# The relationship between the fuel consumption and other properties of cars for early 70's models.

## Abstract

This report presents a study of the relationship between the fuel consumption and other properties of cars, in particular, the transmission. From this study I find that cars with manual transmission have, on average, a lower consumption of fuel (1973-74 models). Precisely, the difference in miles per gallon (MPG) between manual and automatic cars is 7.24 $\pm$ 4.04. When other variables such as weight or horse power, are taken into account, the trend is preserved (i.e. manual cars consume less on average) but the difference is smaller. This and other results are described in detail below.

## The Data

The data used for this study is the mtcars dataset within the datasets package in R. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 cars (1973–74 models). Details of the dataset can be found here (http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html). The only variable that is not self explanatory is perhaps VS, referring to the V engine (0 = V engine, 1 = straight engine).

## Exploratory analysis

An exhaustive exploratory analysis was carried out with the aim of adequately select those variables that have a strong relationship with the MPG. Due to space limitation only a fraction of the code is shown in this document. The entire code can be found at repo (https://github.com/jrzaurin/Regression_Models). Note that if the main purpose of this analysis is to investigate whether automatic or manual transmission is better for MPG, a "simple" t-test comparing the mean between the two groups will directly answer the question:

```
data(mtcars)
am0 <- mtcars$mpg[mtcars$am == 0]; am1 <- mtcars$mpg[mtcars$am == 1]
t.test(am0, am1, paired = FALSE, alternative="two.sided", var.equal=FALSE)
```

From these I find that the difference between the means MPG for manual and automatic cars is 7.24 $\pm$ 4.04, with a p-value of 0.001.

Similarly, t tests, or analysis of variance (ANOVA) tests when there are more than two groups, can be performed for all the other variables in the dataset that can be treated as groups or factors. These are: CYL, VS, GEAR and CARB. The full code can be found in the link provided above. When these variables are consider individually, they are *all* found to be significantly related with MPG. We will later investigate how these relationships are adjusted when more than one variable is taken into account.

At this stage it is important to add a caveat about the nature of the variables in the dataset. While VS and AM are Boolean, with 1 and 0 denoting certain characteristics of cars, CYL, GEAR and CARB are discrete numeric variables. Therefore, these can be treated as discrete numeric or factor variables, and the results of the analysis depend on that "selection". I will briefly discussed this issue in the Discussion section. Throughout the document, these variables are treated as numeric unless otherwise stated.

To explore the relationship between the remaining (continuous) numerical variables, Fig 1 shows a pairwise panel plot. As is clear from the panel, most of the variables are directly correlated with each other. Therefore, one has to be careful when using regression models since the presence of confounders is almost guaranteed.

Once we have a general idea of the potential relationships between the different variables, I proceed with a detailed regression analysis. To that aim I will first use a model comprising *all* variables, and then I will perform a backwards stepwise selection by both forward and backward exact AIC (Akaike Information Criterion (http://en.wikipedia.org/wiki/Akaike_information_criterion)).

```
library(MASS); full.model <- lm(mpg~., data = mtcars)
step <- stepAIC(full.model, direction="both", trace = FALSE); summary(step)
```

Interestingly, the results of this analysis show that the $\sim$ 84% of the variability of MPG can be accounted by a model comprising just WT, QSEC and AM, leaving out variables that are, a priori, important, such as HP or CYL. This is a clear evidence of multiple correlations within the dataset. Due to space limitation I will no discuss this in detail.

To further explore which variables are strongly related with MPG it is possible to use some more "sophisticated" tests. Even though we are bound to use *only* base packages for this work, the following R packages are extremely interesting and helpful for selecting a particular regression model. Therefore, I hope the reader finds them as useful as I did/do.

The following code/analysis performs model selection by exhaustive search, forward or backward stepwise, sequential replacement, and some other metrics (see the package leaps (http://cran.r-project.org/web/packages/leaps/leaps.pdf) and the repo repo (https://github.com/jrzaurin/Regression_Models) for details).

```
data(mtcars)
library(leaps);library(cluster); library(car)
leaps<-regsubsets(mpg ~ . , data=mtcars, nbest=1);summary(leaps)
subsets(leaps, statistic = "adjr2")
```

The results of this analysis are shown in Fig2 in the appendix. The figure shows that, effectively, once WT, QSEC and AM are included in the regression, the increase of the adjusted $R^2$ caused by including more variables is nearly negligible (or even negative), which further reinforces the previous result obtained using stepAIC. Finally, it is possible to perform an additional test that will allow to easily visualize the relative importance of each variable in the dataset, with respect to MPG:

```
library(relaimpo)
relimp <- calc.relimp(full.model,type=c("last","pratt"),rela=TRUE)
boot <- boot.relimp(full.model, b = 100, type = c("last", "pratt"), rank = TRUE, diff = TRUE
, rela = TRUE) ; plot(booteval.relimp(boot,sort=TRUE))
```

The results of this analysis are shown in Fig3 (details for relimpo can be found in relimpo (http://cran.r-project.org/web/packages/relaimpo/relaimpo.pdf)). As expected, is entirely consistent with all the previous tests (emphasize the only reason to include this test is to ease visualization).

