Motor Trend case analysis

Michael Lagunov

8/24/2020

Main part

Introduciton

Motor Trend, a magazine about the automobile industry, is interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

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

Libraries and data loading

```
library(ggplot2)
library(dplyr)
library(wesanderson)
data(mtcars)
```

Exploratory data analysis

Before modeling, provide basic data exploration, to understand it better

```
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 ...
```

From the str function the one can mention, that all the features are numeric, so we will have to transform them into the factors. That is about variables "vs", that stands for the v-shaped and straight engines respectively and "am" for transmission, automatic (0) and manual (1).

Transform them into factors

```
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <-c("Automatic", "Manual")
```

From the boxplot (see the appendix) above manual transmission tends to have more MPG than automatic. Let's see whether the difference is significant by using t.test (see appendix for summary)

```
a_trans <- filter(mtcars, am == "Automatic")
m_trans <- filter(mtcars, am == "Manual")
ttest <- t.test(a_trans$mpg, m_trans$mpg)
```

p-value is less that 0.05, thus we reject the null hypothesis, and state that the difference is significant and manual transmission is better for MPG than automatic.

Regresssion analysis

At first, let's make a regression using only transmission type as an predictor.

```
fit1 <- lm(mpg ~ am-1, mtcars)
```

Residual standard error is 4.9 with the R-squared of 0.36. For the whole summary see the appendix 2.

Now let's try to find a better combination of factors using "step" function, that is using AIC approach.

```
fit2 <- step(lm(data = mtcars, mpg ~ .), trace=0)
```

The new model looks much better, with the formula = mpg ~ wt + qsec + am, residual error reduced by double and multiple R-squared of 0.85. For the whole summary see the appendix 3.

Provide an analysis of variance for the fit1 and fit2 models to see if the second is significantly better (see appendix for summary)

```
ano_res <- anova(fit1,fit2)
```

The second model is significantly much better than the initial. Examine the second model deeper. From the estimates, we can see, that weight has negative effect, whereas all the other features has positive. Per each 1000 lbs of the weight the estimate mpg is reduced by almost 4. Qsec represent 1/4 smile time and for each of the unit mpg is increased by 1.2. And lastly, manual transmission is 2.9 mpg better than the automatic. For the residual plots see the appendix.

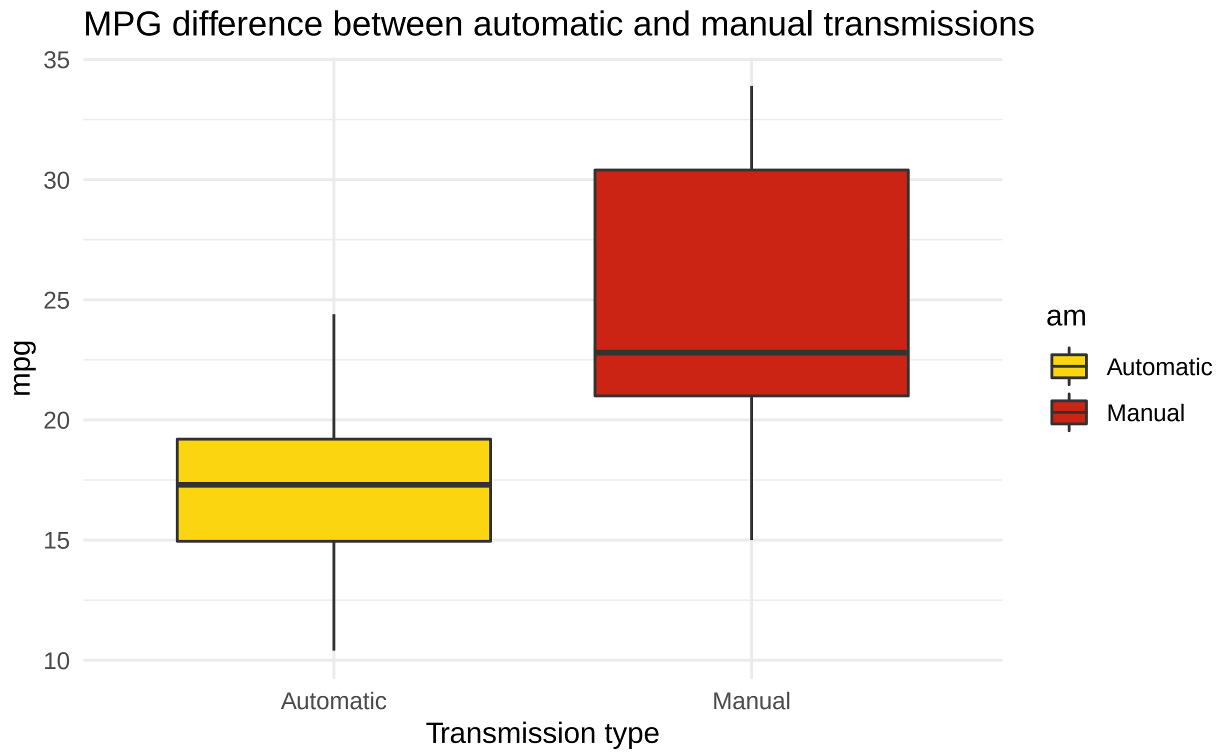
Conclusion

In this analysis, I found, that the manual transmission is significantly better than automatic for MPG with the p-value of 0.0013. The best regression has been received by using the following formula = mpg ~ wt + qsec + am. Manual transmission is 2.9 mpg better than the automatic under this regression model.

