Mileage per Gallon vs Transmission type

Marcelo Guimarães

August 27, 2016

Summary

- This is a report for the Assignment: "Regression Models Course Project". We conducted an exploratory analysis of the data, in which , the data were probed for confounding variables and only the most relevant variables for analysis were retained. A naive regression of **MPG** as a function of **am** (**mpg** ~ **am**) indicates that Manual cars spend on average **7.25** ± **2.89** more MPG over Automatic cars (84% confidence interval, 1 std). However, no significant difference (**p-value** > 0.89) in mileage per gas between manual and automatic cars was found when the analysis account for weight and the number of cylinders as explanatory variables (**adj.** $\mathbf{R^2} = 0.82$, **p-value** << 0.05 for **mpg** ~ **wt+cyl**).

Introduction

The task problem of this assignment consists in determining the relationship between a set of variables and miles per gallon (mpg), specifically it will address two questions: *Is an automatic or manual transmission better for MPG? What is the quantitative MPG difference between the transmissions?*

Section 1 (Exploratory analysis -Heatmap and Dendogram)

The data for analysis is found in a data frame with 32 observations on 11 variables. In order to select the most relevant variables to the analysis a heatmap (see Figure 1) is created using the correlation between the variables as metric of distance.

Observing the dendogram Fig. 1 b), 3 groups can be seen at height 0.5. The first group contains the output variable **MPG** and the third group contains the variable of interest **am**. We select only the first group and the **am** variable for in-depth exploratory analysis.

Section 2 (Exploratory analysis -Plot Matrix of variables)

After selecting the most relevant variables we can summarize the relationship between them in a plot matrix (Figure 2).There is a lot of information in this Figure. The most correlated variables with **MPG** is **wt** (*weight*) and **cyl** (*# of cylinders*). **Note:** *Displacement is highly correlated with mpg, but is also correlated with # of cylinders (90%), we kept the cylinders variable since it is more correlated with mpg than the displacement and horse power.* The boxplot (Fig. 2, last column) for **MPG** for each **am** group indicates that the manual transmission (**am**=1) cars have higher mpg mean than the automatic ones (**am**=0). However, the **am** variable is highly correlated with other explanatory variables. For instance, there is a -0.6924953 correlation between **am** (transmission type) and **wt** (weight). In this case It is expected that most of the variance of **MPG** can be explained by the **wt** and **cyl** variables. That is, the difference in mpg due to **am** should not be significant when other variables are considered.

The amount of cylinders, and weight that defines the mileage. If we maintain these characteristics fixed we should see no mileage difference between automatic and manual transmission cars. To confirm that hypothesis a multivariate regression of the most correlated variables is performed in the following section.

Section 3 (Analysis: Regressions)

A naive regression (mpg ~ am) would indicate a significant (**p-value**=0.0003, **adj. R^2**=0.3384589) mileage difference (7.2) between automatic and manual cars:

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 7.2449392713 1.7644216316 4.1061269831 0.0002850207
```

However, the residue is strongly correlated with the output (see Figure 3).

A plot of the MPG regressed by the most correlated variable ***wt*** and grouped by transmission **am** can be seen at Figure 4. The regression of the whole data (blue line) is inside the confidence interval for the regression of each group data (manual and automatic cars). This indicates that the transmission does not explain the difference in mileage, as It can be seen from the regression statistics (mpg ~ wt+am):

```
##        Estimate Std. Error     t value     Pr(>|t|)
## wt  -5.35281145  0.7882438 -6.79080719 1.867415e-07
## am1 -0.02361522  1.5456453 -0.01527855 9.879146e-01
```

The p-value for the **am** variable is not significant (**p-value**=$0.988 > 0.05$,**adj. R^2**= 0.7357889). It is clear that the residuals are less correlated with MPG when **wt** is added as a regressor, but a separation between groups (**am**) can still be distinguished (Figure 5). The next steps is to include the *number of cylinders* variable in the analysis:

```
##             fit Adjusted_R2
## 1: mpg ~ wt+cyl   0.8185189

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 39.686261  1.7149840 23.140893 3.043182e-20
## wt          -3.190972  0.7569065 -4.215808 2.220200e-04
## cyl         -1.507795  0.4146883 -3.635972 1.064282e-03
```

It can be inferred from that analysis that the weight and number of cylinders alone can predict most of **MPG** data (see Figure 6), which can be confirmed analyzing the residuals (Figure 7). From the adjusted R^2 It can be inferred that the *weight* and *# of cylinders* can explain 81.85% of the variance of the output. Adding the am variable reduces the explained variance per degree of freedom of the fit:

