Regression Models Course Project

Emilio González

Executive Summary

This work explores the relationship between a set of variables related to vehicles and miles per gallon (MPG) using a predefined dataset (mtcars). After an initial exploratory analysis a linear regression model including all the variables as regressors is performed. This model is refined in order to avoid problems of multicollinearity and interaction achieving a model with only three explanatory variables. Finally the work answers the question about which transmission is better in terms of MPG: manual cars seems to have a higher MPG.

Exploratory Analysis

The data set only has 32 cases. We can consider this as a small sample to extract very meaningful conclusions. Each row has 11 variables: the mpg (mileage per gallon) plus some other aspects of the automobile design and performance like number of cylinders, displacement in cu.in., horsepower, rear axle ratio, weight in lb/1000, 1/4 mile time in seconds, V or Straight engine, transmission, number of forward gears and numer of carburetors. The summary of these variables follows:

```
          mpg  cyl   disp     hp  drat     wt   qsec  vs  am gear carb
Min.    10.40 4.00  71.10  52.00 2.760 1.5130 14.500 0.0 0.0 3.00 1.00
1st Qu. 15.35 4.00 120.65  96.00 3.080 2.5425 16.885 0.0 0.0 3.00 2.00
Median  19.20 6.00 196.30 123.00 3.695 3.3250 17.710 0.0 0.0 4.00 2.00
3rd Qu. 22.80 8.00 334.00 180.00 3.920 3.6500 18.900 1.0 1.0 4.00 4.00
Max.    33.90 8.00 472.00 335.00 4.930 5.4240 22.900 1.0 1.0 5.00 8.00
Sd       6.03 1.79 123.94  68.56 0.530 0.9800  1.790 0.5 0.5 0.74 1.62
```

We can appreciate the center and variability data measures: mpg has a great range, there are three values for cylinders (4, 6 and 8), a big disparity in the displacement, horsepower, weight and number of carburetors variables and finally the 1/4 mile time can be considered in a very narrow timeframe compared to the dispersion of the other variables. Out of the 32 cases, only 13 are manual vehicles being the remaining 19 automatic.

The **Figure 1** is a plot of **mpg** versus weight, cylinder, displacement and transmissions in four different boxplots. The continuous variables have been stratified in groups for uniformity reasons. The graphs clearly reveal how the light cars have the highers **mpg** and the negative correlation of cylinders and displacement with **mpg**. Finally when dealing with transmission it is clear that manual vehicles provide higher **mpg** than automatic ones (and the difference is shown)

Regression Models

An initial study is the correlation of each variable with **mpg**, shown here in order of decreasing importance (whereas it is positive or negative). We can see that **wt**, **cyl**, **disp** and **hp** are the variables with a higher correlation with **mpg**. In the appendix in **Figure 2** the complete correlogram for all the variables is shown.

```
    mpg    wt   cyl  disp    hp  drat    vs    am  carb  gear  qsec
1  1.00 -0.87 -0.85 -0.85 -0.78  0.68  0.66  0.60 -0.55  0.48  0.42
```

An initial model will all the variables as regressors is not able to show what explanatory variables are **significantly** related to the response variable:

```
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) 12.303374  18.717884  0.65731 0.518124
cyl         -0.111440   1.045023 -0.10664 0.916087
disp         0.013335   0.017858  0.74676 0.463489
hp          -0.021482   0.021769 -0.98684 0.334955
drat         0.787111   1.635373  0.48130 0.635278
wt          -3.715304   1.894414 -1.96119 0.063252
qsec         0.821041   0.730845  1.12341 0.273941
vs           0.317763   2.104509  0.15099 0.881423
am           2.520227   2.056651  1.22540 0.233990
gear         0.655413   1.493260  0.43891 0.665206
carb        -0.199419   0.828752 -0.24063 0.812179
```

One approach to drop variables and look for an optimal set of explanatory variables is tune fining the model with step-wise selection using `step`. This updated regression model is much improved over the original. This new model accounts for interactions and collinearity including only three regressors.

```
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)   9.6178    6.95959  1.3819 1.7792e-01
wt           -3.9165    0.71120 -5.5069 6.9527e-06
qsec          1.2259    0.28867  4.2467 2.1617e-04
am            2.9358    1.41090  2.0808 4.6716e-02
```

