Analysis of mtcars data

Vladimír Tomeček

Saturday, February 21, 2015

Summary

In this paper, we will analyze Motor Trend US magazine's data and we will try to find out whether the transmission type have impact on fuel consumption.

First we will do some exploratory analysis and discover that cars with manual transmission have significantly lower fuel consumption than cars with automatic transmission.

Later we will try to predict fuel consumption with other variables and we will find out that transmission type have no effect on consumption.

Cars with manual transmission had higher MPG (miles per gallon) due to the fact, that they were mounted in cheaper cars, which had lower weight, power, number of carburetors and cylinders and therefore lower fuel consumption.

Data

For this analysis we will use the data extracted from the 1974 Motor Trend US magazine, which comprises fuel consumption and 10 other aspects of automobile design and performace for 32 automobiles.

It is a standard dataset bundled with R, known as `mtcars`. List of all variables can be obtained by typing `?mtcars` in R.

Exploratory analysis

We will start with comparing automatic vs manual transmission and their respective MPG (figure1). There is a clear distinction between the two, but on the pair plot (figure2), we see that practicaly every variably correlates with each other, so there is a pretty good chance that we will find some confounders.

Stepwise regression

To find some good model that fits the data, we will run stepwise regression. This method tries *many* various models to automatically find the best, so we must first split out dataset to training set and test set to avoid overfitting.

```
# 67% of the sample size
smp_size <- floor(0.67 * nrow(mtcars))
# set some good seed for reproductibility, this number will definitely work
set.seed(666)
train_ind <- sample(seq_len(nrow(mtcars)), size = smp_size)
train <- mtcars[train_ind, ]
test <- mtcars[-train_ind, ]
```

Now we can run stepwise regression on train data. I will use forward approach which starts with intercept only and is continuosly adding variables that improves model the most until the model can't be improved further.

```
null <- lm(mpg ~ 1, data = train)
full <- lm(mpg ~ ., data = train)
forward <- step(null, scope=list(lower=null, upper=full), direction="forward")
```

The automatic procedure finds model with cyl, wt, hp, am, carb to be the best. Now we need to test the model on the test sample to find out if it is good model or not.

```
cor(predict(forward, test), test$mpg)
```

```
## [1] 0.9409623
```

Correlation is 0.94 which coresponds to R^2=0.885. It is even better than prediction on the train data (cor=0.92, R^2=0.848), so it's clear that we did't overfit and our model is good.

Now let's see what will happen if we remove am from our model:

```
fit <- lm(mpg ~ cyl+wt+hp+carb, data = train)
cor(predict(fit, test), test$mpg)
```

```
## [1] 0.9523245
```

It improved our correlation even more to 0.95 (R^2=0.907), which means that transmission type has no effect on fuel consumption.

Note

We had a luck twice with our correlations, thanks to our seed number.

If we hadn't had the luck first time (overfitting the model), we would use expert judgement to identify relevant variables. For example we can say that weight will definitely have an impact on fuel comsumption as well as number of cylinders.

If we hadn't had the luck second time (removing `am` improved prediction performance), we would use coefficients from the lm model to quantify the `am` impact. We would use anova test to decide whether adding `am` improves our model significantly. (P value<0.05 means significant change (probably improvement); our model have P-value 0.1891 on whole dataset)

Conclusion

In this analysis we showed that `mpg` can be predicted very well by car's weight, number of cylinders, horsepower and number of carburetors. These four variables explain ~90% of the `mpg` variance. Car's weight has biggest impact while transmission type has no effect.

I didn't answer the second question because there is no mpg difference between automatic and manual transmission. If there were a difference we can quantify it with the code I included in `all_code.R`. (95% confidence interval)