```
##                fit Adjusted_R2
## 1: mpg ~ wt+cyl+am   0.8121603

##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 39.4179334  2.6414573 14.9227979 7.424998e-15
## wt          -3.1251422  0.9108827 -3.4308942 1.885894e-03
## cyl         -1.5102457  0.4222792 -3.5764148 1.291605e-03
## am1          0.1764932  1.3044515  0.1353007 8.933421e-01
```

Section 4 (Conclusion)

In this report a review of the mtcars data were performed in order to answer the question of how much mileage (if any) does the manual cars spend over automatic ones. A naive regression of **MPG** as a function of **am** indicates that Manual cars spend on average **7.25 ± 2.89** more MPG over Automatic cars (84% confidence interval, 1 std). However, as it was verified, the mileage difference is mainly due to the difference of weight and number of cylinders (**R^2 = 0.82**). An analysis maintaining these variables constant showed that there is no significant change in mileage (**am1=0.176 ;p-value > 0.89**) for cars with different transmission.

Appendix

a) Heatmap of data using correlation as distance metric

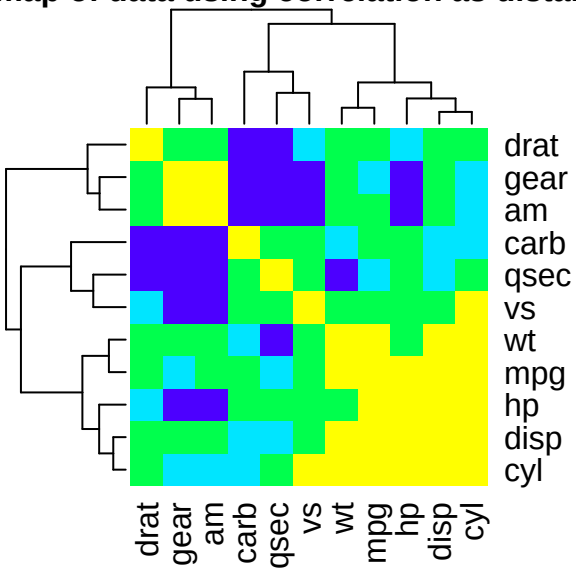