## Results: model Selection

Based on the exploratory analysis described before, it is clear that the simplest model that explains most of the variability of MPG includes WT, QSEC and AM. The increase in percentage of variability explained caused by including any additional variable in the regression is negligible (from ~84% to 86%). Therefore, the *final* model used for this analysis is:

$$H_{final} = \beta_0 + \beta_1 WT + \beta_2 QSEC + \beta_3 1(AM = Manual) + e^*$$

where:

$\beta_0$ = 9.62 $\pm$ 14.25, $\beta_1$ = -3.92 $\pm$ 1.46, $\beta_2$ = 1.23 $\pm$ 0.59 and $\beta_3$ = 2.94 $\pm$ 2.88 (uncertainties are 95% confidence intervals).

It is notable that, among the three variables considered, the weakest correlation, as measured by the corresponding P-value, is that of AM (P-value = 0.047). Fig4, along with Fig1 show that the regression model fulfills all the theoretical requirements, i.e: linearity, normal residuals and constant variability.

## Summary and Discussion

The results of this study show that there is a significant relationship between the fuel consumption (MPG) and the transmission in cars (AM), for 1973–74 models. When analyzing the relationship considering *only* MPG and AM, we find that the difference in MPG between manual and automatic cars is, on average, 7.24 $\pm$ 4.04. Among the remaining variables, weight (WT) and quarter mile time (QSEC), along with AM, are those that explain the highest percentage of variability in MPG. When a model comprising these three variables is considered, I find that, everything else held constant, manual cars can reach, on average, 2.94 $\pm$ 2.88 miles farther per gallon of fueled consumed, compared to automatic cars.

Finally, I would like to briefly mention that the results of this study significantly change if number of cylinders (CYL), number of forward gears (GEAR) and number of carburetors (CARB) are considered as factor variables during the process (rather than discrete numeric). However, the number of observations for this dataset is particularly small (32). Therefore, when such approach is considered (which is actually the adequate approach), the variability of the variable MPG is spread over more regressors. The overall results is that the existing relationships are statistically weaker and therefore, less certain. If a significant number of regressors are considered as factors variables, a larger number of observations is required to adequately study the variability of a certain variable. Nonetheless, the code one would use to carry out such study can also be found at repo (https://github.com/jrzaurin/Regression_Models).

