Regression Models: Cars And MPG

DA

May 27, 2016

Executive Summary

Scenario: you work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:
- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions

This brief analysis considers cars fuel economy and a set of variables, in particular the car transmission (automatic versus manual). We use mtcars dataset made available by Motor Trend Magazine.

We perform exploratory data analysis followed by finding predictors for MPG. The first part of analysis looks into the variables and picks some and evaluates their performance as predictors for MPG. In the second part we use multivariable regression where we allow an automatic selection of the set of best predictors.

Prerequsites

Loading the libraries and the dataset.

```r
library(ggplot2)
library(UsingR)
require(GGally)
library(relaimpo)
data(mtcars)
```

The Data - performing data exploration

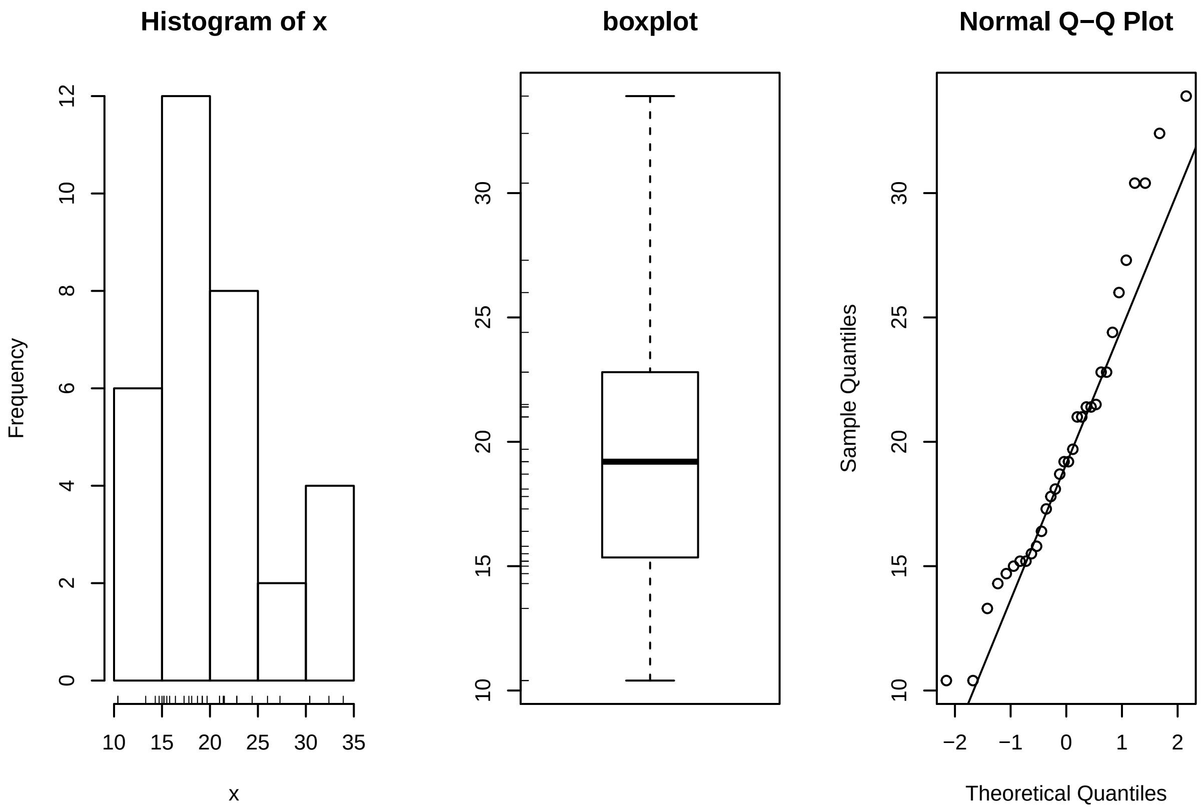
The mtcars dataset, data frame with 32 observations on 11 variables (source: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html).

