Regression Methods: Course Project

Author: Jens Berkmann

Abstract

In this report an investigations on the *mtcars* dataset provided by the R package was done. In particular the relationship between MilesPerGallon and automatic or manual transmission was investigated. It was found that manual transmission is significantly advantageous over automatic transmission with a difference of 7.24 MilesPerGallon. However, also the number of cylinders, the horse power as well as the weight of the car have an effect on the MilesPerGallon. In particular, when keeping the latter variables constant, cars with manual transmission exhibit on average an increase of 1.81 MilesPerGallon compared to automatic transmission.

Exploratory Data Analysis

We first load the data, convert some variables to factor variables and print a summary of the data set.

```r
library(knitr)
data(mtcars)
mtc      <- mtcars
names(mtc)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```r
mtc$cyl  <- as.factor(mtc$cyl)
mtc$vs   <- as.factor(mtc$vs)
mtc$am   <- as.factor(mtc$am)
mtc$gear <- as.factor(mtc$gear)
mtc$carb <- as.factor(mtc$carb)
str(mtc)
```

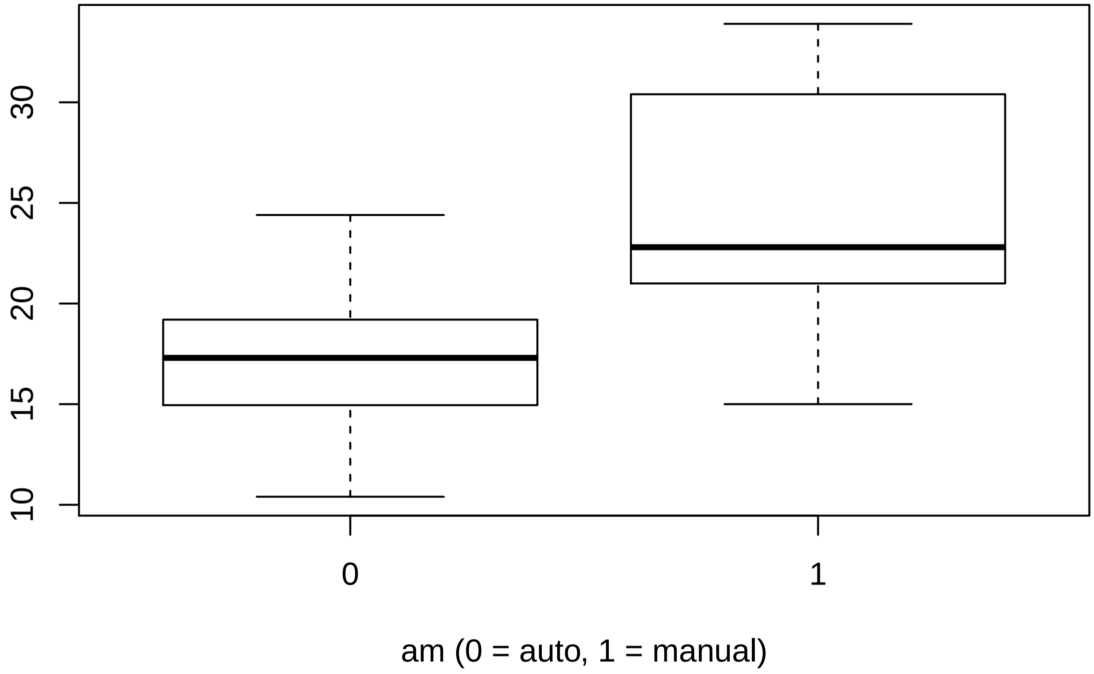
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

The dataframe contains of 32 rows (samples, observations) and 11 columns (variables). According to R's help function the meaning of the variables is as follows:

mpg Miles/(US) gallon; cyl Number of cylinders; disp Displacement (cu.in.); hp Gross horsepower; drat Rear axle ratio; wt Weight (lb/1000); qsec 1/4 mile time, vs V/S; am Transmission (0 = automatic, 1 = manual); gear Number of forward gears; arb Number of carburetors.

The variable *mpg* contains the outcome (Miles/(US)gallon) we are interested in. We are also in particular interested in the column *am* which distinguishes between automatic (am=0) and manual transmission (am=1). As a basic analysis let us plot the outcome *mpg* versus the variable *am*.

```
boxplot(mpg ~ am, data = mtc, xlab = "am (0 = auto, 1 = manual)")
```



It is seen that manual transmission results in higher (better) values of mpg. Moreover, from the plot the separation of the two am-groups are fairly well separated. Still, let's check via t-testing whether the difference in the means is significant.

```
mytest <- t.test(mpg ~ am, data = mtc, var.equal=FALSE)
mytest
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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:
##  -11.28  -3.21
## sample estimates:
## mean in group 0 mean in group 1
##           17.15           24.39
```

From the reported fairly small p-value we conclude that the difference in means of $24.39-17.15=$7.24 is not observed just by accident.

