Regression

thiemo.meeuwissen@gmail.com

Friday, November 21, 2014

This is my analysis for the assignment in the Coursera Regression course by Brian Caffo from the John Hopkins University.

Context

For this assignment I 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). 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"

Data

The data for this assignment is the mtcars data set. So what is this data about? Checking the help provides the required insight.

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

The data format is data frame with 32 observations on 11 variables.

[, 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 (lb/1000)
[, 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

The data source is Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411.

I started by giving the data a quick str to see what the data structure looks like.

```r
data(mtcars)
str(mtcars)
```

It turned out all features have numeric data types. Because from the description it is clear some features, including the "am" feature I'll be studying, are really factors, I'll convert the data types before starting the analysis.

In addition, the unit "Miles per Gallon" is not ideal. Based on the physics involved a better measure would be "Gallon per Miles". This is discussed in more detail on for example http://www.mpgillusion.com. I created a new variable gpm to reflect this. To make the numbers "nicer" the unit will be gallon per 10'000 miles.

```r
mtcars$gpm <- 10000/mtcars$mpg
mtcars$vs <- factor(mtcars$vs, labels=c("V","Straight"))
mtcars$am <- factor(mtcars$am, labels=c("Automatic","Manual"))
```

Now let's look at the transformed data structure.

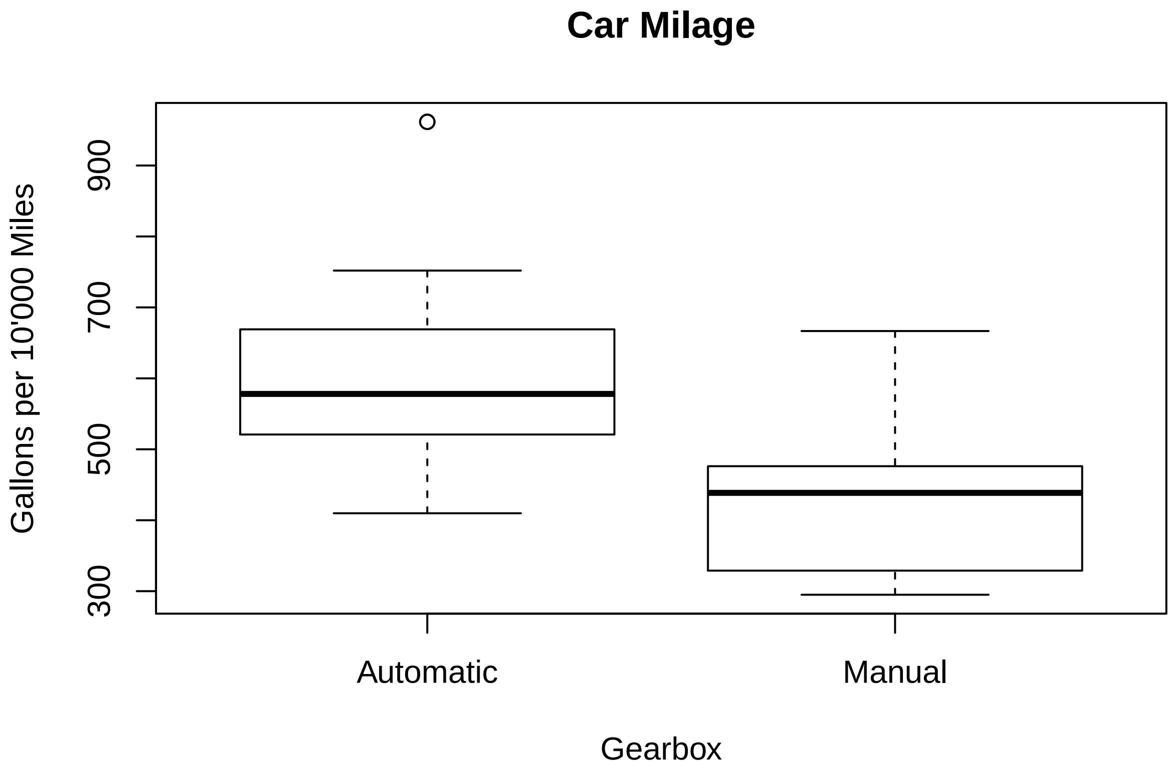
```
## 'data.frame':    32 obs. of  12 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 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 "V","Straight": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
##  $ gpm : num  476 476 439 467 535 ...
```