[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 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

MPG distribution

As MPG is the main subject of interest we are looking into how this variable is distributed using some normal probability plots.

```
simple.eda(mtcars$mpg)
```



We see in the above plots that:
- the cars within 15 to 25 mpg are more frequent (histogram)
- the mpg is symmetrically distributed with a regular tale, as opposed to skewed (boxplot)
- the QQ plot indicates that we can approximate this distribution as normal

Shapiro-Wilk Test - MPG distribution

Using Shapiro-Wilk Normality Test and getting a p-value = 0.1229 which translates in the fact that we cannot reject reject the NULL hypothesis that the samples came from a normal distribution.

```
shapiro.test(mtcars$mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.9476, p-value = 0.1229
```

Exploring correlations

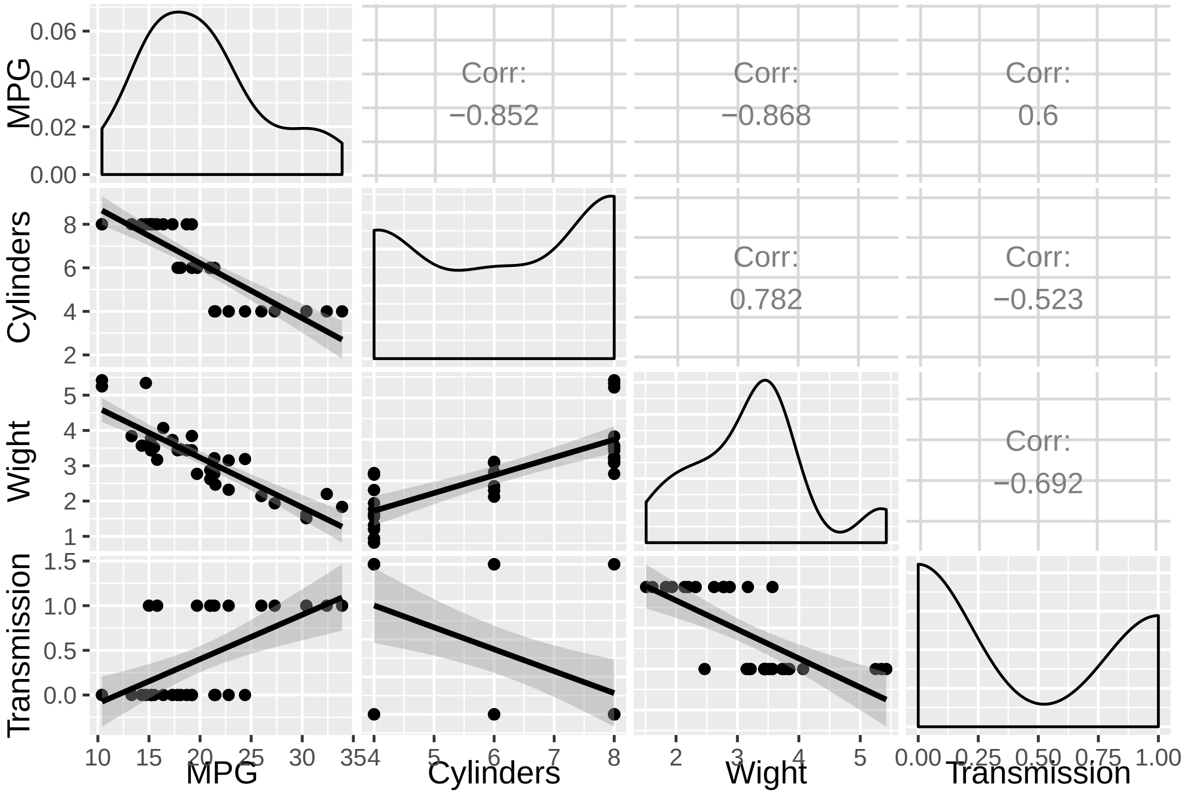
We pick a few variables and look for correlations between them. We ignored displacement, horsepower as they correlate with cylinders. We are mostly into correlations with MPGs.

```
cor(mtcars[,c("mpg", "cyl", "wt","am", "gear")])
```

```
##             mpg        cyl         wt         am       gear
## mpg   1.0000000 -0.8521620 -0.8676594  0.5998324  0.4802848
## cyl  -0.8521620  1.0000000  0.7824958 -0.5226070 -0.4926866
## wt   -0.8676594  0.7824958  1.0000000 -0.6924953 -0.5832870
## am    0.5998324 -0.5226070 -0.6924953  1.0000000  0.7940588
## gear  0.4802848 -0.4926866 -0.5832870  0.7940588  1.0000000
```

