MPG trend research

Mariia Danilenko

Introduction

We are interested in exploring the relationship between a set of variables and miles per gallon. We are 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."

Analysis

Using data

Our research is based on build in R dataset ***mtcars***. We use **am** and **MPG** variables for the first version of calculations and all columns for the second.

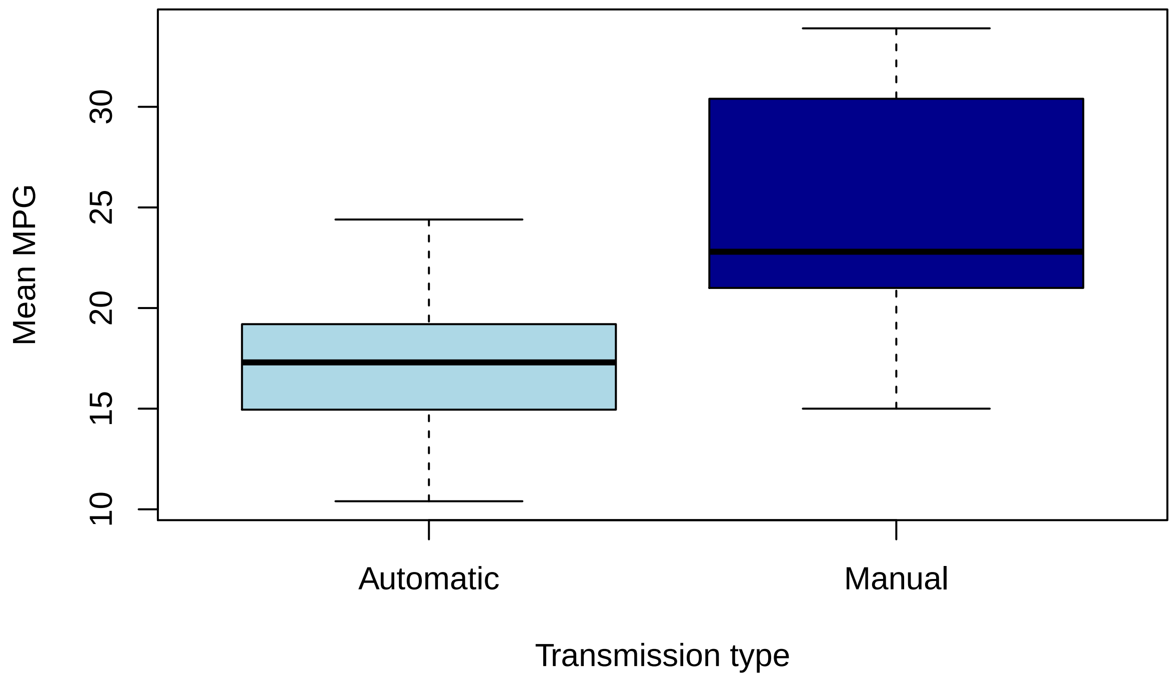
```
library(datasets)
head(mtcars)
```

```
##                    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
```

First model

First linear regression model is based on only two variables **MPG** and **am**. As first step let's take a look at mean values aggregated by transmission type.

```
boxplot(mtcars$mpg~factor(mtcars$am, labels=c("Automatic", "Manual")),
        col=c("lightblue", "darkblue"), xlab="Transmission type",
        ylab="Mean MPG")
```




```
Means<-tapply(mtcars$mpg, factor(mtcars$am), mean)
names(Means)<-c("Automatic", "Manual")
Means
```

```
## Automatic    Manual
##  17.14737  24.39231
```

```
as.numeric(Means[2]-Means[1])
```

```
## [1] 7.244939
```

As we can see, Manual average **MPG** is 7.245 miles per galon higher than Automatic average **MPG**. We got the same result for the difference when constructed linear model on **AM** variable.

```
am_fit<-lm(mpg~am, data=mtcars)
summary(am_fit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

Advance model

The best way to compare **MPG** for Manual and Automatic transmission types is to take a look at car models with same other technical features (CYL, WT and e.c.). However, we don't have this information and will try to explore linear regression model for **MPG** depends on all known variables.
Also, we use function STEPAIC to obtain columns with the best result for different linear models.