## Appendix

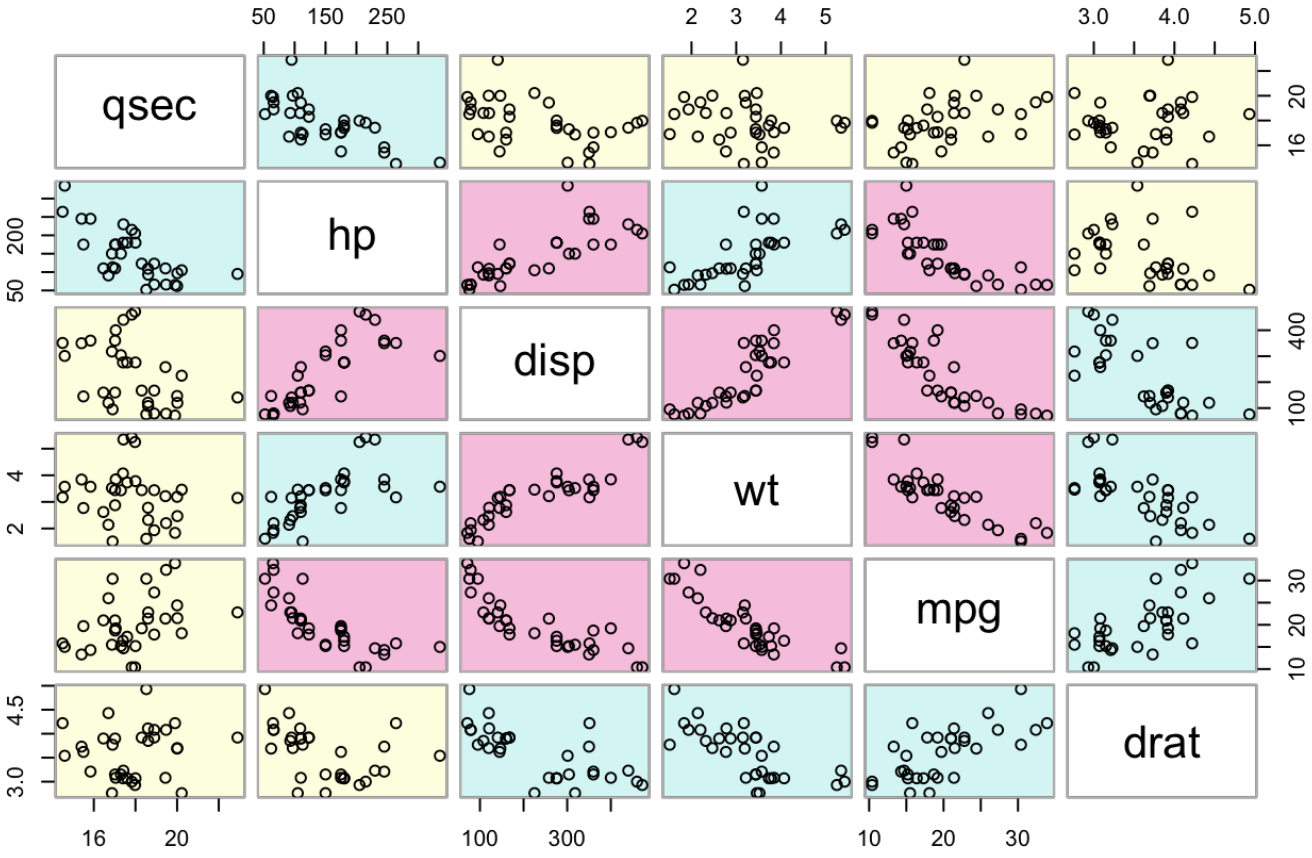


Fig1. pairwise plot for all the (continuous) numerical variables. The colors, and the proximity to the main diagonal code the strength of the correlation. Red: highly correlated. Yellow: weak correlation.

