Mtcars data survey

Data Science / Regression Models / Course Project
Andrey Komrakov
May 27 2016
Source files https://github.com/kolfild26/regmodel.git (https://github.com/kolfild26/regmodel.git)

Abstract

This survey is aimed to explore the Motor Trend Car Road Tests data (*Mtcars*). We are investigating the influence of a transmission type (automatic or manual) on *MPG* - the miles per gallon, taking into account the other car characteristics (*Number of cylinders, Gross horsepower,Rear axle ratio, Weight, Number of forward gears, etc.*).

We are trying to answer the following questions:

1. Is an automatic or manual transmission better for MPG
2. Quantify the MPG difference between automatic and manual transmissions

We will use the ***R*** language for the data processing, getting statistics and the linear model creation.

Exploratory analysis

First, look at the dataset we are going to work with.

```
head(mtcars,5)
```

```
##                    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
```

```
range(mtcars$mpg); mean(mtcars$mpg)
```

We see that the variable of intererst (outcome) **mpg** is of a number type. It varies between [10.4, 33.9] and have a mean of 20.1. And **am** is a factor binary variable - *0* - automatic, *1* - manual.

Now, let's making *t-test* diagnostics to compare two trends (manual / automatic transmission) and find out, if there is any significant difference between them.

```
t.test(mtcars[mtcars$am==1,]$mpg , mtcars[mtcars$am == 0,]$mpg, alternative="two.sided")[3]
```

```
## $p.value
## [1] 0.001373638
```

A small *p-value* (typically ≤ 0.05) indicates strong evidence against the null hypothesis. Since we get *p-value* = 0.001, we can assume the significant influence of a transmissiom type on the miles per gallon characteristic. Also, the same can be seen from the plot (see Appendix *picture 1.*).

Let's go further and check this hypothesis based on the fact that we have more than one variable which can change an outcome.

Multivariable modeling

First, find the variables which have a significant (greather than) correlation with ***mpg***:

```
corcoeff <- cor(mtcars$mpg, mtcars)
corcoeff[ ,abs(corcoeff) > 0.5][-1]
```

```
##        cyl       disp         hp       drat         wt         vs
## -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594  0.6640389
##         am       carb
##  0.5998324 -0.5509251
```

So, ***cyl, disp, hp, drat, wt, vs, am, carb*** might be a basis for a linear model.
Check that the other variables do not tell us more about ***mpg*** variance. We do this through the ***anova()*** function which can compare the different linear models. based on their impacts in the variance explanation.

```r
fit01 <- lm(mpg ~ cyl + disp + hp + drat + wt + vs + factor(am),data = mtcars)
fit02 <- lm(mpg ~ cyl + disp + hp + drat + wt + vs + factor(am) + qsec + gear + carb,data = mtcars)
anova(fit01, fit02)
```

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ cyl + disp + hp + drat + wt + vs + factor(am)
## Model 2: mpg ~ cyl + disp + hp + drat + wt + vs + factor(am) + qsec + 
##     gear + carb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     24 158.65                           
## 2     21 147.49  3     11.16 0.5296 0.6668
```

According to the *p-value* interpretation (> 0.05) we can reject H_0, and conclude that ***qsec,gear,carb*** are not significant in terms of variance explanation

Remember, ***anova()*** implies the normatily of the residuals.

```r
c(shapiro.test(fit01$residuals)$p,shapiro.test(fit02$residuals)$p)
```

```
## [1] 0.1764675 0.2261489
```

The ***p*** values confirm normality of both model residuals, hence the anova results are comprehended.

Now, when we found the scope of parameters which vary ***mpg*** the most, we can test the different factor conbinations to find out whether it's possible to shrink the scope of variables in the model.

Again, through ***anova()*** we see that the model ***lm(mpg ~ factor(am) + wt + cyl,data = mtcars)*** explains the most part of the ***mpg*** variance.

