Peer-graded Assignment: Regression Models Course Project

Thiago Augusto Cardoso

December 23, 2016

Executive Summary

This work intend to find evidence to respond to two questions

```
"Is an automatic or manual transmission better for MPG"
"Quantify the MPG difference between automatic and manual transmissions"
```

Therefore it will start by looking at the data and finding sufficiente evidence to respond the questions above.

1. Load the Dataset and summary statistics

Loading the dataset and checking for basic statistics which are found in Apendix I

```r
data<-mtcars
str(data)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

2. Exploratory Data Analysis

For further understanding of the data let's start checking the variables. So in Apendix I we have figures from the data which support the following:

- manual cars tend to be more economical while having less cylinders
- manual cars tend to have less weigth than automatic and lighter cars, mean less fuel consumption.
- manual cars have lower displacement and the lower displacement the more economical is the vehicle.
- manual cars have lower horsepower and that the lower horsepower the more economical is the vehicle.
- automatic cars have lower rear axle ratio and that the higher the rear axle ratio the more economical is the vehicle.
- manual cars are faster in the quarter mile, and the slower the car in the quarter mile, the more economical it is

To see actual other relations, let's check the correlation between the variables

```
##             cor names
## mpg   1.0000000   mpg
## cyl  -0.8521620   cyl
## disp -0.8475514  disp
## hp   -0.7761684    hp
## drat  0.6811719  drat
## wt   -0.8676594    wt
## qsec  0.4186840  qsec
## vs    0.6640389    vs
## am    0.5998324    am
## gear  0.4802848  gear
## carb -0.5509251  carb
```

To see which factor might be significant, let's create a linear model with all variables, to see if we can knock any variable out

```
##
## 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

## # A tibble: 10 × 3
##    names    pvalues        cor
##    <chr>      <dbl>      <dbl>
## 1    cyl 0.91608738 -0.8521620
## 2   disp 0.46348865 -0.8475514
## 3     hp 0.33495531 -0.7761684
## 4   drat 0.63527790  0.6811719
## 5     wt 0.06325215 -0.8676594
## 6   qsec 0.27394127  0.4186840
## 7     vs 0.88142347  0.6640389
## 8     am 0.23398971  0.5998324
## 9   gear 0.66520643  0.4802848
## 10  carb 0.81217871 -0.5509251
```

The actual table shows that both weight and the type of transmission are the most significant in mpg. Now let's try models that consider both variables, but lets start knocking the least significant variables, without losing explanatory power according to adjust r squared, in this case cylinders.

```
## [1] "vs"

## [1] TRUE
```

Since the adjusted r squared from the original model (all variables) is less than the model above we continue to knock down variables.

Now knocking the least significant, in this case V/S

```
## [1] "carb"

## [1] TRUE
```

Since the new model has a better adusted r-squared we continue to knock down variables. Now knocking the least significant, in this case carb

```
## [1] "gear"

## [1] TRUE
```

Now knocking the least significant, in this case gear

```
## [1] "drat"

## [1] TRUE
```

Now knocking the least significant, in this case drat

```
## [1] "disp"

## [1] TRUE
```

Now knocking the least significant, in this case disp

```
## [1] "hp"

## [1] FALSE
```

Now knocking the least significant, in this case hp

```
## [1] "(Intercept)"