Appendix

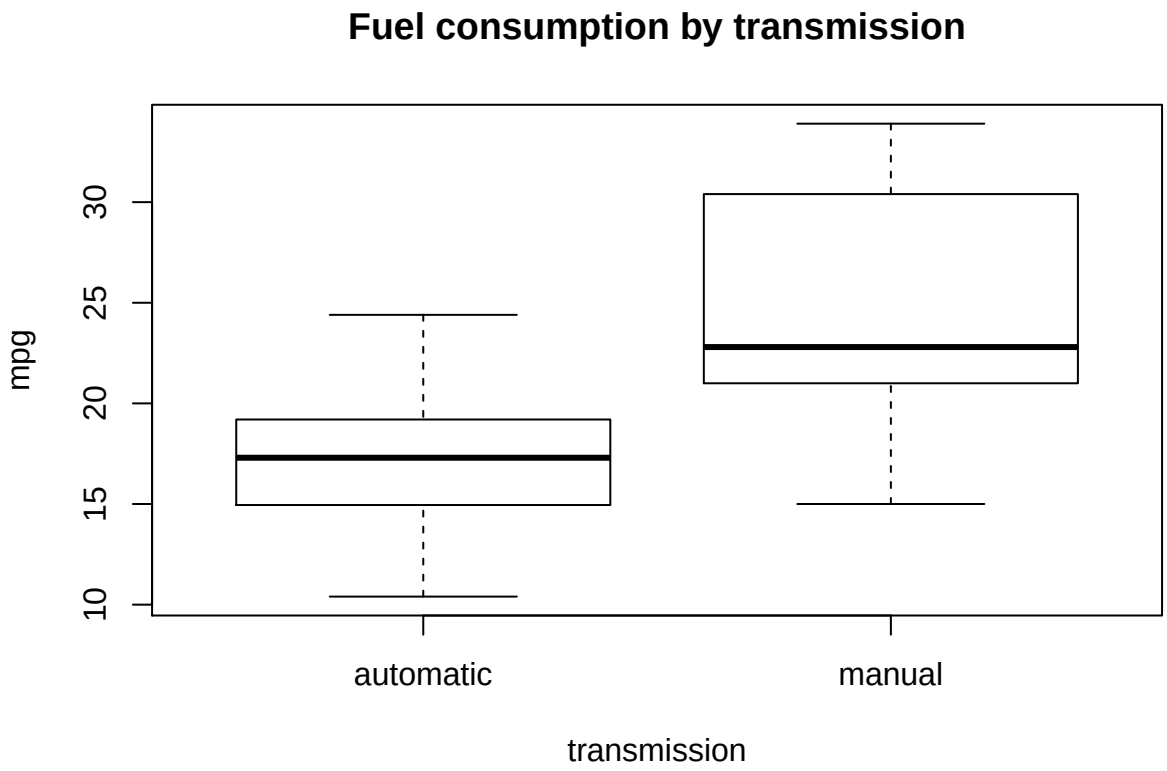

Figure 1: Figure 1 - Difference between automatic and manual transmission

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + hp + carb, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2314 -1.7103 -0.2035  0.7918  6.5483
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.86730    2.54248  14.107 1.91e-10 ***
## cyl         -0.84495    0.86445  -0.977   0.3429
## wt          -2.45103    1.10400  -2.220   0.0412 *
## hp          -0.01519    0.01833  -0.828   0.4196
## carb        -0.24119    0.48671  -0.496   0.6269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.498 on 16 degrees of freedom
## Multiple R-squared:  0.8123, Adjusted R-squared:  0.7654
## F-statistic: 17.32 on 4 and 16 DF,  p-value: 1.153e-05