```r
fit0 <- lm(mpg ~ factor(am) ,data = mtcars)
fit1 <- lm(mpg ~ factor(am) + wt ,data = mtcars)
fit2 <- lm(mpg ~ factor(am) + wt + cyl,data = mtcars)
fit3 <- lm(mpg ~ factor(am) + wt + cyl + hp,data = mtcars)
fit4 <- lm(mpg ~ factor(am) + wt + cyl + hp + disp,data = mtcars)
fit5 <- lm(mpg ~ factor(am) + wt + cyl + hp + disp + drat,data = mtcars)
fit6 <- lm(mpg ~ factor(am) + wt + cyl + hp + disp + drat + vs,data = mtcars)
```

```r
residM <- cbind(resid(fit0),resid(fit1),resid(fit2),resid(fit3),
                resid(fit4),resid(fit5),resid(fit6))
apply(residM,2, function(x) shapiro.test(x)$p.value )
```

```
## [1] 0.85734421 0.10239346 0.06108291 0.07694562 0.12528988 0.08411377
## [7] 0.17646750
```

All the residuals are normally distributed.

```r
anova(fit0,fit1,fit2,fit3,fit4,fit5,fit6)[6]
```

```
##    Pr(>F)    
## 1            
## 2 < 2e-16 ***
## 3 0.00132 ** 
## 4 0.08700 .  
## 5 0.31789    
## 6 0.75009    
## 7 0.45694    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A residuals vs. fits plot (see Appendix *picture 2*) visually confirm that the ***lm()*** function is being applied properly (no visible unexplained variance).

```r
summary(fit2)$r.squared
```

According to R^2 criteria, our model explains 0.83% of a total variance.

From the model summary it can be easily seen that the difference between the manual and automatic transmision in their influence in ***mpg*** is significant in framework of our model.

```r
summary(fit2 <- lm(mpg ~ factor(am) + wt + cyl - 1 ,data = mtcars))$coeff
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## factor(am)0 39.417933  2.6414573 14.922798 7.424998e-15
## factor(am)1 39.594427  1.8721428 21.149255 9.322776e-19
## wt          -3.125142  0.9108827 -3.430894 1.885894e-03
## cyl         -1.510246  0.4222792 -3.576415 1.291605e-03
```

