Regression Models Course Project

Artem Braun

11th November 2017

Analysis of the MPG (miles per gallon) measure in regards to automatic and manual transmissions

1. Executive summary

Motor Trend, a magazine about the automobile industry, is interested in exploring the data set of a collection of cars. Particularly, they asked to explore the relationship between a set of variables and miles per gallon (MPG) measure. We have to give elaborated answers to the following questions:

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

In order to achieve this task we performed regression model selection and diagnosed it.

Based on provided analysis we conclude that manual transmission is more effective in terms of MPG, and the difference in effectiveness is 2.9. With manual transmission one can drive 2.9 miles more with the same amount of gas (gallon) than with automatic transmission.

2. Data set exploratory analysis

One of the best and concise ways to get acquaintance with a data set is 'str' function.

```r
data("mtcars")
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 ...
```

We need to understand what does each variable define

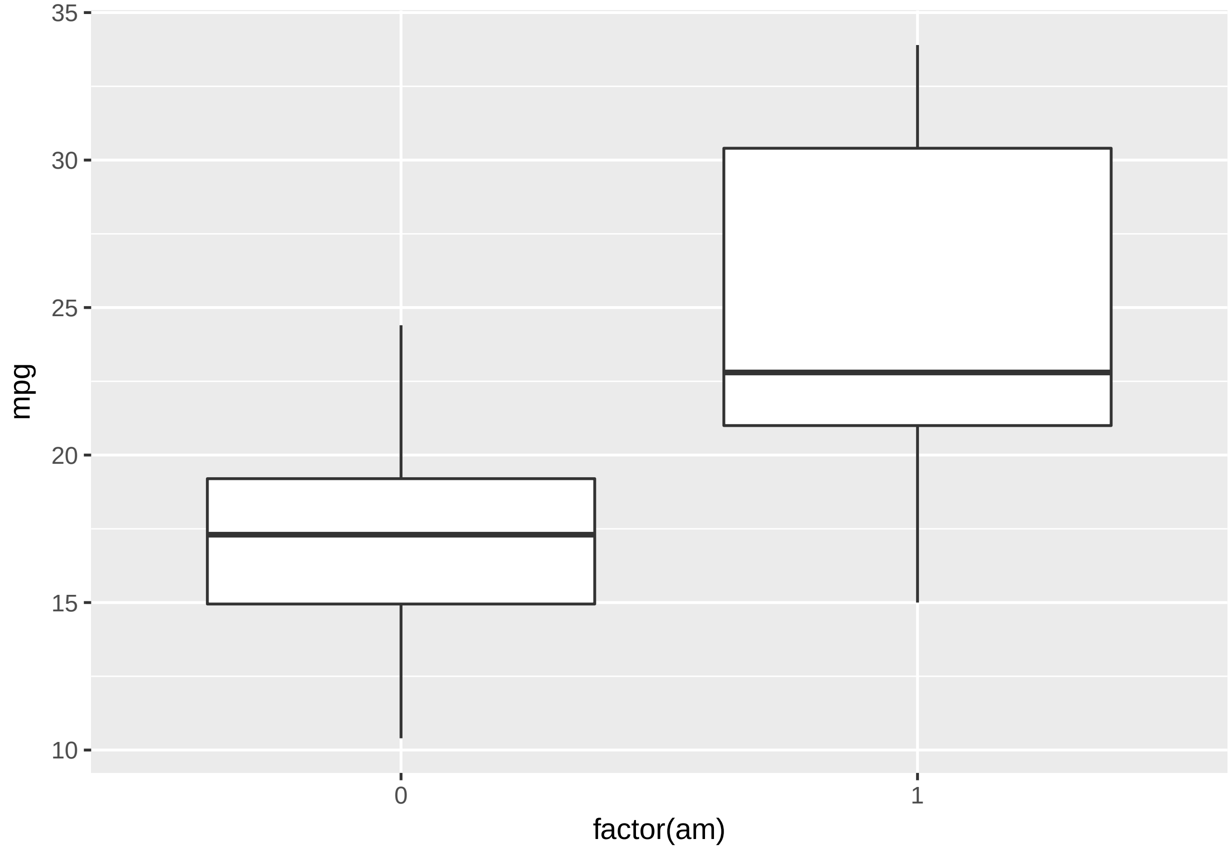
Variable - Description
[, 1] & mpg & Miles/(US) gallon \{}
[, 2] & cyl & Number of cylinders \{}
[, 3] & disp & Displacement (cu.in.) \{}
[, 4] & hp & Gross horsepower \{}

