How does transmission affect MPG: a data exploration on mtcars in R

Chun Fang

11/7/2016

The data analytic question

Given mtcars data, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). In particular, we answer the following two questions.

- Is an automatic or manual transimission better for MPG?
- Quantify the MPG difference between automatic and manual transimssions.

Checking the data

The mtcars data is collected in R datasets and one can obtain its basic information by running *?mtcars* in the R console or going to the following link:

https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html

In short, the data mtcars was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). To fit our questions, note that for transmission (coded as am) variable, 0 means automatic and 1 means manual.

Exploratory analysis

Recall that our purpose is to investigate the relationship between transimission and MPG. Note that each observation of the data records a value of mpg with respect to a SINGLE value of transimission for a car. However, we are not clear the comparision of mpg values under different transimissions of the same car. So it might be missleading if we only fit a linear model mpg ~ am to answer even the first question, because a third regressor can distort the analysis. As a result, we decide to choose a better model among relations mpg ~ me + others beased on Akaike Information Criterion (AIC). In R, we realize it by using function *step()*.

First, let have a look at the structure of the data.

```r
data(mtcars)
str(mtcars)
```

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

```r
unique(mtcars$carb)
```

```
## [1] 4 1 2 3 6 8
```

Having aims in the mind, we tidy the mtcars data so that its variables am, cyl, vs, gear are factors. Here, we didn't convert the variable carb to factor, because it has 6 values while we have only 32 obs..

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
mtcars <- mutate(mtcars, am = as.factor(am), cyl = as.factor(cyl),
                         vs = as.factor(vs), gear = as.factor(gear))
```

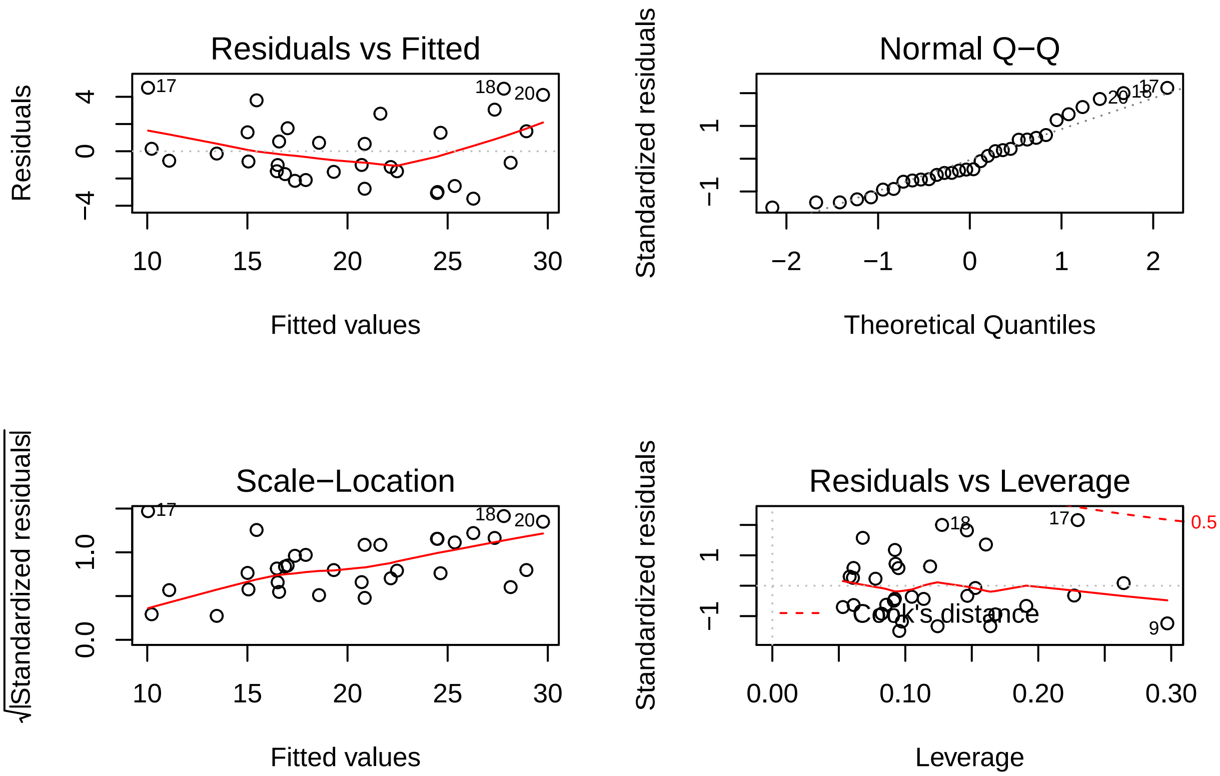
Statistical modeling and inference

```r
fit <- lm(mpg ~ ., data = mtcars)
stepModel <- step(fit, trace = 0)
summary(stepModel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am1           2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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, let us have a look at the residual plot.

```r
par(mfrow = c(2,2))
plot(stepModel)
```



From residual plot, we don't find a special pattern, which is good.

