Analyze the 'miles per gallon' characteristic in cars

By: Oleg Rozenshtain

This report is written as a solution to the "Regression Models" course project, as part of the "data science" specialization by Johns Hopkins University on Coursera.

Executive summary

This report provides an analysis of the relationship between a set of variables and miles per gallon (mpg). The method of analysis used is multivariable regression. Results of data analyzed show that weight, 1/4 mile time and transmission type parameters are best to model and understand mpg. As we are mostly interested in the effect of the transmission type, the report finds that manual transmission adds about 3 miles per gallon.

Data overview

The data include 11 car characteristic for 32 automobiles (1973–74 models). The 11 characteristics are:

1. mpg - Miles/(US) gallon
2. cyl - Number of cylinders
3. disp - Displacement (cu.in.)
4. hp - Gross horsepower
5. drat - Rear axle ratio
6. wt - Weight (lb/1000)
7. qsec - 1/4 mile time (seconds)
8. vs - Cylinders form (0 = straight line, 1 = V form)
9. am - Transmission (0 = automatic, 1 = manual)
10. gear - Number of forward gears
11. carb - Number of carburetors

Exploratory data analyses

The first row in Fig.1 in the appendix shows us the relation between each individual parameter and the mpg. It is notable from Fig.1 that cyl, vs, am and gear may be treated as factor variables, so it will be more convenient to turn them from numeric to factor (Fig.2: box plot of mpg by each of the factors). We can notice straight ahead that cyl, disp, hp, drat, wt, vs, am, are potential regression variables for mpg (this can be also seen in the correlation matrix). Our main concern is the transmission effect so it is good that the types distributed nicely among the records (19-automatic, 13-manual).

Model

Trying to fit a linear model using all parameters results in insignificant coefficients in all parameters. That makes sense because probably not all parameters have relation with mpg, and some of the parameters are highly correlated between themselves which causes multicollinearity and high variance inflation. So using *step()* function in R, I determine which predictors fit the best regression model. The *step()* function uses AIC measure to exclude iteratively parameters that hurt the quality of the model. The result is that there are 3 predictors: weight, 1/4 mile time and transmission type. The model coefficients are:

```
> summary(lm(mpg ~ wt + qsec + am , data = newmtcars))$coeff
             Estimate    Std. Error    t value      Pr(>|t|)
(Intercept)  9.617781    6.9595930     1.381946    1.779152e-01
wt          -3.916504    0.7112016    -5.506882    6.952711e-06
qsec         1.225886    0.2886696     4.246676    2.161737e-04
am1          2.935837    1.4109045     2.080819    4.671551e-02
```

