# Regression Models Course Project

Daniel Arturo Lopez Sanchez

8/16/2020

## Overview

This is the Regression Models Course Project. The instructions are the following: You work for *Motor Trend*, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They 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

## Enviroment

Loading the libraries and data set

```
library(datasets)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

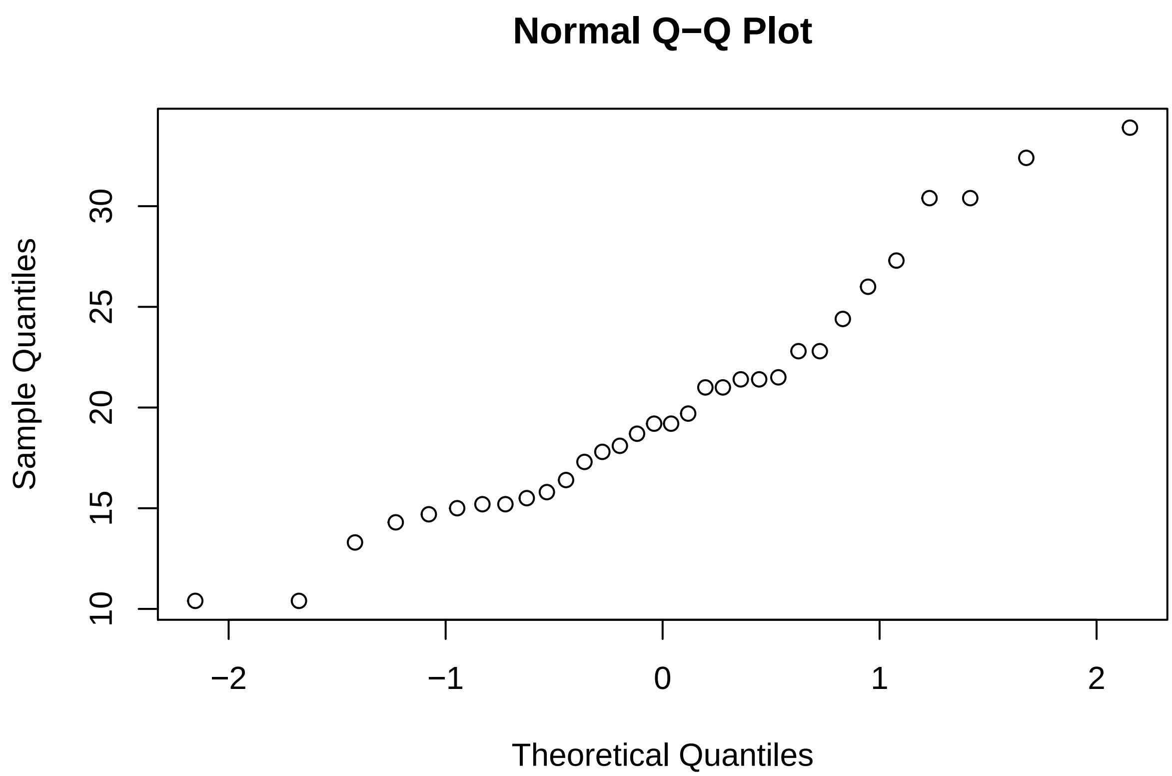
```
data("mtcars")
```

## Exploratory Data Analysis

```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

```r
qqnorm(mtcars$mpg)
```




with this plot we can assume our mpg variable to be normal and work on it with no trouble.

#### Converting to factor the non numeric variables

```r
mtcars$cyl <- factor(mtcars$cyl) # number of cylinders
mtcars$vs <- factor(mtcars$vs) # Engine type (V-shaped or straight)
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual")) # Transmission
mtcars$gear <- factor(mtcars$gear) # Number of forward gears
mtcars$carb <- factor(mtcars$carb) # Number of carburetors
```

#### Automatic vs Manual Transmission

We want to visually see the difference in mpg if whether the car's transmission type is Automatic or Manual

