Motor Trend

MATIAS, V. A.

30/03/2021

Summary

This report is the final project of the Regression Models Course offered by Johns Hopkins University in the Coursera Plataform. The paper's objective is to review a collection of cars and explore the relationship that explains quantitatively 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 has 32 observations with 11 numeric attributes for each. This paper analyzes the available data to understand if automatic transmissions are better than the manual for mpg and quantify the difference. The Conclusions section shows the answers. To know more about the data set and the variables use the command `?mtcars`.

Simple Linear Regression

Some exploratory analysis about the transmissions shows that automatic transmissions have lower mpg values (appendix, figure 1). To see if this relation follows a pattern, we'll do a simple linear regression with mpg (Miles/US gallon) as the outcome and am (transmission) as the 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 $\pm$ 1.125 and the cost to change the transmission for manual is equal to 7.245, getting around 24.392 $\pm$ 1.764 for manual transmission. The result of the t-test results in a P-value very low for transmission, which is good. The results of Multiple R^2 and Adjusted R^2, says that the linear model can explain around of one-third of the variation in the data, so *am* alone brings some information about the mpg, but it's better to add more variables.

Multivariable Linear Regression

First, we need to find the main variables to put on a new model. One way is using variables with significant correlation with mpg, but well uncorrelated between themselves. The correlation table can draw some insights to understand the pattern (appendix, figure 2), else can be done a model with all variables and, with him, extract that with lower p-values on the 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 influenceable variables (lower p-values) are wt (Weight in 1000 lbs), qsec (1/4 mile time) and hp (Gross horsepower), plus am. Now we'll build one model with these four variables and another without hp. Then, the ANOVA is used to find the model with a 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 approx. 0.85 and 0.83, respectively.

The residuals (appendix, figure 3) showed no relevant problems and no relevant pattern compared with the fitted values. They yet tended to a linear tendency in normal QQ plot, although a small sinusoidal pattern can be detected, it isn't anything impactful. The other residual plots show did not show an abnormal trend, having expected behavior.

Conclusions

As can be derived by the boxplot analysis of the data and the pattern **with other variables** showed before, automatic transmission tends to be better than manual transmissions for mpg. Analyzing the mpg trend just with the fact to be manual or automatic don't show information with significant 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. Following the trend of the chosen model, the automatic transmission starts with 9.6178 for mpg and increases/decreases with wt and qsec values. Manual transmissions, on another side, have a start increased in 2.9358 mpg, so, starts with 12.5536 and changes the value by the values of qsec and wt.