All coefficients are significant except the intercept. Furthermore the intercept doesn't have any meaning, so I want to move it so its value will be the average mpg for average weighted average 1/4 mile time, automatic cars.

```
> fit <- lm(mpg ~ I(wt-mean(wt[am == 0])) + I(qsec-mean(qsec[am == 0])) + am, data = newmtcars)
> summary(fit)
Coefficients:
                              Estimate  Std. Error   t value    Pr(>|t|)
(Intercept)                    17.1474     0.5641     30.398     < 2e-16 ***
I(wt - mean(wt[am == 0]))      -3.9165     0.7112     -5.507    6.95e-06 ***
I(qsec - mean(qsec[am == 0]))   1.2259     0.2887      4.247    0.000216 ***
am1                             2.9358     1.4109      2.081    0.046716 *

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497,        Adjusted R-squared:  0.8336
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Now all the coefficients are significant. The qsec parameter didn't seem to have any linear correlation with mpg in the exploratory analysis, but further investigation shows interaction between qsec and am (Fig.3). Exploring the residuals plot (Fig.4) shows good linearity in the Normal Q-Q plot which indicates the deviations from the model (everything we didn't measure) are normally distributed. Furthermore the Residuals vs. Fitted plot is pretty scattered which indicates a good model fit. Each additional regression parameter to the model would hurt the models significance. Characteristics such as number of cylinders and displacement, which