Regression models. Assingment

tonuta

December 4, 2016

Regression Models Course - *Project*

Clean the environment.

```r
rm(list = ls())
```

Install some R packages and upload libraries. install.packages("knitr") install.packages("markdown") library(knitr) library(markdown)

Synopsis.

This project of the Regression Models course is an analysis the *mtcars* data set and an investigation of the relationship between a set of variables and miles per gallon (MPG). The data is 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).

Project assignment.

The main objectives of this investigation are:

a) to compare cars with automatic and manual transmission and determine which car transmission is better for MPG.
b) to quantify how different is the MPG between automatic and manual transmissions.

Step 1: Perform the data exploration.

```r
### Upload data from R repositories.
data(mtcars)
```

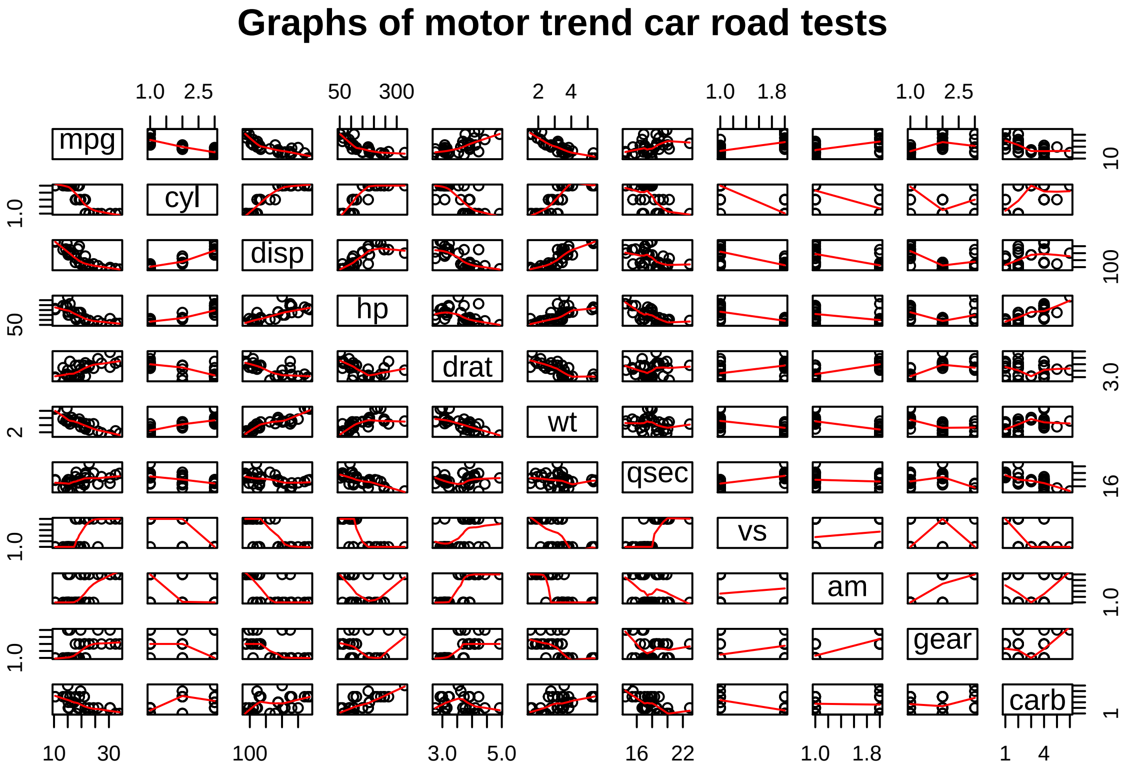
```r
### Take a look at the uploaded data.
str(mtcars) # results are hidden
```

```r
### Make the necessary variables as factors.
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am, labels = c('Automatic','Manual'))
```

```r
### Take a look at the data after factorization of the variables.
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 : 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 "Automatic","Manual": 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 ...
```

