Regression models. Peer assignment.

Andrej Kulunchakov

In this project we analyse a data set of a collection of cars. We are interested in exploring the relationship between a set of variables and miles per gallon (MPG). Especially, we are interested in the following two questions:

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

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

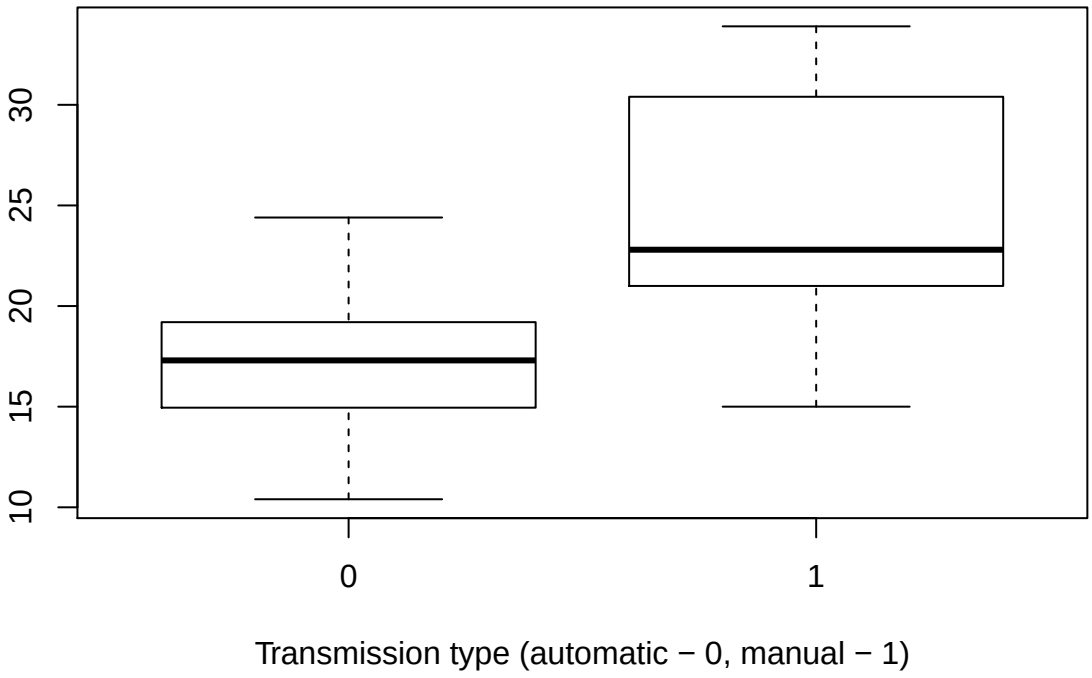
Exploratory data analysis

Retrieve the data and use some plots to analyse it.

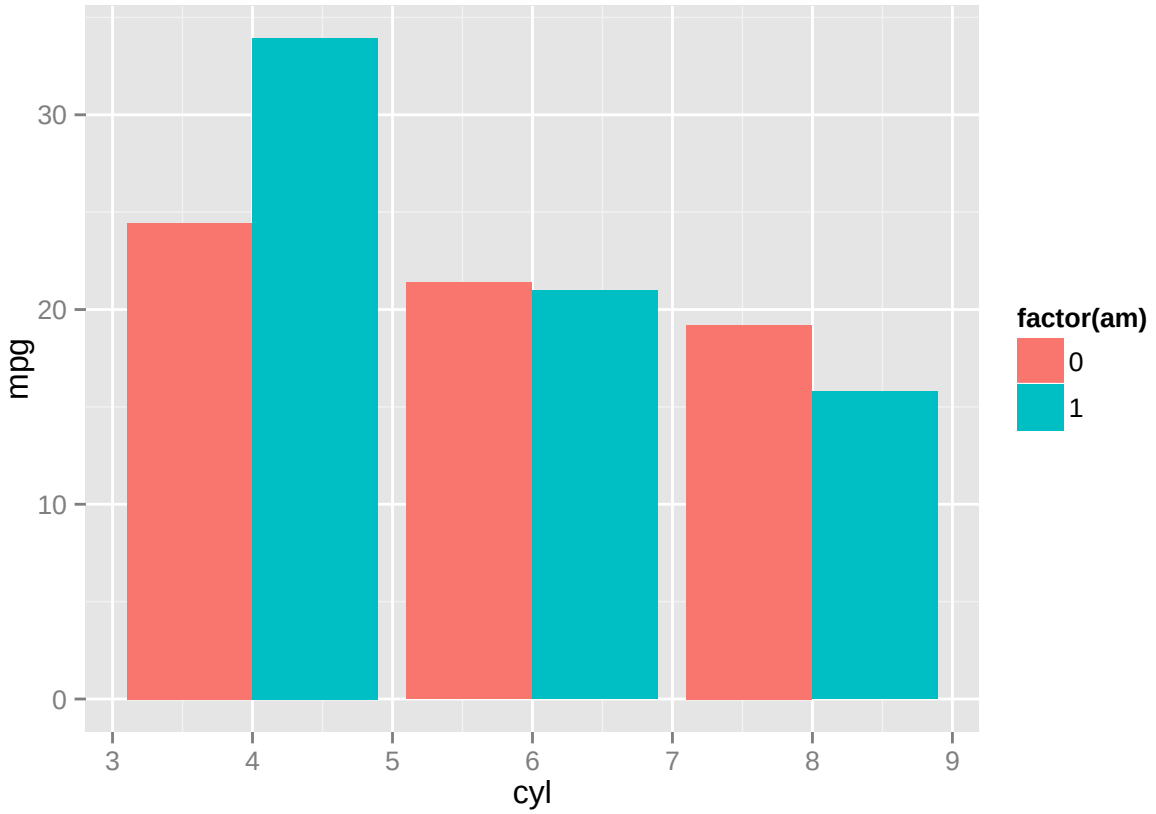
```
dt <- mtcars
head(dt,4)
```

```
##                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
```

```
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission type (automatic - 0, manual - 1)")
```



```
ggplot(dt,aes(x=cyl,y=mpg,fill=factor(am)))+geom_bar(stat="identity",position="dodge")
```



