Influence of transmission type on MPG

Lukasz Konczyk

13 maja 2018

1. Explotatory data analysis

Introduction

In this report, we are focused on which type of transmission: automatic or manual, has better influence on MPG. Data which are used to analysis comes from *mtcars* dataset from basic R library.

Data loading and brief summary

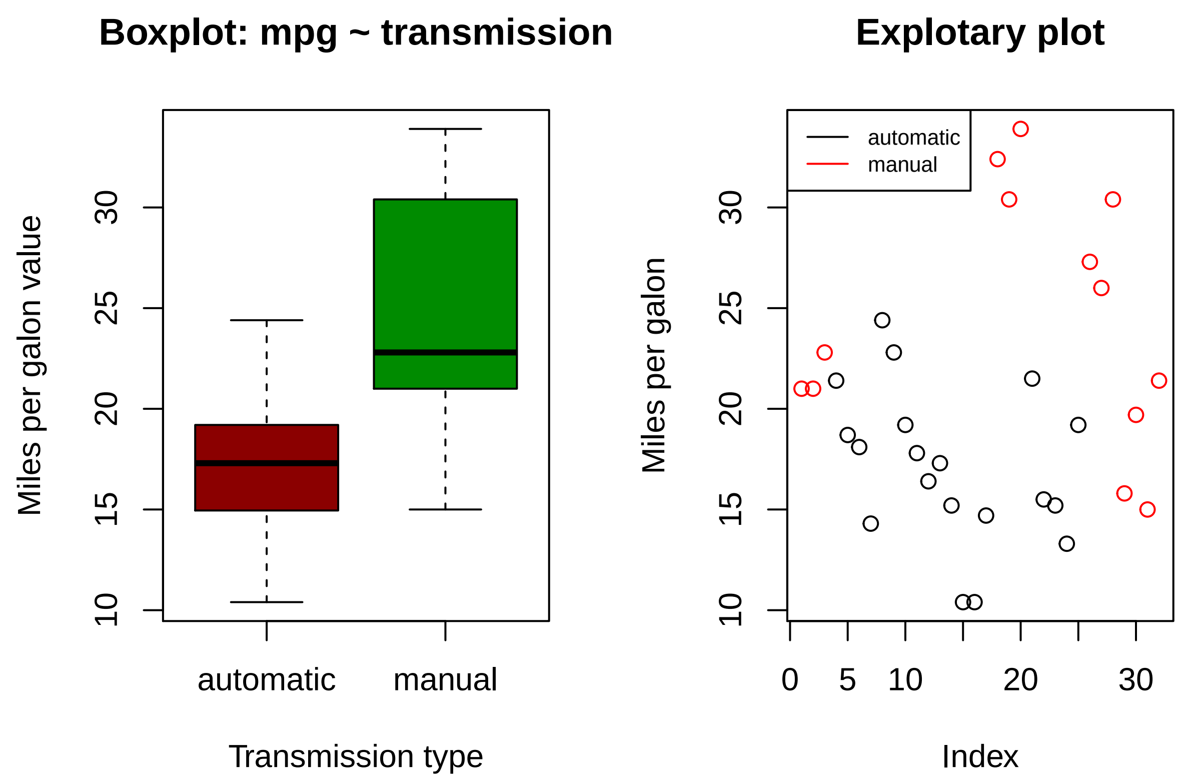
First thing to do, is loading dataset and briefly look at this.

```
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 ...
```

We are trying to answer the question if an automatic or manual transmission is better for MPG. In our data we will used variables **mpg** and **am** in first basic model. Variable **am** will be used as factor with two values: 0 for automatic transmission and 1 for manual transmission. Variable **mpg** is continous variable with positive values. Let's have a look on plots below to see any pattern.

```
par(mfrow=c(1,2))
with(mtcars,boxplot(mpg~factor(am,labels=c("automatic","manual")),col=c("red4","green4"),xlab="Transmis
                    ylab="Miles per galon value",main="Boxplot: mpg ~ transmission"))
with(mtcars, plot(mpg,col=am+1,ylab="Miles per galon",main="Explotary plot"))
legend("topleft", legend=c("automatic","manual"), col=1:2,cex=0.7,lty=1)
```



Looking at plots above, we can observe that manual transmission cause larger use of fuel expressed by miles per galon (MPG). But if there exist any linear relationship between fuel usage and transmission type? We will try answer this question using linear regression models. Also we will try what is difference in average mpg value for automatic and manual transmission.