In **Figure 3** different plots diagnose the regression, including QQ Plot for normality test and a density plot of residuals (see the legend). Finally with an analysis of variance we can test for the difference between both models verifying we are not losing effectiveness.

```
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     21 148
2     28 169 -7     -21.8 0.44   0.86
```

Coefficient Interpretation

The intercept gives the starting point. For each regressor (**holding the other ones constant**) the interpretation is the following: the **wt** coefficient tells us that **mpg decreases** by 3.9165 miles/gallon as the weight increases in 1000lbs (negative correlation); the **qsec** coefficient tells us that **mpg** increases in 1.22589 miles/gallon as 1/4 mile time increases by 1 second (slowest vehices have higher mpg) and finally the **am** coefficient tells us that manual cars enjoy 2.93584 miles/gallon more than automatic ones.

Question of interest: Which transmission is better for MPG? Quantification

The positive sign of the coefficient for **am** indicates a positive correlation (manual cars provide better mpg). Also looking in the appendix for **Figure 1**, the boxplot comparison for manual and automatic cars gives clear evidence of automatic ones suffering in terms of mpg. On average, the difference is **7.24494** Miles/(US)gallon in favour of manual cars. Automatic and manual cars are clearly two leagues respect **mpg**. In any case it is important to notice that the number of cases in the data set is small and hence uncertainty is big. Also as show in Cook's Distance plot of **Figure 3** some cases are influential observations having disproportional impact on the values of the model parameters.

Apendix. Figures

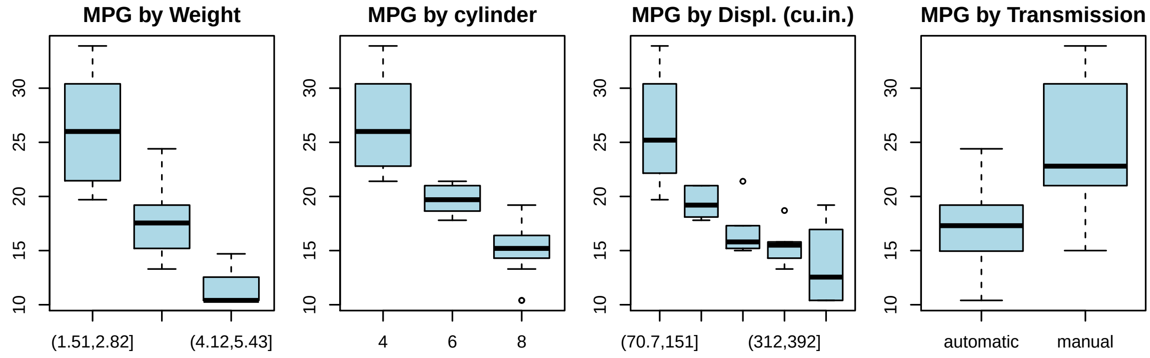

Figure 1. Boxplot multiple for variables weight, cylinder, displacement and am versus mpg.

