Analysing car performance

Oleksand Fialko

December 8, 2016

Looking at a data set of a collection of cars, it is interesting to explore the relationship between a set of variables and miles per gallon (MPG) (outcome). I am particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Preliminary exploratory data analysis

The data

```
library(ggplot2)
data("mtcars")
head(mtcars)
```

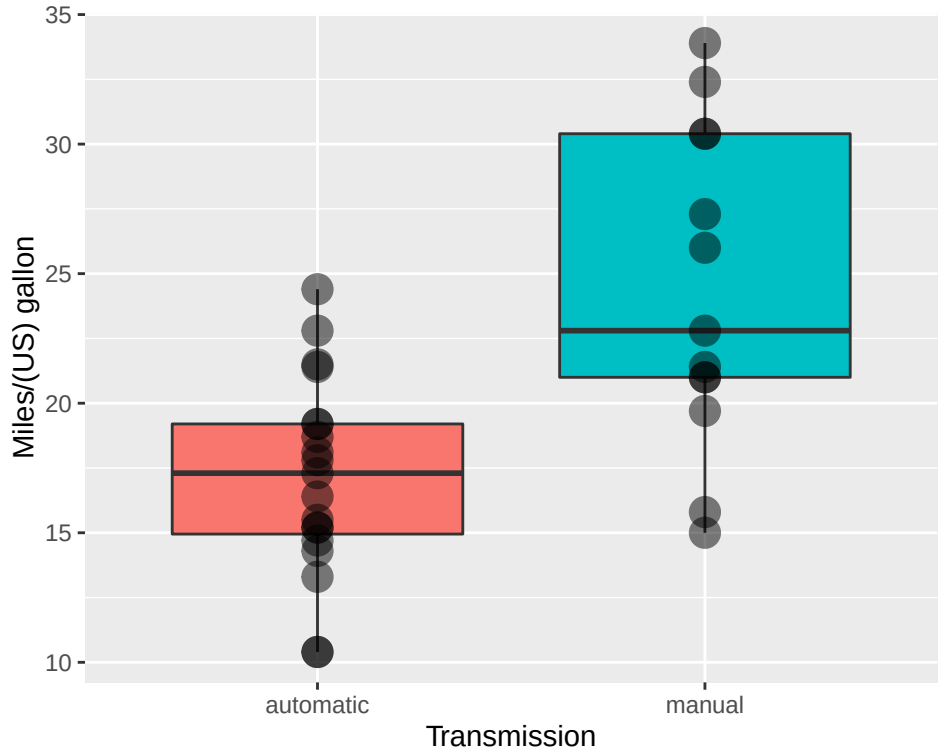
```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Three variables are of particular interest:

1. mpg: Miles/(US) gallon
2. cyl: Number of cylinders
3. wt: Weight (1000 lbs)

Simple box-plot suggests that cars with manual transmission are more efficient:

```
g<- ggplot(data = mtcars,aes(x=factor(am),y=mpg,fill=factor(am)))
g<- g   + geom_boxplot()+ geom_point(size=5,alpha=0.5)
g<- g + xlab('Transmission') + ylab('Miles/(US) gallon')
g<- g + scale_x_discrete(labels=c('automatic','manual'))
g<- g + theme(legend.position='none')
g
```



I create two vectors containing data for `mpg` of cars with automatic and manual transmissions. I perform t-test to see if their means are different:

```
data_auto <- subset(mtcars,am==0)$mpg
data_manu <- subset(mtcars,am==1)$mpg
t<- t.test(data_manu,data_auto)
t$p.value
```

```
## [1] 0.001373638
```

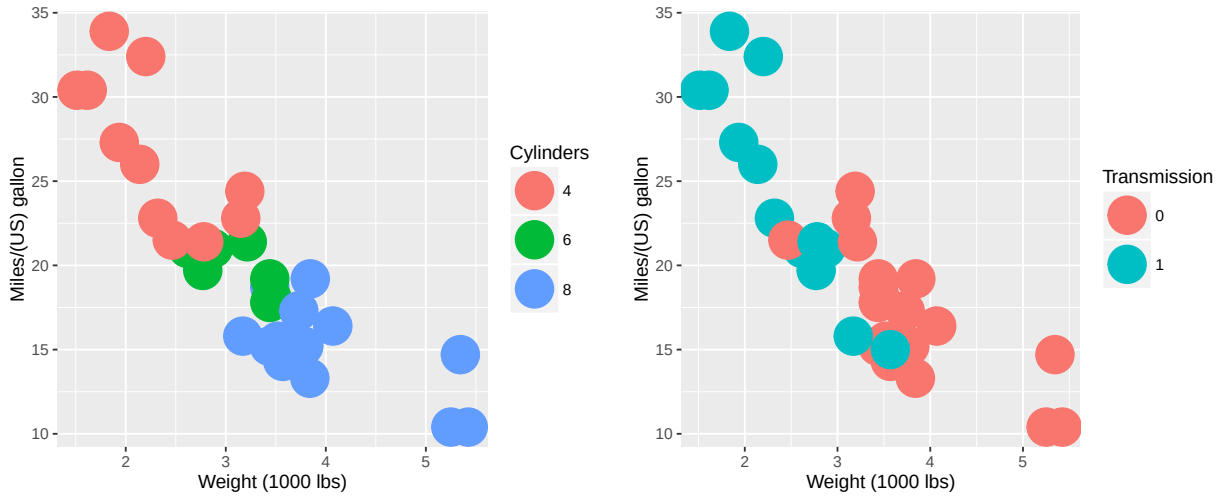
The p-value is significant (<0.05), so the means are different.

Fitting linear regression model

We need to be careful when fitting linear regression model, since it appears that weight is related to both the number cylinders and transmission as can be seen from the two plots below. Heavy cars have larger number of cylinders and have mostly automatic transmission. We will thus disregard `wt` from out analysis.

