# Coursera Regression Models

*Daniel Resende*

*April 25, 2015*

## Executive Summary

In this project we are going to analyse the `mtcars` dataset to explore the relationship among `miles per gallon` consumption, as outcome, and `manual` or `automatic` cars, as predictor.

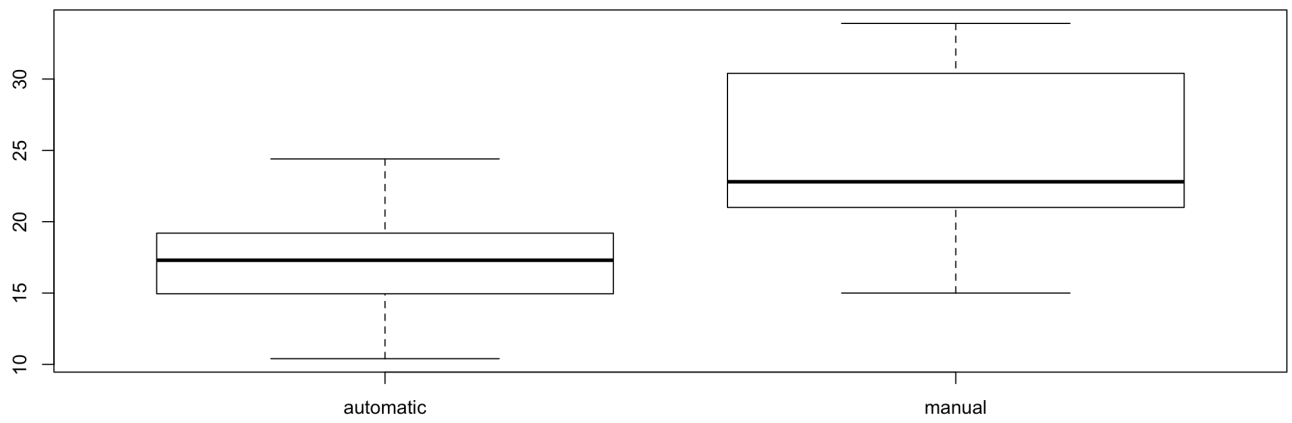
We are particularly interested in the following two questions: * Is an automatic or manual transmission better for MPG * Quantify the MPG difference between automatic and manual transmissions

The `mtcars` has 11 variables. We will have to make a multivariate regression to understand the influence of each one and of our target `am` .

## Specific Regression

As we want to check the `am` influence in mpg, the most straightfoward way to look is to make a regression considering only this variable.

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




At first sight, manual transmissions causes an increase of 7.24 in `mpg` . It seems very significant, the mean estimated for manual is 4.11 t deviations away (p-value of 2.8510^{-4}).

But there are several other variables affecting the model and this increase might note be true.

## Complete Regression

After making the analysis of just one variable, we’re going to throw them all in the model.

Considering all variables, the effect of `am` decreases drastically. Now it only increases the `mpg` by 2.52. Not only the influence, but the significance of this varible decreaeses too. The t value is only 1.23 deviances away (p-value of 0.234).

# Selective Regression

To make a better model and reduce the noise, we should exclude some more insignificant variables. To do that we are going to look to Correlation and Variance Inflation.

Below it's possible to see the correlation among `am` and all other variables. A better looking plot is available on appendix.

```
##    am  gear  drat    wt   mpg  disp   cyl    hp  qsec    vs  carb
##  1.00  0.79  0.71 -0.69  0.60 -0.59 -0.52 -0.24 -0.23  0.17  0.06
```

Another criteria to exclude variables is the Variance Inflation. Below, each variable impact is listed.

```
##     disp      cyl       wt       hp     carb     qsec     gear       vs
## 4.649757 3.920948 3.894212 3.135608 2.812249 2.743712 2.314617 2.228424
##       am     drat
## 2.156035 1.837014
```