2. Linear regression model fitting

Model with only two variable

Let's look on the most basic model.

```
fit<-lm(mpg~factor(am),data=mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## factor(am)1    7.245      1.764   4.106 0.000285 ***
```

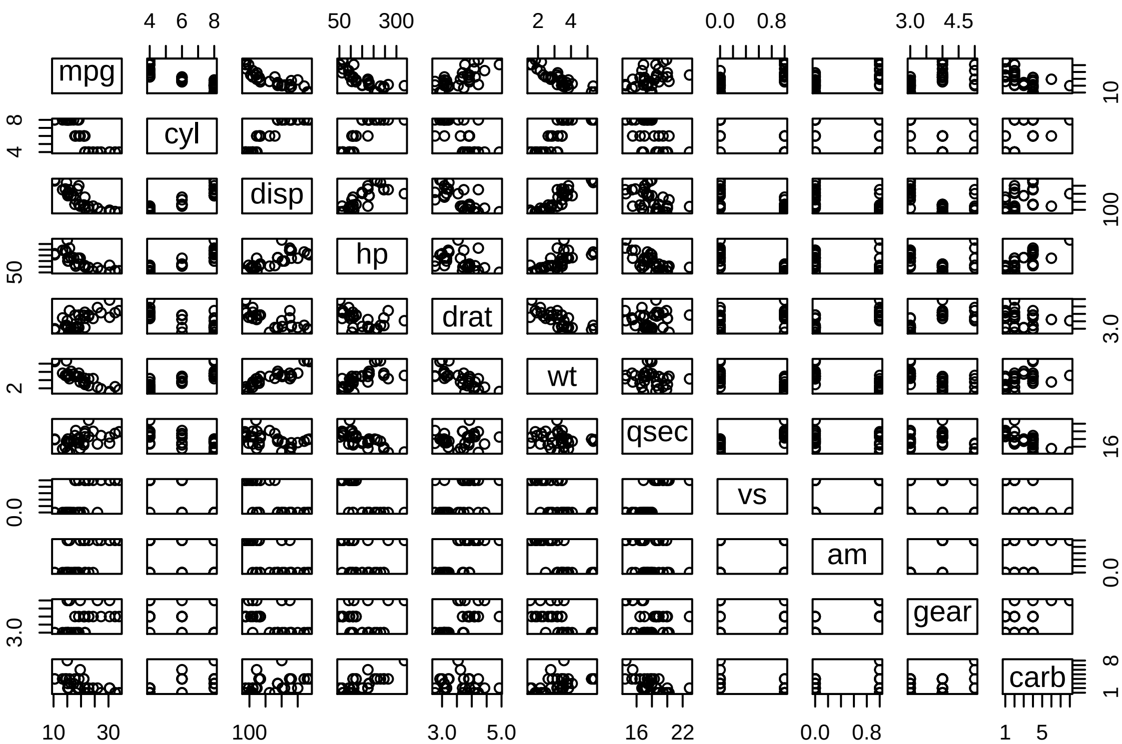
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Quick look on coefficients. Intercept coefficient is equal 17.1473684 - and according to linear model theory it is mean value of **mpg** for automatic transmission. Scope coefficient is equal 7.2449393 and it is equal of increase in average **mpg** for manual transmission, so average **mpg** for manual transmission is sum of coefs: 24.3923077. P-values in summary for coefficients is significant less than 0.05 border and in both cases hypothesis of 0 value of coef is rejected. We can see that R squared value, which tell us how good model explains variance in model, is equal 0.3597989. It is a little small value - only third part of variance is explained by model. We will try to find better model with multivariable models.

Multivariable regression model

At the begining we plot comparision plots to see which variables are linear with **mpg** variable

```r
with(mtcars, plot(mtcars))
```



Looking at the plot we can see that there exist strong linear relationship between **mpg** and **cyl**, **disp**, **hp**, **wt** and **am** variables and we will use they to new model.

```r
fit1<-lm(mpg~factor(cyl)+disp+hp+wt+factor(am), data=mtcars)
summary(fit1)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ factor(cyl) + disp + hp + wt + factor(am),
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.864276   2.695416  12.564 2.67e-12 ***
## factor(cyl)6 -3.136067   1.469090  -2.135   0.0428 *
## factor(cyl)8 -2.717781   2.898149  -0.938   0.3573
## disp          0.004088   0.012767   0.320   0.7515
## hp           -0.032480   0.013983  -2.323   0.0286 *
## wt           -2.738695   1.175978  -2.329   0.0282 *
## factor(am)1   1.806099   1.421079   1.271   0.2155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

Let's do some diagnostics. First of all, we see if there are some outliers or influential observation. We use leverages and Cook's distance respectively.

