Regression Models Course Project

Alexander Cherkashin

Tuesday, May 19, 2015

Summary

In this paper we'll look at the `mtcacars` - built into R data set and explore the relationship betweeen set of variables and miles per gallon (MPG). We'll try to answer is automatic (A/T) or manual (M/T) transmission is better for MPG and if it is then for how much better (higher MPG is better). We'll find that with `mtcars` data we can't be sure that transmission type has any influence.

Exploratory data analysis

The dimensions of `mtcars` are 32, 11. We have 32 different cars and 11 variables. Variables descrition is available by `?mtcars`. The mean MPG for A/T is 17.15 and for M/T is 24.39. Looks like M/T is better for MPG, but there are other variables that could affect MPG and thus adjust the transmission effect.

Relationship between variables and MPG. Model selection.

The simpliest regression model would be just take in account transmission factor variable (`am`): `lm(mpg ~ am, mtcars)` which yields us the interception coefficients for M/T and A/T. They are just the mean values from previous paragrath. This model is like two horizontal parralell lines at their intersepts with the difference `lm(mpg ~ am, mtcars)$coef[2]` = 24.39 - 17.15 = 7.24.

Let's look at other correlated to MPG variables. The correlation coefficients between MPG and other variables are: cyl: -0.85, disp: -0.85, hp: -0.78, drat: 0.68, wt: -0.87, qsec: 0.42, vs: 0.66, am: 0.6, gear: 0.48, carb: -0.55

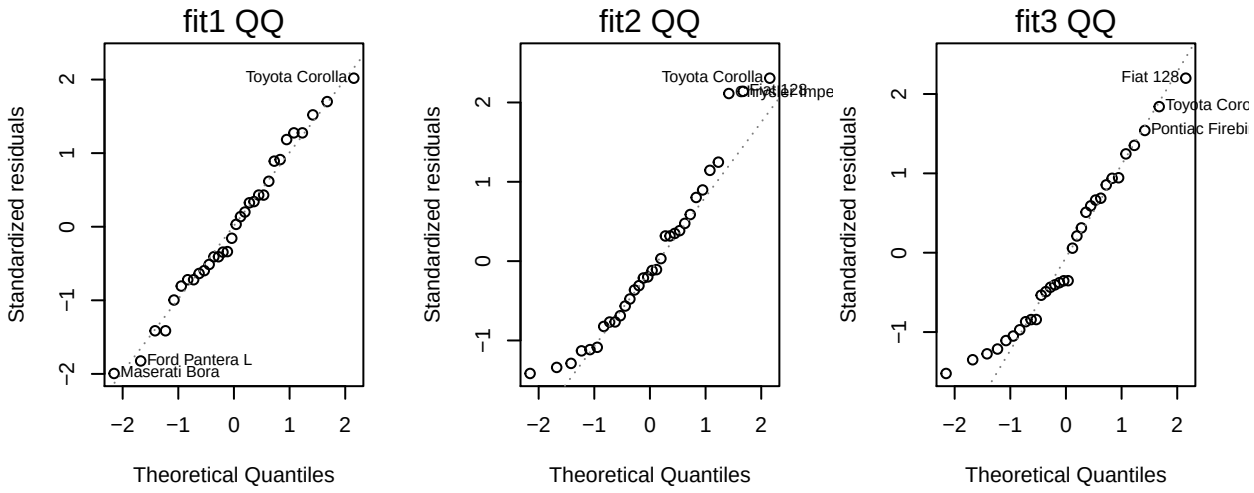
Clearly there are highly correlated variables. But to prevent variance inflation in beta coefficients with no positive overall effect we should carefully select them. The fist best candidate for a regressor is weight (`wt`). It is higly correlated with MPG and at the same time it should affect it from expirience: you need more energy to move more weight.

There should be another variable that reflects the effectivines of such moving of one point of weight, something with the engine that provides power. And we have some candidates: number of cyllinders (`cyl`) and gross hp (`hp`) of the engine are highly correleated with MPG. But there is high correlation between them: 0.83 and it seems that both of them could describe mostly the same part of mpg variance. Correlation between `wt` and `hp` is: 0.66, and it is less then between `wt` and `cyl` is: 0.78. We'll choose only `hp` for our models for that reasons.

We'll fit three models and compare them by ANOVA test: ordinary model with just `am` variable, model with `wt` and `hp` regressors, model with interaction between `wt` and `hp` in case their interdependancy is not linear for MPG.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
## Model 3: mpg ~ am + wt * hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 56.261 2.348e-10 ***
## 3     27 129.72  1     50.57 10.526   0.00313 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that `fit3` model has the lowest residuals squares sum (RSS) and the p-value is small enought that we could be confident in this result. To be sure we must check residuals normality. Lets, look at QQ plot's and perform `shapiro.test` on residuals:



The results of `sapiro.test`:

```
##         fit1   fit2   fit3
## p.value "0.86" "0.11" "0.16"
```

Residuals destributions appears to be close to normal.

Is an automatic or manual transmission is better for MGP

To answer this question we choose `fit3` model. Let's look at `summary(fit3)$coef`:

```
##                Estimate  Std. Error     t value     Pr(>|t|)
## (Intercept) 49.45224079 5.280730731  9.36465866 5.694894e-10
## am           0.12510693 1.333430965  0.09382333 9.259423e-01
## wt          -8.10055755 1.789325217 -4.52715777 1.084926e-04
## hp          -0.11930318 0.026549992 -4.49352965 1.187315e-04
## wt:hp        0.02748826 0.008472529  3.24439879 3.130390e-03
```

From `mtcars` dictionary `am` = 0 mean A/T and `am` = 1 means M/T.

`am` coefficient is 0.125 and it should indicate that car with M/T could travel further then a car with A/T by 0.125 miles for each gallon if both cars has the same weight and horse power. But the p-value of `am` coefficient is 0.926 high and thus **we can't reject the null hypothesis that the true impact (coefficient) of `am` iz zero**.

Answering the question of this paper we can't say that either A/T or M/T is better for MPG. It looks like that weight and horse power have the most and sustainable impact and **true influense of transmission could be zero with p-value 0.926**. On a given data with `fit3` model.

For reproduceability purpose all the code is availible at github: https://github.com/amchercashin/RM_CourseProject.