Regression Models - Course Project

Jens Hooge

17.9.2014

Executive Summary

The goal of this analysis is tho answer the question, whether automatic or manual transmission cars have a significant influence on fuel consumption. In particular we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). There fore we will **quantify the MPG difference between automatic and manual transmissions** and try to answer the question, **whether an automatic or manual transmission is better for MPG.**

Using a number of linear modeling techniques together with hypothesis testing, we concluded, that on average cars with manual transmission run **7.3 miles per gallon** farther than cars with automatic transmission. We could show, that this difference is mostly influenced by the car's **displacement, horsepower and weight**.

Note: Due to spacial constraints, raw R code won't be displayed here. Please refer to GitHub.

Parts of analysis:

- Exploratory Data Analysis
- Model Selection and Uncertainty Quantification
- Regression Diagnostics and Variable Importance

Exploratory Analyis

First let's have a look at the difference in fuel consumption between cars with manual and automatic transmission.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.4    15.0    17.3    17.1    19.2    24.4

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0    21.0    22.8    24.4    30.4    33.9

##
##  Welch Two Sample t-test
##
## data:  man_data$mpg and auto_data$mpg
## t = 3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.21 11.28
## sample estimates:
## mean of x mean of y
##     24.39     17.15
```

The mean difference of **7.3** between cars with automatic and manual transmission is significant [Fig.: 1] with about **5.5 mpg [p = 0.0014]** and a **95% confidence interval of [3.2097, 11.28]**, with a lower fuel consumption for cars with manual transmission. It is unclear however, which features are responsible for this difference. Before we answer this question, using a number of different linear models, we will compute the correlation of each of the features provided in the mtcars dataset with the response.

```
##          cyl    disp      hp   drat      wt   qsec    vs     am   gear
## [1,] -0.8522 -0.8476 -0.7762 0.6812 -0.8677 0.4187 0.664 0.5998 0.4803
##         carb
## [1,] -0.5509
```

The number of cylinders, displacement, horsepower and the weight of the cars show a highly negative correlation with the output/response variable, which indicate a strong influence on the fuel consumption. The pairwise correlation between all variables [Fig.: 3] have been computed and show **strong colinearities between the variables disp, wt and cyl**, indicating equal importances on the fuel consumption.

Model Selection and Uncertainty Quantification

To answer the question which and how many variables have an influence on the fuel consumption, a number of linear models have been fit to the data. To ensure robustness of the results, we performed a bootstrapped cross-validation scheme. Besides a naive model that included all variables, forward, backward and stepwise selection methods have been applied and their performance [Fig.: 4] and similarity [Fig.: 5] have been compared.

```
## Loading required package: leaps
```

```
##
## Call:
## summary.resamples(object = resamps)
##
## Models: normal, backward, forward, stepwise
## Number of resamples: 25
##
## RMSE
##          Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## normal   1.55    2.77   2.99 3.12    3.72 4.53    0
## backward 1.55    2.77   2.99 3.12    3.72 4.53    0
## forward  1.36    2.26   2.80 2.80    3.21 4.51    0
## stepwise 1.74    2.36   3.02 3.02    3.47 4.53    0
##
## Rsquared
##           Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's
## normal   0.531   0.672  0.794 0.779   0.895 0.944    0
## backward 0.531   0.672  0.794 0.779   0.895 0.944    0
## forward  0.641   0.761  0.852 0.827   0.886 0.987    0
## stepwise 0.594   0.754  0.809 0.801   0.855 0.962    0
```

The best performance could be achieved with the stepwise selection approach, which **explained 80.89%** of the variance and resulted in a **median RMSE of 3.021**. Even though explained variance in the forward selection approach was higher, this was accompanied by a higher RMSE. Note that, due to the prediction on the training data, during cross-validation, these values might be optimistic and more samples would be needed for more realistic performance measures.

Since models are fit on the same versions of the training data, it makes sense to make inferences on the differences between models. In this way we reduce the within-resample correlation that may exist. We can compute the differences, then use a simple t-test to evaluate the null hypothesis that there is no difference between models.

