Miles per gallon for manual vs automatic transmission

Piotr Turski

October 24, 2015

Executive summary

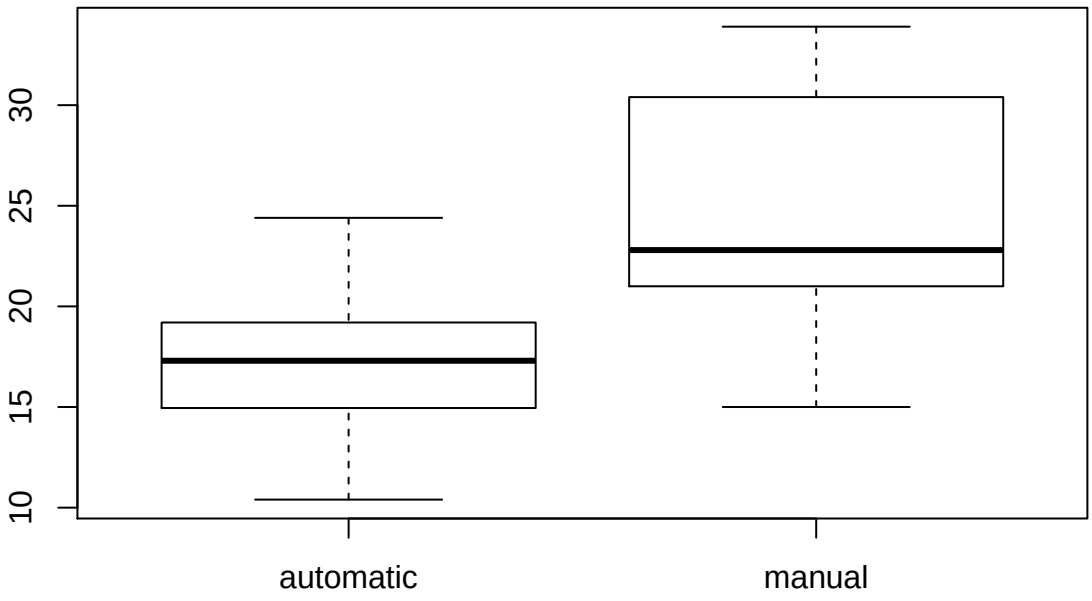
Based on car parameters given in R dataset `mtcars` we conclude that change in miles per gallon for transmission type depends on car's weight. For cars lighter than 3400lbs manual transmission is better. For heavier cars automatic transmission is better. More precisely: for a car with weight x, the average mpg increase for manual transmission compared to automatic transmission is about 14.08 - 4.14x.

Exploratory data analisys

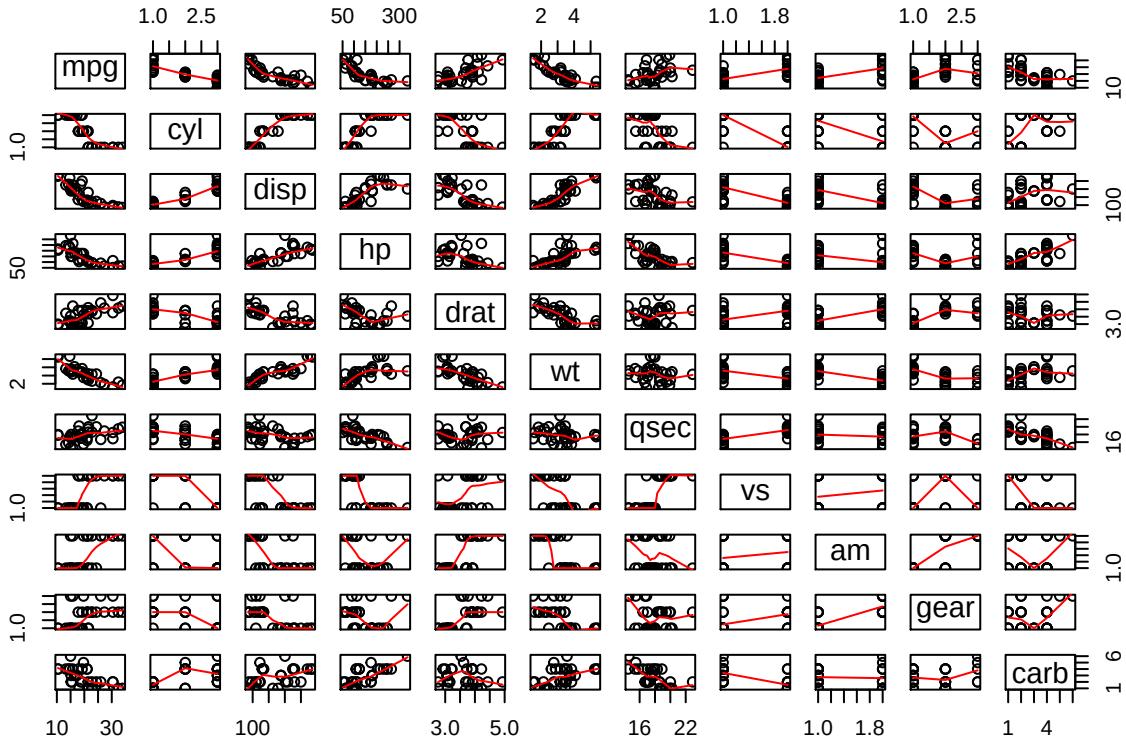
```
data(mtcars)
str(mtcars)
```

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

```
library(dplyr)
cars <- mtcars %>% mutate_each(funs(factor), vs, cyl, gear, carb) %>%
        mutate(am = factor(am, labels=c('automatic', 'manual')))
rownames(cars) <- rownames(mtcars)
boxplot(mpg ~ am, cars)
```




```
pairs(cars, panel = panel.smooth)
```