Appendix

```
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)")
```

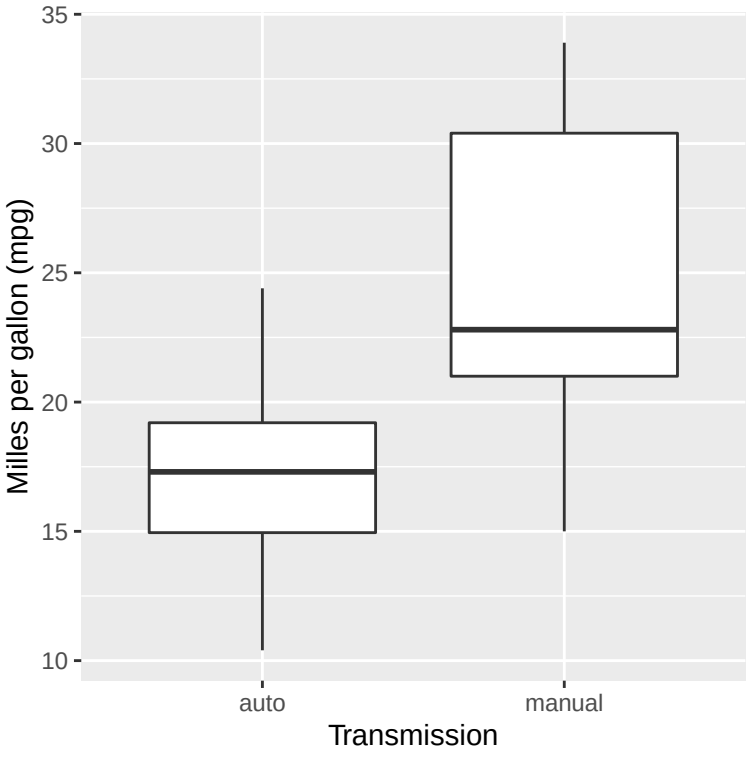

Figure 1: mpg X Transmission Boxplot

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

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