we would initially expect to be part of the model aren't included because of high correlation with the weight which causes multicollinearity. It can be seen by calculating the variance inflation factors and see their high values.

Conclusions

The calculated regression model explains about 85% (R^2) of the variation in mpg which is pretty high, leaving only about 15% of uncertainty. Each additional regression parameter will improve a little this measure, but also will add a lot of variance inflation. The final model is:
mpg = 17.15 – 3.92*wt + 1.23*qsec + 2.94*(am==1)
The intercepts value was explained earlier. wt coefficient: for each lb/1000 increase in the cars weight, the car would go 3.92 miles less per gallon, if all other parameters held fixed. qsec coefficient: for each second increase in the cars 1/4 mile time, the car would go 1.23 miles more per gallon, if all other parameters held fixed. The signs of the coefficients make perfect sense. The main parameter in interest is the transmission type, so we can see that the type indeed affect the mpg, and in particular a change from an automatic to a manual transmission will increase miles per gallon by 2.94.

Appendix

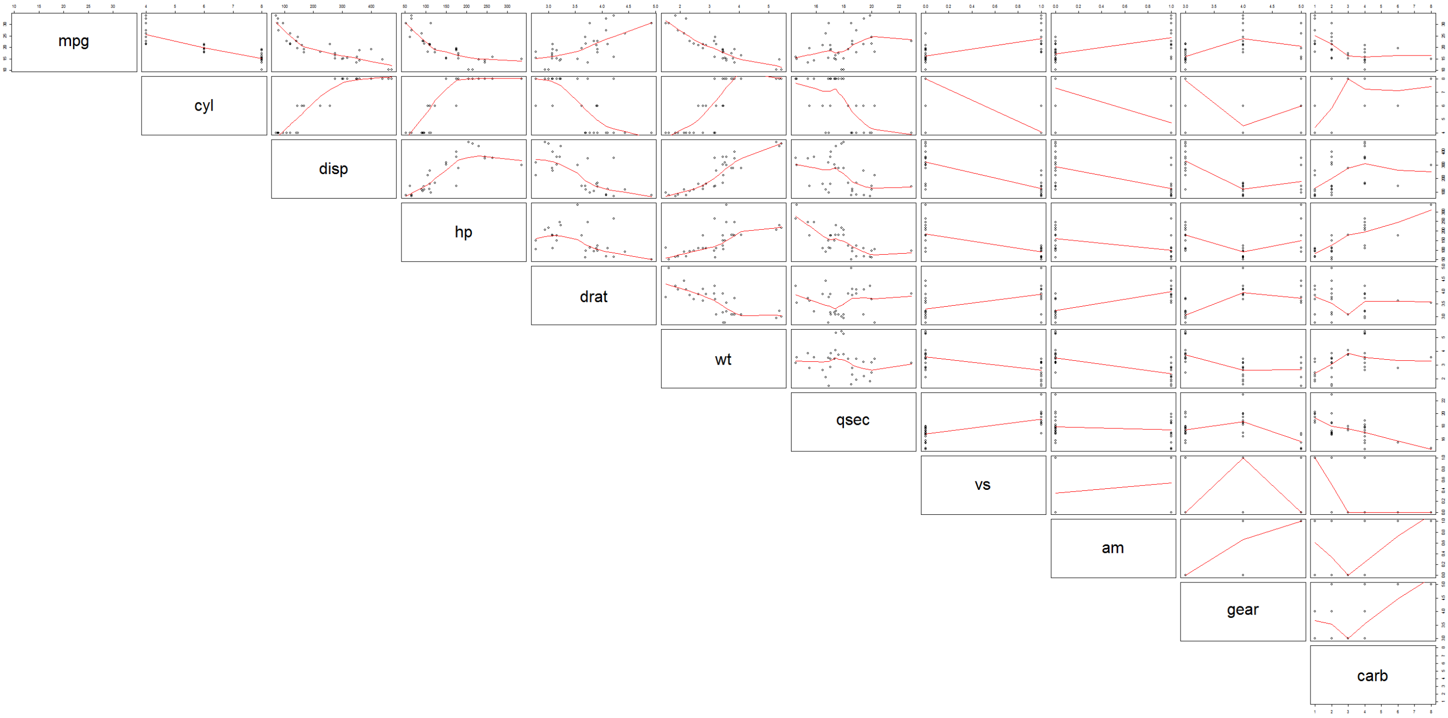

Fig1: scatter plot of all pairs of parameters

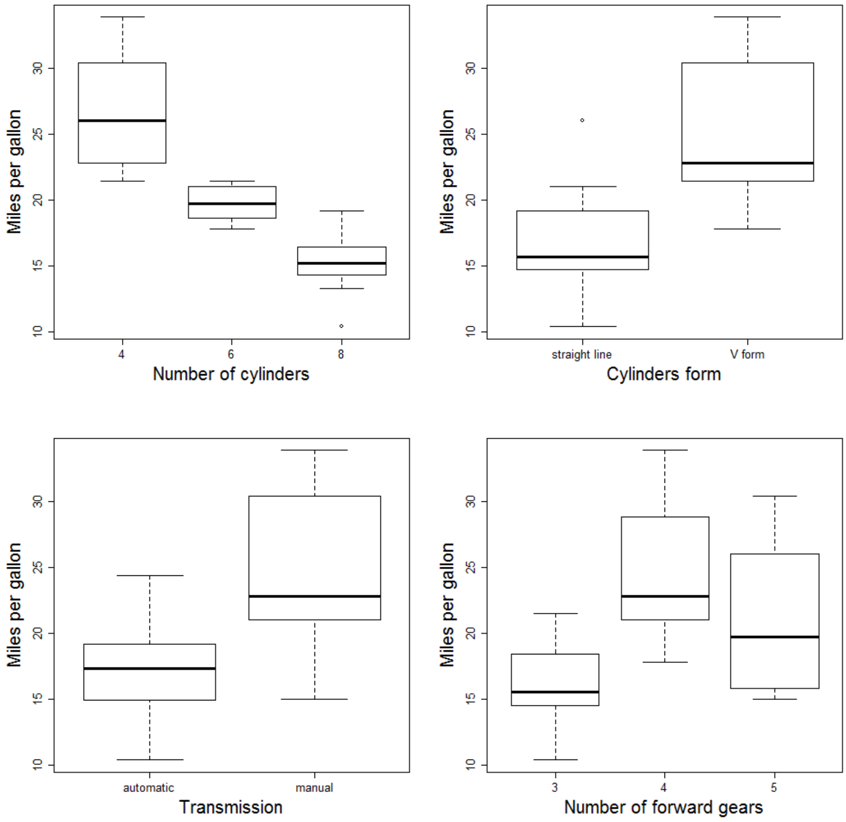

Fig2: box plot of mpg by each of the factors: cyl, vs, am, gear

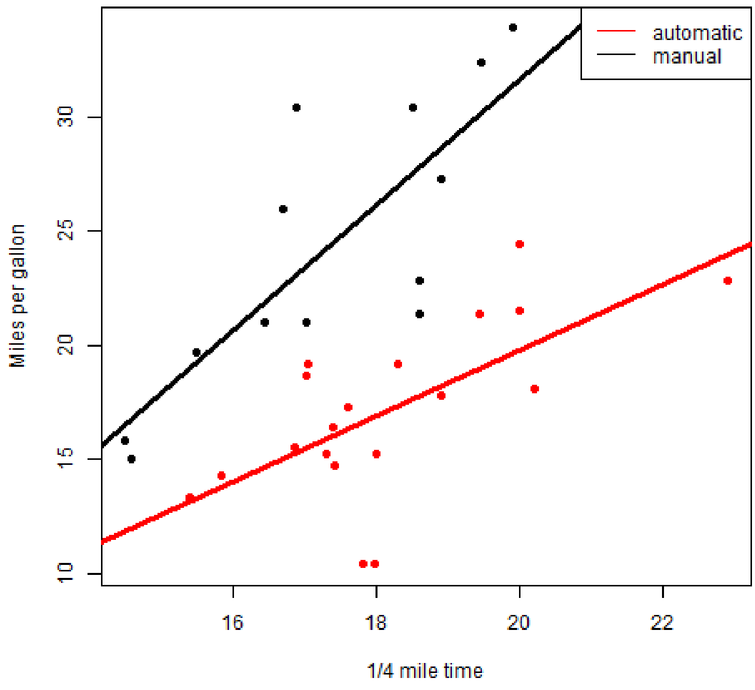

Fig3: interaction between qsec and am

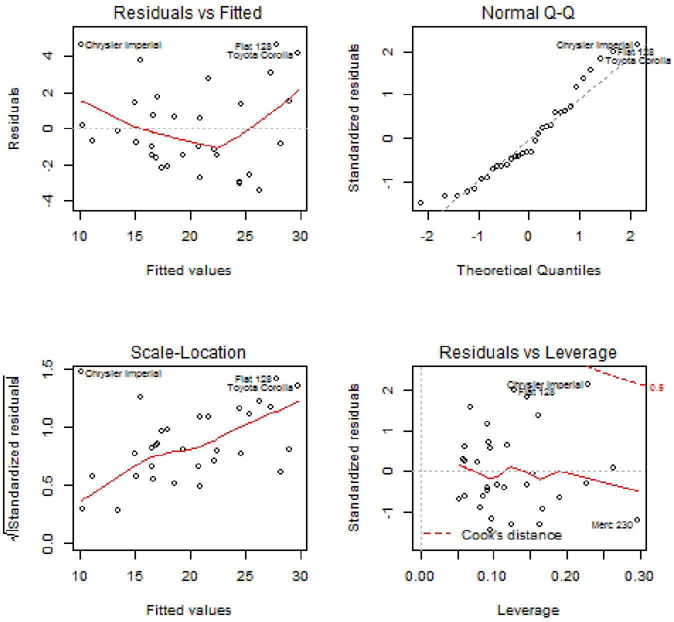

Fig4: residuals plot