JHU/Coursera Regression Models course project

Ilya Semenov

29/05/2016

Summary

Project goal is to explore MTCARS dataset and to name and quantify key automobile parameters that have influence on fuel consumption expressed in MPG that is miles per gallon consumption. Two particular questions of interest are:

- Is an automatic or manual transmission better for MPG?
- Is it possible to quantify the MPG difference between automatic and manual transmissions?

Exploratory data analysis, statistical inference and regression modelling were used to perform the analysis.

According to the **mtcars** dataset automobile fuel consumption mostly defined by No of cylinders and car weight. There is no enough proof found that transmission type has influence on MPG.

Data loading and processing

Since **mtcars** dataset is preinstalled in R, we just use *data* command to load the data.

```
data(mtcars)
head(mtcars, 3)
```

```
##                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
```

Exploratory data analysis

```
dim(mtcars)
```

```
## [1] 32 11
```

The dataset consists of 32 observations of 11 variables. An overview of **mpg** for auto and manual transmissions plot (see Fig. 1 in Appendix section) suggests that manual transmissions cars have higher fuel consumption level. We will check this assumption with further analysis. Also pairwise scatterplots (see Fig. 2 in Appendix) suggest noticeble correlations between **mpg** and **cyl**, **disp**, **hp**, **vs**, **gear** and **carb** variables.

Statistical inference for MPG with different transmission types

Let check if the difference in fuel consumption between auto and manual transmission cars is statistically significant. Our null hypothesis is that fuel consumption for auto and manual transmission cars is the same.

```
ttest <- t.test(mtcars$mpg ~ mtcars$am)
ttest$p.value
```

```
## [1] 0.001373638
```

```
ttest$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Since P-value obtained is small we reject the null hypothesis and can claim that according to **mtcars** data manual transmission cars have higher mpg than the auto ones with the means difference about 7. But we still need to check the influence of other variables.

Regression analysis

Firstly, let check the influence level of all predictor variables.

```
infl <- aov(mpg ~ ., data = mtcars)
summary(infl)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          1  817.7   817.7 116.425 5.03e-10 ***
## disp         1   37.6    37.6   5.353  0.03091 *
## hp           1    9.4     9.4   1.334  0.26103
## drat         1   16.5    16.5   2.345  0.14064
## wt           1   77.5    77.5  11.031  0.00324 **
## qsec         1    3.9     3.9   0.562  0.46166
## vs           1    0.1     0.1   0.018  0.89317
## am           1   14.5    14.5   2.061  0.16586
## gear         1    1.0     1.0   0.138  0.71365
## carb         1    0.4     0.4   0.058  0.81218
## Residuals   21  147.5     7.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to p-values obtained No of cylinders and weight are the most influential on **mpg**. **am** is only the 5th in mpg-influence rate. Let build a linear model for top-5 mpg-influence variables.

```
fit <- lm(mpg ~ cyl+wt+disp+drat+am, data = mtcars)
summary(fit)$coef
```

```
##                 Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 41.296379972 7.53839376  5.47814047 9.558465e-06
## cyl         -1.793995156 0.65053958 -2.75770330 1.050796e-02
## wt          -3.587040698 1.21049961 -2.96327292 6.433129e-03
## disp         0.007374511 0.01231949  0.59860537 5.546159e-01
## drat        -0.093627682 1.54877978 -0.06045255 9.522575e-01
## am           0.172981155 1.53004335  0.11305638 9.108543e-01
```

Due to high p-values we can remove **drat** and **disp** variables from the model and leave **am** just for the sake of initial task. The final linear model is as follows:

```
fit <- lm(mpg ~ cyl+wt+am, data = mtcars)
summary(fit)$r.squared
```

```
## [1] 0.8303383
```

```
summary(fit)$coef
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 39.4179334  2.6414573 14.9227979 7.424998e-15
## cyl         -1.5102457  0.4222792 -3.5764148 1.291605e-03
## wt          -3.1251422  0.9108827 -3.4308942 1.885894e-03
## am           0.1764932  1.3044515  0.1353007 8.933421e-01
```

The model explains about 83% of data variability and since **am** p-value is high (**am** p-value = 0.89) we fail to reject the hypothesis that coefficient of **am** variable in a linear model equals to zero. For final testing we will consider model residuals (see Fig. 3 in Appendix section). There is no clear patterns in Residuals vs Fitted graph, Normal Q-Q plot suggests the model fits normality criteria, Scale-Location plot shows no abnormal variation jumps, Residuals vs Leverage shows all residuals in acceptable ranges.

