Motor_Trend

Pankaj Sharma

12 November 2015

This is a prject assignment of the Coursera specialization named `Regression Models`.In this project I have used `mtcars` dataset.We build regression models and exploratory data analysis to find how transmissions ie `automatic` or `manual` effects `MPG`

The main objective of this research is

- Is an automatic or manual transmission better for MPG?
- Quantifying how different is the MPG between automatic and manual transmissions?

Data Processing and Transformation

```
library(ggplot2)
data(mtcars)
str(mtcars)
```

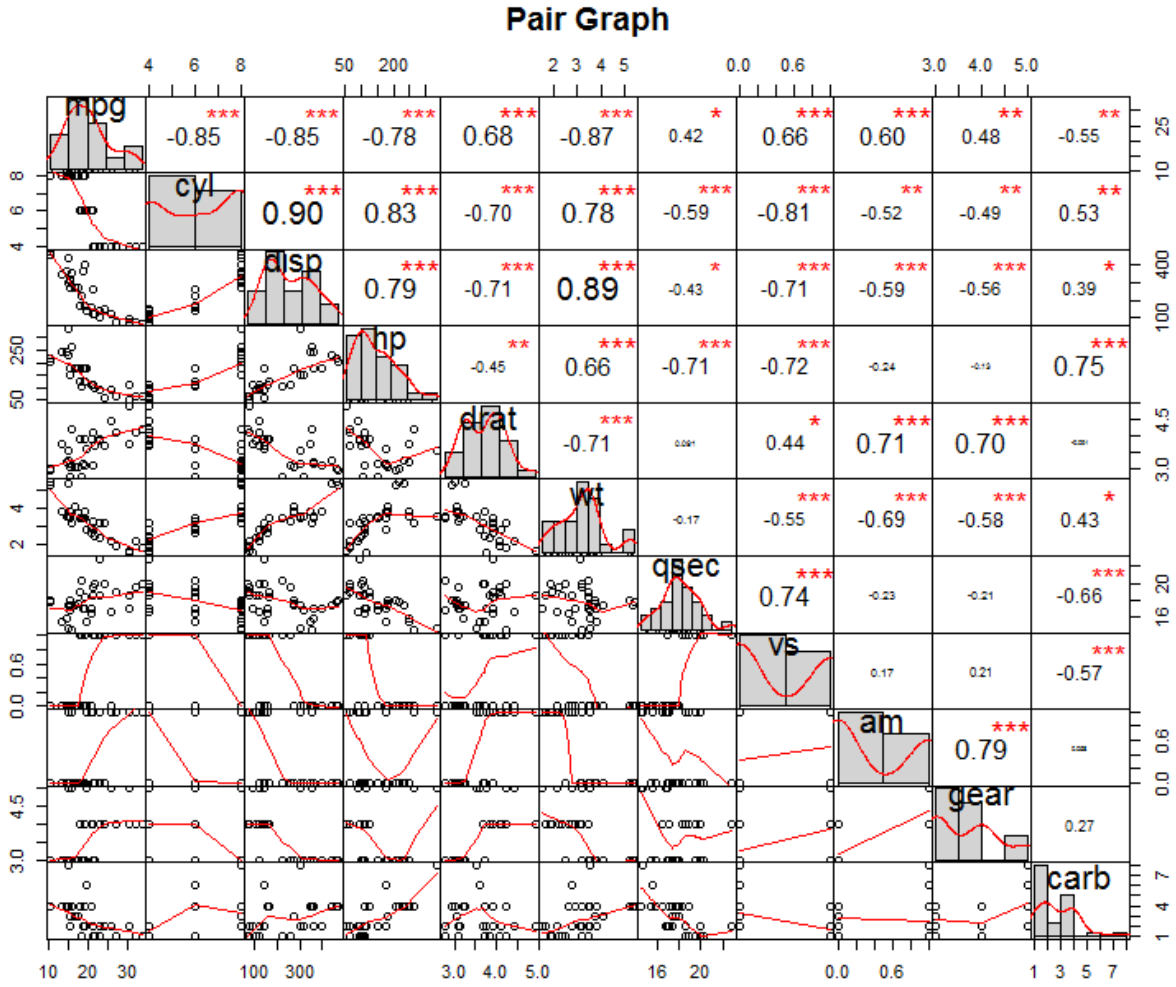
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 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 ...