Exploratory Data Analysis (EDA)

For convenience we create a dataframe where we use gpm as outcome and remove mpg. This dataframe will be called fmtcars.

```r
fmtcars <- subset(mtcars, select=-mpg)
```

Next, we check the data distribution for the first question to be answered "Is an automatic or manual transmission better for mpg".



The mean gpm of cars with manual transmission is about 200 gpms lower than that of cars with automatic transmission. Let's run a t-test to check if the difference is significant. As null hypothesis we formulate that cars with automatic transmission are having equal mileage than cars with manual transmission.

```
t <- t.test(fmtcars[fmtcars$am=="Automatic",]$gpm,
            fmtcars[fmtcars$am=="Manual",]$gpm)
```

```
##
##  Welch Two Sample t-test
##
## data:  fmtcars[fmtcars$am == "Automatic", ]$gpm and fmtcars[fmtcars$am == "Manual", ]$gpm
## t = 3.6912, df = 29.493, p-value = 0.0009018
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   79.31344 276.09229
## sample estimates:
## mean of x mean of y
##  614.4642  436.7613
```

```
## The p-value is 0.0009018232 so we and reject the null hypothesis and conclude that within the dataset
```

However, to conclude that the difference is actually caused by the tranmission type is not clear yet. To This conclusion would be true only if all other characteristics are the same. For example, cars with automatic transimission should have the same weight and horsepower distribution and relation as cars with manual transmission). As can be seen from the scatter matrix plot in the appendix this is not the case here.

Correlation Analysis

To get an idea of the relation between gpm and the other features we have a look at the linear correlation between gpm and other features.

```r
sort(cor(fmtcars[11], fmtcars[c(1:6,9:10)])[1,])
```

```
##       drat       gear       qsec       carb         hp        cyl 
## -0.6380538 -0.4792352 -0.3858480  0.5263402  0.7629477  0.8137493 
##       disp         wt 
##  0.8798217  0.8898927
```

The correlations confirm what we learned from the scatter plots in the appendix. Increased power, weight, displacement, cylinders and carburators are correlated with increased fuel consumption. Quarter mile time, number of forward gearsand rear axle ratio are correlated with decreased fuel consumption. Moreover, it looks like weight, displacement and horsepower are the most relevant features. What we do not see is the effect of the categorical features.

Multivariate Linear Regression

First let's just create a model using all features without interactions.

```r
fit <- lm(gpm~., data = fmtcars)
summary(fit)
```

```
## 
## Call:
## lm(formula = gpm ~ ., data = fmtcars)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -170.499  -33.109    4.737   38.263  117.856 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 575.1712   517.9898   1.110    0.279
## cyl         -18.2354    28.9195  -0.631    0.535
## disp          0.3999     0.4942   0.809    0.427
## hp            0.2876     0.6024   0.477    0.638
## drat         -6.3192    45.2565  -0.140    0.890
## wt           86.2705    52.4251   1.646    0.115
## qsec        -11.4173    20.2250  -0.565    0.578
## vsStraight    9.0979    58.2392   0.156    0.877
## amManual     18.3750    56.9148   0.323    0.750
## gear        -46.3261    41.3238  -1.121    0.275
## carb         19.1364    22.9345   0.834    0.413
## 
## Residual standard error: 73.34 on 21 degrees of freedom
## Multiple R-squared:  0.8649, Adjusted R-squared:  0.8006 
## F-statistic: 13.45 on 10 and 21 DF,  p-value: 5.15e-07
```

