Regression Models Assignment - Car data analysis

Fabien Nugier

11/17/2019

Introduction

This project exploits data from the R dataset `mtcars` and aims at exploring the relationship between a set of variables and the outcome variable of miles per gallons (MPG). More precisely, we want to know which of automatic or manual transmission is better for MPG and to quantify the difference between the two categories of cars. We will then explore further fits as an application of the course material.

Data Processing

We load the data and display its first rows:

```r
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

From the help page of the dataset (`?mtcars`), we get that the data consists of a data frame with 32 observations on 11 (numeric) variables:

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

The two columns that interest us the most are `mpg` and `am`. Let us display summary information about them:

```r
summary(mtcars[,c("mpg","am")])
```

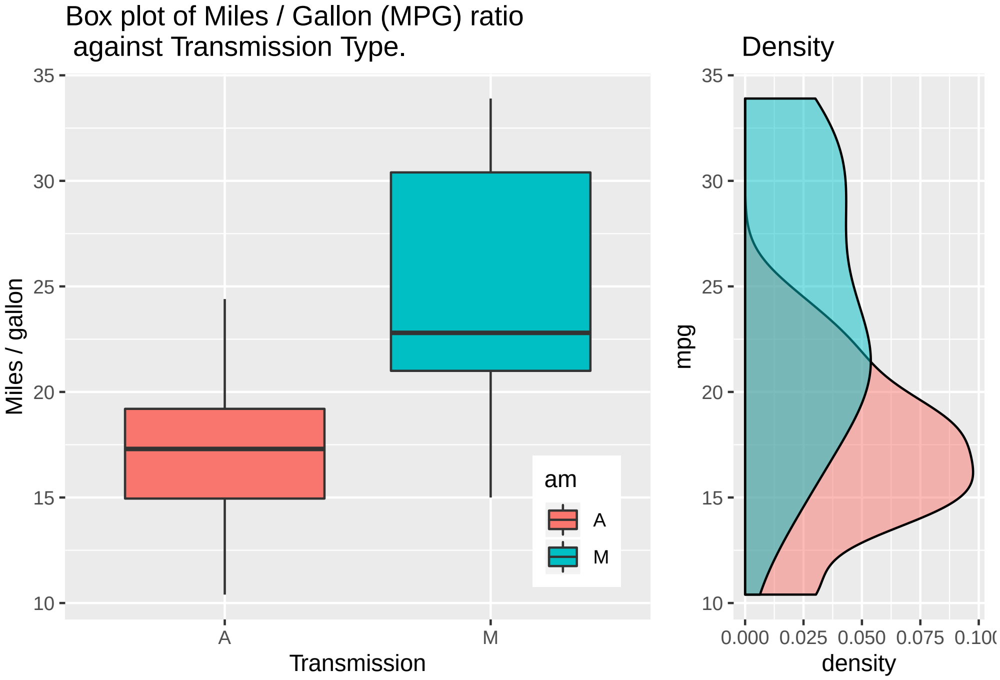
```
##       mpg              am        
##  Min.   :10.40   Min.   :0.0000  
##  1st Qu.:15.43   1st Qu.:0.0000  
##  Median :19.20   Median :0.0000  
##  Mean   :20.09   Mean   :0.4062  
##  3rd Qu.:22.80   3rd Qu.:1.0000  
##  Max.   :33.90   Max.   :1.0000  
```

Let us make the `am` variable a factor variable:

```
mtcars$am <- factor(mtcars$am, levels=c(0,1), labels=c("A","M"))
```

and we can plot the MPG output agains the transmission type as a box plot:

```
library(ggplot2)
library(gridExtra)
g1 = ggplot(data=mtcars, aes(x=am,y=mpg, fill=am)) + geom_boxplot()
g1 = g1 + scale_x_discrete("Transmission") + scale_y_continuous("Miles / gallon")
g1 = g1 + ggtitle("Box plot of Miles / Gallon (MPG) ratio \n against Transmission Type.")
g1 = g1 + theme(legend.position=c(0.95,0.3), legend.justification=c(1,1))
g2 = ggplot(mtcars, aes(x=mpg, fill=am)) + geom_density(alpha=0.5)
g2 = g2 + coord_flip() + theme(legend.position="none") + labs(title="\n Density")
grid.arrange(g1,g2,ncol=2,nrow=1,widths=c(4,2))
```



Data fitting

We can do a regression of MPG with the transmission taken as a factor variable. The regression is done as follows:

```
fit <- lm(mpg~factor(am), data=mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## factor(am)M    7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

where we can see that the average MPG is 17.147 for automatic transmission and 17.147+7.245=24.392 for manual transmission. This suggests, as the boxplot, that **a manual transmission allows more milage per gallon of gas**. The fit also gives us the standard error for each category of cars, so we have:

- $MPG(A) = 17.147 \pm 1.125$ miles per gallon
- $MPG(M) = 24.392 \pm 1.764$ miles per gallon

The residual standard error is 4.902 while $R^2 = 0.3598$. Since this value is far from 1, this suggests that a better fit can be obtained (as we will explore after).