Model selection

Let's compare a few simple models. Let's start with smallest and full. Looking on the pairs diagram and common sens, good predictors for mpg may be wt, cyl, hp, qsec. Let's see how much variation can we explain using some simple, reasonable models:

```
sapply( c(
  mpg ~ am, mpg ~ ., mpg ~ cyl, mpg ~ hp, mpg ~ qsec, mpg ~ wt,
  mpg ~ cyl + wt + qsec + am, mpg ~ cyl + qsec + am, mpg ~ wt + qsec + am,
  mpg ~ cyl + wt + am, mpg ~ hp + qsec + am
), function(formula){summary(lm(formula, cars))$adj.r.squared})
```

```
##  [1] 0.3384589 0.7790215 0.7140090 0.5891853 0.1478062 0.7445939 0.8311946
##  [8] 0.7342857 0.8335561 0.8134405 0.7654448
```

We got best score using `mpg ~ wt + qsec + am`. Based on `pairs` plot, `wt` and `am` are somehow correlated. Let's see if we can improve the model by adding interaction between those parameters:

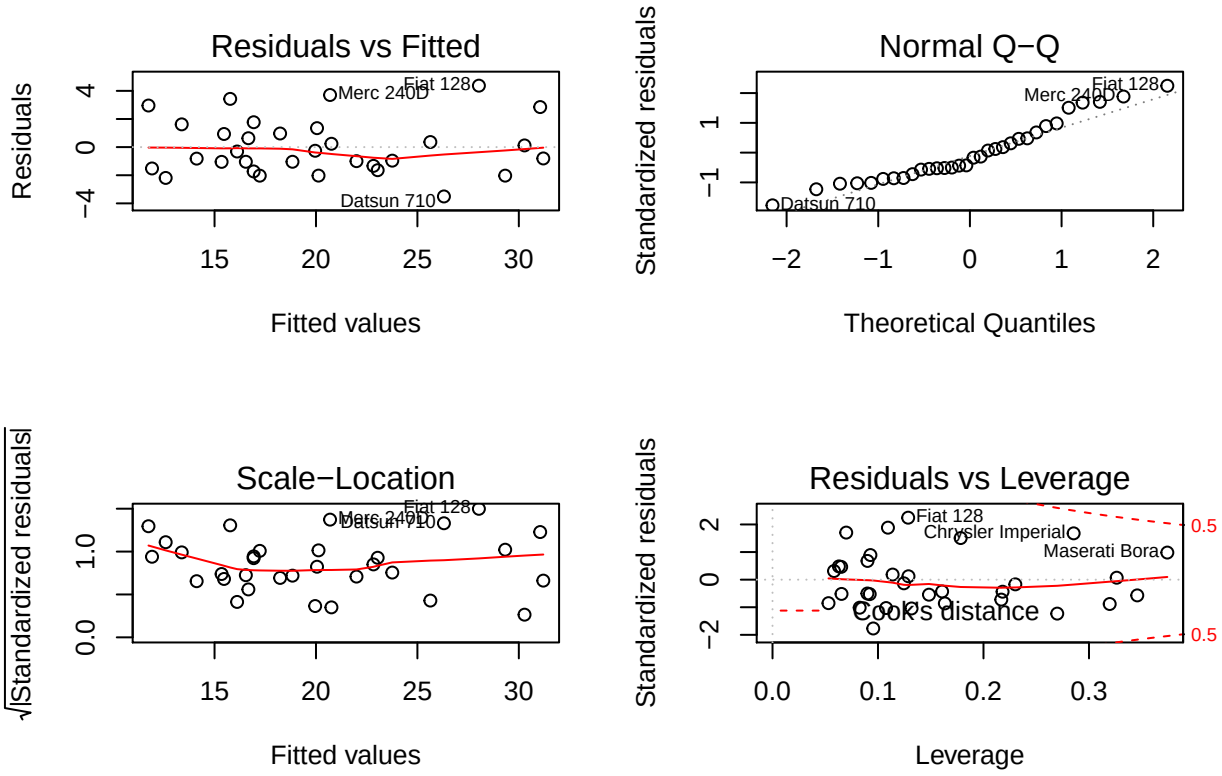
```
fit <- lm(mpg ~ qsec + am * wt, cars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ qsec + am * wt, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## qsec           1.017      0.252   4.035 0.000403 ***
## ammanual      14.079      3.435   4.099 0.000341 ***
## wt            -2.937      0.666  -4.409 0.000149 ***
## ammanual:wt   -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

88% seems good enough and have all coefficients statistically significant.

Diagnostics

```
par(mfrow = c(2,2))
plot(fit)
```



Based on the diagnostic plot, residuals seem normally distributed without any additional patterns.

Conclusion

```
summary(fit)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## ammanual    14.079428  3.4352512  4.098515 0.0003408693
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## ammanual:wt -4.141376  1.1968119 -3.460340 0.0018085763
```

Interpretation: For a car with weight x, the average mpg increase for manual transmission compared to automatic transmission (having other parameters constant) is about 14.08 - 4.14x. So for cars lighter than 3400lbs manual transmission gives better mpg. For heavier cars, automatic transmission is better.

Uncertainty: All the coefficients are statistically significant for the 0.99 significance level.