

# 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.

## Data

The data set used can be retrieved in the base R software by the lines below:

```
library(dplyr)

data("mtcars")
mtcars <- as_tibble(mtcars)

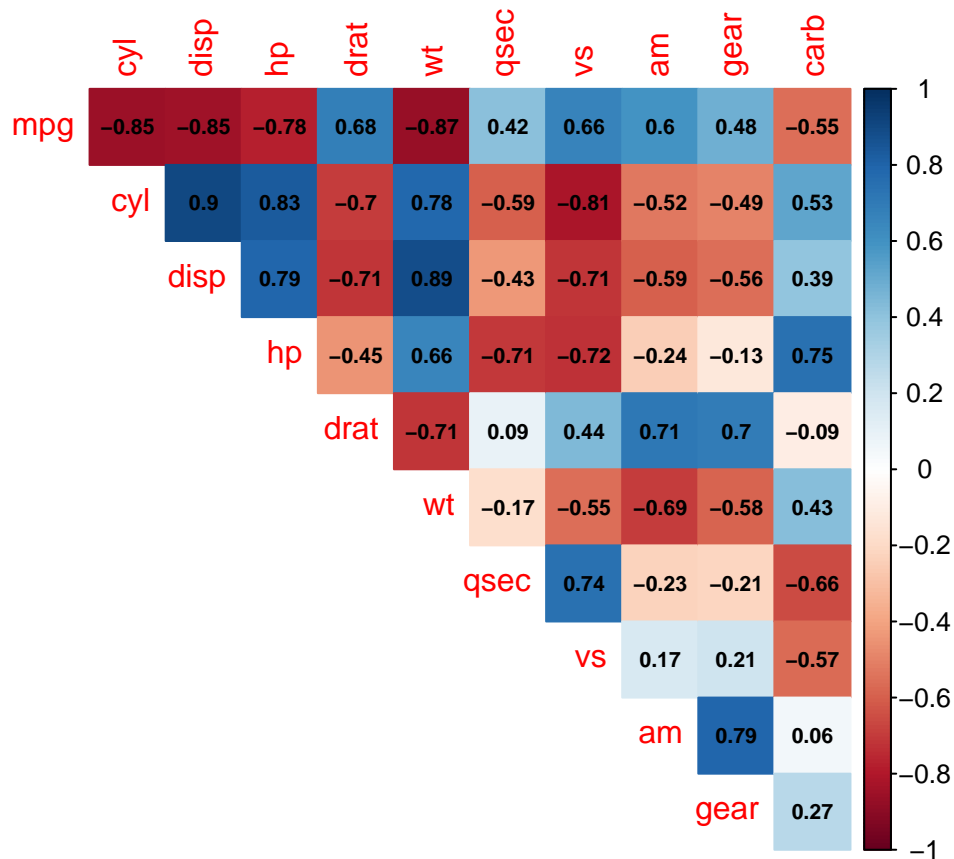
mtcars
```

```
## # A tibble: 32 x 11
##   mpg   cyl  disp    hp  drat    wt  qsec    vs  am  gear  carb
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  21     6  160   110  3.9   2.62  16.5    0    1    4     4
## 2  21     6  160   110  3.9   2.88  17.0    0    1    4     4
## 3 22.8     4  108    93  3.85  2.32  18.6    1    1    4     1
## 4 21.4     6  258   110  3.08  3.22  19.4    1    0    3     1
## 5 18.7     8  360   175  3.15  3.44  17.0    0    0    3     2
## 6 18.1     6  225   105  2.76  3.46  20.2    1    0    3     1
## 7 14.3     8  360   245  3.21  3.57  15.8    0    0    3     4
## 8 24.4     4  147    62  3.69  3.19   20     1    0    4     2
## 9 22.8     4  141    95  3.92  3.15  22.9    1    0    4     2
## 10 19.2     6  168   123  3.92  3.44  18.3    1    0    4     4
## # ... with 22 more rows
```

So we have 32 observations with 11 numeric attributes for each. Let's see the correlation between the attributes.

```
library(corrplot)

corrplot(cor(mtcars),
  method = "color",
  type = "upper",
  addCoef.col = TRUE,
  diag = FALSE,
  number.cex = 0.7)
```



The figure shows that mpg has a strong negative correlation for cyl, disp, hp and wt (more intense red). Positive correlation can be found, mainly, with drat, vs and am.