```r
sort(hatvalues(fit1),decreasing = T)[1:5]
```

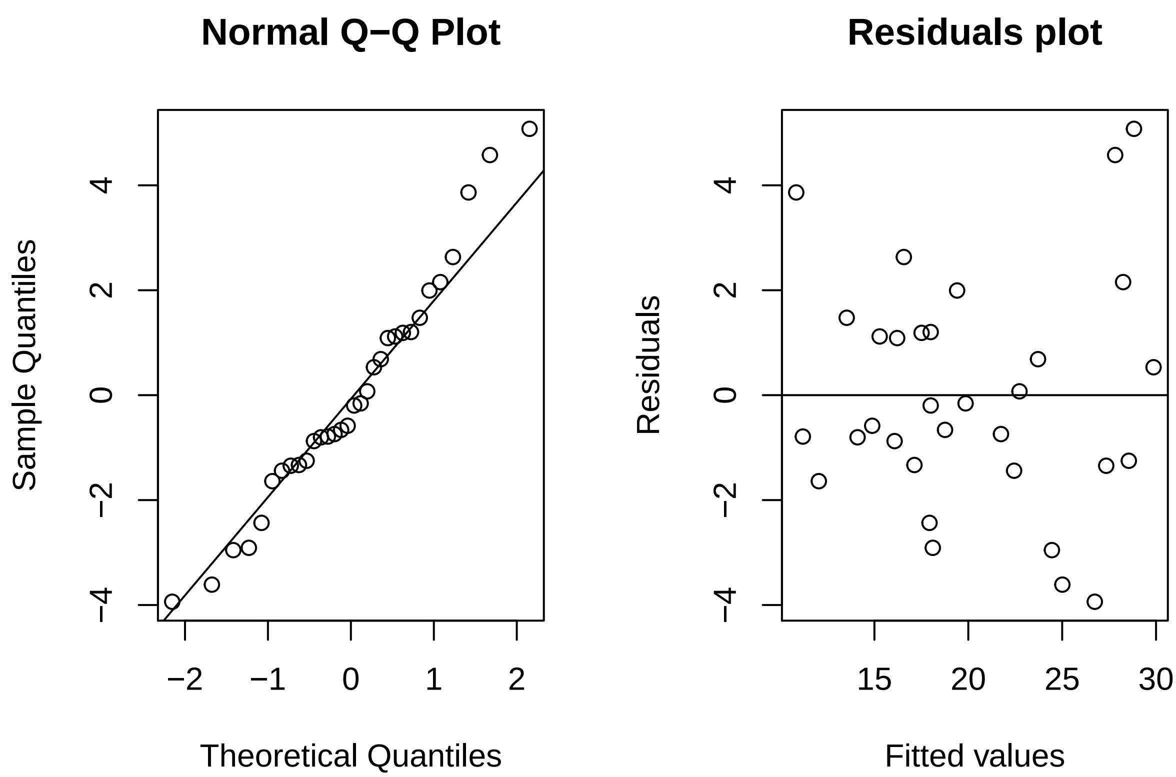
```
##       Maserati Bora      Hornet 4 Drive Lincoln Continental
##           0.5120823           0.3180707           0.3070164
##   Cadillac Fleetwood       Toyota Corona
##           0.3054138           0.2781398
```

```r
sort(cooks.distance(fit1),decreasing=T)[1:5]
```

```
## Chrysler Imperial     Maserati Bora     Toyota Corona    Toyota Corolla
##        0.17043519        0.11111419        0.11058080        0.10370427
##          Fiat 128
##        0.09777152
```

Analysing the biggest leverages values, we could indicate one obesrvation which could be the outlier but hat values is quit similar to others. There are no influential observation according to Cook's distance. Now we try examinate how residuals like and normality of residuals.

```r
par(mfrow=c(1,2))
qqnorm(y=resid(fit1))
qqline(resid(fit1))
plot(fit1$fitted, resid(fit1), main="Residuals plot", xlab="Fitted values", ylab="Residuals")
abline(h=0)
```



We can say, that the residuals' variance is constant without any pattern. Also we can say that they are normal but it could be provide with Shapiro Test.

```
shapiro.test(resid(fit1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit1)
## W = 0.971, p-value = 0.5274
```

P value is greater that 5% border, so we can find that residuals are normal. At the end of diagnostic, we check if multivariable model is better than one variable model.

```
anova(fit,fit1)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(cyl) + disp + hp + wt + factor(am)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     25 150.41  5    570.49 18.965 8.637e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test gives p value less than 0.05 and cause that "reacher" model with additional variables better describes relationship between MPG and transmission type.

3. Summary

Analysis above proves that to describe on which variables depends MPG, transmission type is not sufficient. Much better effects are with additional variables. Diagnostic methods proves, that model is fitted properly. Both models point that automatic transmission is better for fuel use. Less MPG indicates less pollution in environment and more kilometers driven with full tank - only adventages. Automatic transmission is better customize by computer than any human and this is cause of better results with automatic transmission.