The graphs show that there is a trend for the dependence between mpg and am. Cars with manual transmission mainly have bigger mpg. Moreover, bigger values for cyl demonstrate bigger mpg for cars with automatic transmission. Lets now check these trends by regression model.

Regression model and residuals/diagnostic

We'll analyse an impact of am on mpg by regression model which contains am as a feature and have good quality (for example, adjusted R-squared).

Firstly, try to build a regression containing all features of given data:

```
model <- lm(data = dt, mpg~.)
#p-values of the coefficients:
coeftest(model)[,1]
```

```
## (Intercept)         cyl        disp          hp        drat          wt
## 12.30337416 -0.11144048  0.01333524 -0.02148212  0.78711097 -3.71530393
##        qsec          vs          am        gear        carb
##  0.82104075  0.31776281  2.52022689  0.65541302 -0.19941925
```

As one can see from the model summary, all coefficients are insignificant (significance = 0.05). It may appear because of excessive parameters in the model. So, for example, significant parameters appear in the model without interception:

```
coeftest(aov(data = dt, mpg~.))[,1]
```

```
## (Intercept)         cyl        disp          hp        drat          wt
## 12.30337416 -0.11144048  0.01333524 -0.02148212  0.78711097 -3.71530393
##        qsec          vs          am        gear        carb
##  0.82104075  0.31776281  2.52022689  0.65541302 -0.19941925
```

Then we iteratively filter insignificant parameters (with p-value more than 0.05) and build new regressions (now with an interception coefficient). The final model is

```
model <- lm(data = dt, mpg ~ cyl + wt + am)
coeftest(model)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 39.41793    2.64146 14.9228 7.425e-15 ***
## cyl         -1.51025    0.42228 -3.5764  0.001292 **
## wt          -3.12514    0.91088 -3.4309  0.001886 **
## am           0.17649    1.30445  0.1353  0.893342
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model have all coefficients significant (except for am, which can not be eliminated as it's important for the project purpose). Now it's interesting to compare final regression model with the simple one:

```
coeftest(lm(data = dt, mpg ~ am))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  17.1474     1.1246 15.2475 1.134e-15 ***
## am            7.2449     1.7644  4.1061  0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