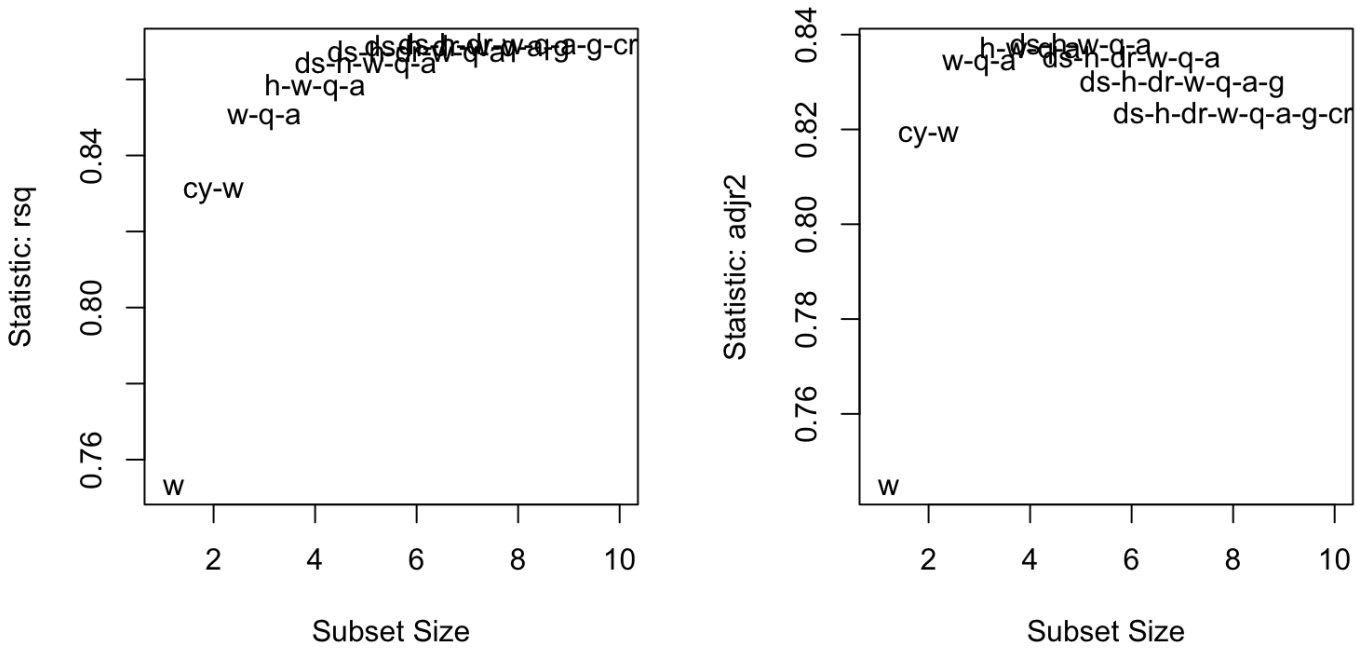
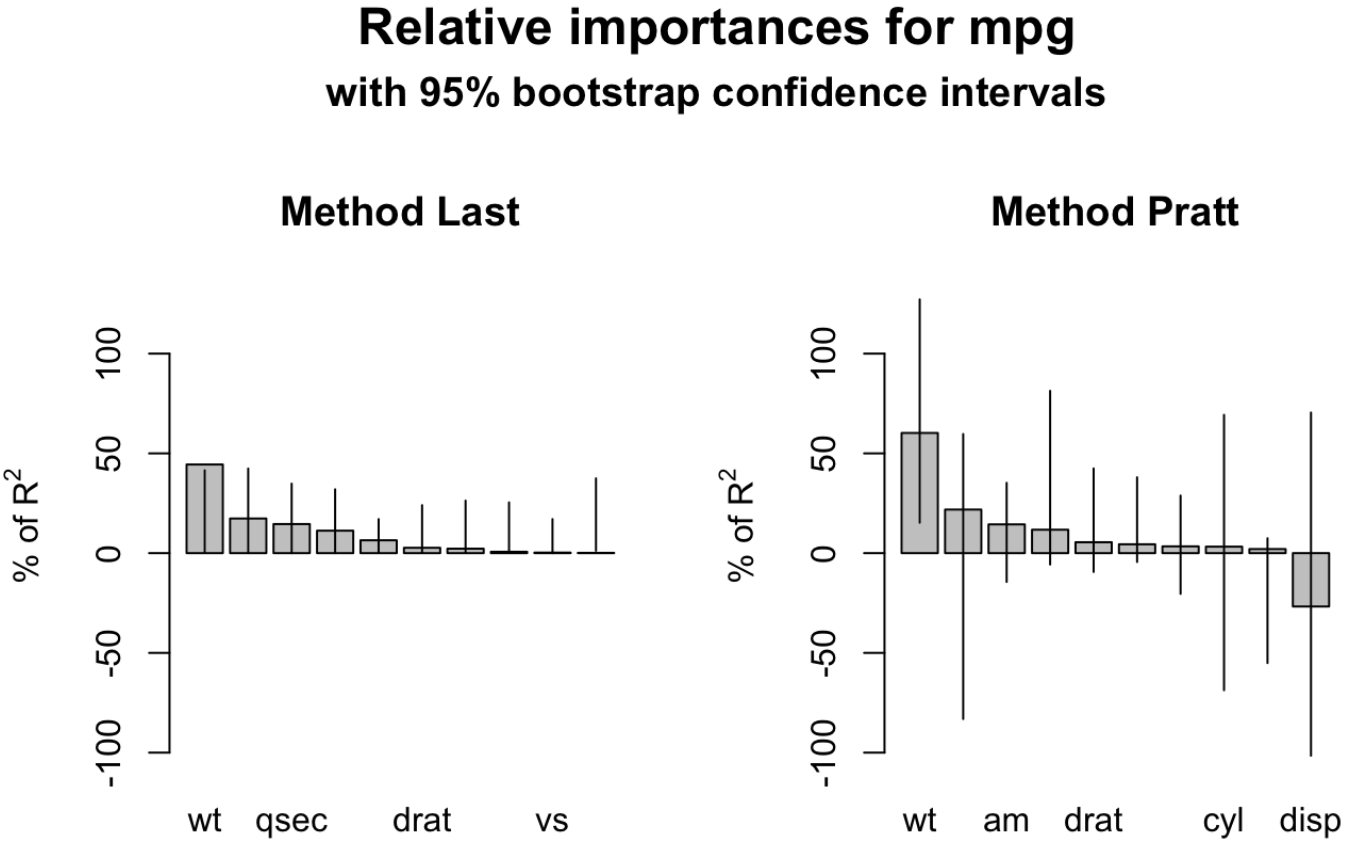



Fig2. The r-squared and adjusted r-squared statistic for the best fitting model comprising an increasing number of variables up to a maximum of 8. Given its exploratory nature, the figure is not intended to be aesthetically perfect. Nonetheless the abbreviations are straightforward and the results are clear, i.e. a model comprising WT+QSEC+AM explains most of the variability of the data.