To select the “best model” we will use the step method which runs lm multiple times and select the best variables. The command and result are shown below.

```
bestfit <- step(fit, direction="both")
```

```
summary(bestfit)
```

```
##
## Call:
## lm(formula = gpm ~ disp + wt + carb, data = fmtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -157.996  -30.783    1.104   40.677  113.025
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 123.7867    49.3593   2.508  0.01822 *
## disp          0.5452     0.2085   2.615  0.01420 *
## wt           75.7992    26.8343   2.825  0.00863 **
## carb         17.3636     8.1373   2.134  0.04176 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.11 on 28 degrees of freedom
## Multiple R-squared:  0.8537, Adjusted R-squared:  0.838
## F-statistic: 54.45 on 3 and 28 DF,  p-value: 8.307e-12
```

Now interestingly enough, the transmission is NOT included in the final model. SO we could conclude that transmission is not significant in preduction fuel consumption.

The model explains 85% of the variance in gallons per mile (gpm). Moreover, we see that weight is the main feature related to milage.

Further Discussion

To double check the obtained result we create and alternative model where we add transmission (am) to the best model and check if there is indeed no signinficatn improvement.

```
altfit <- lm(gpm ~ disp + wt + carb + am, data = fmtcars)
```

```
anova(bestfit, altfit)
```

```
## Analysis of Variance Table
##
## Model 1: gpm ~ disp + wt + carb
## Model 2: gpm ~ disp + wt + carb + am
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 122364
## 2     27 122276  1    88.624 0.0196 0.8898
```

The p-value is 0.8898, so we accept the null hypothesis and claim that transmission is indeed not significant in predicting fuel consumption.

As discussed above, miles per gallon is actually not a good measure of fuel consumption which is why we used gallons per 10'000 miles instead. But, what would happen if we stick to strickly answering the questions posed for the assignment.

- "Is an automatic or manual transmission better for mpg"
- "Quantify the mpg difference between automatic and manual transmissions"

Following the same analysis as above and just skipping straight to the final result yields.

```r
data(mtcars)
fit <- lm(mpg~., data = mtcars)
bestfit <- step(fit, direction="both")
```

```r
summary(bestfit)
```

```
##
## 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
```

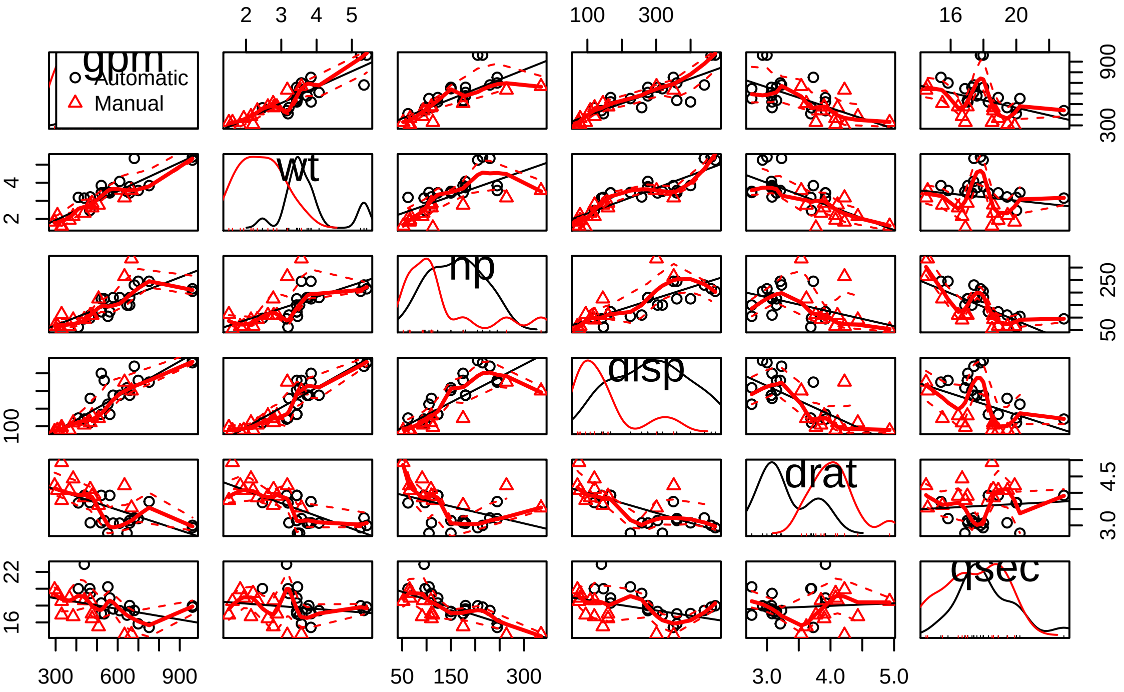
So now manual transmission cars have 2.9 mpgs more than automatic transmission cars. However this effect is much less strong than weight and acceleration (qsec). However, as shown before, if we try to answer the underlying question if automatic transmission increases fuel consumption the answer is no.