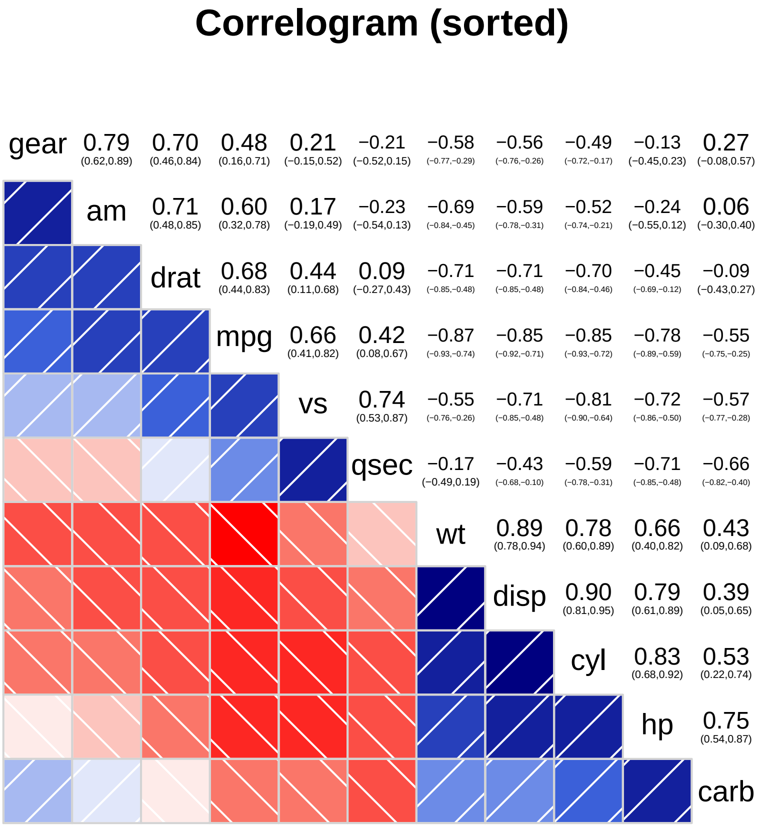

Figure 2. Correlogram of variables. On the lower-left part the correlation is shown easily in a visual way: intense darker colour shows high correlation (blue= positive, red=negative). On the upper-right part numeric correlations are shown

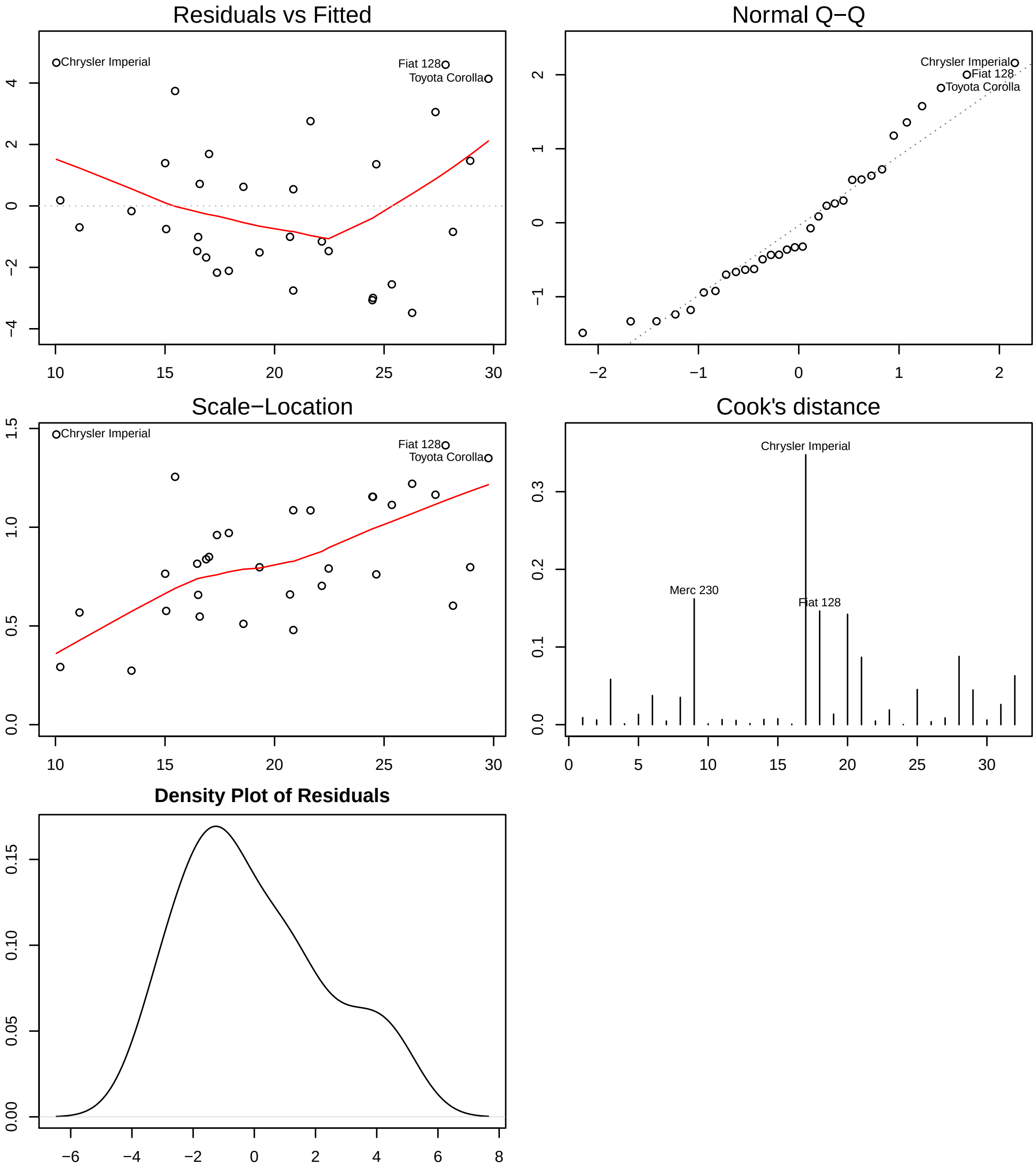

Figure 3 Regression Diagnostics and plot of residuals. The QQ Plot tells us we have violated the normality assumption (the points are not in the straight line). Linearity is also not met (Residuals vs Fitted). The Cook's distance shows us we have several influential observations that have great impact on the model.