Variable - Description
[, 5] & drat & Rear axle ratio \{}
[, 6] & wt & Weight (1000 lbs) \{}
[, 7] & qsec & 1/4 mile time \{}
[, 8] & vs & V/S \{}
[, 9] & am & Transmission (0 = automatic, 1 = manual) \{}
[,10] & gear & Number of forward gears \{}
[,11] & carb & Number of carburetors

We see that variable 'am' (which we are interested in) is a factor one. It is binomial: it could be 0 or 1. Therefore, we cannot use simple correlation analysis here.

The easiest way to visualize factor variables is a box plot:

```r
library(ggplot2)
ggplot(mtcars, aes(x=factor(am), y=mpg)) + geom_boxplot()
```



Now we have something to work with. Apparently, we can infer that manual transmission provides more miles with the same amount of gas. This relation could be causative, since automatic transmission consumes more power than mechanical one. But this is just an idea, and visualization could be insignificant due to random chance. In other words - plot is not enough. This is the part where test statistics comes in.

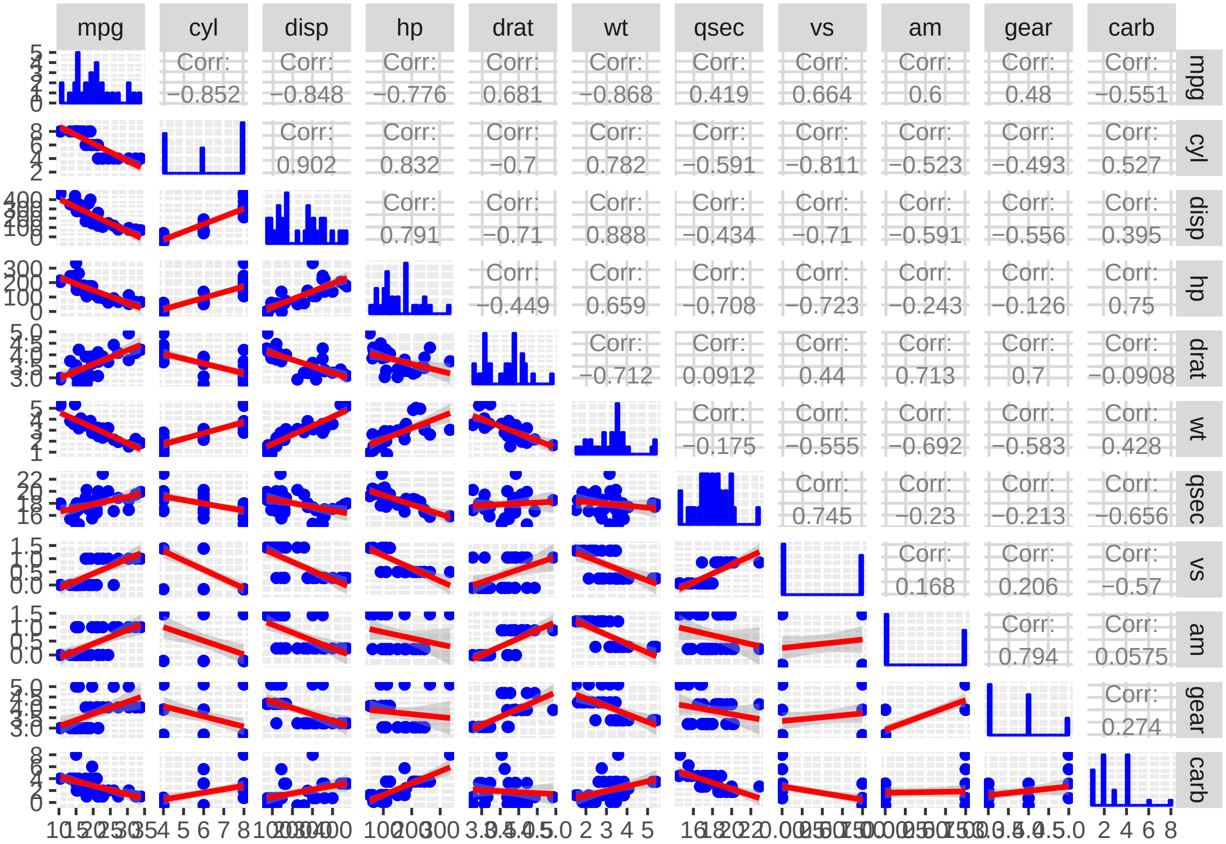
```r
test <- t.test(mpg ~ am, data= mtcars, var.equal = FALSE, paired=FALSE ,conf.level = .95)
test$p.value
```

```
[1] 0.001373638
```