So our final claim is as follows: according to the **mtcars** dataset automobile fuel consumption mostly defined by No of cylinders and car weight. There is no enough proofs found that transmission type has influence on mpg.

Appendix

```
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission type",
        ylab = "MPG")
title(sub = "Figure 1. MPG plots for different transmission types")
```

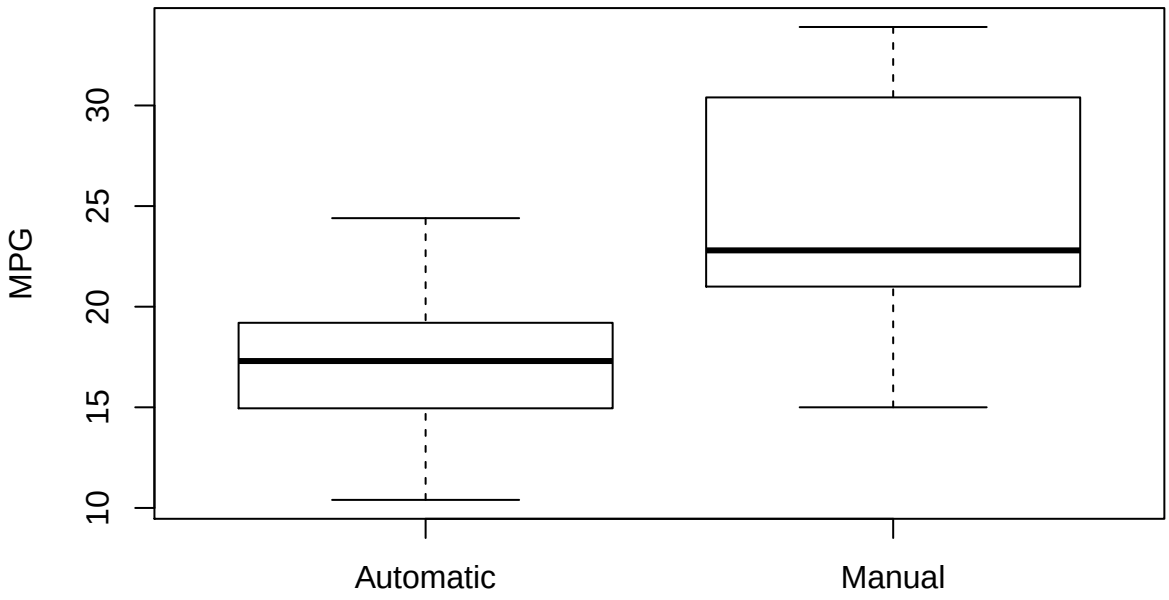


Figure 1. MPG plots for different transmission types

```
pairs(mpg ~ ., data = mtcars)
title(sub = "Figure 2. Scatterplots of *mtcars* variables pairs")
```