```
##
## Call:
## summary.diff.resamples(object = difValues)
##
## p-value adjustment: bonferroni
## Upper diagonal: estimates of the difference
## Lower diagonal: p-value for H0: difference = 0
##
## RMSE
##          normal backward forward stepwise
## normal           0.000    0.314   0.096
## backward NA               0.314   0.096
## forward  0.562  0.562            -0.218
## stepwise 1.000  1.000    1.000
##
## Rsquared
##          normal backward forward stepwise
## normal           0.0000  -0.0478 -0.0212
## backward NA              -0.0478 -0.0212
## forward  0.617  0.617             0.0266
## stepwise 1.000  1.000    1.000
```

The models show high similarities [Fig.: 5], with difference estimates between **-0.1797 (normal - forward)** and **0.2756 (forward -stepwise)**.

Diagnostics and Variable Importance

Given the model performance measures and similarity estimates, the **model that best explaines the mpg variable is the stepwise feature selection model** [Fig.: 3]. It includes all variables other than gear and the **three most important variables to explain fuel consumption, are displacement, weight and horsepower**[Fig.: 6].

Appendix

Fig 1 - Automatic and Manual Transmission

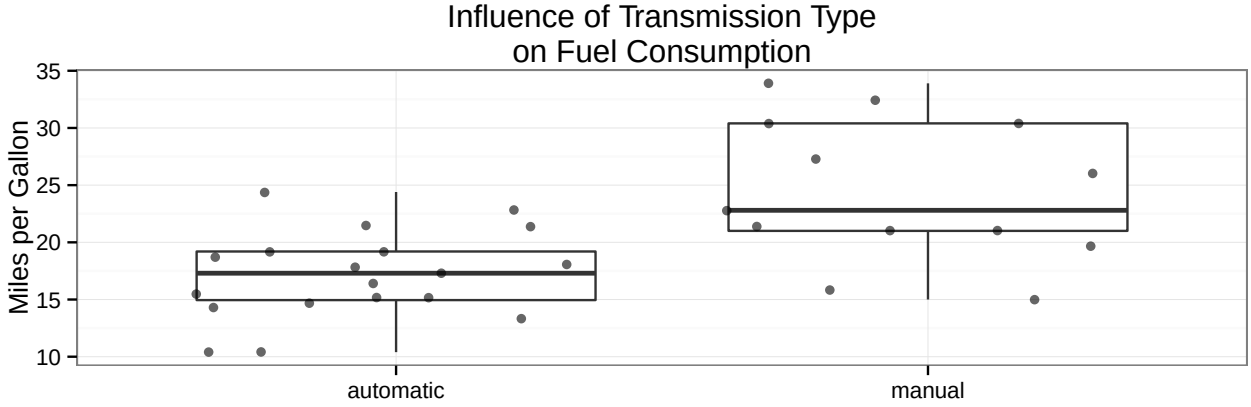

Fig 2 - Pairwise Correlation

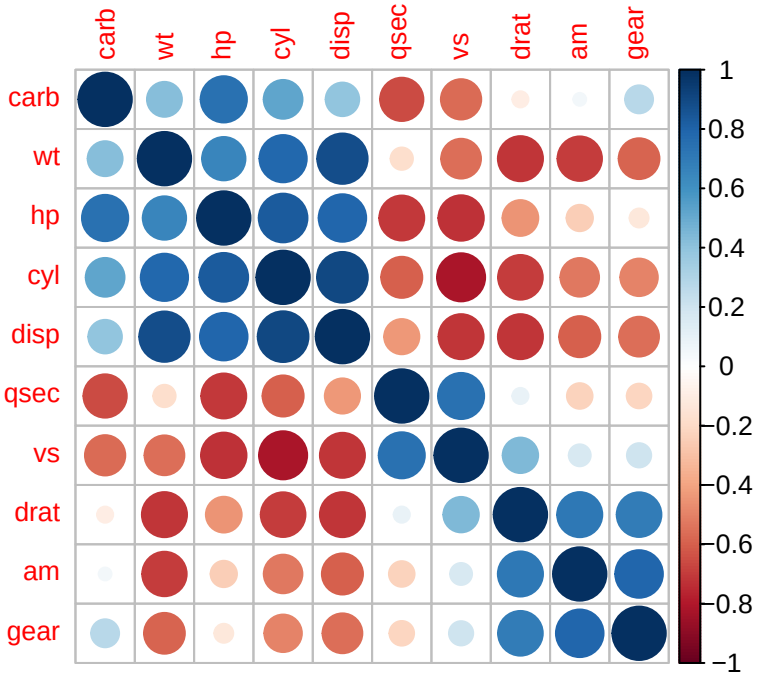

Fig 3 - Diagnostic Plots

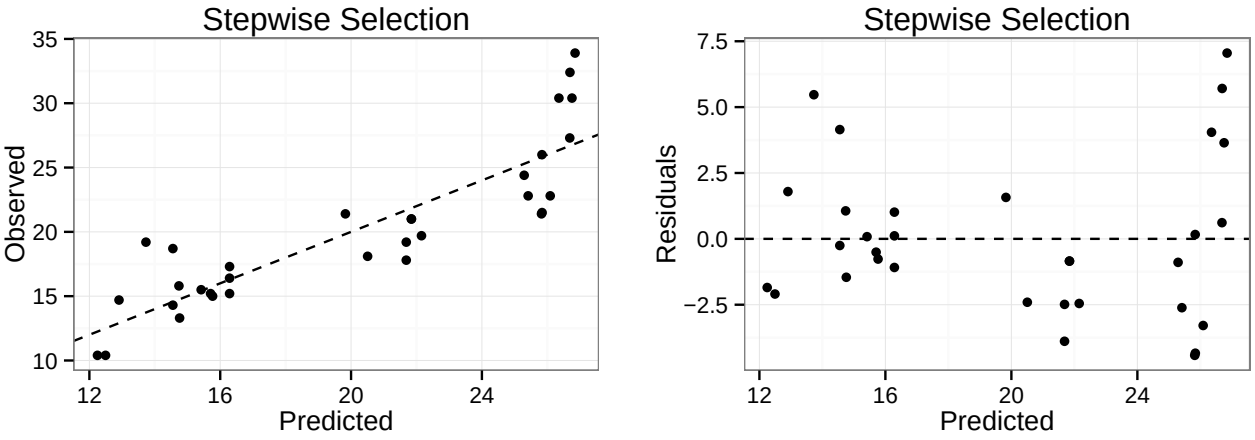

Fig 4 - Model Performance

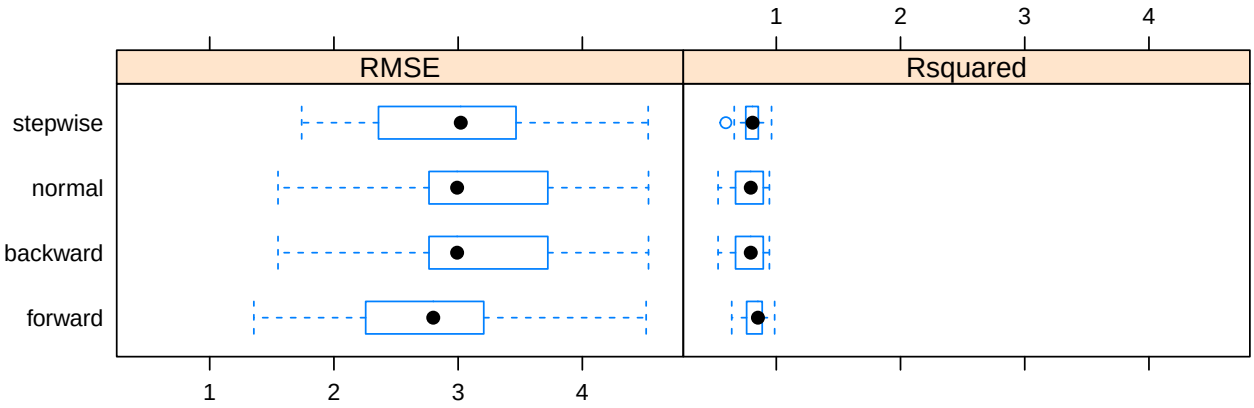

Fig 5 - Model Similarity

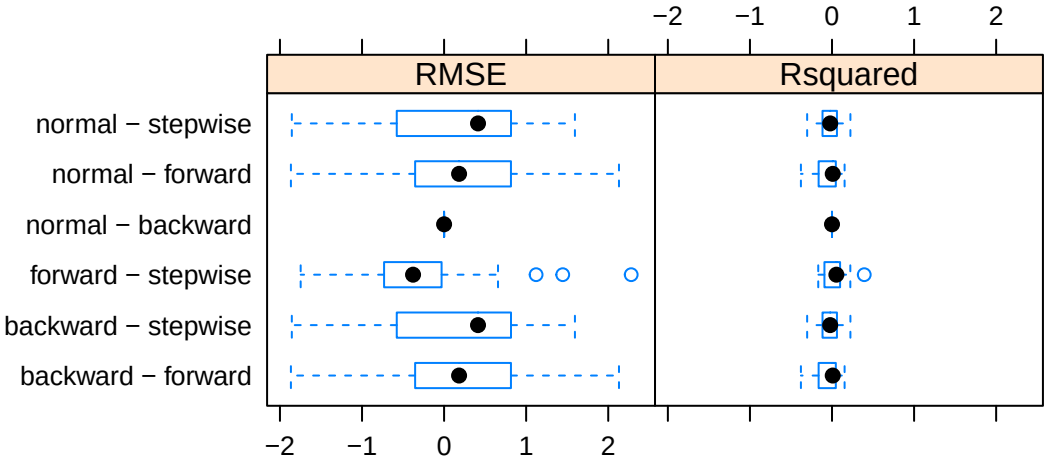

Fig 6 - Variable Importance