Peer-graded Assignment: Regression Models Course Project

Alejandro Osorio

June 29, 2018

Choosing the Model

As

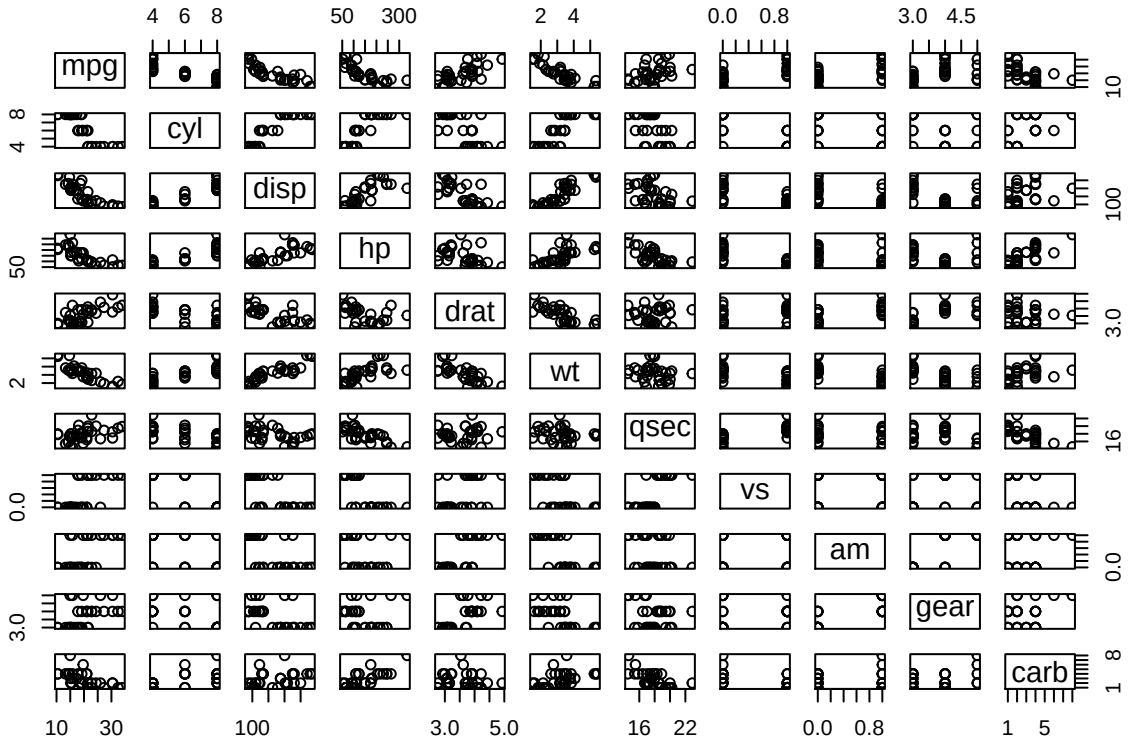
Appendix 1: Preliminary Data Analysis

Basic properties of the 'mtcars' dataset:

```
## '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 ...
```

So, we're talking of a small sample of only 32 observations.

Additionally, visual correlations among pairs of potential variables can be seen as follows:



Appendix 2: Model Selection

Given the nature of the analysis (continuous outcome, obtained from discrete and continuous regressors), the size of the dataset (only 32 observations), plus the visual correlations observed between 'mpg' (the outcome) and its potential regressors, the model to be used wil be linear (lm type), with 'am' as a not-so-dummy binary-factor variable.

Appendix 3: Choosing Preliminary Regressors

A preliminary analysis considered all mtcars variables as potential regressors. With variables 'cyl', 'vs', 'am', 'gear' and 'carb' as factors, the following lm function was analysed:

```
fitAll <- lm(mpg ~ ., mtcars2)
summary(fitAll)$coef
```

```
##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 23.87913244 20.06582026  1.19004018 0.25252548
## cyl6        -2.64869528  3.04089041 -0.87102622 0.39746642
## cyl8        -0.33616298  7.15953951 -0.04695316 0.96317000
## disp         0.03554632  0.03189920  1.11433290 0.28267339
## hp          -0.07050683  0.03942556 -1.78835344 0.09393155
## drat         1.18283018  2.48348458  0.47627845 0.64073922
## wt          -4.52977584  2.53874584 -1.78425732 0.09461859
## qsec         0.36784482  0.93539569  0.39325050 0.69966720
## vs1          1.93085054  2.87125777  0.67247551 0.51150791
```

```
## am1          1.21211570  3.21354514  0.37718957 0.71131573
## gear4        1.11435494  3.79951726  0.29328856 0.77332027
## gear5        2.52839599  3.73635801  0.67670068 0.50889747
## carb2       -0.97935432  2.31797446 -0.42250436 0.67865093
## carb3        2.99963875  4.29354611  0.69863900 0.49546781
## carb4        1.09142288  4.44961992  0.24528452 0.80956031
## carb6        4.47756921  6.38406242  0.70136677 0.49381268
## carb8        7.25041126  8.36056638  0.86721532 0.39948495
```

With a P-value based criteria, no variable makes the cut (with 'hp' and 'wt' the closest, though). Therefore, the first conclusion was that some industry research was required in order to determine the best regressor candidates for 'mpg' outcome.

After some web research (such as http://www.driverside.com/auto-library/top_10_factors_contributing_to_fuel_economy-317 and https://www.quora.com/On-what-factors-does-mileage-of-a-vehicle-depend), the main variables suggested (and therefore candidates for main regresors), were: displacement ('disp'), power ('hp'), aerodynamics, weight ('wt') and number of forward gears ('gear').

Besides aerodynamics (not available in the dataset), the visual correlations between them and 'mpg' are pretty clear, as seen in Appendix 1 (such as the inverse relation between 'mpg' and both 'hp' or 'wt' -both of which had the lowest P-values when including all variables in the model, as seen before-). Therefore, including 'am' (required in order to answer the main question), they were selected as candidates for regression modeling in the next appendix.

Finally, given the lowest P-value results obtained by 'wt' and 'hp' regressors (in comparison with the rest), plus well known Weight/Power KPI (Key Performance Indicator) from the industry, in further analysis both 'wt' and 'hp' will be considered primary potential regressors, while 'disp' and 'gear' will be considered secondary potential regressors.

Appendix 4: Regression Modeling

For the regression analysis, two types of models were used.

The first one with different intersection points but the same slope for all regression lines (no interaction between 'am' -the binary factor regressor- and the rest of the regressors), such as:

```
## lm(formula = mpg ~ hp + wt + factor(am), data = mtcars)
```

The second one, with different intersects AND slopes for regression lines (factor(am):regressor type interactions), such as:

```
## lm(formula = mpg ~ hp + wt + factor(am) + factor(am):hp + factor(am):wt,
##     data = mtcars)
```

Additionally, ANOVA comparisons (incrementally adding regressors from the preliminary set, defined in Appendix 3) were used in order to determine the best regressor mix for each model. The first regressor candidates, given

Different intersection points, with same slopes

Different intersection points and slopes