Regression on Miles per Gallon Car Performance

Andres Camilo Zuñiga Gonzalez

26/5/2020

As a first step, it is necessary to load the packages and set the working directory.

```
setwd('./Regression_Models')
```

```
library(ggplot2) #Plotting system
library(cowplot) #Panel for ggplot2
```

In this report we are going to review the performance of cars measured in miles per gallon. The idea is to create a regression model that explains which other variables explain better this metric. One of the main variables we are going to look at is the transmission type. As a first measure we are going to do a exploratory data analysis an then fit several models checking several variables. At the end the bet model will be found using the Akaike Information Criterion.

First, the dataset must be loaded. Then we are going to see the structure of it.

```
data("mtcars")
str(mtcars)
```

Seeing the description on the dataset using `?mtcars` and the result of the previous chunk, it is possible to determine that some of the variables are actually factors, even though they appear to be numerical. So the following step is to convert them to factors. See **Figure 1** in the *Appendix*. Notice that `gear` and `cyl` are not necessarily factors since they are a measure of quantifiable aspect of the cars, but for the purpose of this analysis they are going to be treated as factors.

```
mtcars$vs <- factor(mtcars$vs, labels = c('v-shaped', 'straight')) #Engine
mtcars$am <- factor(mtcars$am, labels = c('automatic', 'manual')) #Transmission
mtcars$gear <- factor(mtcars$gear, labels = c('3', '4', '5')) #Number of forward gears
mtcars$cyl <- factor(mtcars$cyl, labels = c('4', '6', '8')) #Number of cylinders
```

Secondly, we are going to evaluate the milles per gallon `mpg` variable using a t.test for independant samples, according to their transmission `am`. For this exercise it will be assumed that both groups are normally distributed and their variances are equal.

```
t.test(mpg ~ am, data = mtcars)
```

Using a significance level of 5%, it is possible to conclude that both groups are different (p = 0.001374), a fact seen in the *Figure 2* in the *Appendix*.

Next, knowing that they are different we are going to build two linear models to explain the variance in `mpg`, one where all the variables are included as explanatory, and another where only `am` is the explanatory variable.

```
fit_all <- lm(mpg ~ ., data = mtcars)
summary(fit_all)
fit_am <- lm(mpg ~ am, data = mtcars)
summary(fit_am)
```

From the results above, it is possible to determine that the model with all the variables explains better the variance in `mpg` given an adjusted R^2 of **0.8116**, while the one with only `am` as explanatory variable only explains **0.3385**. We see this significant difference using ANOVA on the fitted models.

```
anova(fit_all, fit_am)
```

Despite concluding that the model with all variables is significantly better than the one with only `am` (p « 0.05), we are not certain if this is the best model we can fit, therefore we will use the *Akaike Information Criterion (AIC)*, a method that measures each model and assesses how good they are. This can be done with the `step()` on the model with all the variables, in this case `fit_all`, as shown below. Notice when you run this command, there will be a huge output evaluating each possible model.

```
fit_best <- step(fit_all, direction = 'both')
```

This is the summary of the best model according to AIC.

```
summary(fit_best)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, 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
## wt           -3.9165     0.7112  -5.507 0.00000695 ***
## qsec          1.2259     0.2887   4.247   0.000216 ***
## ammanual      2.9358     1.4109   2.081   0.046716 *
## ---
## 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: 0.0000000000121
```

According to this, the best model includes `am`, `wt` (weight in 1000 lbs) and `qsec` (1/4 mile time) as explanatory variables and the adjusted R^2 is **0.8336**, which is slightly better than the model with all the variables. Now we are going to compare both using ANOVA.

```
anova(fit_all, fit_best)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     19 130.05
## 2     28 169.29 -9   -39.235 0.6369 0.7524
```

As expected by their adjusted R^2 values, they are not different in explaining `mpg`, however the one found using AIC is preferred since it contains less variables, hence more parsimonious.

Finally, we evaluate some graphical diagnostics of the best model as seen in **Figure 3** in the *Appendix*:

- The residuales seem to be randomly distributed, and no evidence of heteroscedasticity

- According to the QQ-plot the residuals appear to be normally distributed, although slightly skewed
- There are some evident outliers in the dataset or leverages in the Residuals vs Leverage plot

In conclusion, when only evaluating transmission, mannual has an increase of **7.245**, but when other variables are included, namely 1/4 mile time and weight, that value is reduced to **2.9358**.