We see a strong negative (-0.87) correlation between MPG and the weight of the car (as the car becomes heavier it gets lesser MPG). Number of gears doesn't seem to be an interesting variable.
Another way to look into correlations:

```
cp1 = ggpairs(mtcars, columns =c("mpg", "cyl", "wt","am"),columnLabels = c("MPG", "Cylinders","Wight",
             continuous = "smooth",combo = "facetdensity", diag = NULL)
)
cp1
```



We have a correlation factor of -0.852 between MPG and the Number of Cylinders which we can interpret as a strong downhill linear relationship which makes sense as a greater number of cylinders means generally a

bigger engine consuming more gas.
MPG and transmission correlate with a coefficient of 0.6 which is a moderate positive relationship. The data set models the manual transmission as 1 and automatic as 0, hence one can say that a manual transmission tends to provide a better MPG.

T-test - comparing the means of MPG, transmission and number of cylindres

```
t.test(mtcars$mpg, mtcars$am, var.equal=TRUE, paired=FALSE)
```

```
##
##  Two Sample t-test
##
## data:  mtcars$mpg and mtcars$am
## t = 18.4126, df = 62, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  17.54734 21.82141
## sample estimates:
## mean of x mean of y
##  20.09062   0.40625
```

```
t.test(mtcars$mpg, mtcars$cyl, var.equal=TRUE, paired=FALSE)
```

```
##
##  Two Sample t-test
##
## data:  mtcars$mpg and mtcars$cyl
## t = 12.5116, df = 62, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11.68184 16.12441
## sample estimates:
## mean of x mean of y
##  20.09062   6.18750
```

We get a very small p-value which translates in the fact that the difference in means is not by chance but indeed there is clear difference in these two populations. So we reject the NULL hypothesis: “these two samples have the same means.”

Regression Models

Single Variable Models

Starting by building some simple (non multivariable) regression model: mpg function of cylinders, weight and transmission

```
model.1 = lm(mpg~cyl, data = mtcars)
coef(model.1)
```

```
## (Intercept)         cyl
##    37.88458    -2.87579
```

```
model.2 = lm(mpg~wt, data = mtcars)
coef(model.2)
```

```
## (Intercept)          wt
##   37.285126   -5.344472
```

```
model.3 = lm(mpg~am, data = mtcars)
coef(model.3)
```

```
## (Intercept)          am
##   17.147368    7.244939
```

We see intercepts of 37, 37 and 17 respectively. The slope is negative for the first two models and positive for the third that we interpret as follows:
- for every additional cylinder we lose 2.9 MPGs
- for every 1000 pounds of additional weight we lose 5.3 MPGs
- manual transmission provides a gain of 7 MPGs

Residual Analysis

It is considered that residual standard error is a better approximation of the model goodness than the R-squared.

```
summary(model.1)$sigma
```

```
## [1] 3.205902
```

```
summary(model.1)$r.squared
```

```
## [1] 0.72618
```

```
summary(model.2)$sigma
```

```
## [1] 3.045882
```

```
summary(model.2)$r.squared
```

```
## [1] 0.7528328
```

```
summary(model.3)$sigma
```

```
## [1] 4.902029
```

```r
summary(model.3)$r.squared
```