$R^2$ = 86.9%, metrics are normalized to sum 100%.

Fig3. Relative importance of each variable in the dataset. Again, WT, QSEC and AM are the variables with the highest relative importance.

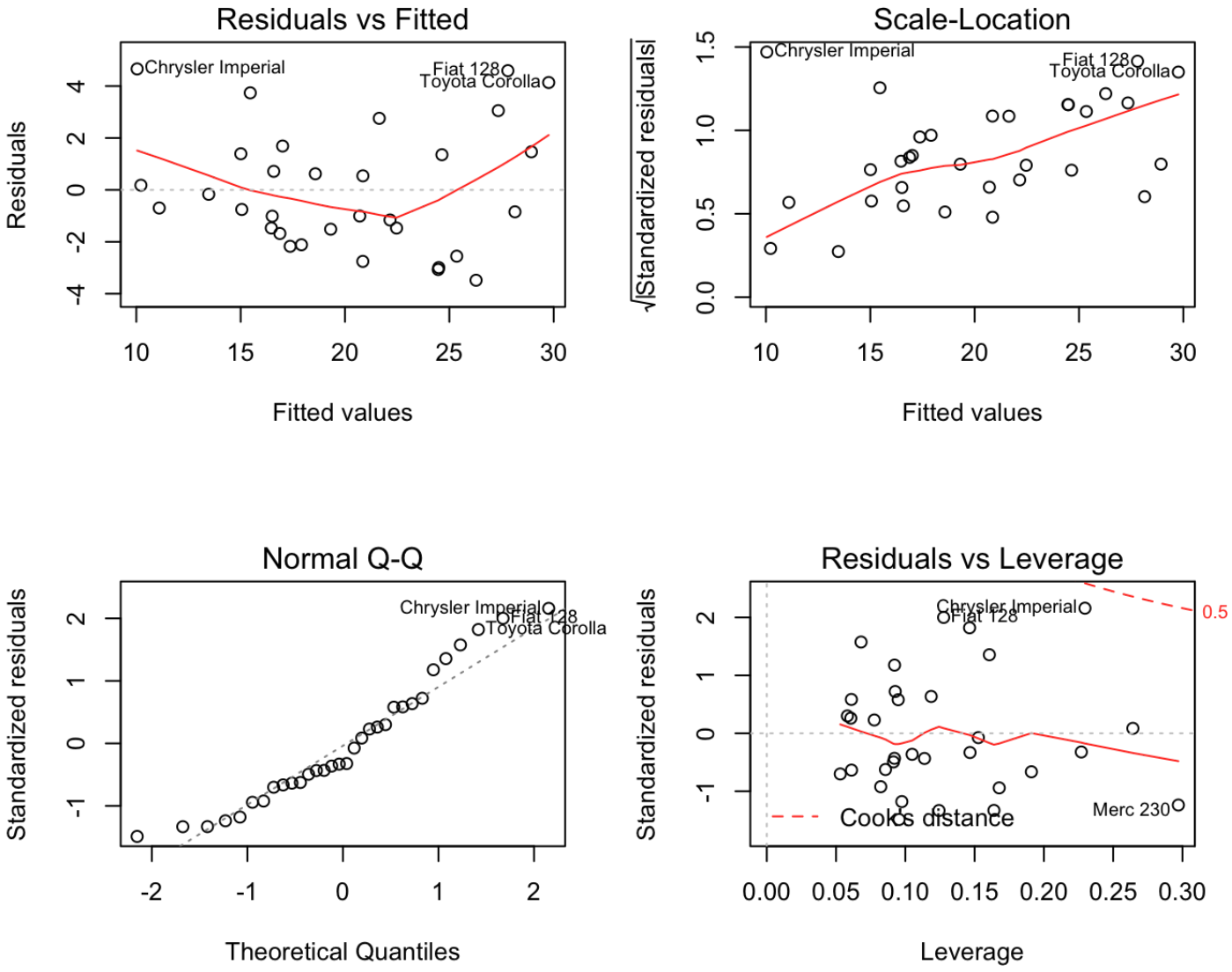



Fig4. The statistics showing that the behavior of the regression model is adequate in terms of normal residuals and constant variability. In addition, the leverage is also “controlled” within reasonable upper limits.