We've decided to drop the `gear` and `drat`, because they have high Correlation with the `am`. We also decide to drop `vs` as it appears to have low significant impact and was just producing noise.

```
fit$selective <- lm(mpg ~ am + disp + cyl + wt + qsec + hp, mtcars)
anova(fit$specific, fit$selective, fit$complete)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + disp + cyl + wt + qsec + hp
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     25 150.99  5    569.91 16.2284 1.357e-06 ***
## 3     21 147.49  4      3.50  0.1245     0.972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Conclusion

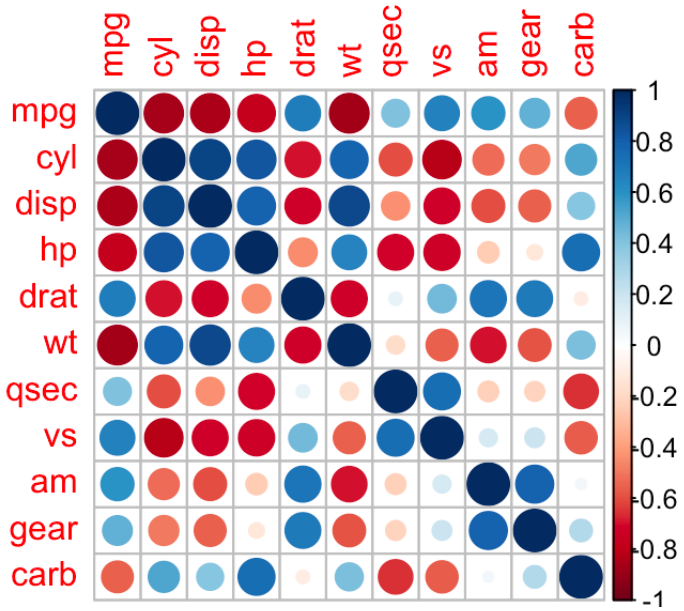
`Manual` cars might have more autonomy than `automatic`. They can run around 20.05 miles per gallon more.

But these numbers are not significant as they are 1.51 t deviations away from the mean of `automatic` (p-value of 0.1443). The conclusion is no conclusion as almost aways.

More important is that you're more likely to drop coffee on your leg while driving `manual` car. The p-value is 0.000000000001. Source: I was a `manual` car owner.

# Apendix

## Corelation




## R Squared

The selective model capture most of the variance,almost the same of the complete model (0.8659 against 0.869). The model with just `am` as predictor captures only 0.3598 of the variation.

It makes the selective model pretty reasonable.