I conclude that tranmission type is not relevant for fuel consumption prediction, at least within the data available in the mtcars dataset. However, tranmission type is relevant for miles per gallon prediction. More material to fuel the miles per gallon illusion discussion!

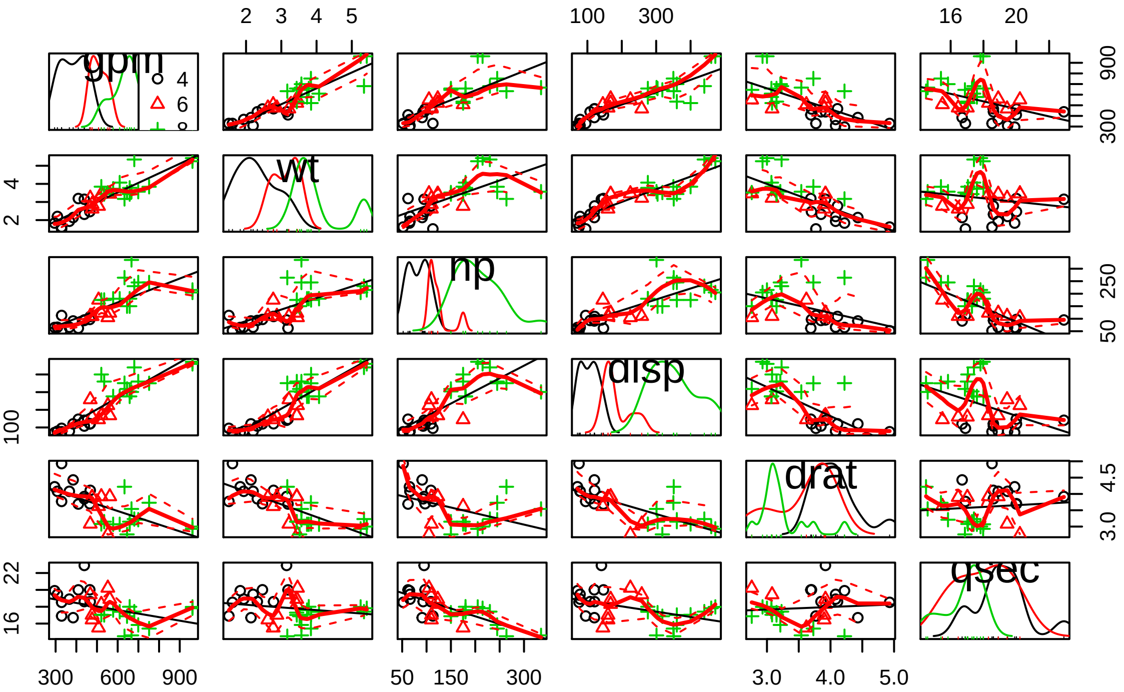
Appendix

```
## Loading required package: car
```

Car Milage by Transmission



Car Milage by Cylinders


gpm
wt
hp
disp
drat
qsec