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

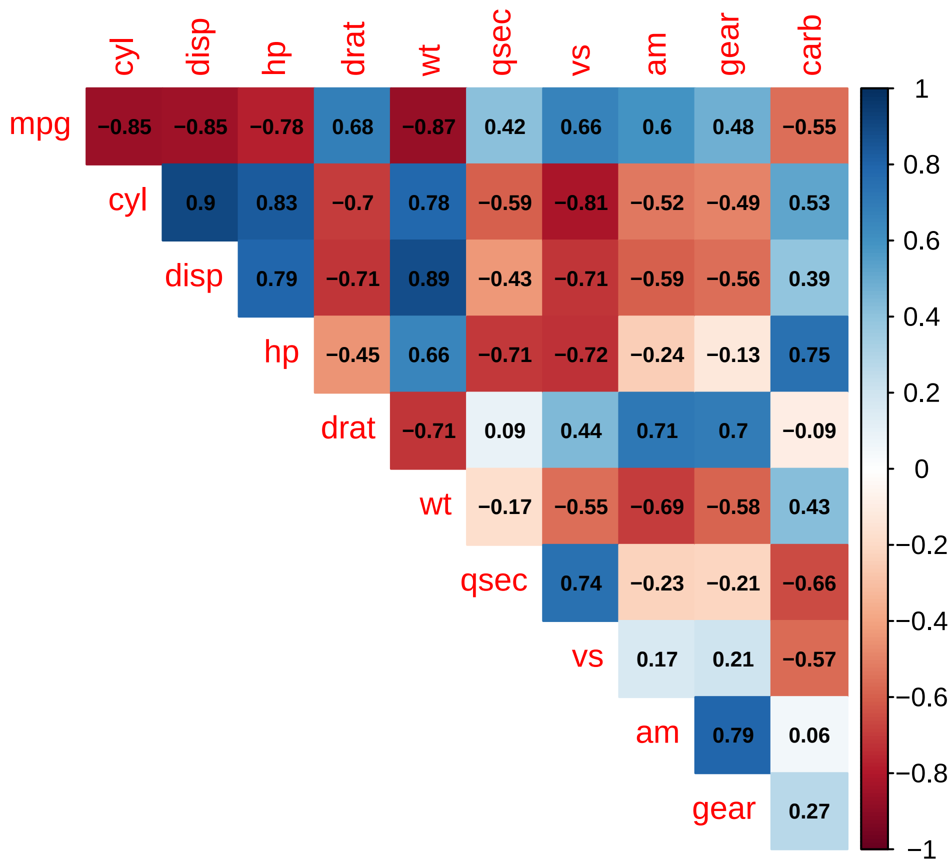
The data set used can be retrieved in the base R software by the lines below:

```
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.

```r
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

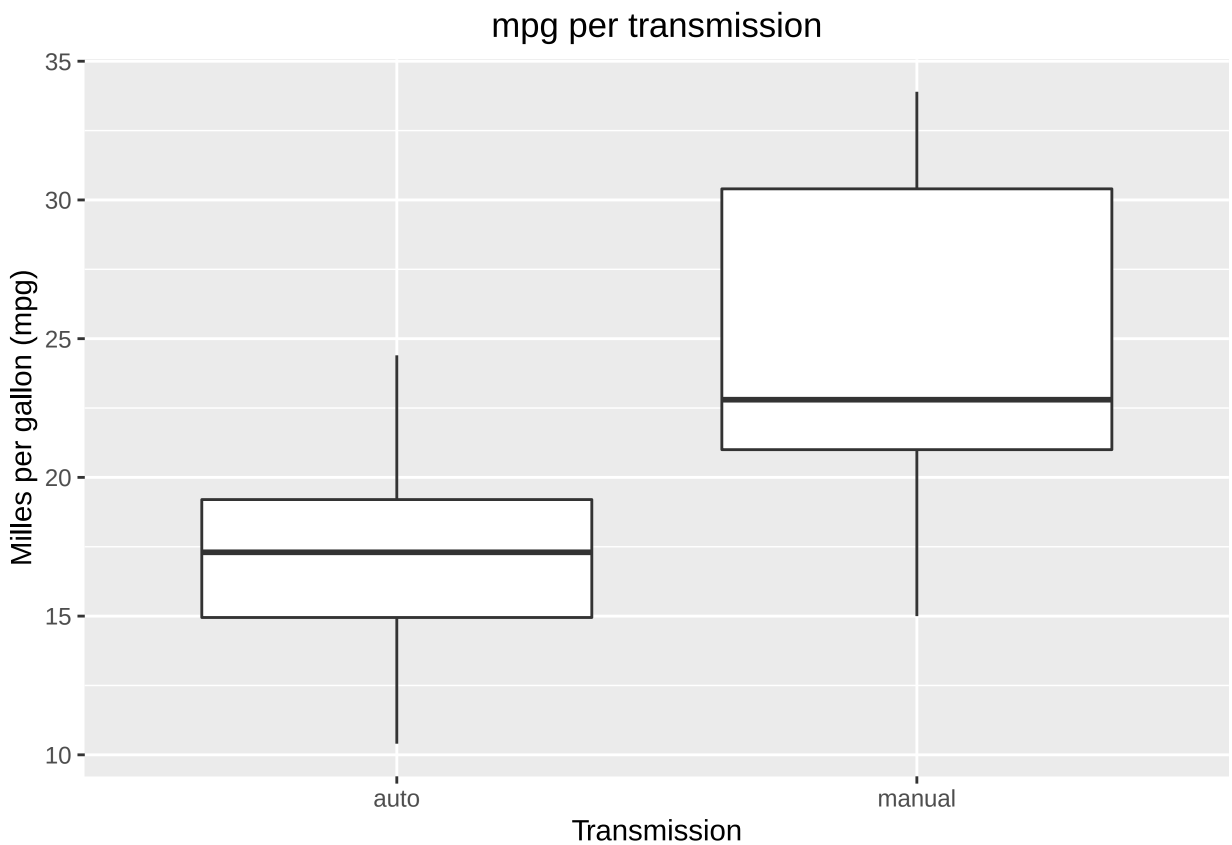
This paper analyzes the available data and answer two questions with regression models. The topics below treat the questions to then answer them. To keep in mind, the two questions are "Is an automatic or manual transmission better for MPG?" and "Quantify the MPG difference between automatic and manual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

```r
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.

```r
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 ± 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 ± 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.

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, else can be done a model with all variables and, with him, extract that with lower p-values on t-test.

```
summary(lm(mpg ~., mtcars))$coefficients
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
```

```
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## ammanual     2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

As we can see, the most influencible variables are wt, qsec and hp (plus am). Now we'll build two models, one with these 4 variables and another without hp. They will be compared to the first fit to see if exists some significance adding hp or not.

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

anova(fit, mult_fit, mult_fit_hp)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt + qsec
## Model 3: mpg ~ factor(am) + wt + qsec + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 46.5228 1.787e-09 ***
## 3     27 160.07  1      9.22  1.5551    0.2231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model without hp has higher F, so lower p-value. Let's show summary of the model:

```
summary(mult_fit)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.6178     6.9596   1.382 0.177915
## factor(am)manual   2.9358     1.4109   2.081 0.046716 *
## wt                -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec               1.2259     0.2887   4.247 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

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. The p value from F-statistic was too very low.

About the residuals, 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.

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 increasedn om 2.9358 mpg, so, starts with 12.5536 and change the value by the values of qsec and wt.

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.

Apendix

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