```r
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.3.3
```

```r
all_fit<-lm(mpg~., data=mtcars)
summary(all_fit)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

```r
best_fit <- stepAIC(lm(mpg~., data=mtcars), trace=0)
summary(best_fit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

Comparing three models:

```r
anova(am_fit, best_fit, all_fit)
```

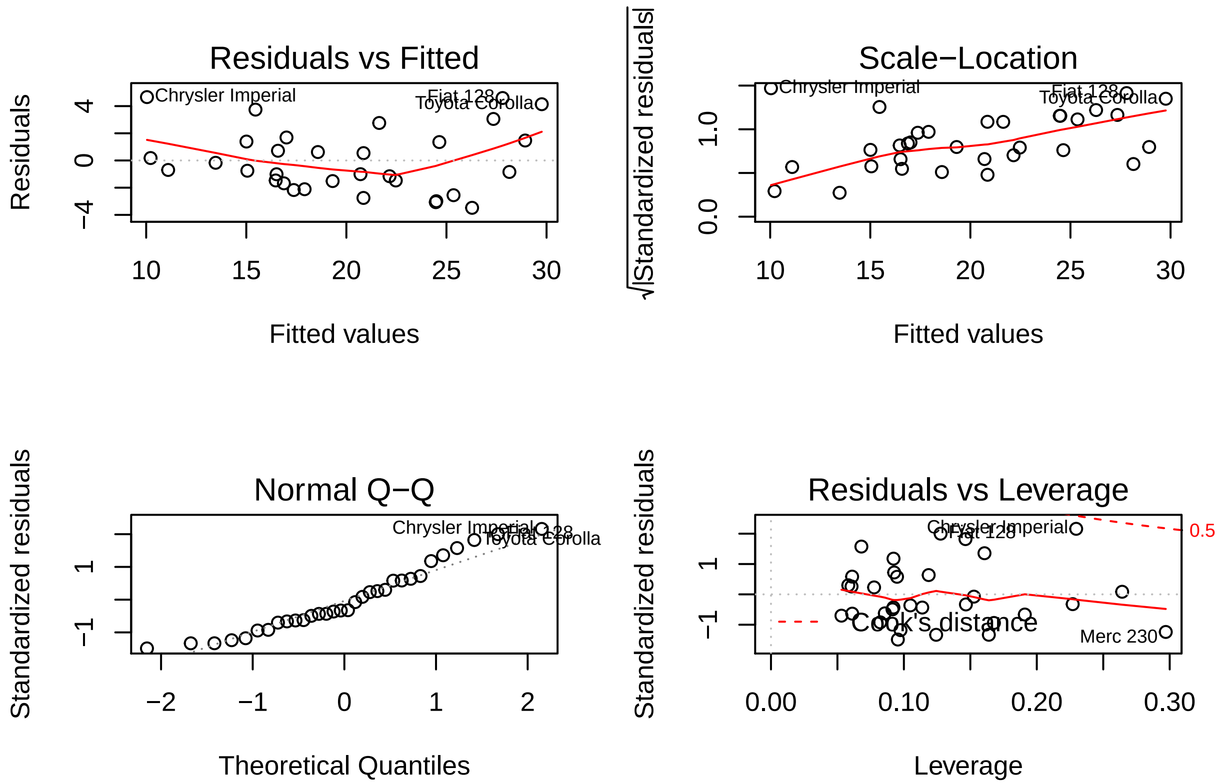
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 39.2687 8.025e-08 ***
## 3     21 147.49  7     21.79  0.4432    0.8636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

S statistics says that adding **qsec** and **wt** is necessary, so we select **best_fit**. If we think about interpretation of variables, choosing *Weight* and *Transmission type* looks logically, but **QSec** is not. However, **QSec** has to depend on **Cyl** and **HP**, hence **MPG** depends on **Cyl** and **HP** too.

Residuals for selected model

Let's plot residualts for the best_fit model:

```r
par(mfcol=c(2,2))
plot(best_fit)
```



There are no patterns at first, second and last plots. Q-Q plot looks normal. Based on these results we approbate **best_fit** model.

Conclusion

MPG definetely depends on **am**. The rate differs according to a constructed model, but it is always will be positive: when you switch Automatic to Manual MPG grows. For the **best_fit**, where predictors are: **wt**, **am**, **qsec**, this rate is 2.9 mpg.