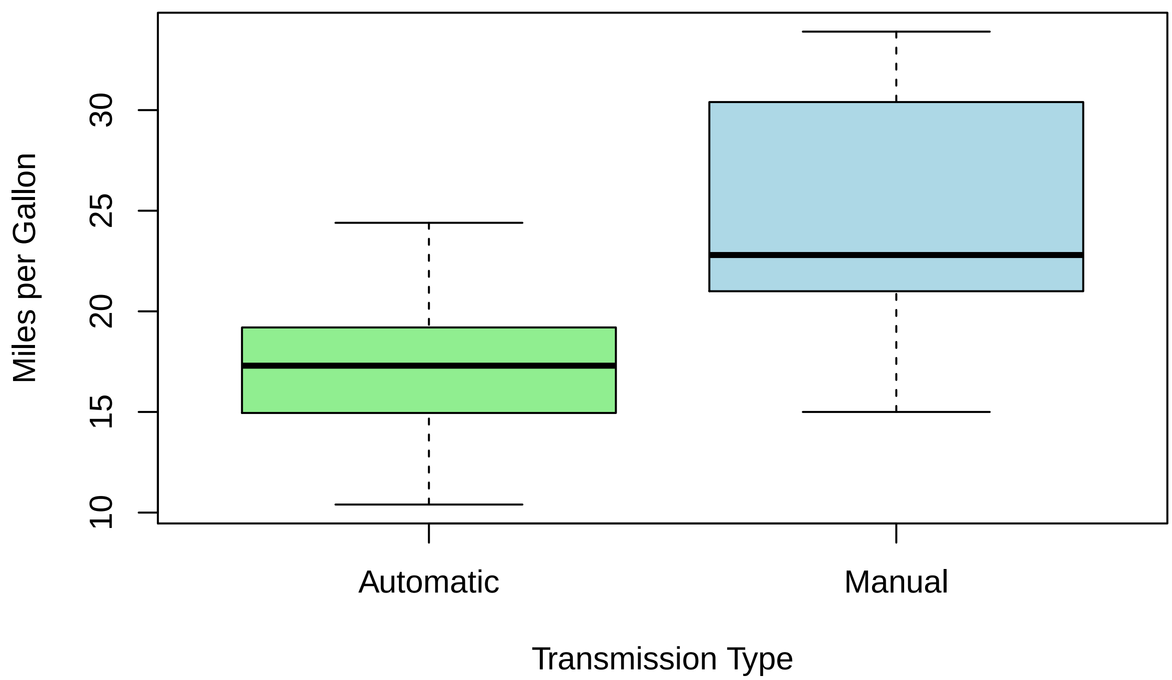
```r
mean_transmission <- mtcars %>% group_by(am) %>% summarise(average=mean(mpg))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
head(mean_transmission)
```

```
## # A tibble: 2 x 2
##   am         average
##   <fct>        <dbl>
## 1 Automatic     17.1
## 2 Manual        24.4
```

```r
boxplot(mpg~am, data = mtcars, xlab = "Transmission Type",
        ylab = "Miles per Gallon", col = c("light green", "light blue"))
```




Now that we've seen the difference in mpg on automatic and manual transmission, The want to fit a model to evaluate if this difference is statistically significant, And also to check if there are some other confounding variables that affect directly the mileage.

### Simple Linear Regression

We want to fit a linear model with only one regressor, which is "am" - Transmission type.

```r
fit1 <- lm(formula = mpg~am, data = mtcars)
summary(fit1)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ 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 ***
## amManual       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
```

We can see that the difference in Miles per Gallon on transmission type is indeed highly significant with a p-value of nearly 0.0003. Seeing this fit, we can think of saying that the average Miles per gallon of a given car increases by 7.25 if the transmission is Manual, holding everything else constant. However, if we analyze the Adjusted R-squared value, we can observe that our model is just explaining one third of the variance in Miles per Gallon. We can do a better fitted model including some other variables with some correlation to mpg.

### Multivariate Linear Regression

To know which variables to include in our model we are going to perform an Analysis of variance including all of our predictors.

```
variance_analysis_mpg <- aov(formula = mpg~., data = mtcars)
summary(variance_analysis_mpg)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
## disp         1   57.6    57.6   7.181   0.0171 *
## hp           1   18.5    18.5   2.305   0.1497
## drat         1   11.9    11.9   1.484   0.2419
## wt           1   55.8    55.8   6.950   0.0187 *
## qsec         1    1.5     1.5   0.190   0.6692
## vs           1    0.3     0.3   0.038   0.8488
## am           1   16.6    16.6   2.064   0.1714
## gear         2    5.0     2.5   0.313   0.7361
## carb         5   13.6     2.7   0.339   0.8814
## Residuals   15  120.4     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the number of cylinders, displacement and weight are quite significant for our dependent variable.

We are going to fit a new model adding this 3 regressors.

