What impacts miles per gallon - analysis of mtcars data set

Dorota

21 June 2015

Executive Summary

Based on below analysis, we can coclude that manual transmissions are better than automatic in terms of miles per gallon. A change from automatic to manual transmission increased the MPG by 7.245. However, transmission type only explained 36% (R2) of the variation in mpg. Other variables like weight, number of cylinders and horsepower are better indicator of miles per gallon.

Data Processing

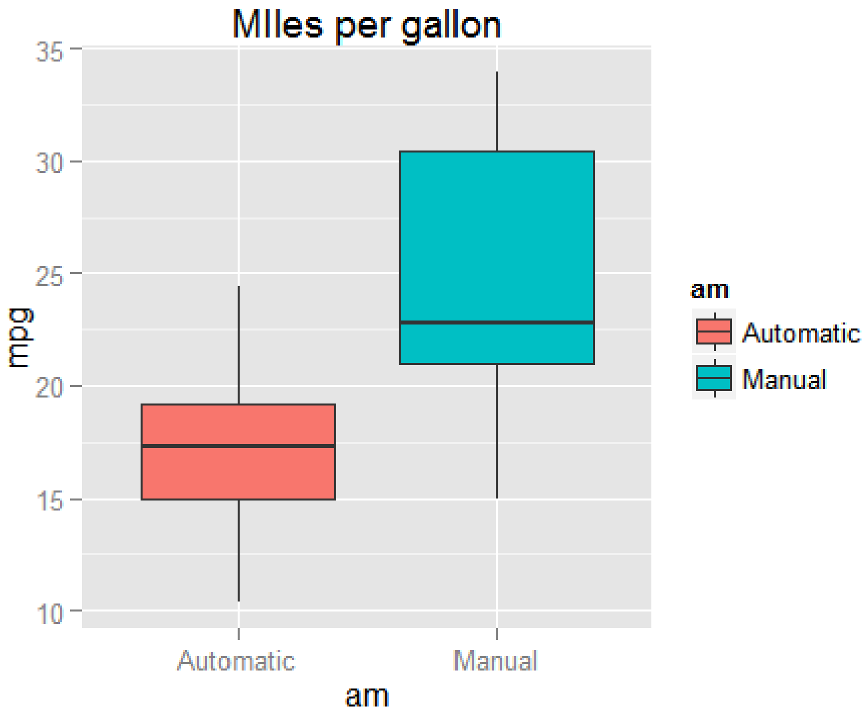
After uploading the data, some of the vriables were coded as factor

```
mtcars<-mtcars
mtcars$cyl=factor(mtcars$cyl)
mtcars$vs=factor(mtcars$vs)
mtcars$gear=factor(mtcars$gear)
mtcars$carb=factor(mtcars$carb)
mtcars$am=factor(mtcars$am,labels=c('Automatic','Manual'))
```

Exploratory Data Analysis

Below boxplot shows that there might be a difference between automatic and manual transmission

```
library(ggplot2)
ggplot(mtcars, aes(x=am, y = mpg, fill=am)) +
  geom_boxplot() + labs(title = "MIles per gallon")
```



Additionally, running t test shows that this difference is significant (t-test results in p-value 0.0013736)

Building the model

We wil start building the model, buy using only tranmission variable, later on we will use stepwise approach to find the best model (based on AIC criterion)

```
library(MASS)
fit<- lm(mpg ~ am, data=mtcars)
fit2<-lm(mpg ~ ., data=mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ 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 ***
## amManual       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
```

```
step = stepAIC(fit2, scope=list(lower=~am),direction="both",trace=FALSE)
summary(step)
```

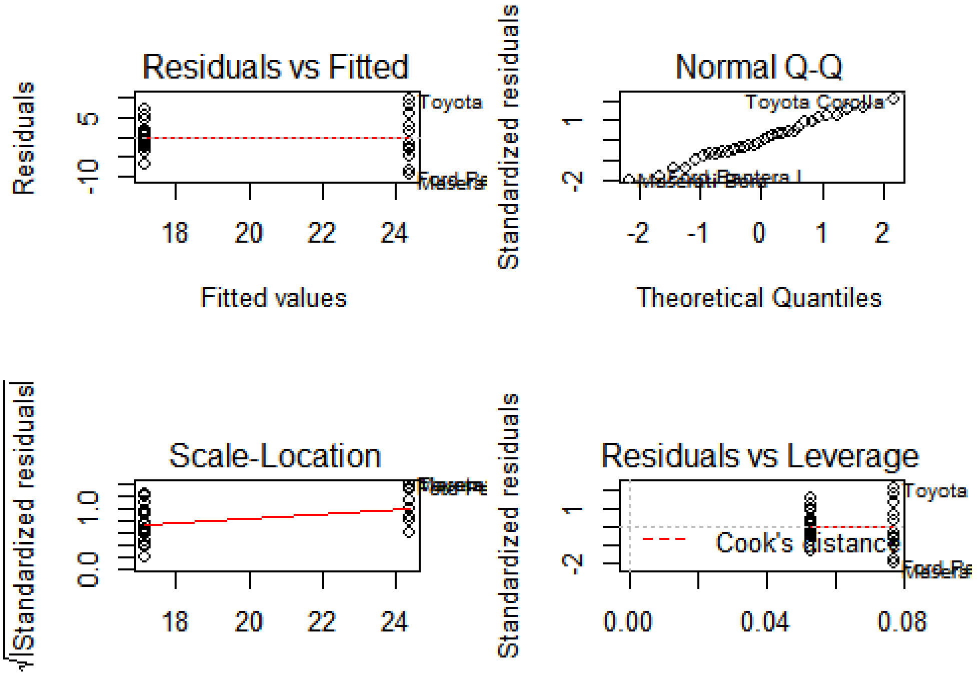
```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Adding number of cylinders, weight and horsepower to the model improves how much variance can be explained by the model (from 36% to 87%). Also coefficient changed from 7.245 to 1.8

Diagnostics

We can also check residuals

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



Conclusion

There is a difference in mpg in relation to transmission type, however, transmission type does not appear to be a very good explanatory variable on its own.