```
library(gridExtra)
h1<- ggplot(data = mtcars,aes(x=wt,y=mpg))
h1<- h1 + geom_point(size=10, aes(col=factor(cyl))) + xlab('Weight (1000 lbs)') + ylab('Miles/(US) gall
h1<-h1+ guides(col=guide_legend(title="Cylinders"))

h2<- ggplot(data = mtcars,aes(x=wt,y=mpg))
h2<-h2 + geom_point(size=10, aes(col=factor(am)))+ xlab('Weight (1000 lbs)') + ylab('Miles/(US) gallon'
h2<-h2+guides(col=guide_legend(title="Transmission"))
grid.arrange(h1,h2,ncol=2,nrow=1)
```



We will fit several linear models using `am` and `cyl` as predictors and `mpg` as outcome:

```
fit1 <- lm(mpg~factor(am),data=mtcars)
fit2 <- lm(mpg~factor(am)+factor(cyl),data=mtcars)
fit3 <- lm(mpg~factor(am)*factor(cyl),data=mtcars)
```

Analysis of variances allows us to choose the best model

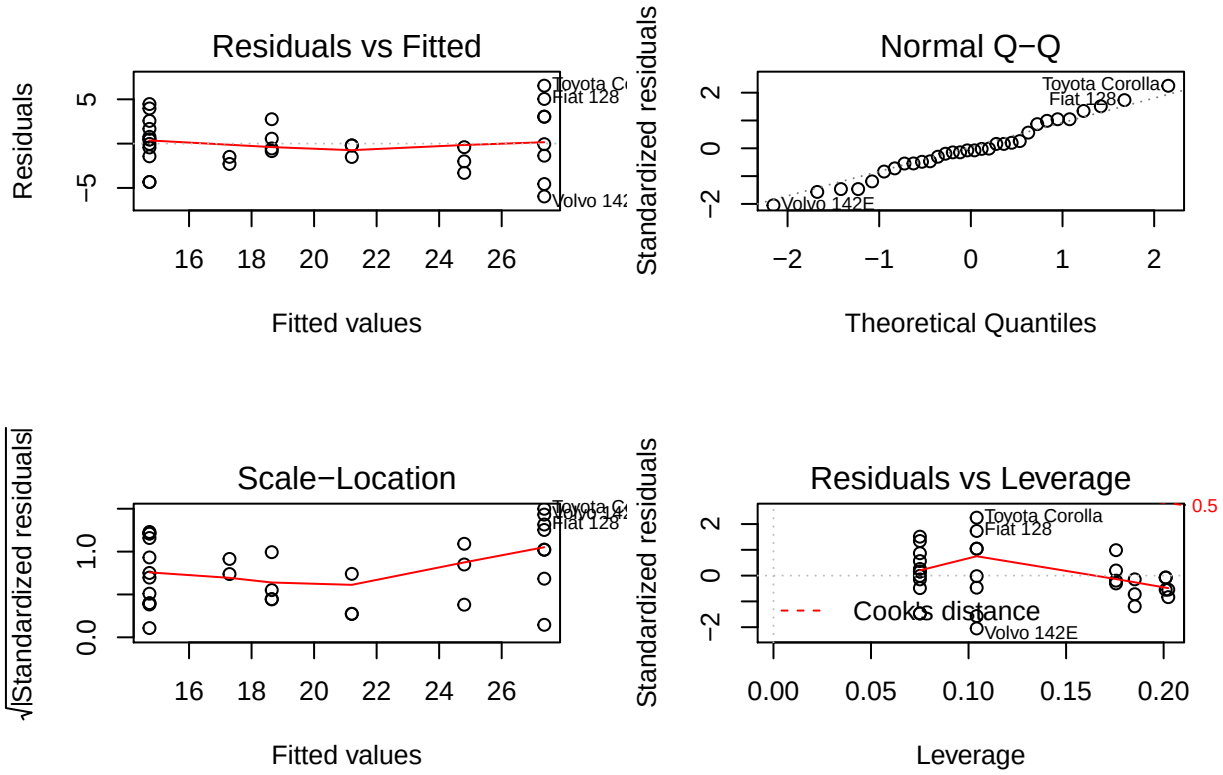
```
anova(fit1,fit2,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + factor(cyl)
## Model 3: mpg ~ factor(am) * factor(cyl)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 24.8190 9.355e-07 ***
## 3     26 239.06  2     25.44  1.3832    0.2686
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Apparently, the second `fit2` model is the best one among those three models.

Let's have a look at its residuals. They look good: they have near-zero mean (1-st plot) and follow approximatelly normal distribution (2nd plot):

```
par(mfrow=c(2,2))
plot(fit2)
```



Model coefficients:

```
summary(fit2)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   24.801852   1.322615 18.752135 2.182425e-17
## factor(am)1    2.559954   1.297579  1.972869 5.845717e-02
## factor(cyl)6  -6.156118   1.535723 -4.008612 4.106131e-04
## factor(cyl)8 -10.067560   1.452082 -6.933187 1.546574e-07
```

As can be seen from the `Pr(>|t|)` values all the coefficient are significant.

Summary

Cars with manual transmission are more efficient than cars with automatic transmission. For a given number of cylinders, cars with manual transmission covers on average 2.56 miles per gallon more than the cars with automatic transmission.