Summary

From above analysis, we conclude that the model mpg ~ am + wt + qsec explained 84.97% variation of the data. The p-value is 1.21e-11. While we are not satisfy the Pr(>|t|) value of Intercept.

Based on obtained result on mtcars, we conclude that automatic transimission is better for MPG, since under the same conditions, automatic transimission can save 2.9358 (US) gallon per mile than manual transimission.

Appendix

Pairwise plot matrix for revised data mtcars

```r
library(GGally)
```

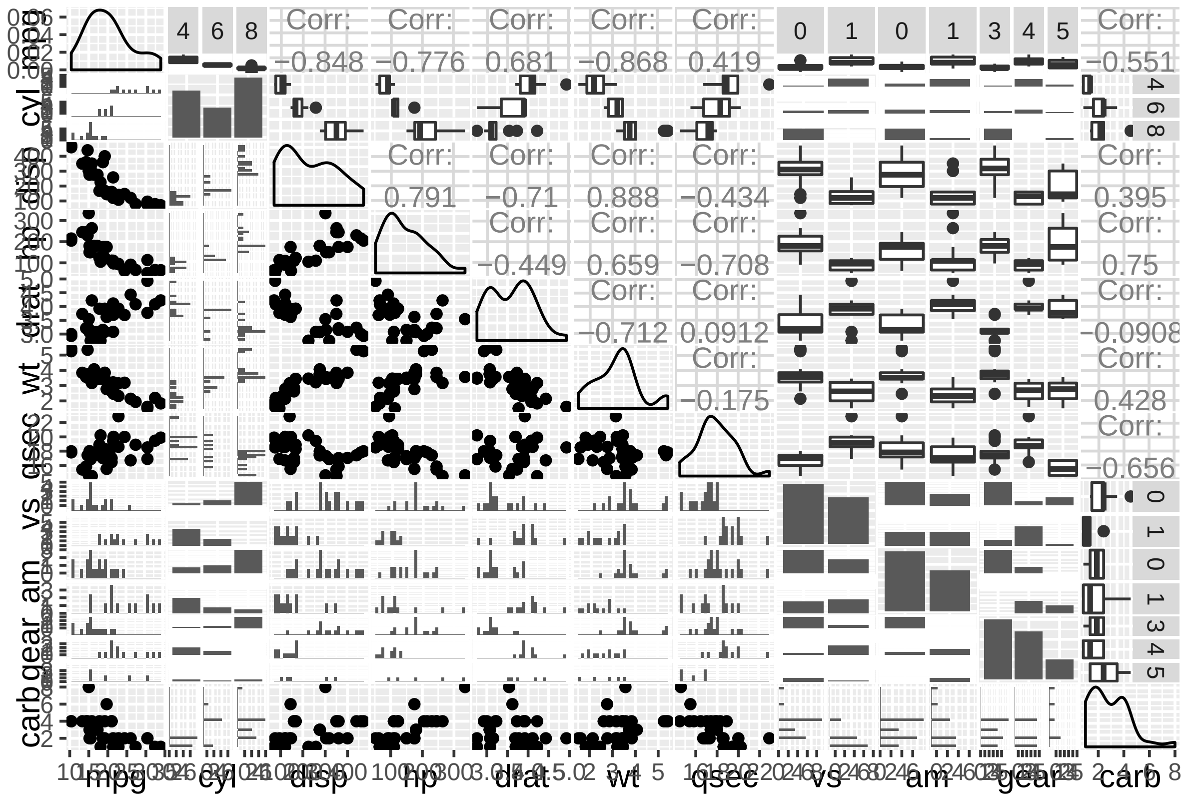
```
##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

```r
ggpairs(mtcars)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
Corr:
−0.848
−0.776
0.681
−0.868
0.419
−0.551
0.791
−0.71
0.888
−0.434
0.395
−0.449
0.659
−0.708
0.75
−0.712
0.0912
−0.0908
−0.175
0.428
−0.656