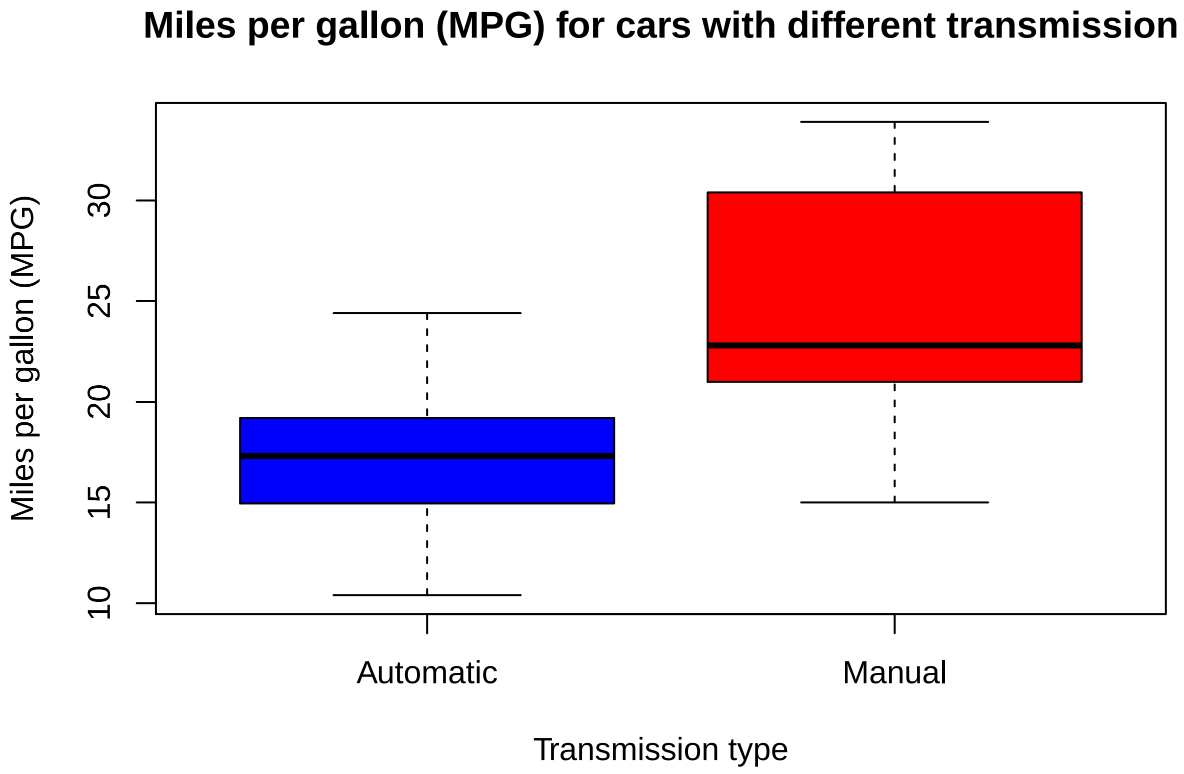
```
### Make pair plots of the $mtcars$ data for a better visual data exploration.
pairs(mpg ~ ., data = mtcars, main = "Graphs of motor trend car road tests",
      panel = panel.smooth)
```



From these pair plots, we notice that some variables (*cyl*, *disp*, *hp*, *drat*, *wt*, *vs* and *am*) display strong correlation with MPG.

The project is related in particular to the study of the MPG as a function of the type transmission. The plot related to this particular dependence is shown downwards:

```
boxplot(mpg ~ am, data = mtcars, col = (c("blue","red")), ylab = "Miles per gallon (MPG)",
        xlab = "Transmission type",
        main = "Miles per gallon (MPG) for cars with different transmission")
```



From the above plot, one can say that thre is an increase in the MPG when using a manual transmission.

Step 2: Perform the regression analysis.

Step 2.1: Detail a strategy for model building and selection.

The strategy for model building should be based first on a linear model that includes all variables as predictors of MPG as follows:

```
overall.model <- lm(mpg ~ ., data = mtcars)
summary(overall.model)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407  
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997  
## vs1          1.93085    2.87126   0.672   0.5115  
## amManual     1.21212    3.21355   0.377   0.7113  
## gear4        1.11435    3.79952   0.293   0.7733  
## gear5        2.52840    3.73636   0.677   0.5089  
## carb2       -0.97935    2.31797  -0.423   0.6787  
## carb3        2.99964    4.29355   0.699   0.4955  
## carb4        1.09142    4.44962   0.245   0.8096  
## carb6        4.47757    6.38406   0.701   0.4938  
## carb8        7.25041    8.36057   0.867   0.3995  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779 
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

One can see that adjusted $R - squared\ value$ is 0.78 (again, the maximum adjusted $R - squared\ value$ obtained considering all combinations of variables. The conclusion is than 78% of the variance of the MPG variable is explained by the model that includes the following variables: cyl, $disp$, hp, $drat$, wt, vs and am. Also, all of the coefficients are NOT significant at 0.05 significant level.

A selection of the best model with significant predictors must be chosen from this first overall linear model. The selection of the best model is done by using a stepwise model selection as follows:

```
model.for.analysis <- step(overall.model, direction = "both") # ... using both forward selection and ba
```

```
summary(model.for.analysis) # description of the best linear model.
```

```
## 
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9387 -1.2560 -0.4013  1.1253  5.0513 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *  
## cyl8        -2.16368    2.28425  -0.947  0.35225    
## hp          -0.03211    0.01369  -2.345  0.02693 *  
## wt          -2.49683    0.88559  -2.819  0.00908 ** 
## amManual     1.80921    1.39630   1.296  0.20646    
## ---
## 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:  0.8659, Adjusted R-squared:  0.8401 
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

One can see that adjusted $R - squared\ value$ is 0.84 (again, the maximum adjusted $R - squared\ value$ obtained considering all combinations of variables. The conclusion is than 84% of the variance of the MPG

variable is explained by the model that includes the following variables: *cyl*, *hp*, *wt*, and *am*. Also, all of the coefficients are significant at 0.05 significant level.

