Regression models course project

Maurizio Murino

25 February 2016

# 1. Introduction

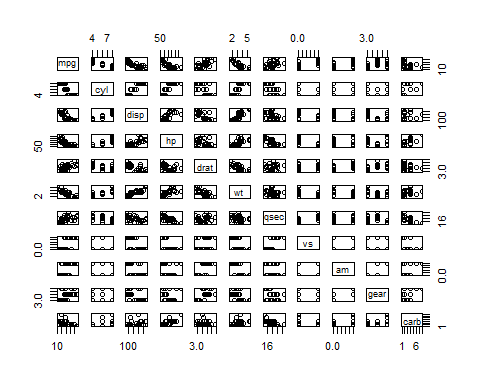
You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

“Is an automatic or manual transmission better for MPG” "Quantify the MPG difference between automatic and manual transmissions"

# 2. Exploration

mtcars counts 32 observations on 11 variables. pairs() allows us to sketch a rapid idea of the relations between the variables that we would like to explore in the second part of the analysis.

pairs(mtcars)



Operativelly, we explore the relationship between miles-per-gallon (MPG) and other variables in the mtcars data set.

dim(mtcars)

## [1] 32 11

str(mtcars)

## '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 : 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 : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

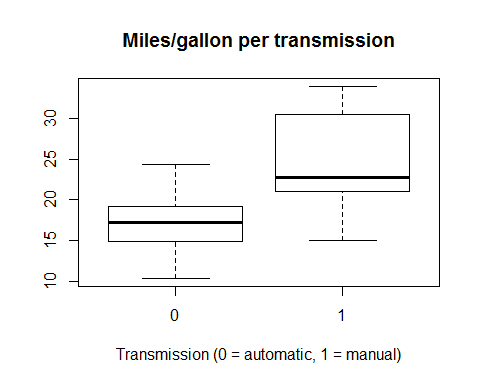
summary(mtcars)

## mpg cyl disp hp   
## Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0   
## 1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5   
## Median :19.20 Median :6.000 Median :196.3 Median :123.0   
## Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7   
## 3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0   
## Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0   
## drat wt qsec vs   
## Min. :2.760 Min. :1.513 Min. :14.50 Min. :0.0000   
## 1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000   
## Median :3.695 Median :3.325 Median :17.71 Median :0.0000   
## Mean :3.597 Mean :3.217 Mean :17.85 Mean :0.4375   
## 3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000   
## Max. :4.930 Max. :5.424 Max. :22.90 Max. :1.0000   
## am gear carb   
## Min. :0.0000 Min. :3.000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000   
## Median :0.0000 Median :4.000 Median :2.000   
## Mean :0.4062 Mean :3.688 Mean :2.812   
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :1.0000 Max. :5.000 Max. :8.000

# 3. Analysis

As told before, we focus on the relationship between mpg (Miles/(US) gallon) and am (Transmission).

data(mtcars)  
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission (0 = automatic, 1 = manual)", main = "Miles/gallon per transmission")



Manual transmission has, in appearance, a role in favorably increase the average vehicle consumption.

To have a further confirm, we have to have an idea of the other predictors of the dataset. An ANOVA model can turn in use.

anova1 <- aov(mpg ~ ., data = mtcars)  
summary(anova1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## cyl 1 817.7 817.7 116.425 5.03e-10 \*\*\*  
## disp 1 37.6 37.6 5.353 0.03091 \*   
## hp 1 9.4 9.4 1.334 0.26103   
## drat 1 16.5 16.5 2.345 0.14064   
## wt 1 77.5 77.5 11.031 0.00324 \*\*   
## qsec 1 3.9 3.9 0.562 0.46166   
## vs 1 0.1 0.1 0.018 0.89317   
## am 1 14.5 14.5 2.061 0.16586   
## gear 1 1.0 1.0 0.138 0.71365   
## carb 1 0.4 0.4 0.058 0.81218   
## Residuals 21 147.5 7.0   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Because of the low p-value (below 0.05), we will consider the variables cyl, disp, wt, drat, am as more interesting predictor variables.

lm1 <- lm(mpg ~ cyl + disp + wt + drat + am, data = mtcars)  
summary(lm1)

##   
## Call:  
## lm(formula = mpg ~ cyl + disp + wt + drat + am, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3176 -1.3829 -0.4728 1.3229 6.0596   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 41.296380 7.538394 5.478 9.56e-06 \*\*\*  
## cyl -1.793995 0.650540 -2.758 0.01051 \*   
## disp 0.007375 0.012319 0.599 0.55462   
## wt -3.587041 1.210500 -2.963 0.00643 \*\*   
## drat -0.093628 1.548780 -0.060 0.95226   
## am 0.172981 1.530043 0.113 0.91085   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.692 on 26 degrees of freedom  
## Multiple R-squared: 0.8327, Adjusted R-squared: 0.8005   
## F-statistic: 25.88 on 5 and 26 DF, p-value: 2.528e-09

drat and disp has a really high coefficient, they could be of some disturb. We try to make the approach more precise by cutting uit from the model.

lm2 <- lm(mpg ~ cyl + wt + am, data = mtcars)  
summary(lm2)

##   
## Call:  
## lm(formula = mpg ~ cyl + wt + am, 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 \*\*\*  
## cyl -1.5102 0.4223 -3.576 0.00129 \*\*   
## wt -3.1251 0.9109 -3.431 0.00189 \*\*   
## am 0.1765 1.3045 0.135 0.89334   
## ---  
## 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

The adjusted r-squared is 0.83. We cannot reject the hypothesis that the coefficient of am is 0.

# 4. Diagnosis

par(mfrow = c(2, 2))  
plot(lm2)

Appearently, there is not a relevant pattern found according to upper left graph. The normal Q-Q suggests the model mets the normality assumption. Scale-Location shows constant variance assumption are satisfied. We can conclude that weight and number of cylinders play important role to determination of mpg.