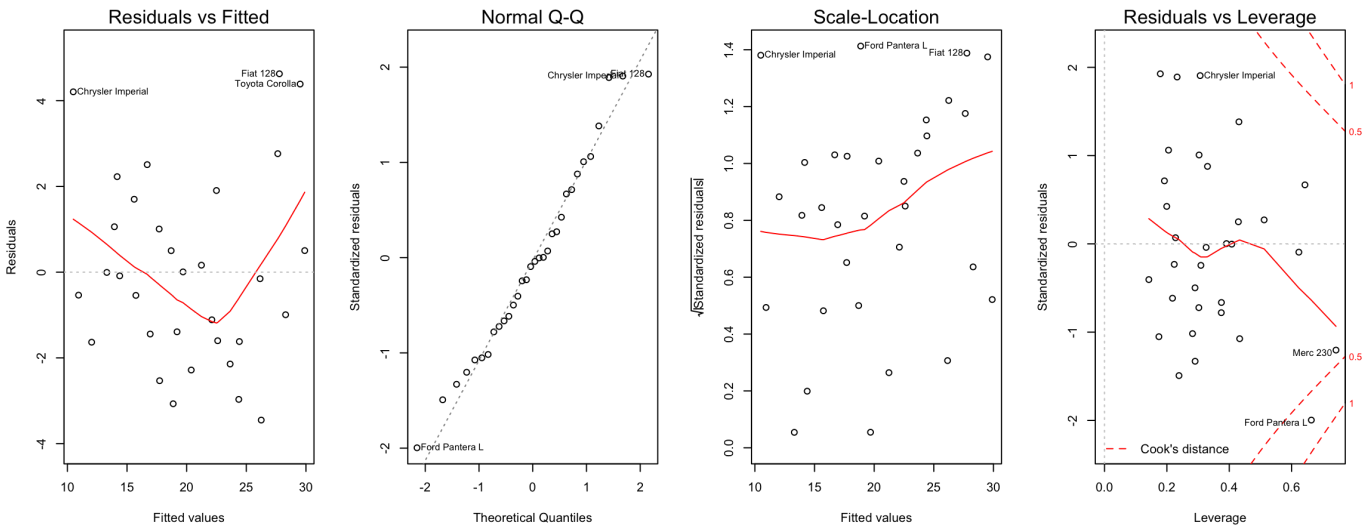
## Residual diagnostics

The `Residuals vs Fitted` looks independent in the selective model. It might indicate that there are no significant variable out of the model.

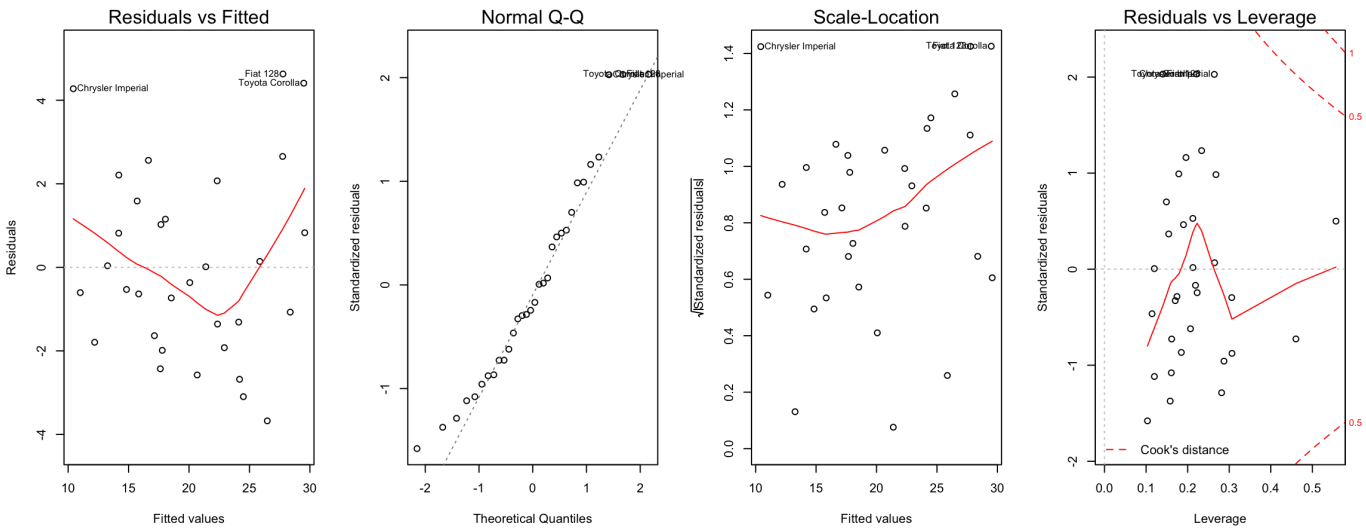
The residuals looks normaly distributed, as the `Normal Q-Q` indicates.

The data does not have any significant outlier, the y axis of the scale location show all points are less than 1.5 standard deviations away. They look normal, as QQ plot show.

```
## [1] "complete"
```



```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
##
## [1] "selective"
```



```
##
## Call:
## lm(formula = mpg ~ am + disp + cyl + wt + qsec + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6755 -1.6757 -0.4477  1.2615  4.6289
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.05170   13.30486   1.507  0.14432
## am           2.94075    1.71810   1.712  0.09935 .
## disp         0.01396    0.01155   1.209  0.23802
## cyl         -0.50207    0.78882  -0.636  0.53025
## wt          -3.99773    1.21564  -3.289  0.00299 **
## qsec         0.81018    0.57171   1.417  0.16879
## hp          -0.01956    0.01489  -1.314  0.20088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.458 on 25 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8337
## F-statistic: 26.91 on 6 and 25 DF,  p-value: 9.29e-10
```