Low p-value indicates that the probability of chance relationship between MPG and type of transmission is very low (<0.05). We can conclude that **manual transmission is better for MPG**. This inference is useful, but MPG might be dependent on any of other 10 variables that we have in a data set. Therefore, difference in MPG between transmission' types might depend on some other variables as well. In order to calculate the exact difference ratio we have to fit rigorous statistical model.

To get initial broad picture of possible dependencies among variables we can use pairwise scatterplot.

```r
library(GGally)
lowerFn <- function(data, mapping, method = "lm", ...) {
        p <- ggplot(data = data, mapping = mapping) +
             geom_point(colour = "blue") +
             geom_smooth(method = method, color = "red", ...)
        p
}
ggpairs(mtcars, lower = list(continuous = wrap(lowerFn, method = "lm")),
                diag = list(continuous = wrap("barDiag", colour = "blue")),
                upper = list(continuous = wrap("cor", size = 3))
)
```



We can see that many relationships have distinct linear nature and might be causative. Some of predictors may be correlated and variance of the model may be inflated. So, there is a possibility that we will have redundancies in case of fitting multivariable linear model with all 10 regressors.

3. Model selection

Since we observed linear relationships among many variables, we have to select only some of those ones which predict MPG the best. At first, we look at p-values of all predictors

```
MPG_fit <- lm(mpg ~ ., data = mtcars)
knitr::kable(summary(MPG_fit)$coeff)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3033742	18.7178844	0.6573058	0.5181244
cyl	-0.1114405	1.0450234	-0.1066392	0.9160874
disp	0.0133352	0.0178575	0.7467585	0.4634887
hp	-0.0214821	0.0217686	-0.9868407	0.3349553
drat	0.7871110	1.6353731	0.4813036	0.6352779
wt	-3.7153039	1.8944143	-1.9611887	0.0632522
qsec	0.8210407	0.7308448	1.1234133	0.2739413
vs	0.3177628	2.1045086	0.1509915	0.8814235
am	2.5202269	2.0566506	1.2254035	0.2339897
gear	0.6554130	1.4932600	0.4389142	0.6652064
carb	-0.1994193	0.8287525	-0.2406258	0.8121787

Not all of regressors are statistically significant at 95% confidence level. For selecting regressors we will use automated model selection by AIC in a stepwise algorithm. There is the function **'step'** for that in R. We will not limit the scope. If scope operand is missing, the initial model is used as the upper model - that is what we need.

```
selection <- step(MPG_fit, data=mtcars, k=2, trace = FALSE)
```

4. Conclusions

Three variables were selected as the best predictors:
- Weight (wt)
- 1/4 mile time (qsec)
- Transmission type (am)

In order to quantify the MPG difference between automatic and manual transmissions we just need 'am' coefficient in our selected model.