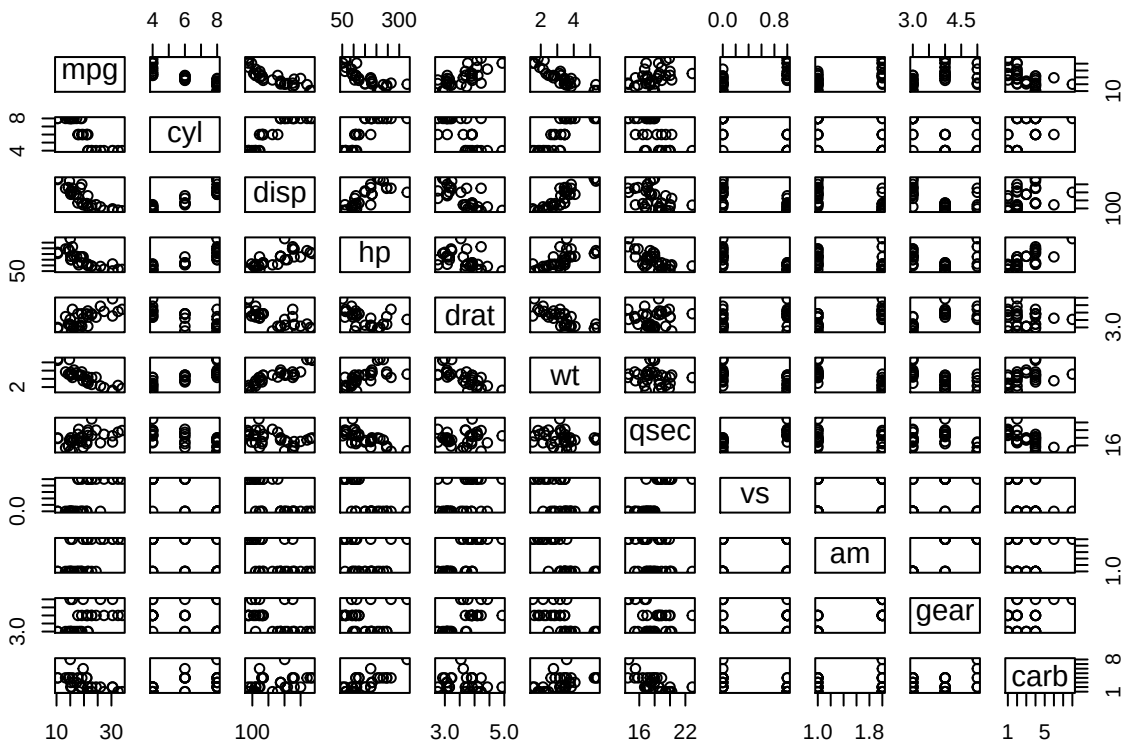
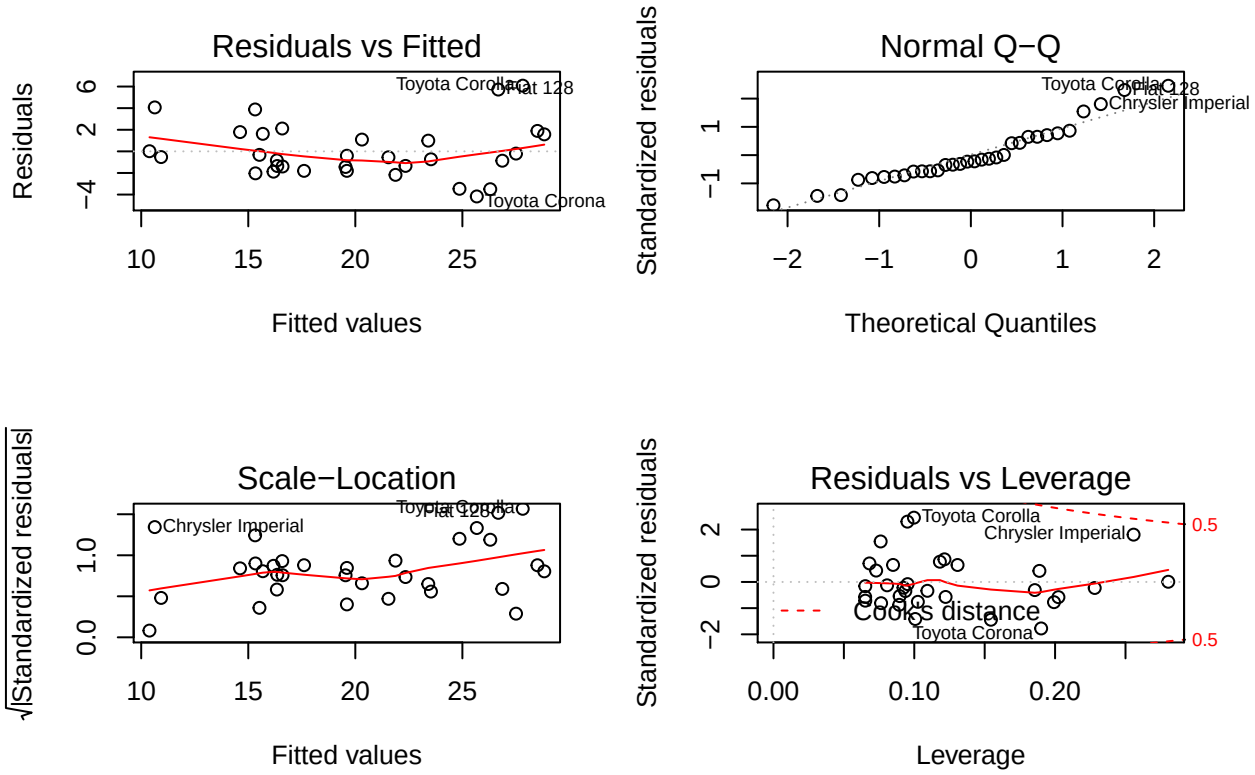


Figure 2. Scatterplots of *mtcars* variables pairs

```
pp <- par(mfrow = c(2,2))
plot(fit)
```




```
par(pp)
```

Figure 3. Final linear model verification