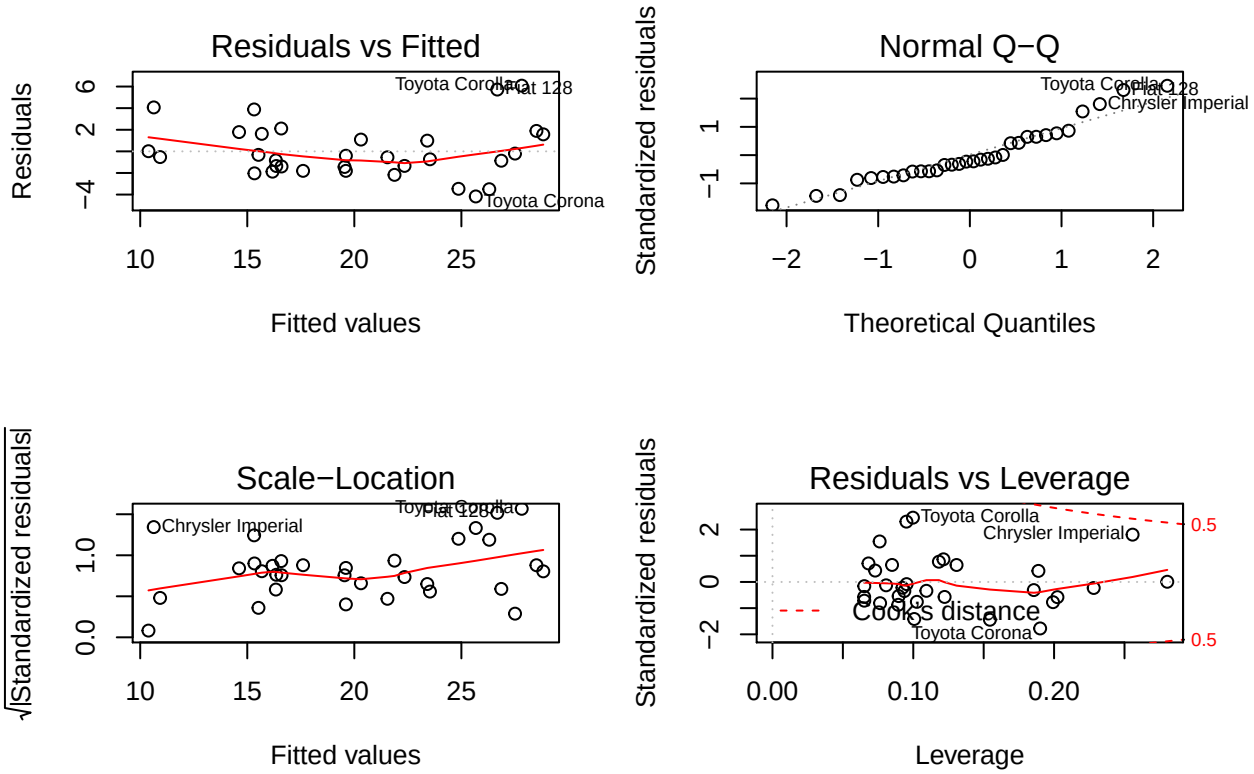
```
anova(model, lm(data = dt, mpg ~ am))$"Pr(>F)"
```

```
## [1]           NA 8.428366e-09
```

As the p-value is highly significant, we can not assume the models to discribe outcome similarly. Our final model is better and we'll use it for the further conclusions.

But firstly we'll analyze the residuals and do some diagnostics:

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



Lets discribe each plot:

- Residuals vs. Fitted plot shows that the points are randomly scattered. So it verifies the independence of observations.
- Distribution of points on Normal Q-Q plot is well described by linear model. So we can say that the residuals are distributed normally.
- From the Scale-Location plot one can see that residuals have constant variance.
- Residuals vs. Leverage shows us some noize in the data. The following text is about finding these leverage points.

```
levPoints <- hatvalues(model)
tail(sort(levPoints),3)
```

```
##  Cadillac Fleetwood   Chrysler Imperial Lincoln Continental
##           0.2281436           0.2557513           0.2803808
```

```
levPoints <- dfbetas(model)
tail(sort(levPoints[,4]),3)
```

```
##          Fiat 128 Chrysler Imperial     Toyota Corona
##         0.3226482         0.3953935         0.6741522
```

This result is in a good agreement with the last plot.

Conclusions

Firstly, we should say that despite our first conclusions (from plots) the am-coefficient is insignificant. The most plausible explanation for it is strong dependence between am and the other coefficients.

Assume that am is distributed normally. We perform a t-test and see that transmission types are significatively different:

```
t.test(mpg ~ am, data = dt)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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 in group 0 mean in group 1
##        17.14737        24.39231
```

So, mpg seems to depend on transmission type. With 95% confidence we can claim that the difference between mpg means of cars with different transmission lies in the interval (3.2, 11.3). And cars with manual transmission have the biggest mpg. But we actually don't know if there is a truly dependence between mpg and am. Because according to the linear regression this dependence is well described in terms of other features (cyl, wt).