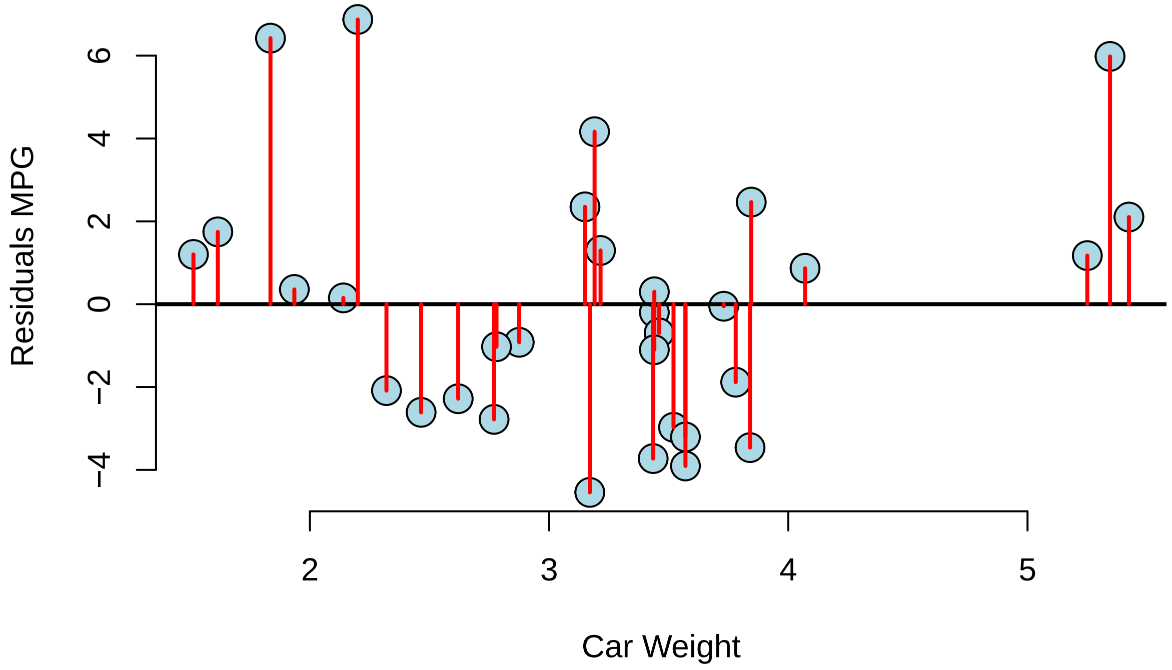
```
## [1] 0.3597989
```

We see above the the second model (vehicle weight as independent variable) performs best as MPG predictor with a residual standard error of 3 and an R-squared of 0.75 while the third one (transmission) is the worst with an error of 4.9 and a R-squared of 0.36.

Plotting Residuals

We plot as example the model using the weight as independent variable.

```r
plot(mtcars$wt, resid(model.2), xlab = "Car Weight", ylab = "Residuals MPG", bg = "lightblue", col = "bl
abline(h = 0, lwd = 2)
for (i in 1 : length(mtcars$wt))
        lines(c(mtcars$wt[i], mtcars$wt[i]), c(resid(model.2)[i], 0), col = "red" , lwd = 2)
```



We see a pattern less plot.

Comparing the models

We use Anova to compare the three models.

```r
anova(model.1, model.2, model.3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl
## Model 2: mpg ~ wt
## Model 3: mpg ~ am
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     30 308.33
## 2     30 278.32  0     30.01
## 3     30 720.90  0   -442.57
```

The second model with an RSS of 278 is the winner.

Multivariable Regression

Selecting the Right Predictors

Building a regression model using Backward Stepwise Regression (starting with all predictors and removes the ones that are not statistically significant).

```r
initial.model <- lm(mpg ~., data= mtcars)
best.model <- step(initial.model, direction = "both")
```