Appendix

MPG difference between automatic and manual transmissions

```
ggplot(mtcars, aes(am, mpg, fill = am)) +
    geom_boxplot() +
    scale_fill_manual(values=wes_palette(n=2, name="BottleRocket2")) +
    ggtitle("MPG difference between automatic and manual transmissions") +
    xlab("Transmission type") +
    theme_minimal()
```



Summary of the t.test

```
ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  a_trans$mpg and m_trans$mpg
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

Summary of the fit1 model

```r
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ am - 1, 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|)
## amAutomatic   17.147      1.125   15.25 1.13e-15 ***
## amManual      24.392      1.360   17.94  < 2e-16 ***
## ---
## 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:  0.9487, Adjusted R-squared:  0.9452
## F-statistic: 277.2 on 2 and 30 DF,  p-value: < 2.2e-16
```

Summary of the fit2 model

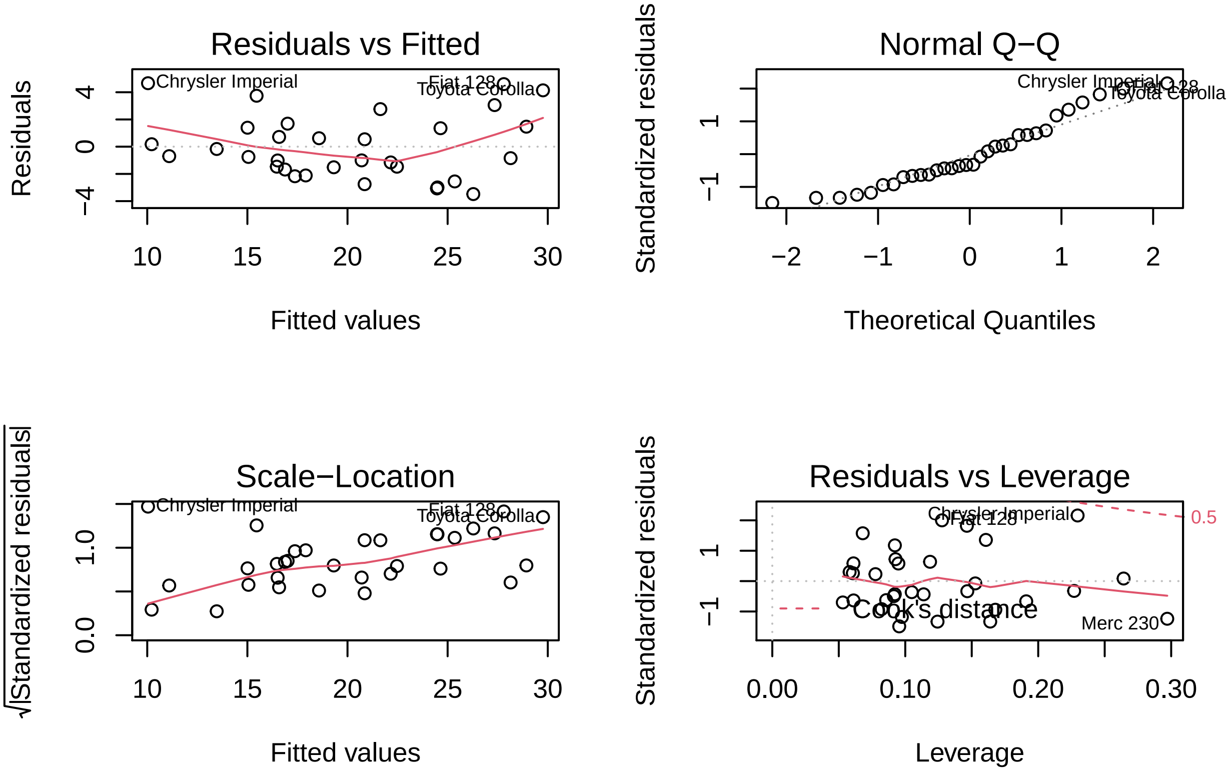
```r
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Regression residual and diagnostic plots

There is some correlation in the plot1, and the right part of the qq plot is dipped. I may be improved by using more data

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



Summary of the analysis of models variance

```
ano_res
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am - 1
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```