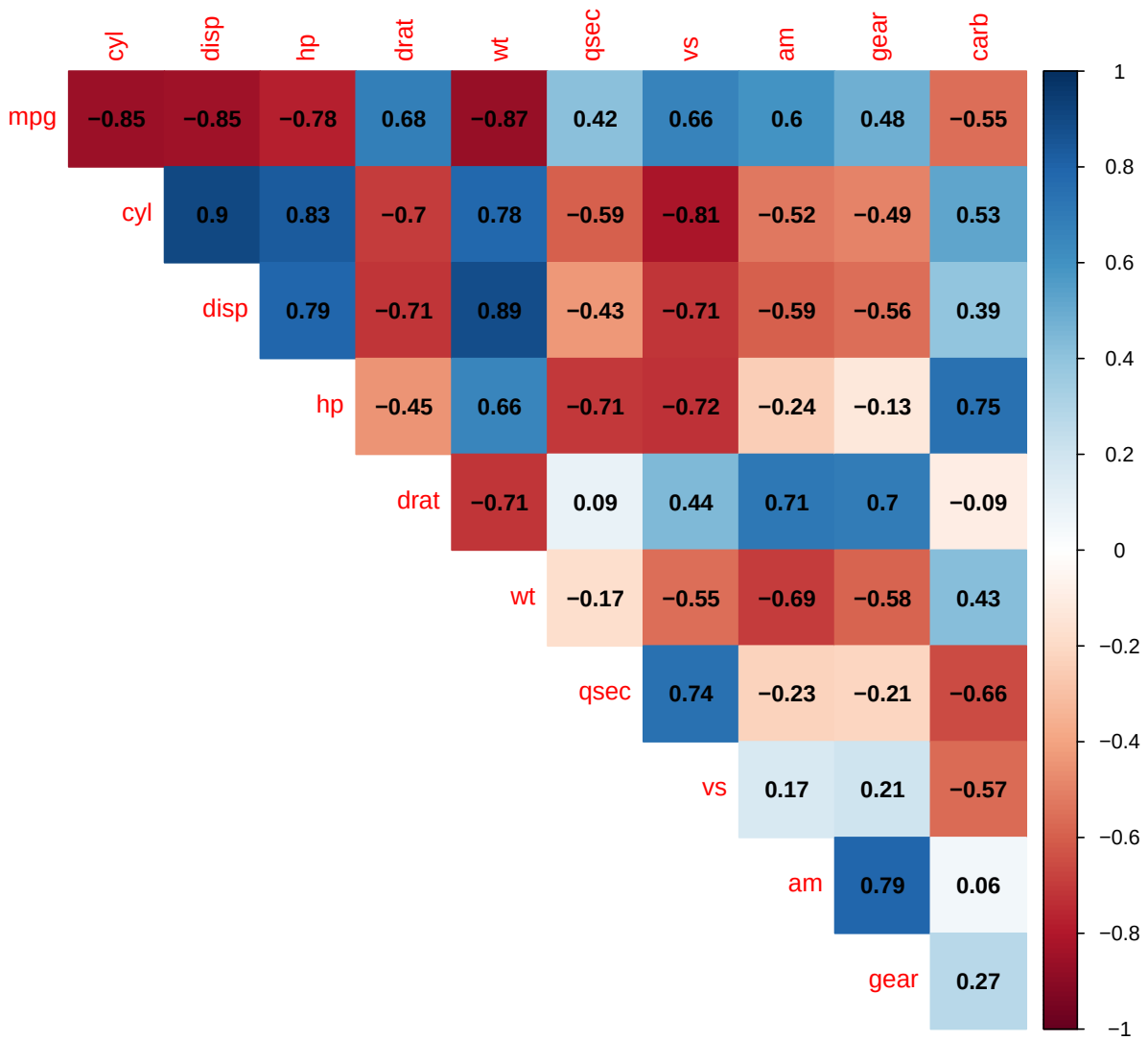

Figure 2: Correlation Table

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

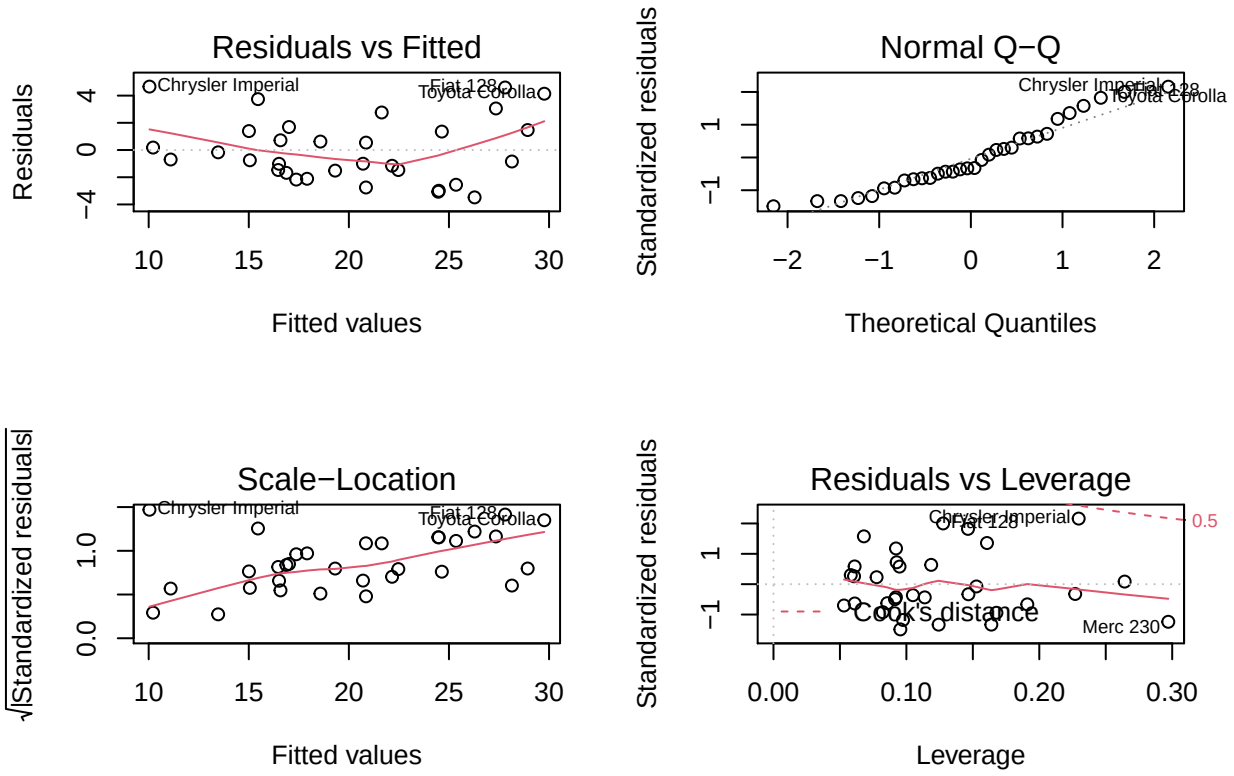

Figure 3: Residuals plots (multivariable model)