Appendix

Figure 1. Variable relationship

```
pairs(mtcars)
```

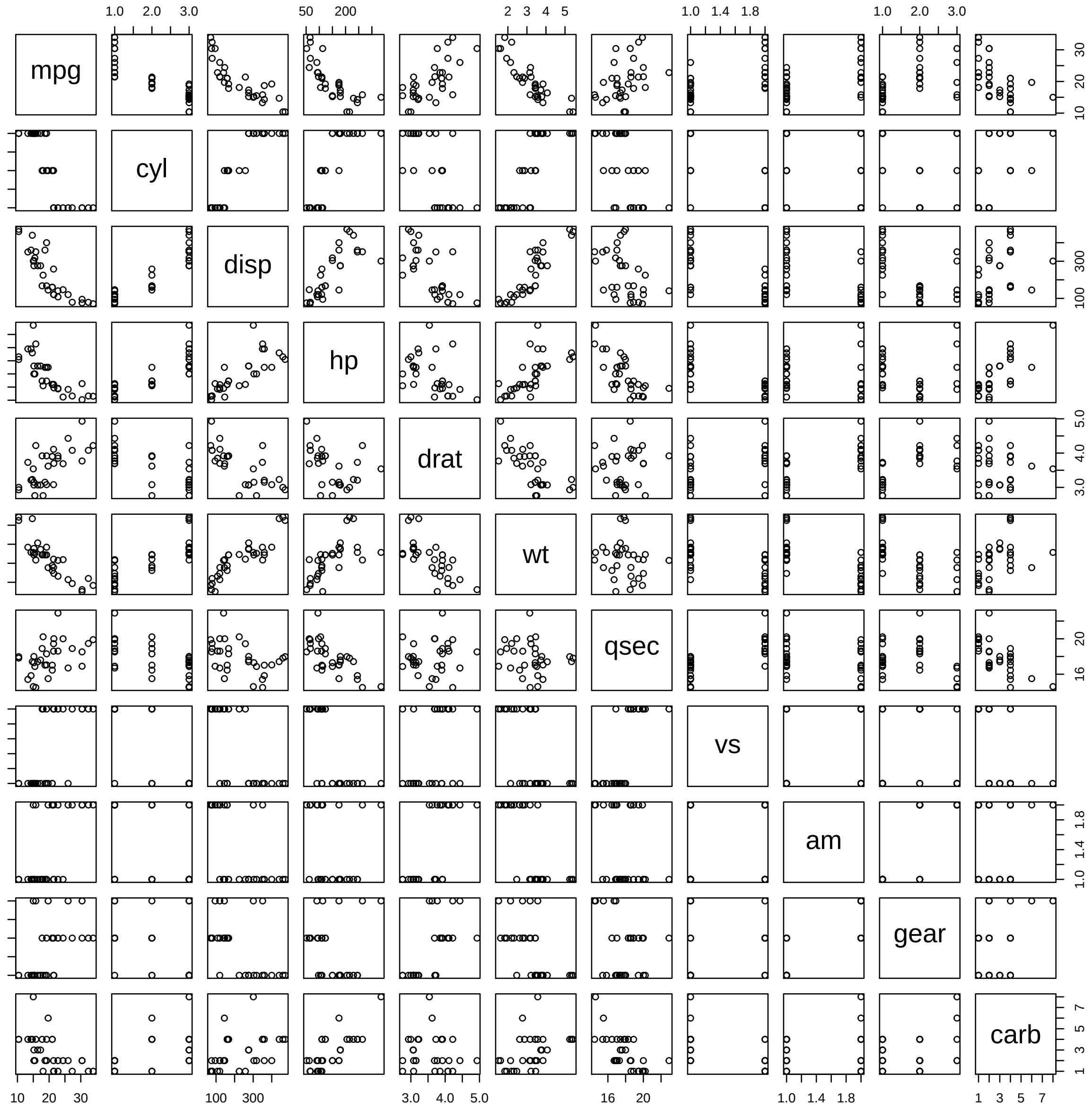


Figure 2. Miles per Gallon versus Transmission

```
ggplot(mtcars, aes(x = am, y = mpg)) +
    geom_boxplot(aes(fill = am)) +
    labs(x = 'Transmission', y = 'Miles per Gallon') +
    theme_bw() + theme(legend.position = 'none', panel.grid.major.x = element_blank())
```

