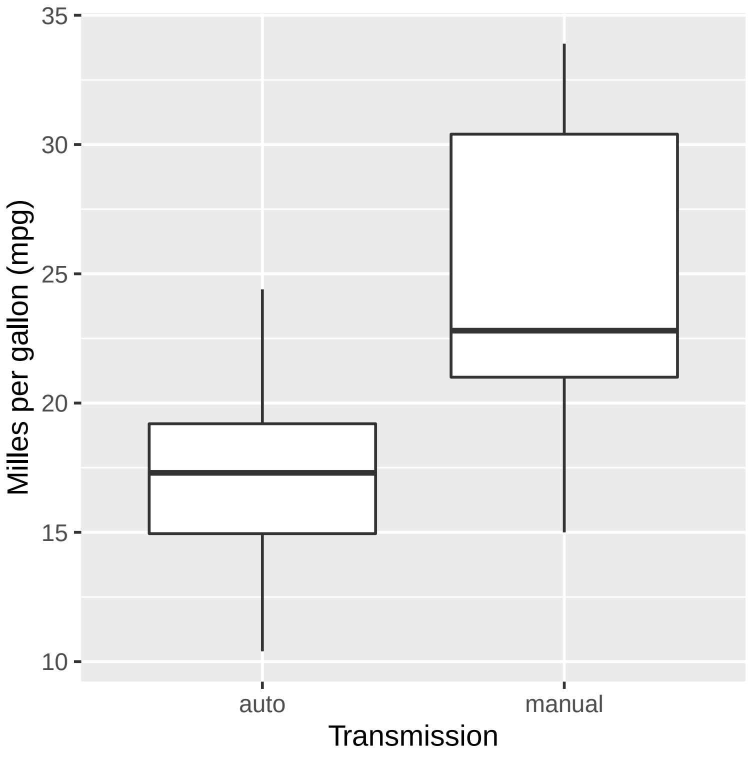
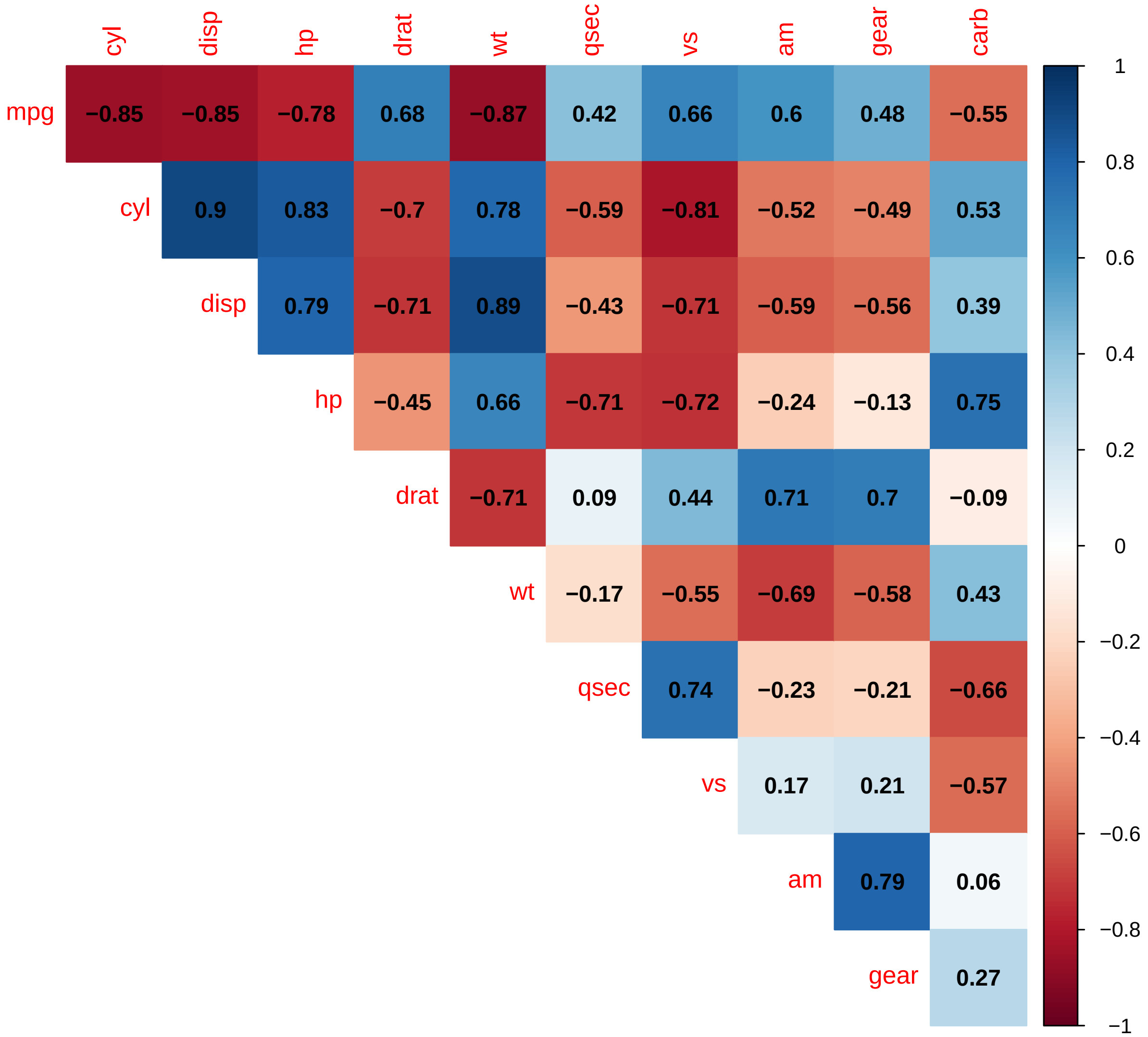