Figure 1. a)

b) Dendogram cut at height=0.5

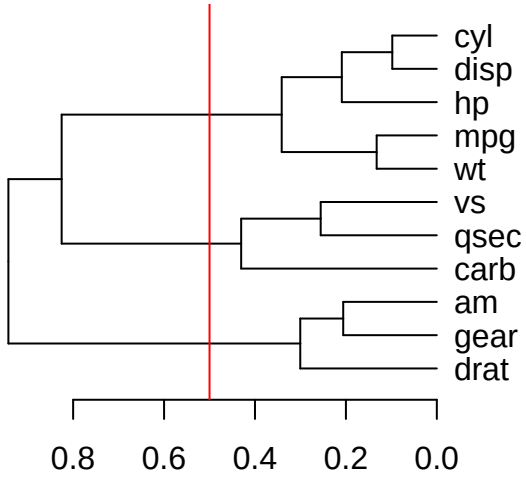


Figure 1. b)

Figure 2. Plot matrix of most correlated variables

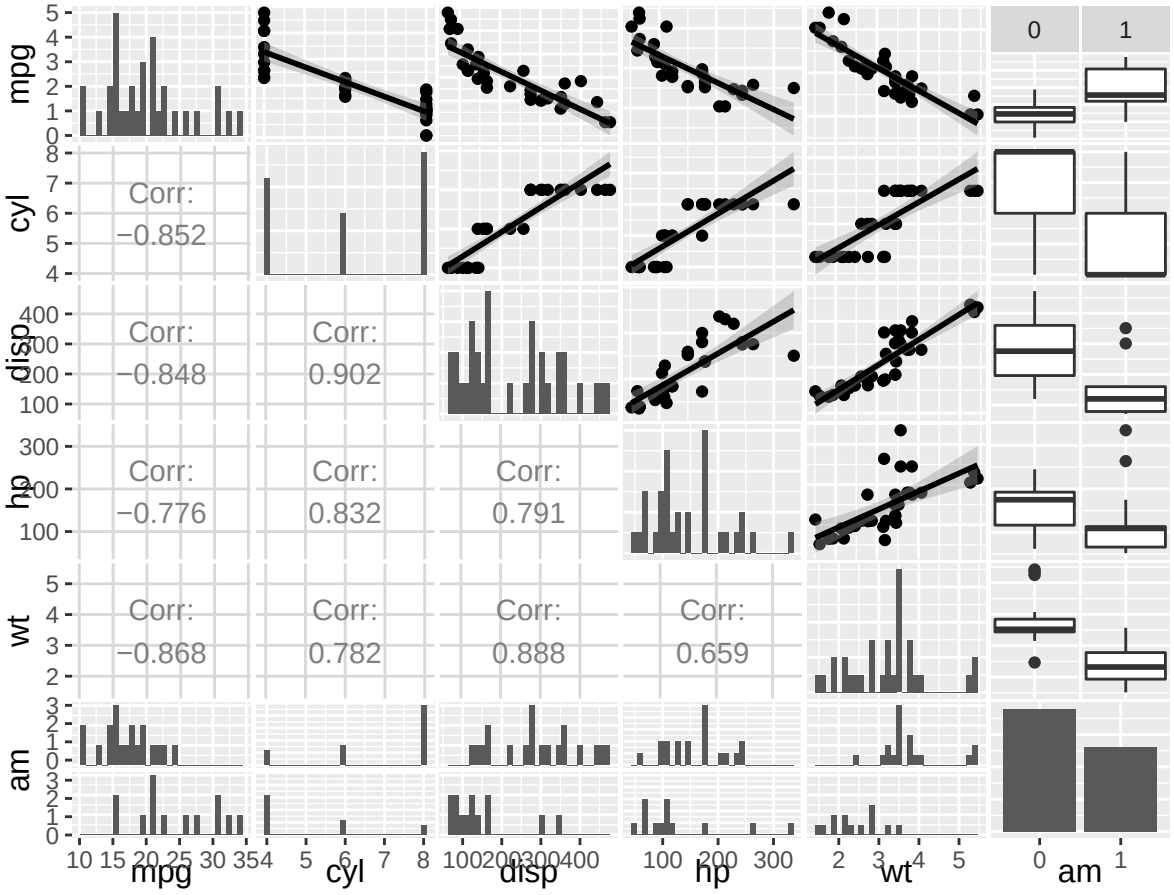


Figure 3. Residuals(mpg ~ am) vs MPG

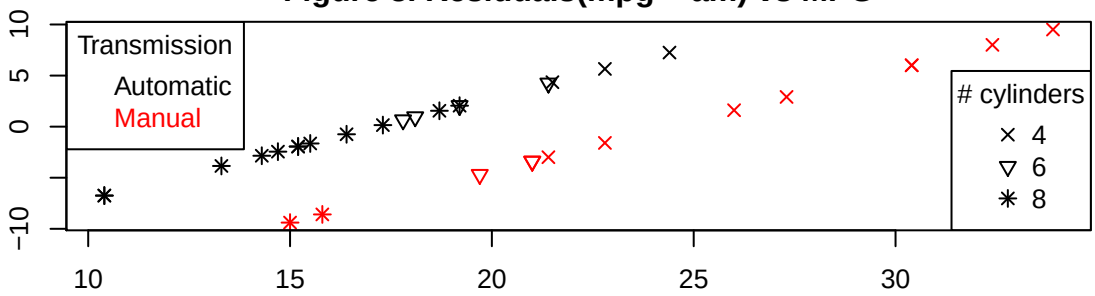


Figure 4. MPG vs wt grouped by am

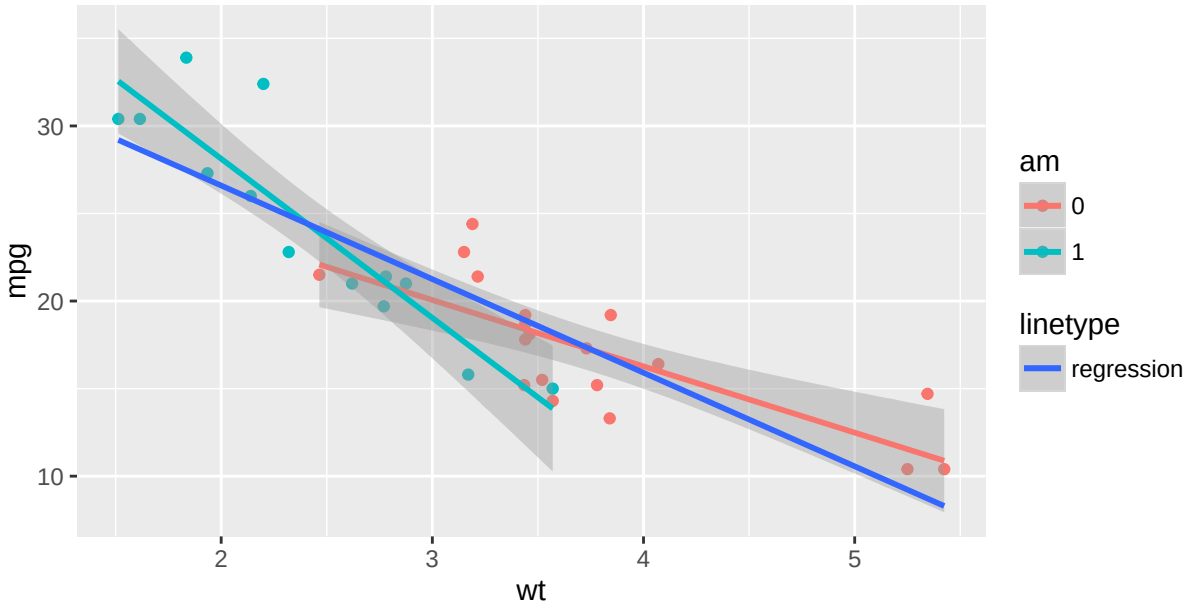
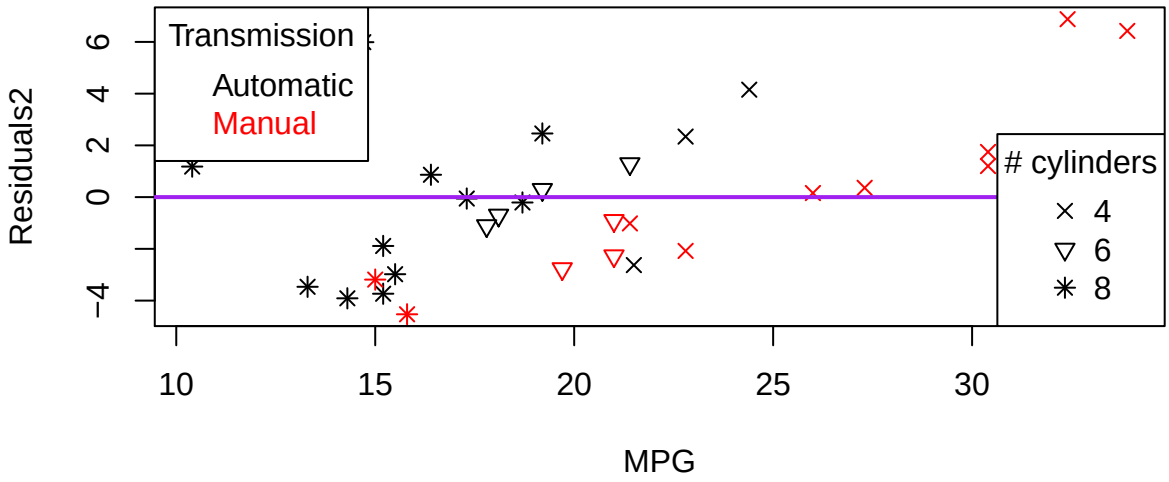
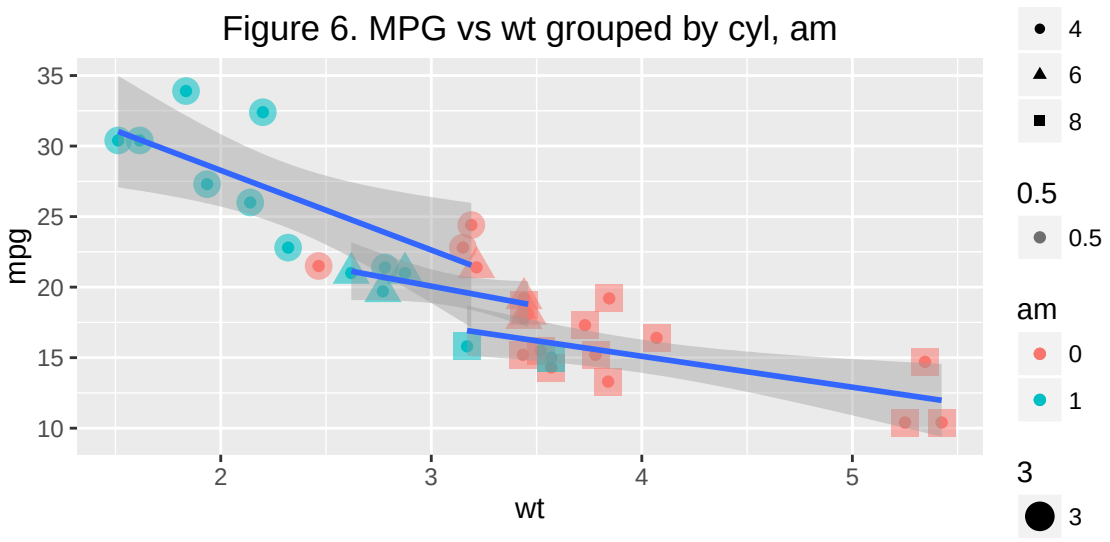


Figure 5. Residuals (mpg ~ wt+am) vs MPG






```
## [1] "Figure 7 Residual plot: plot(lm(mpg ~ wt+cyl,data=dt))"
```