## Trends: MPG and Transmission

This paper analyzes the available data and answer two questions with regression models. The topics below treat the questions to then answer them. To keep in mind, the two questions are “Is an automatic or manual transmission better for MPG?” and “Quantify the MPG difference between automatic and manual transmissions”

### Simple Linear Regression

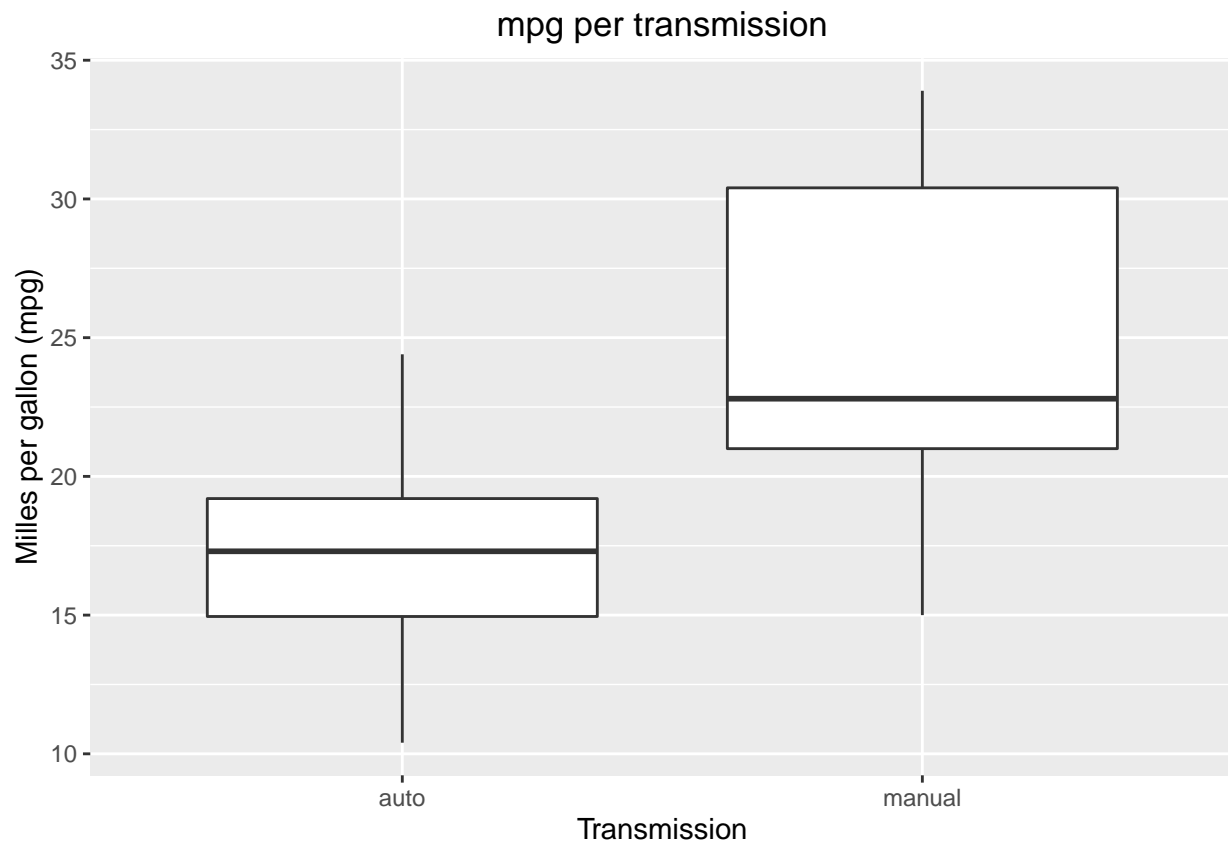
We want to understand if has one trend to milles per gallon (mpg) when compared to cars that has automatic transmission and those how has manual transmission. The mtcars data set utilizes the value 0 to cars with automatic transmission and the value 1 to manual transmission cars in the column *am*.