## [1] FALSE
```

3. Model used

```
{ r model} fit<-lm(mpg~wt+am+qsec ,data)
```

This last model, every variable is significant and there is no loss of explanatory power according to the adjusted r squared. That seems logical since the quarter mile is correlated with cylinders, displacement and all the other dropped variables.

4. Is an automatic or manual transmission better for MPG?

According to the model tested, a manual transmission implies a better mpg

5. Quantify the MPG difference between automatic and manual transmissions

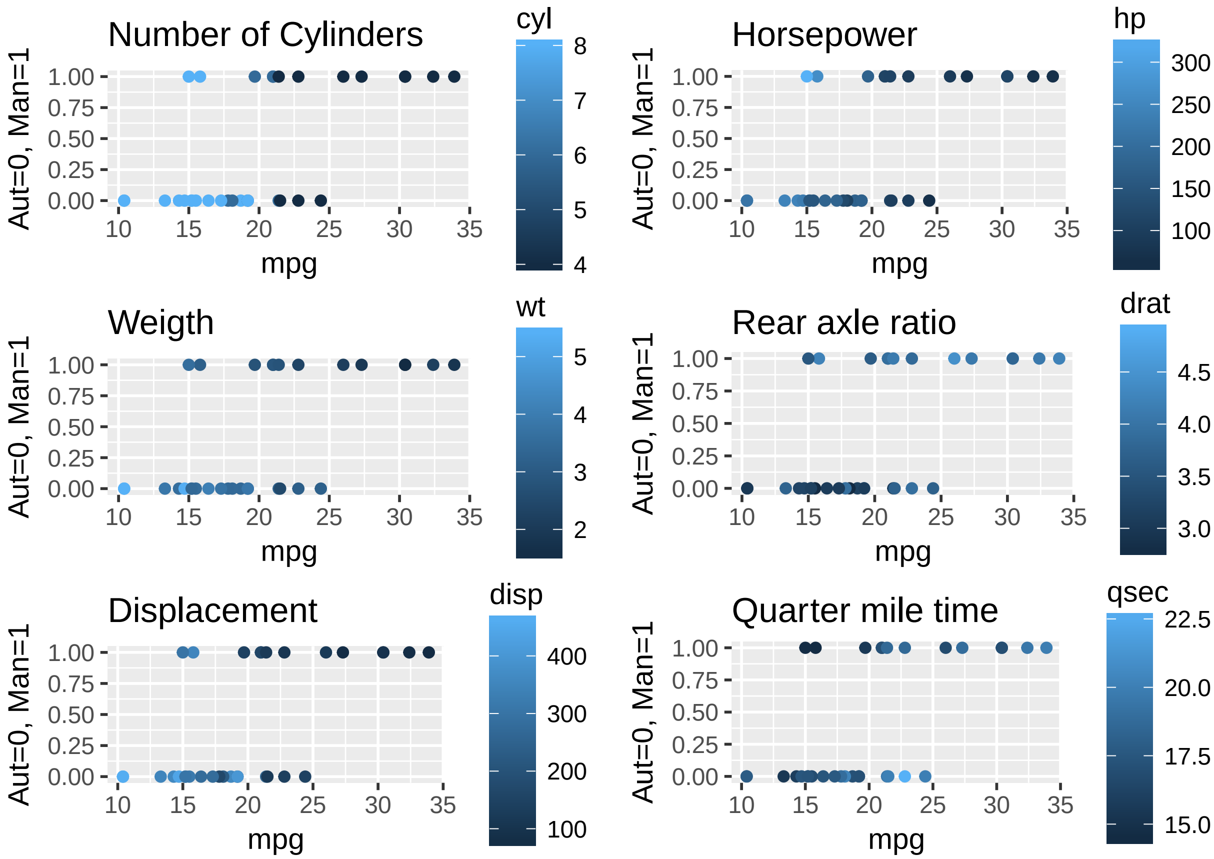
According to the model tested, ceteris paribus, a manual transmission implies an average of 2.94 mpg more of economy, in comparisson with the automatic.

6. Residual plots

The residual plots are show in Apendix II

Observing the residuals, it seems reasonable that the residuals are normal.

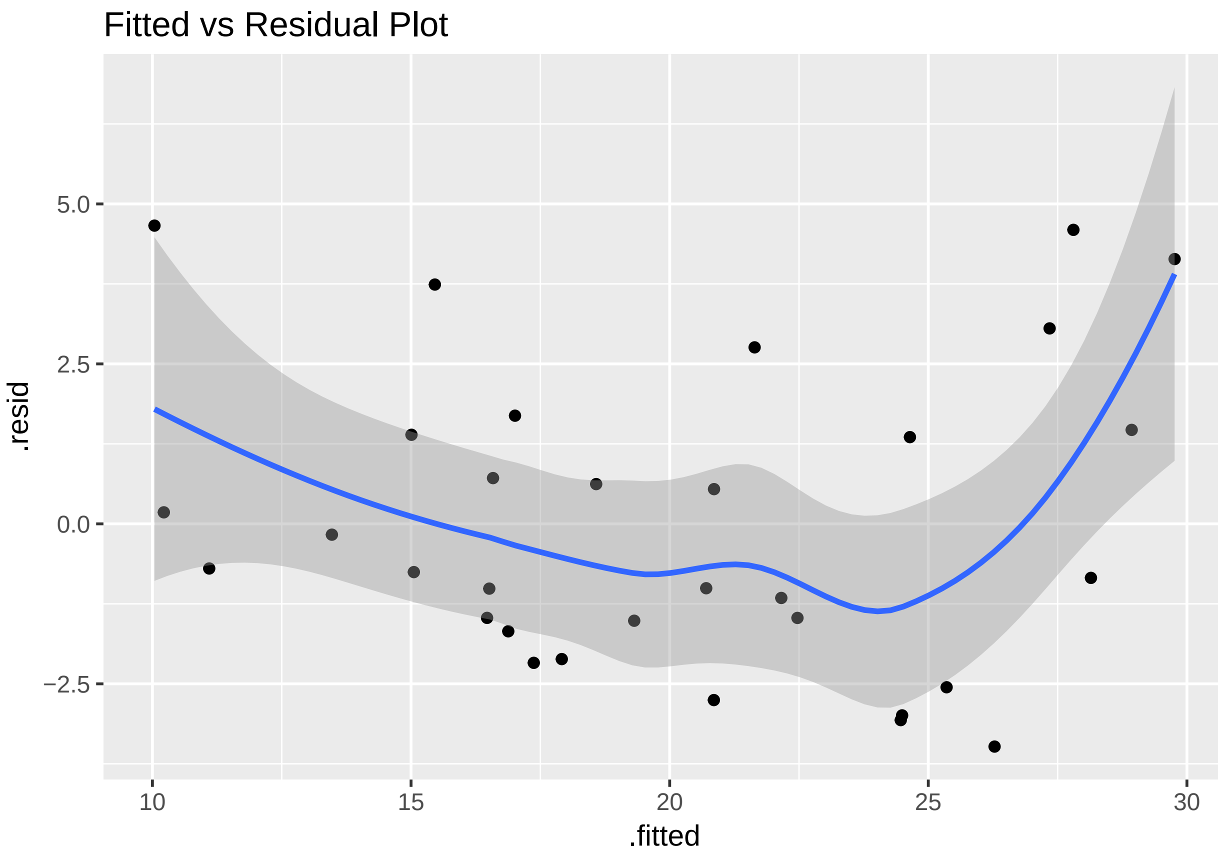
7. Apendix I - Graphing the data



8. Apendix II - Residuals from the model

```
library(broom)
library(ggplot2)
fit<-lm(mpg~wt+am+qsec ,data)
fitf <- fortify(fit)
g<-ggplot(fitf, aes(x = .fitted, y = .resid)) + geom_point()+geom_smooth()+labs(title='Fitted vs Residu
g
```

```
## `geom_smooth()` using method = 'loess'
```


Fitted vs Residual Plot
5.0
2.5
0.0
−2.5
.resid
10
15
20
25
30
.fitted