```
MPG_fit <- lm(mpg ~ wt + qsec + am, data = mtcars)
knitr::kable(summary(MPG_fit)$coeff)
```

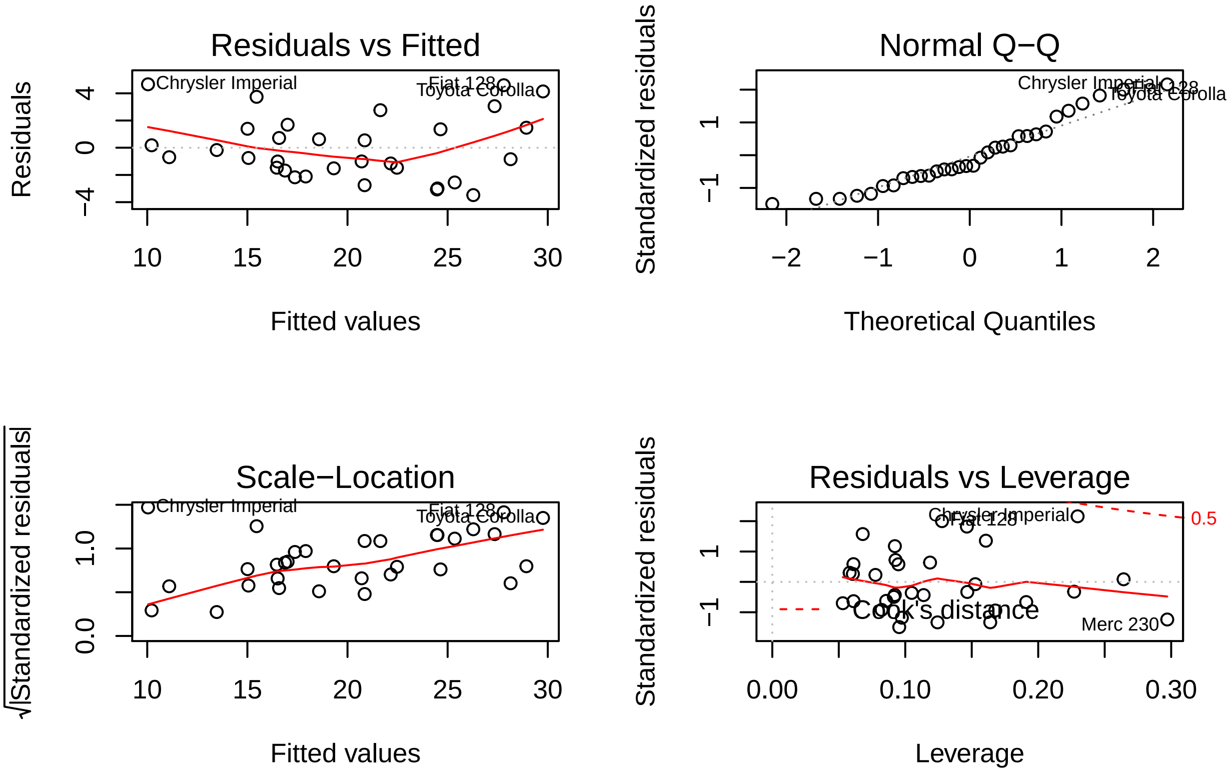
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
am	2.935837	1.4109045	2.080819	0.0467155

At 95% confidence level (p-value for "am" < 0.05) we may conclude that the MPG difference between automatic and manual transmissions is 2.9. With manual transmission one can drive 2.9 miles more than with automatic transmission with the same amount of gas (gallon).

5. Diagnostics of the model

In order to be completely comfortable with our selected model we have to perform additional diagnostic tests. To evaluate the fit and residuals of a linear model generated by R (selection), we can use the plot(MPG_fit) to produce a series of 4 diagnostic plots:

```
par(mfrow = c(2, 2)); plot(MPG_fit)
```



Looking at these plots we can be sure of the following:

- there are no patterns for missing variables or heteroskedasticity
- there are no unusual patterns in residuals
- errors are normal
- there are no points that have substantial influence on our regression model

We can conclude that our regression model is statistically significant.