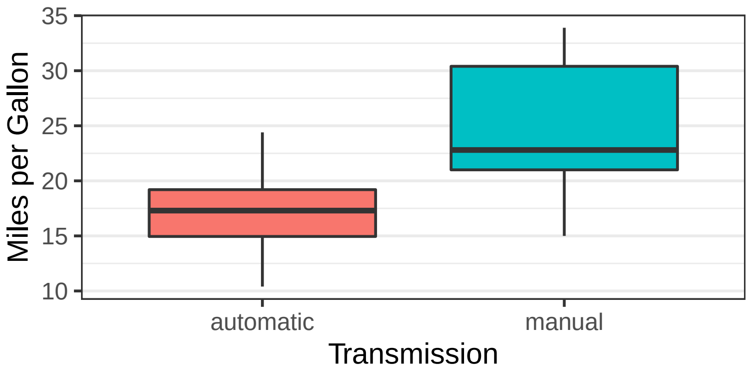
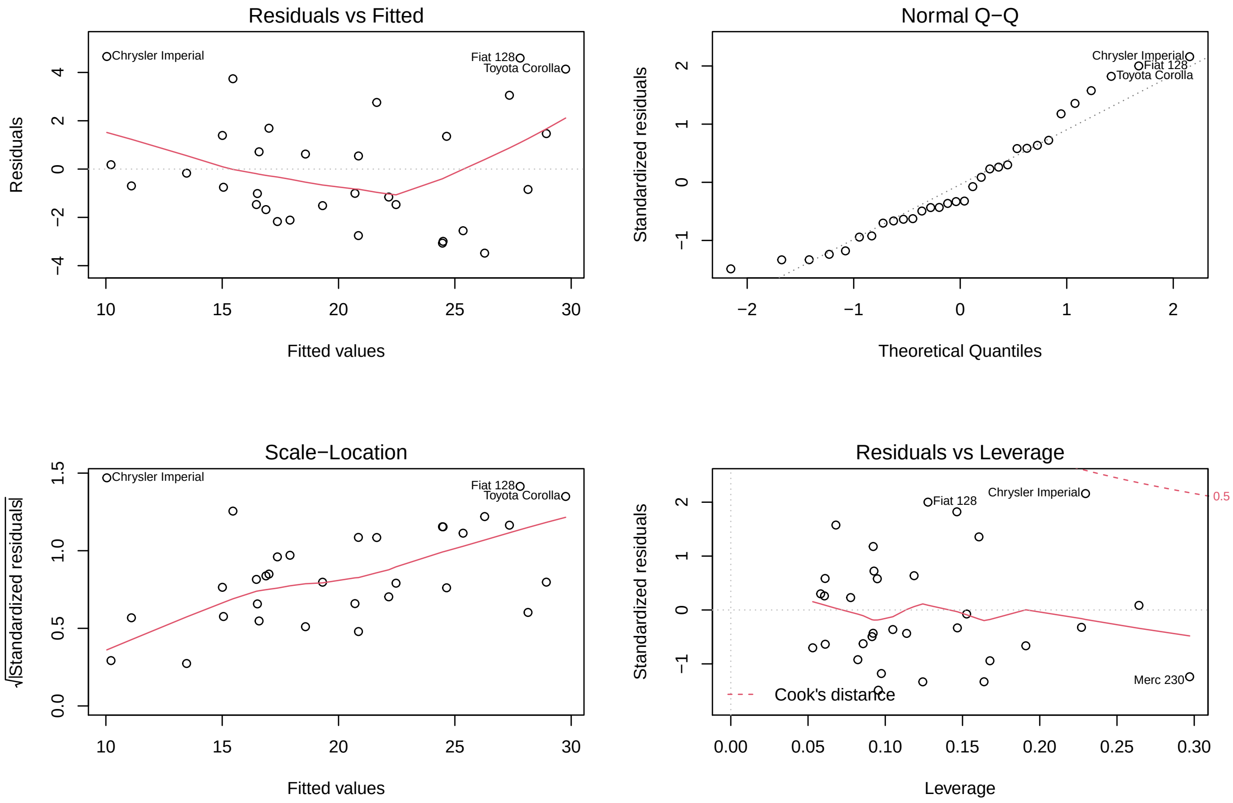


Figure 3. Graphical Diagnostics of the best model

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


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
4
2
0
−2
−4
10
15
20
25
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corolla
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
0.0
0.5
1.0
1.5
Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Chrysler Imperial
Merc 230
0.5
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30