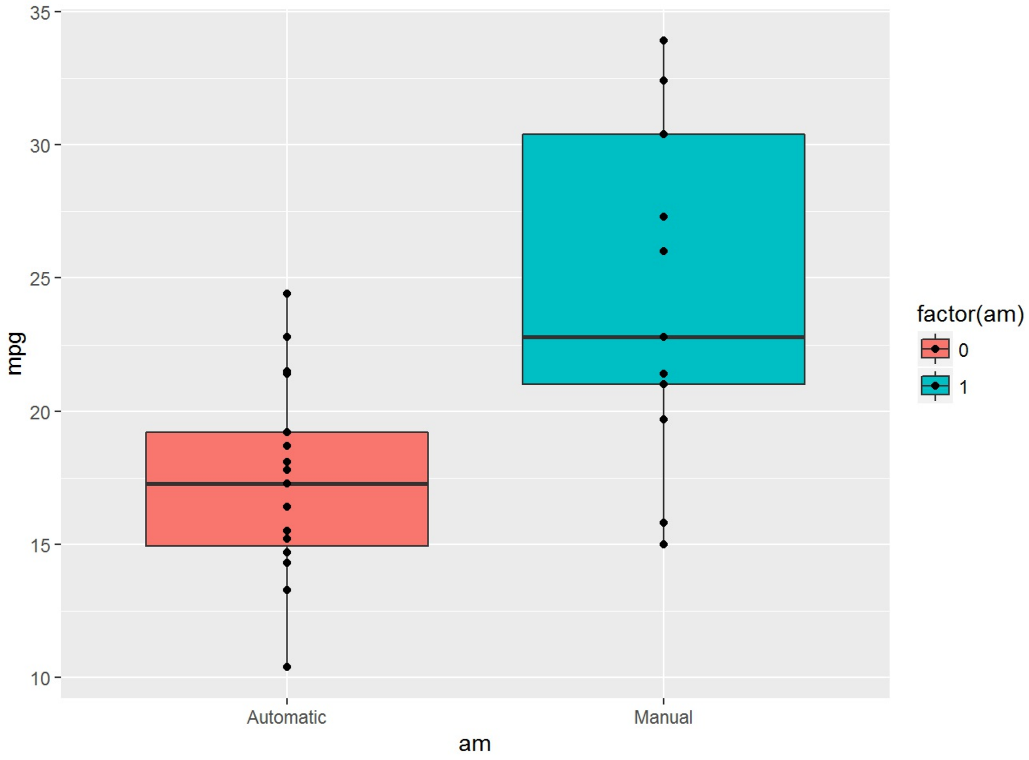
Conclusion

Having proceed the comparison between two types of transmission we detected a possible difference in thier influence on ***mpg***. Futher, the linear model approved that the same takes place even in presence of the other factors.
The ***lm()*** analysis also gave as an estimation of that difference. ***Mpg*** for the cars with manual transmission is greater on **0.18** than for the cars with automatic transmission. Thus manual transmission is better than automatic.

Appendix

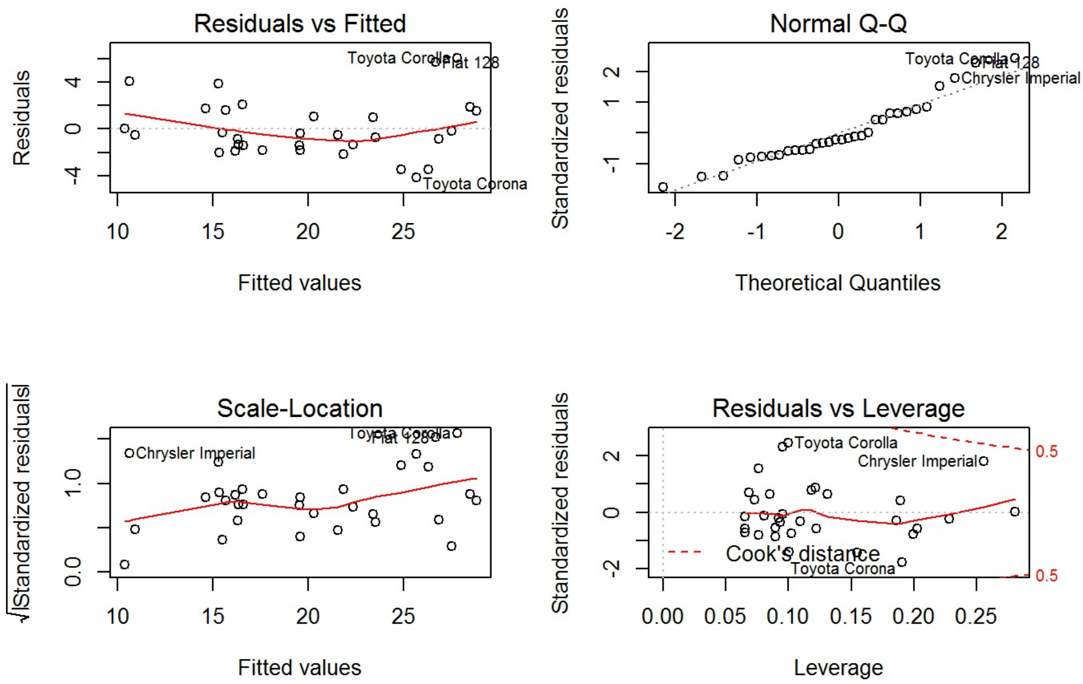
picture 1.

```
ggplot(mtcars, aes(am, mpg, fill = factor(am)) ) +
     geom_boxplot(data = mtcars, aes(x=factor(am, labels = c("Automatic", "Manual")), mpg)) +
     geom_point  (data = mtcars, aes(x=factor(am, labels = c("Automatic", "Manual")), mpg))
```



picture 2.

```
par(mfrow = c(2,2))
plot(lm(mpg ~ am + wt + cyl,data = mtcars))
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
0.5