## Analysis of Variance Table
##
```

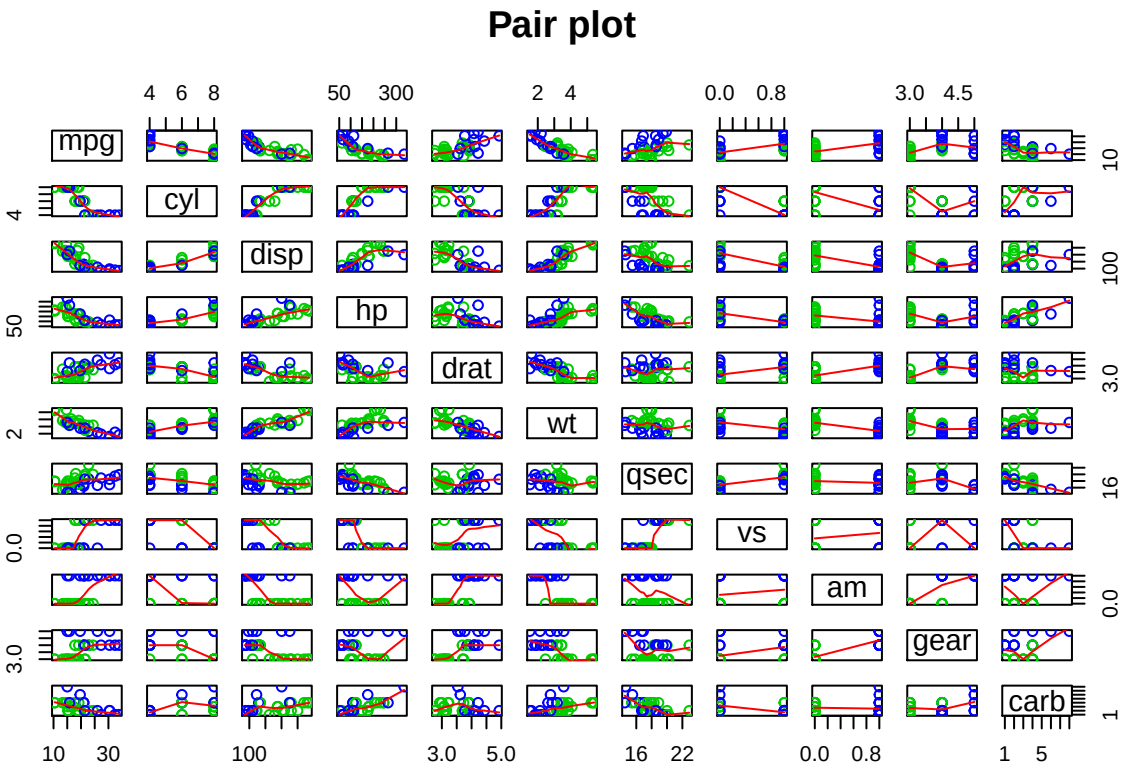

Figure 2: Figure 2 - Pair plot

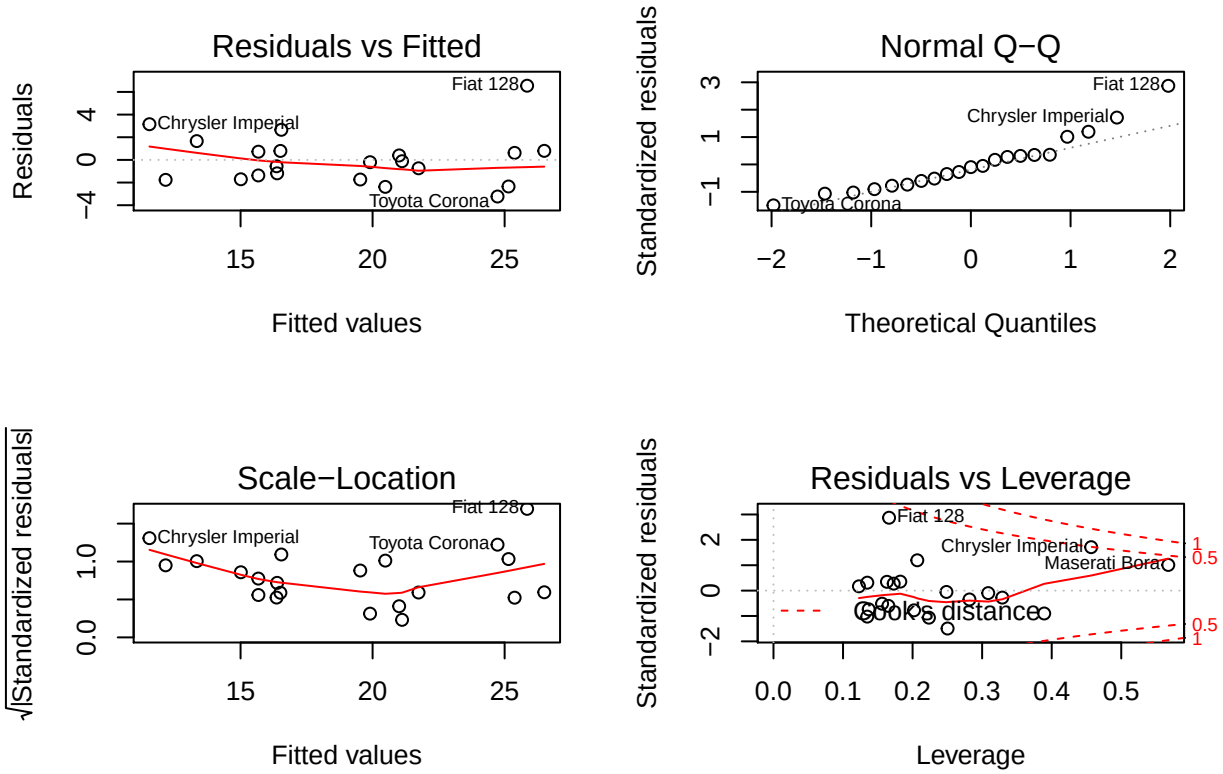

Figure 3: Figure 3 - Best model (cyl + wt + hp + carb) performance on training set

```
## Model 1: mpg ~ cyl + wt + hp + carb
## Model 2: mpg ~ cyl + wt + hp + carb + am
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 174.10
## 2     26 162.72  1    11.385 1.8191 0.1891
```

Anova - value 0.18 (it is > 0.05) doesn't mean an improvement