Regression Models

Let us now turn to the task of (regression) model selection. Via R's regsubsets function we find the best model starting from the model using all predictor variables.

```
library(leaps)
simple <- lm(mpg ~ am, data = mtc)
full   <- lm(mpg ~ .,  data = mtc)
best   <- step(full, data=mtc, direction="both", trace=0)
summary(best)
```

```
## 
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtc)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -3.939 -1.256 -0.401  1.125  5.051 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  33.7083     2.6049   12.94  7.7e-13 ***
## cyl6         -3.0313     1.4073   -2.15   0.0407 *  
## cyl8         -2.1637     2.2843   -0.95   0.3523    
## hp           -0.0321     0.0137   -2.35   0.0269 *  
## wt           -2.4968     0.8856   -2.82   0.0091 ** 
## am1           1.8092     1.3963    1.30   0.2065    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.84 
## F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

The best linear model is found to explain 84% of the data variance and uses apart from the variable *am* also the variables *cyl*, *hp* and *wt* which is also meaningful from a common sense perspective. Let us compare now the simple model (mpg~am) versus the model found by the last optimization method and check its significance. We do this by calling the anova-funcion of R.

```
anova(simple,best)
```

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)    
## 1     30 721                              
## 2     26 151  4       570 24.5 1.7e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the reported tiny P-value we conclude that the optimized model *best* explains the data better than the simple model not just by accident. However, when looking at the high P-value of the *beta*-coefficient in the just found model

```r
summary(lm(mpg ~ am+cyl+hp+wt, data = mtc))
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.939 -1.256 -0.401  1.125  5.051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7083     2.6049   12.94  7.7e-13 ***
## am1           1.8092     1.3963    1.30   0.2065
## cyl6         -3.0313     1.4073   -2.15   0.0407 *
## cyl8         -2.1637     2.2843   -0.95   0.3523
## hp           -0.0321     0.0137   -2.35   0.0269 *
## wt           -2.4968     0.8856   -2.82   0.0091 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.84
## F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

we have doubts that including *am* is significant. Therefore we perform a last verification step by checking whether excluding the variable *am* from the found model is significant or not.

```r
testfit <- lm(mpg ~ cyl+hp+wt, data = mtc)
anova(best,testfit)
```

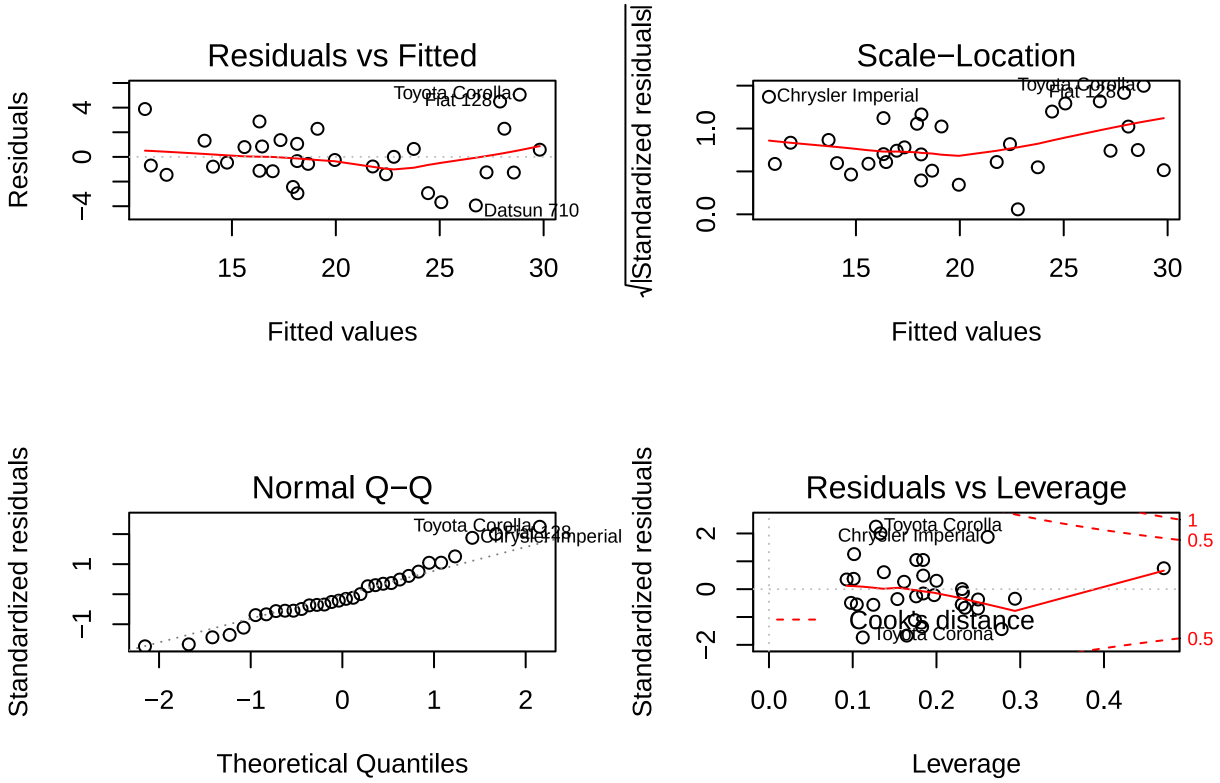
```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ cyl + hp + wt
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     26 151
## 2     27 161 -1     -9.75 1.68   0.21
```

From the observed P-value we reject the exclusion of the prediction variable *am*. The final model is therefore selected to be $mpg = 33.71 - 3.03 \cdot cyl_6 - 2.16 \cdot cyl_8 - 0.03 \cdot hp - 2.5 \cdot wt + 1.81 am_{manual}$.

Residual Analysis

We finally do a residual analysis on the selected model *best*.

```r
layout(matrix(c(1,2,3,4),2,2))
plot(best)
```



No suspicious patterns are found in the upper left figure. Moreover, from the lower left plot the normality assumption of the data seems to be sufficiently given.