Another part of the model selection strategy is to compare the previous model with the simplest model containing only *am* as a predictor.

```
basic.model <- lm(mpg ~ am, data = mtcars)
anova(basic.model, model.for.analysis)
```

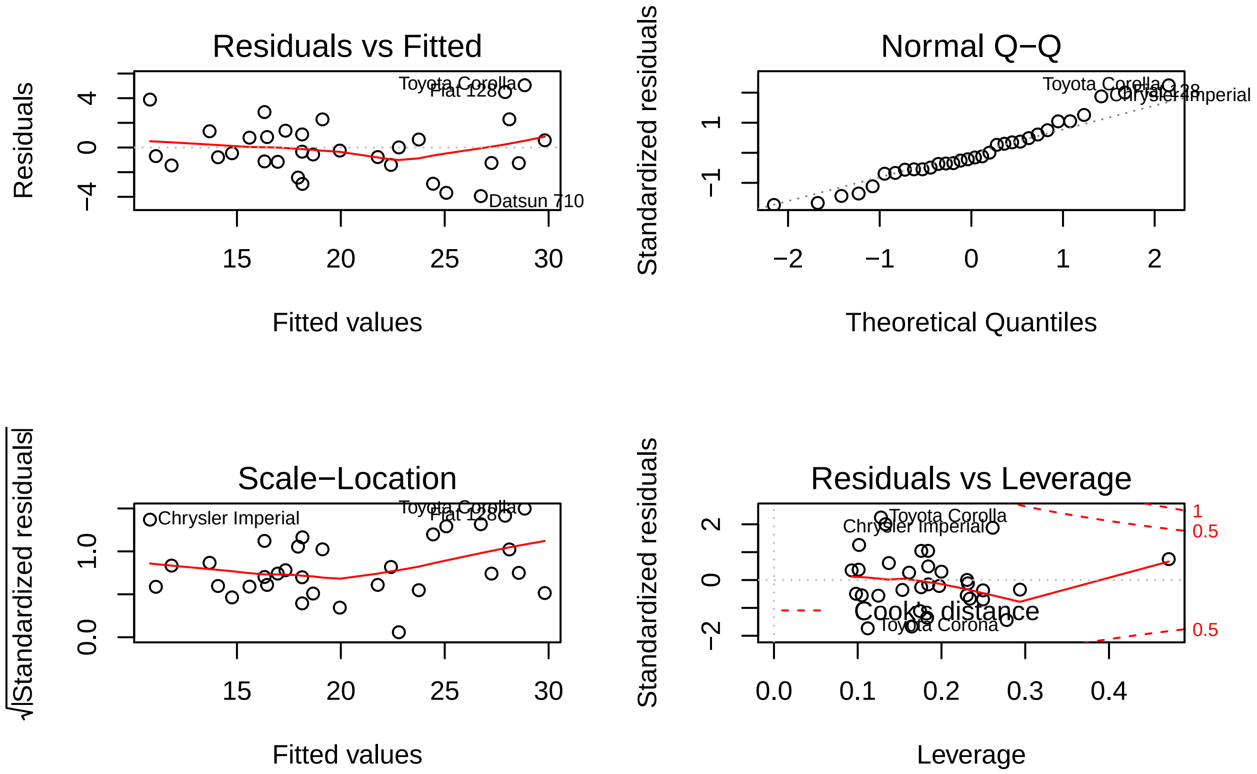
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is highly significant and we reject the null hypothesis that the confounder variables cyl, hp and wt don't contribute to the accuracy of the model.

Step 2.2: Perform residual analysis and diagnostics.

One of the reasons of investigated the residuals is to find leverage points and any related potential problems with the linear model. The residual plots of our regression model along with computation of regression diagnostics for our liner model are shown downwards:

```
par(mfrow=c(2, 2))
plot(model.for.analysis)
```



One can verify the following assumptions, according with the residual plots of our regression model.