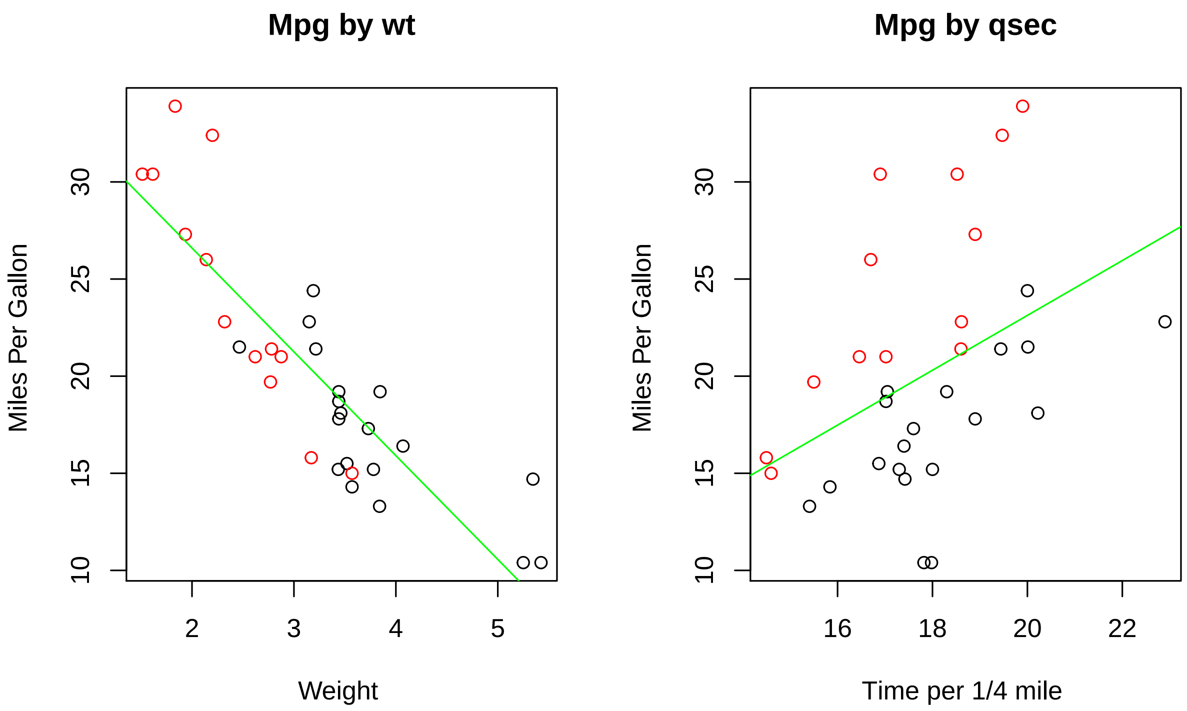

Figure 4 Indepent Scatterplot for the most important regressors with individual regression line but showing the difference between automatic (black) and manual (red) cars. It is easy to appreciate light cars are manual and with a better mpg while the heavier ones are principally automatic and provide less mpg. When looking at the 1/4 mile time vs mpg the most interesting finding is that when 2 vehicles have the same **qsec** always the manual one has a higher mpg (most of the manual cars are above the regression line).

```
# LM model with all the variables as regressors
fitall <- lm(mpg ~ ., data=mtcars)

# Step-wise refined model
fitstep <- step(fitall, direction="both", trace=0)

# steps taken in the search for the most effective model as variables are removed
# from the original model.
fitstep$anova

# Anova tset for comparison of both models
anova(fitall, fitstep)

# Computing the average in mpg between manual and automatic cars
mean(dplyr::filter(mtcars, am==1)$mpg)-mean(dplyr::filter(mtcars, am==0)$mpg)
```

Listing 1 R code for the key analysis performed