Figure 1: mpg X Transmission Boxplot

```
library(corrplot)
data("mtcars")

corrplot(cor(mtcars),
         method = "color",
         type = "upper",
         addCoef.col = TRUE,
         diag = FALSE,
         number.cex = 0.9)
```



	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	−0.85	−0.85	−0.78	0.68	−0.87	0.42	0.66	0.6	0.48	−0.55
cyl		0.9	0.83	−0.7	0.78	−0.59	−0.81	−0.52	−0.49	0.53
disp			0.79	−0.71	0.89	−0.43	−0.71	−0.59	−0.56	0.39
hp				−0.45	0.66	−0.71	−0.72	−0.24	−0.13	0.75
drat					−0.71	0.09	0.44	0.71	0.7	−0.09
wt						−0.17	−0.55	−0.69	−0.58	0.43
qsec							0.74	−0.23	−0.21	−0.66
vs								0.17	0.21	−0.57
am									0.79	0.06
gear										0.27

Figure 2: Correlation Table

```
par(mfrow = c(2,2))
plot(mult_fit)
```

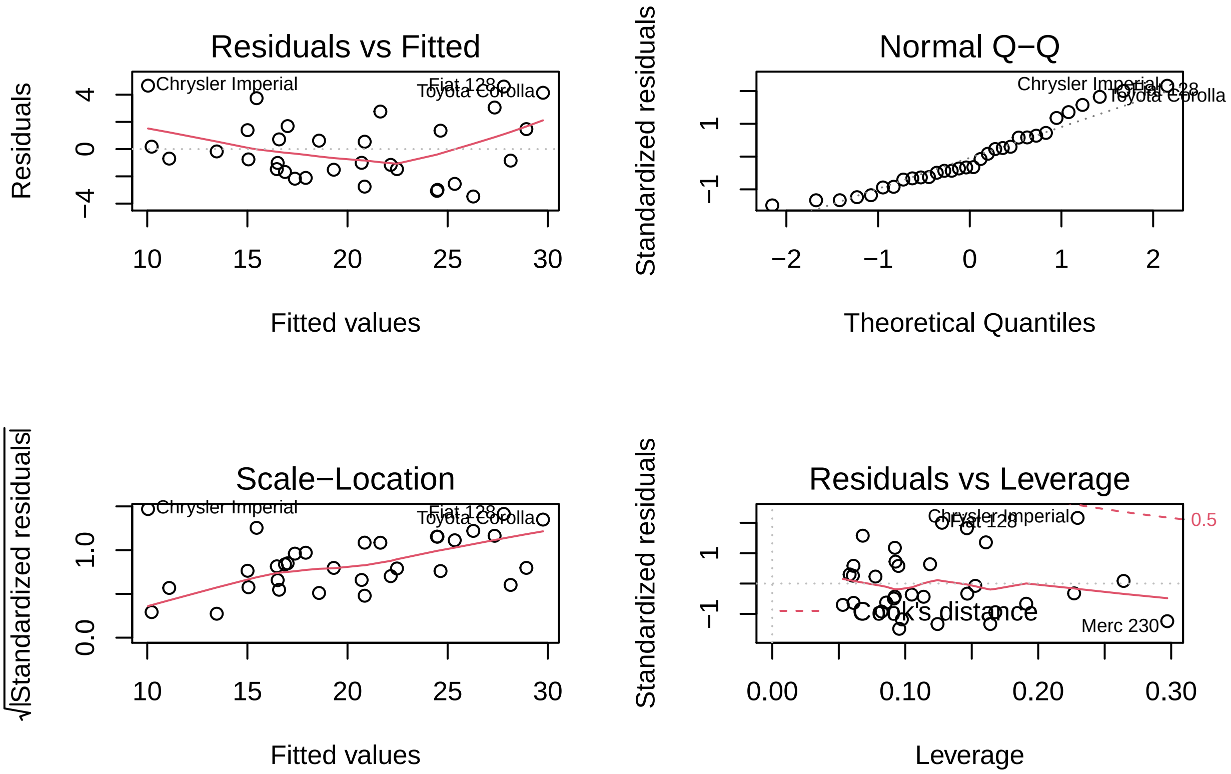


Figure 3: Residuals plots (multivariable model)