In addition, we can perform a T-test on the data:

```
t.test(mpg~factor(am),data=mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by factor(am)
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group A mean in group M
##        17.14737        24.39231
```

We thus get that the difference in the means of samples A and M is different from zero, the null hypothesis ("difference equals zero") being rejected with a p-value less than 1%. This is a clear indication that the transmission plays a significant role in predicting MPG values.

Exploring further fits

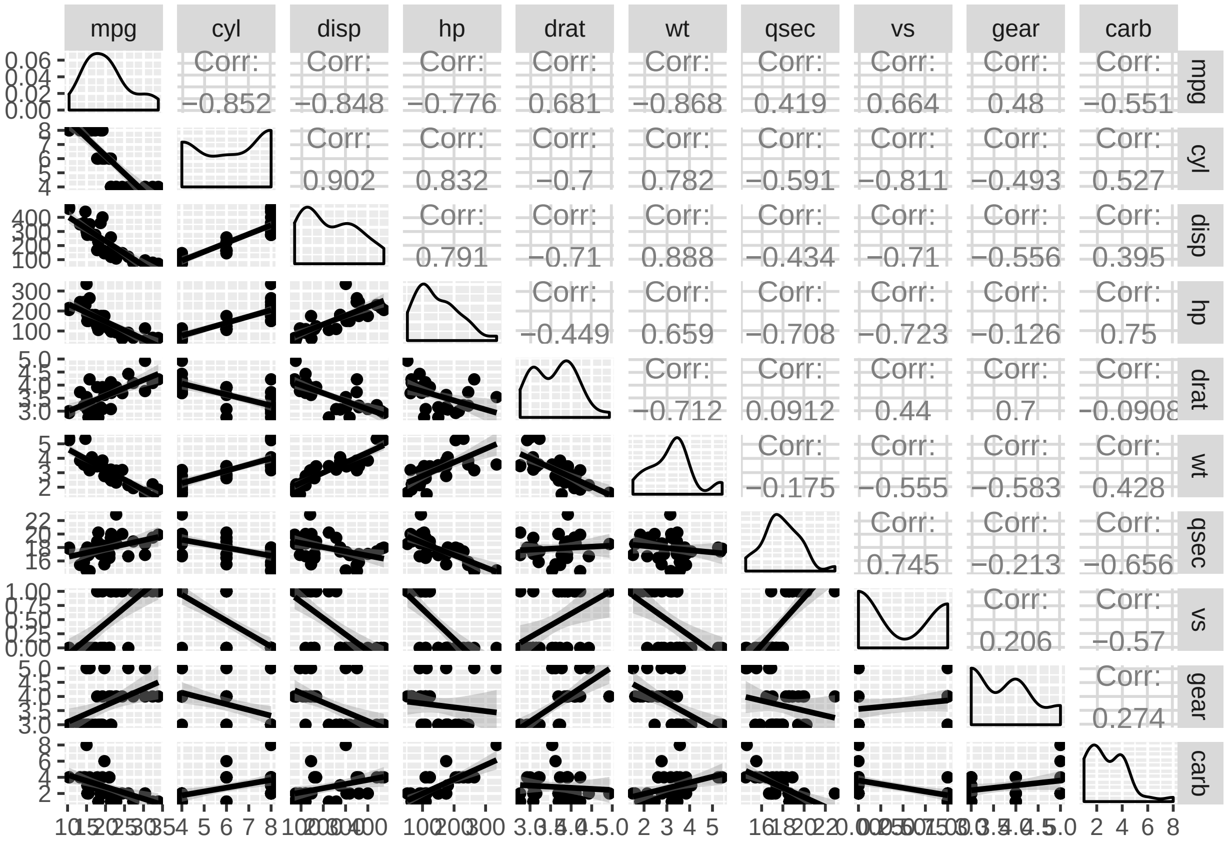
Let us try to improve the fit since the R^2 value was pretty low. We first look into the correlations between the different variables at our disposal, excluding the am variable:

```
require(GGally)
```

```
## Loading required package: GGally
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(mtcars[, !(names(mtcars) %in% c("am"))] , lower=list(continuous="smooth"))
```



From the first row we can see that `mpg` is strongly anti-correlated with `cyl`, `disp` and `wt`, respectively the number of cylinders, the displacement of the cylinders and the weight of the car. Let us first include cylinders into the linear regression:

```
fit2 <- lm(mpg ~ factor(am) + cyl, data=mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6856 -1.7172 -0.2657  1.8838  6.8144
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.5224     2.6032  13.262 7.69e-14 ***
## factor(am)M   2.5670     1.2914   1.988   0.0564 .
## cyl          -2.5010     0.3608  -6.931 1.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.059 on 29 degrees of freedom
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.7424
## F-statistic: 45.67 on 2 and 29 DF,  p-value: 1.094e-09
```

As we can see, considering the number of cylinders increased the quality of the fit significantly, with $R^2 =$

0.759 now. We now have a linear dependence with `cyl` described by the slope of -2.5010 MPG units / cylinder. The sign can be understood from the fact that the more cylinders a car motor has, the more consumption there is from this motor (in general).