```

Converting some features into factors for processing

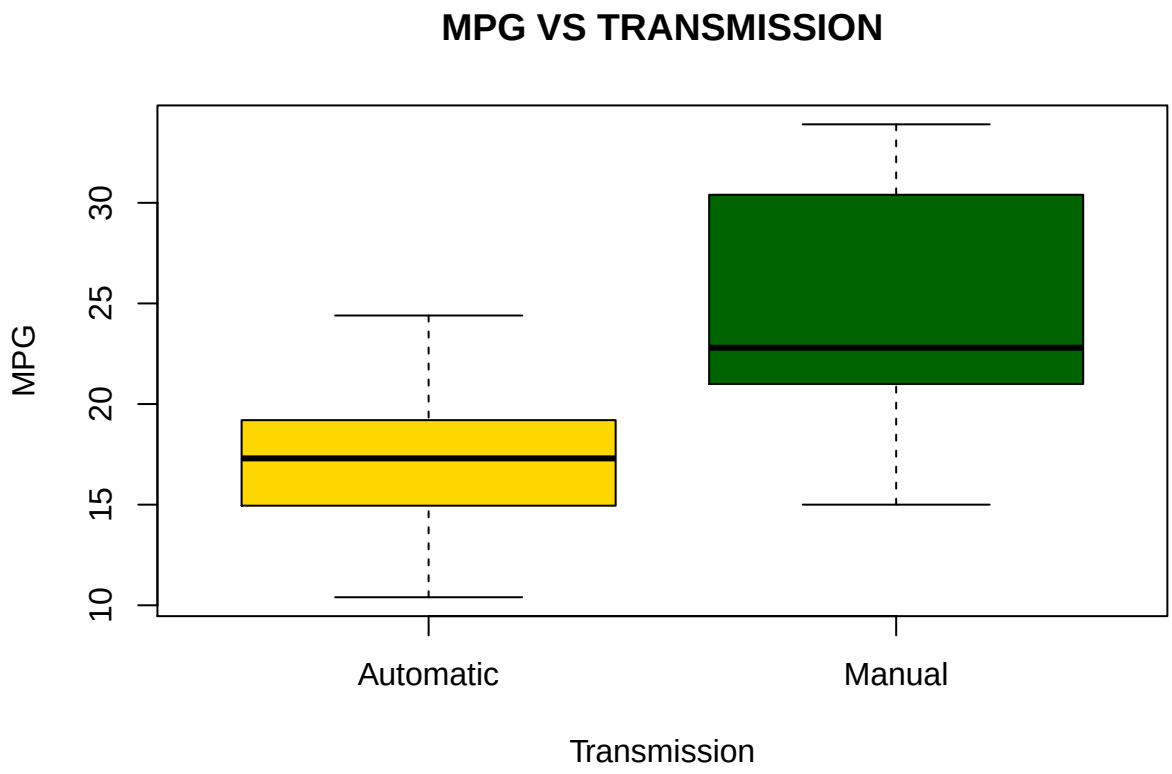
```
mtcars[,2]<-as.factor(mtcars[,2])
mtcars[,8]<-as.factor(mtcars[,8])
mtcars[,10]<-as.factor(mtcars[,10])
mtcars[,11]<-as.factor(mtcars[,11])
mtcars[,9]<-factor(mtcars[,9],labels=c('Automatic','Manual'))
```

Exploratory Data Analysis

- In this section we use exploratory analysis like boxplot and pairs graph.From the plot we notice that the variables cyl, disp,hp and wt are highly correlated with each other.
- By Box Plot between mpg and am we found out that manual transimssion yields higher values of mpg