```r
fit2 <- lm(formula = mpg~am+cyl+disp+wt, data = mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5029 -1.2829 -0.4825  1.4954  5.7889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.816067   2.914272  11.604 8.79e-12 ***
## amManual     0.141212   1.326751   0.106  0.91605
## cyl6        -4.304782   1.492355  -2.885  0.00777 **
## cyl8        -6.318406   2.647658  -2.386  0.02458 *
## disp         0.001632   0.013757   0.119  0.90647
## wt          -3.249176   1.249098  -2.601  0.01513 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.652 on 26 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8064
## F-statistic: 26.82 on 5 and 26 DF,  p-value: 1.73e-09
```

We can see by fitting this new model that we are now explaining almost 84% of the mpg variance with this variables.

## Evaluating our model

To evaluate the performance of our model we are going to do the process of model selection, using ANOVA.

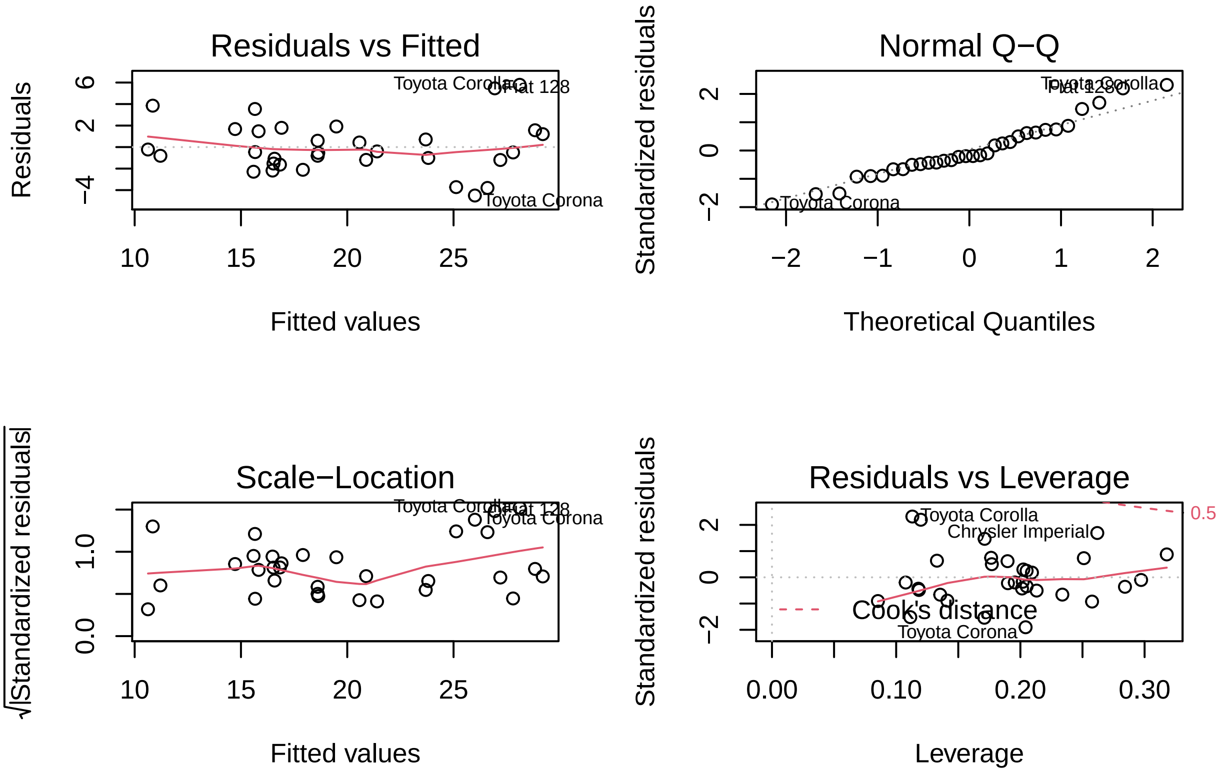
```r
anova(fit1,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 182.87  4    538.03 19.124 1.927e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can now select Model 2 as the best for explaining our independent variable "miles per gallon"

## Residual Plots

```
par(mfrow = c(2, 2))
plot(fit2)
```




In our Residuals vs Fitted plot, there seems like there isn't any patterns, which is good. Also our residuals are normally distributed.

## So.. Is an automatic or manual transmission better for MPG?

At first we saw a statistically significant difference (with a 0.05 level of significance) between Automatic and Manual Transmission. However, there are other variables that we can consider more significant, like the number of cylinders or the weight. We can see that the average miles per gallon of a given car decreases by 6.32, being this an 8-cylinder car; holding everything else constant.

But, why did we observed earlier that Transmission Type on a vehicle is highly significant on Miles per gallon? We can answer this by analyzing the correlation between the predictors.

```
cat("Correlation between transmission and weight: ", cor(x = as.numeric(mtcars$am), y = as.numeric(mtca
```

```
## Correlation between transmission and weight:  -0.6924953
```

```
cat("Correlation between transmission and the number of cylinders: ", cor(x = as.numeric(mtcars$am), y =
```

```
## Correlation between transmission and the number of cylinders:  -0.522607
```

The correlation coefficient values are not that low. The significance of our very first model can be attributed to this correlation.