- The Residuals vs. Fitted plot shows no consistent pattern, supporting the assumption of data independence.
- The Normal Q-Q plot shows that the residuals are normally distributed because the points lie closely to the line.
- The Scale-Location plot conforms the constant variance assumption, as the points are randomly distributed.
- The Residuals vs. Leverage plot shows that there are no outliers (all values fall within the 0.5 bands).

```
sum((abs(dfbetas(model.for.analysis))) > 1)
```

```
## [1] 0
```

A measure of how an observation affects an estimiate of the regression model is the sum of *dfbetas*. This summation of *dfbetas* is zero, thusthe performed analysis meets all the assumptions of a linear regression.

Step 3: Perform statistical inference.

Assuming that the transmission data has a normal distribution, a t-test on the two subsets of mpg data: manual and automatic transmission can be performed. The t-test can help accept/reject the null hypothesis that they come from the same distribution.

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

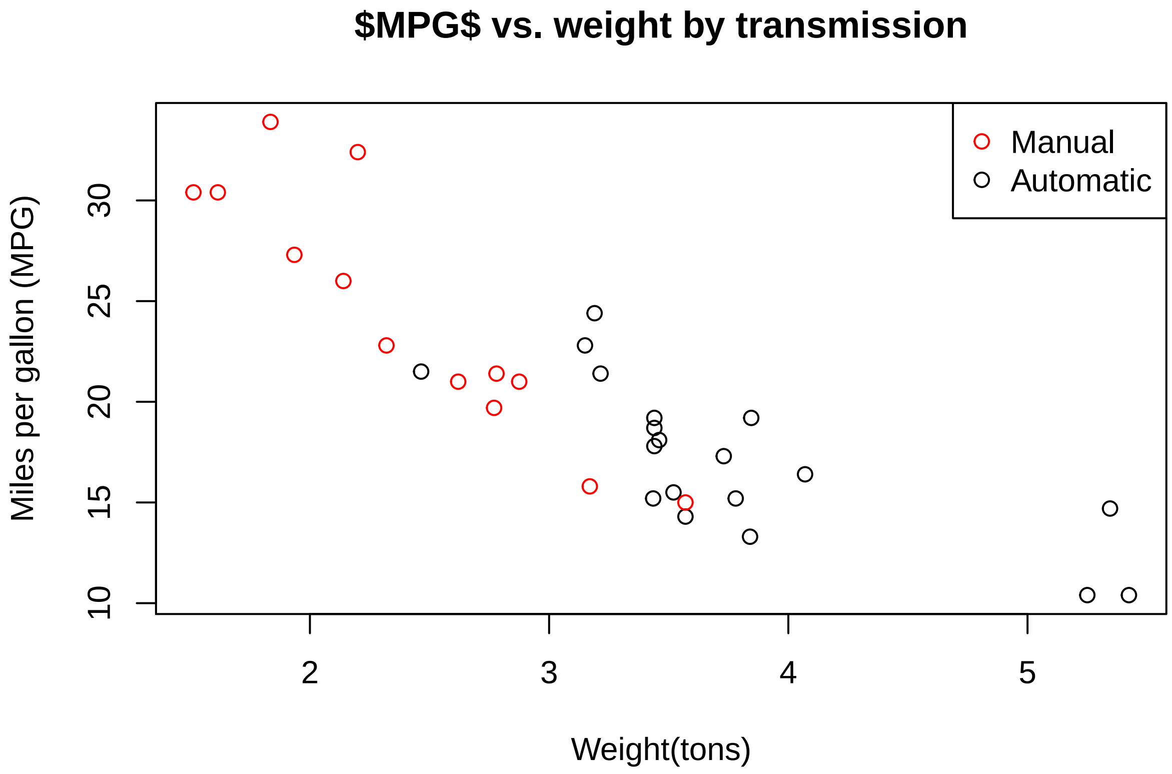
```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group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

The t-test results shows that one rejects the null hypothesis, Thus the $MPG$ distributions for manual and automatic transmissions are the same.

Step 4: Conclusions.

The performed analysis shows that cars with manual transmission get 1.8 more miles per gallon compared to cars with automatic transmission. (1.8 adjusted for hp, cyl, and wt). A plot of the $MPG$ vs. weight by transmission could help having a better insite in drawing a better coclusion about the influence of other variables when analysing the transmission type on $MPG$.

```
plot(mpg ~ wt, col = am, data = mtcars, ylab = "Miles per gallon (MPG)",
     xlab = "Weight(tons)", main = "$MPG$ vs. weight by transmission")
legend("topright", legend = c("Manual", "Automatic"), col = c("red", "black"),
       pch = c(1,1))
```


$MPG$ vs. weight by transmission
○ Manual
○ Automatic
Miles per gallon (MPG)
Weight(tons)