Analyzing the selected predictors

We are looking into the independent variables picked.

```r
summary(best.model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

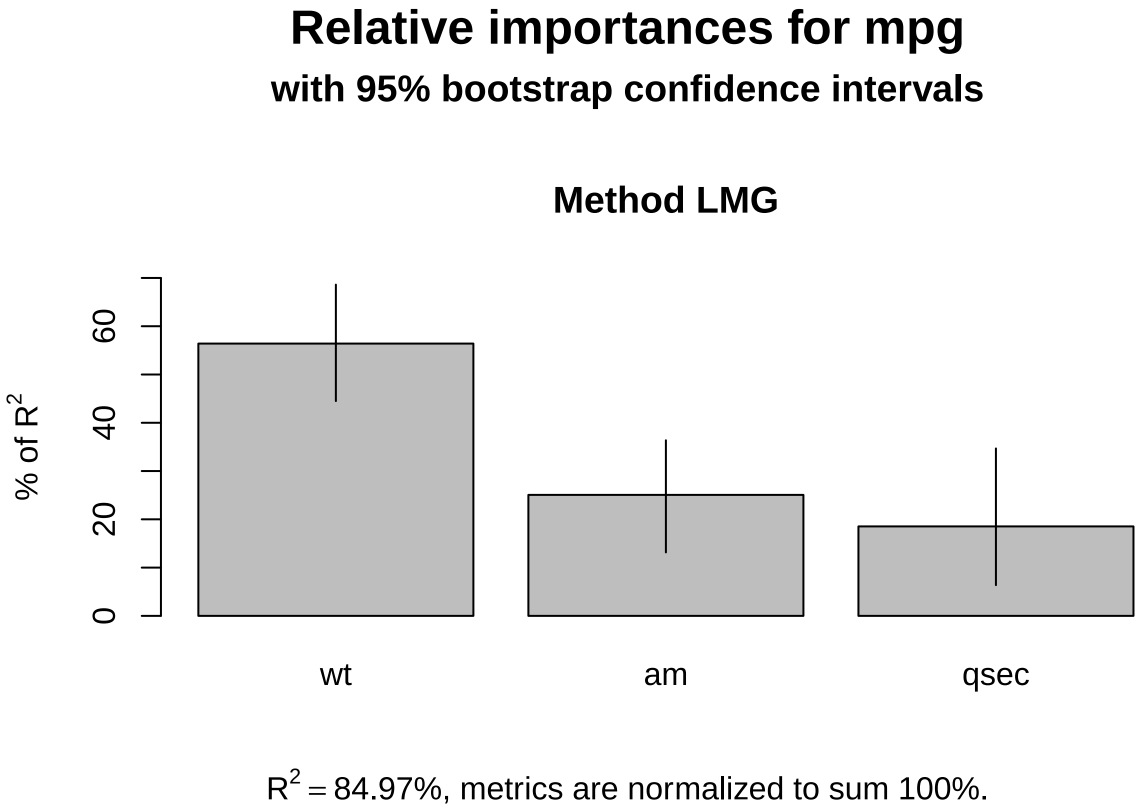
We see (p value) the weight of the car and qsec making a bigger difference than the transmission.
We also calculate the relative importance of each variable using the relaimpo library.

```
calc.relimp(best.model)$lmg
```

```
##        wt      qsec        am
## 0.4792448 0.1574791 0.2129397
```

We can find above the r-squared averaged over orderings among regressors - this time weight is again most important regressor. The transmission come in second followed by qsec.
We can also plot the relative importance of each independent variable by bootstrapping using 500 samples.

```
boot <- boot.relimp(best.model, b = 500, rank = TRUE, diff = TRUE, rela = TRUE)
plot(booteval.relimp(boot,sort=TRUE))
```



Sumarizing the best regression model

The most significant variables are identified as being:
- the weight of the car (wt)
- the 1/4 mile time (qsec) The transmission seems to be of a lower importance as predictor of MPG.

We see a really small p-value (1.21e-11 much smaller than 0.05) meaning a good model. We also see an decent R-squared: 0.8497 which translates to outcome variance explained by this model.

```r
summary(best.model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```r
m1 = lm(mpg ~ wt + disp + cyl+gear+am, data = mtcars);
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## wt         1 847.73  847.73 120.8346 2.852e-11 ***
## disp       1  31.64   31.64   4.5099   0.04338 *
## cyl        1  58.19   58.19   8.2944   0.00786 **
## gear       1   2.65    2.65   0.3779   0.54409
## am         1   3.44    3.44   0.4898   0.49022
## Residuals 26 182.41    7.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```