First, one boxplot of mpg X automatic/manual

```
library(ggplot2)

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)",
       title = "mpg per transmission") +
  theme(plot.title = element_text(hjust = 0.5))
```



By the boxplot above, automatic transmission seems have less mpg spent than manual transmissions, being better.

Let's see what a simple linear regression says about the relationship. Here will use mpg as outcome and am as explanatory variable.

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

##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      17.147      1.125   15.247 1.13e-15 ***
## factor(am)manual    7.245      1.764    4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

There is just two values for x-axis: Automatic or manual. The model says that our intercept is 17.147 and the slope is equal to 7.245, so, is expected that the values of mpg for automatic transmission be around  $17.147 \pm 1.125$ , being this our first factor variable. The cost to change the transmission for the next type (manual) is equal to 7.245, so we increase this value from 17.147 and get that the expected value by the regression linear model be around  $24.392 \pm 1.764$  for manual transmission.

Note that we have 30 degrees of freedom, so is better to use a T-test than a Z-test, because it's a good approximation to normal curve with a longer tail. The result of t-test results in a P-value very low for manual and automatic transmissions (both smaller than 0.0005).

A model trained with a small number of data has your problems. As we can see, the Residual Standard Error (RSE) is equal to 4.902, what is smaller than the slope (big difference between the manual and automatic transmission), but is even higher than the standard error of both. This means that not all observations can be well fitted, but is so (as we saw in the p-value) a good linear approximation.

The power of explain the data by the model, otherwise, can be analyzed with the value of Multiple R-squared and Adjusted R-squared, with both with values next to 0.33 (0.3598 to the first and 0.3385 to the second), so, the linear model can explain round of one third of the variation in the data. This means that *am* alone can brings some information about the mpg, but is not a good idea to generalize the model, being that we have more variables to increase the power of explanation of the model, even with a small number of observations.

So, we can't assume that automatic transmissions is better than manual transmissions, being that round 33% of the data can be explained by this relationship.

## Multivariable Linear Regression

How one explanatory variable can't bring relevant information, we will analyze one new model with more than one explanatory variable.

We'll create three models, the first with *am* and the 4 variables with more correlation with mpg

The second with *am* and two variables with high correlation with *mpg* and lower correlation between them (that is, explain mpg, but tries to reduce the correlation between explanatory variables)

The third with all variables

```
high_cor_fit <- lm(mpg ~ factor(am) + cyl + wt + disp + hp, data = mtcars)
some_high_cor_fit <- lm(mpg ~ factor(am) + cyl + wt, data = mtcars)
all_fit <- lm(mpg ~., data = mtcars)
```

Doing the Analysis of Variance (ANOVA)

```
anova(fit, all_fit, high_cor_fit, some_high_cor_fit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 3: mpg ~ factor(am) + cyl + wt + disp + hp
## Model 4: mpg ~ factor(am) + cyl + wt
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      30 720.90
## 2      21 147.49  9    573.40 9.0711 1.779e-05 ***
## 3      26 163.12 -5    -15.63 0.4449    0.8121
## 4      28 191.05 -2    -27.93 1.9881    0.1619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Appendix

mtcars attributes [, 1] mpg Miles/(US) gallon [, 2] cyl Number of cylinders [, 3] disp Displacement (cu.in.)  
[, 4] hp Gross horsepower [, 5] drat Rear axle ratio [, 6] wt Weight (1000 lbs) [, 7] qsec 1/4 mile time [, 8] vs  
Engine (0 = V-shaped, 1 = straight) [, 9] am Transmission (0 = automatic, 1 = manual) [,10] gear Number  
of forward gears [,11] carb Number of carburetors