Let us try to include the 2 other variables into a new fit:

```
fit3 <- lm(mpg ~ factor(am) + cyl + disp + wt, data=mtcars)
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + cyl + disp + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## factor(am)M  0.129066   1.321512   0.098  0.92292
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## disp         0.007404   0.012081   0.613  0.54509
## wt          -3.583425   1.186504  -3.020  0.00547 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

As we can see the fit increased in quality again, but the low value of the slope for the variable `disp` and it's high probability of equating zero, as suggested by the t-value probability which is very large. Hence we can remove this variable from the fit:

```
fit4 <- lm(mpg ~ factor(am) + cyl + wt, data=mtcars)
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + cyl + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## factor(am)M   0.1765     1.3045   0.135  0.89334
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## wt           -3.1251     0.9109  -3.431  0.00189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

We get comforted in our choice by looking at R^2 which almost did not change, going from 0.8327 to 0.8303, and by the fact that the standard error got reduced, going from 2.642 to 2.612. This shows that `disp` was a redundant variable here.

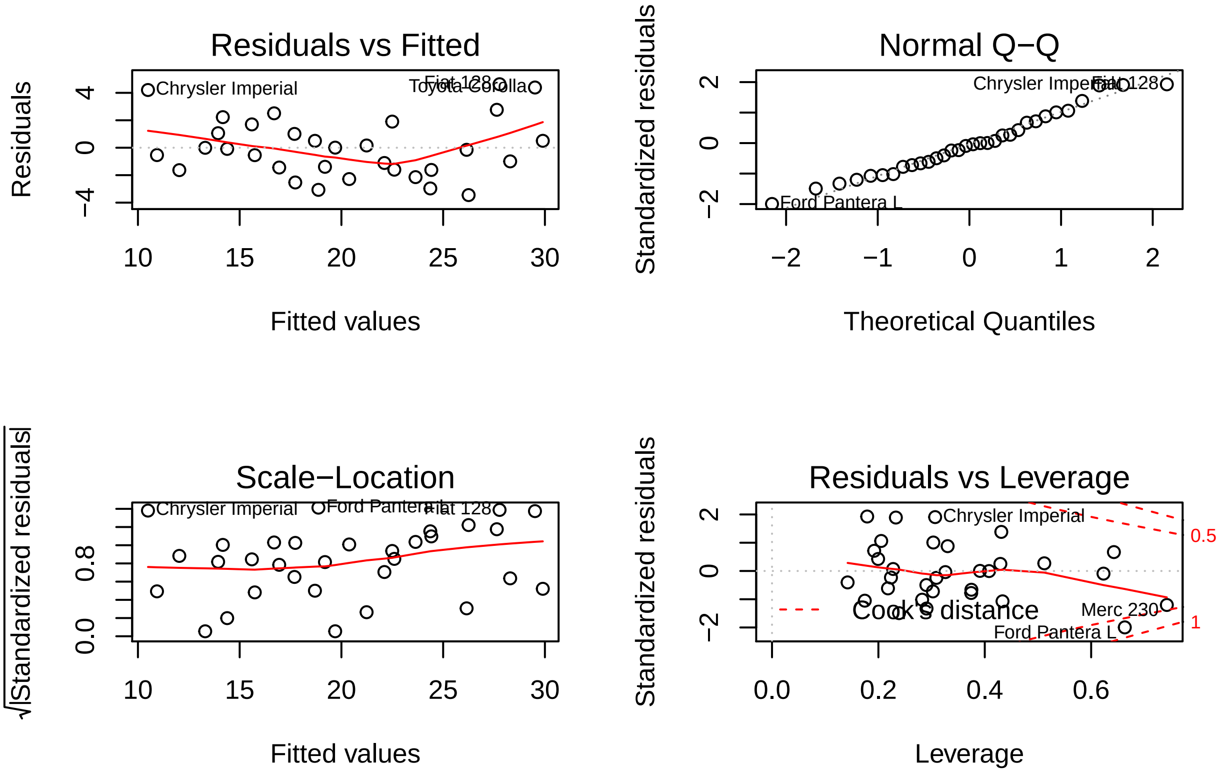
Fitting MPG against all variables

Let us try a last fit in which we ignore our last conclusion and fit MPG against all the other variables:

```
par(mfrow=c(2,2))
fit5 <- lm(mpg ~ ., data=mtcars)
summary(fit5)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## amM          2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```
plot(fit5)
```



As we can see the R^2 value increased to 0.869, but not so much compared to our last fit. The standard deviation also increased, which has a negative impact on predictions from the model. Finally, we can see from the residual plots that this dataset does not seem to contain outliers.

Conclusions

We have analysed a car dataset called `mtcars` and did a first regression against a categorical variable which is the transmission type. We have shown that a clear relation exists, meaning that transmission is definitely an important discriminant in the milage per gallon ratio of a car. We then included few other parameters and analysed how the regression improved, with some variables being more useful than others (typically those which are orthogonal to each other). Finally, we included all the variables in the linear regression and saw that such an expensive fit is not so relevant in terms of simplicity and quality of the regression.