```
boxplot(mtcars$mpg~mtcars$am, data=mtcars, notch=FALSE,
  col=(c("gold","darkgreen")),
  main="MPG VS TRANSMISSION", xlab="Transmission",ylab="MPG")
```



Hypothesis

At this step we take a `NULL` hypothesis that `MPG` of `Automatic` and `Manual` transmission are from the same population.We take a two sample T-test to show it.

```
result<- t.test(mpg ~ am,mtcars)
result$p.value
```

```
## [1] 0.001373638
```

```
result$estimate
```

```
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

- As p-value<0.01 it is significant and our hypothesis is wrong.
- Also estimates show that mean in group manual is about 7 more than that of group automatic which shows that they are from different poplulations.

Regression Analysis

First we build linear regression models with all the regressors

```r
fit<- lm(mpg ~.,mtcars)
summary(fit)
```

```
##
## 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
```

This model has a

- residual error of 2.833 on 15 degrees of freedom
- None of variables are marked significant at significant level 0.05

So we need to search a better formula to model.First we go for both forward and backward selection of variables

We form a new model

Variable Selection

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

```r
summary(new_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
```

- The model find out formula = mpg ~ wt + qsec + am as the best formula for reducing residual standard error from 2.833 to 2.459 on 28 degrees of freedom.

Now we form a base model with only am as a predictor and mpg as a dependent.

Base Model

```r
base_model <- lm(mpg ~ am, data = mtcars)
```

```r
summary(base_model)
```

```
##
## Call:
## lm(formula = mpg ~ 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 ***
## amManual       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
```

- It shows that on average a car has 17.147 mpg with automatic transmission and if manual transmission 7.245 mpg</span is increased.
- Adjusted R-squared is as low as 0.3385 which indicates that we need to add more variables to the model.

Comparing the Models

Now we do a Anova test to compare best model and the base model

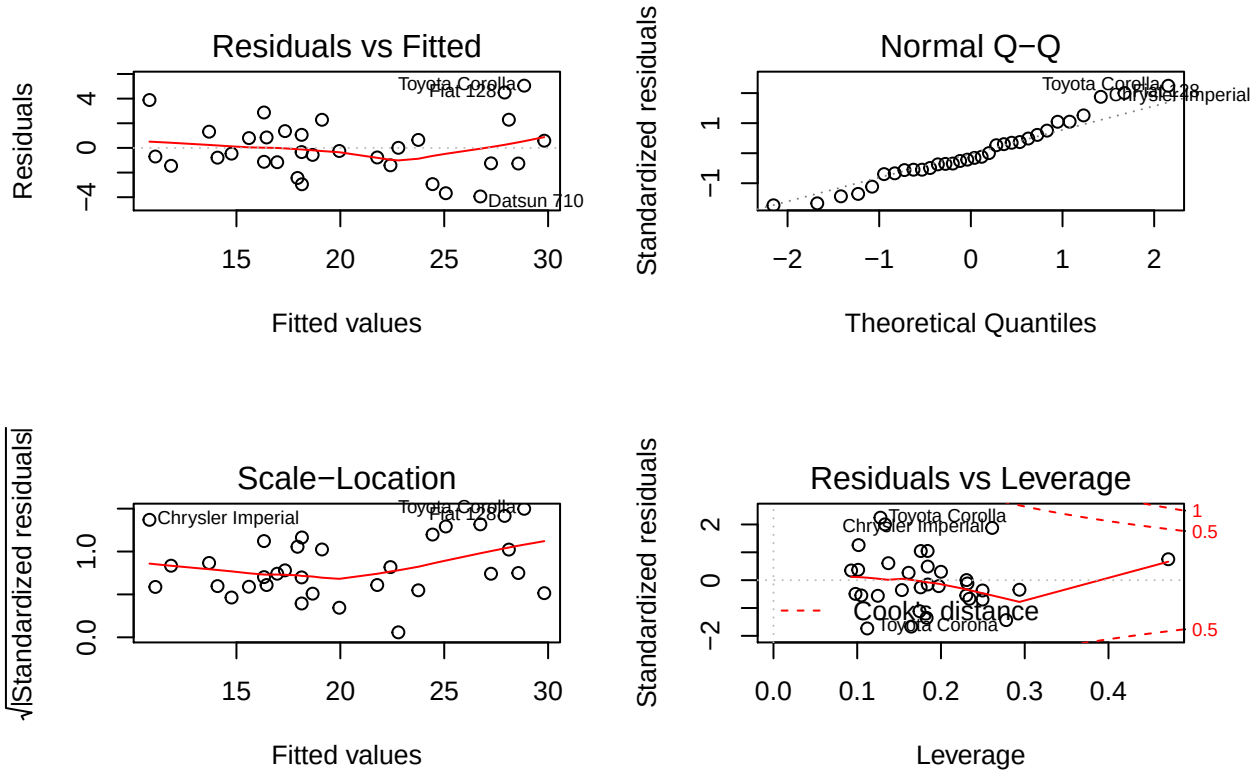
```
anova(base_model,new_model)
```

```
## 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
```

Looking at the results **p-value** obtained for the new model is highly significant.

Residuals and Diagnostics

```
par(mfrow=c(2,2))
plot(new_model)
```



Observations

From the above plots we can make the following observations.

- The points in the **Residuals vs. Fitted plot** seem to be randomly scattered on the plot and verify the independence condition.
- The **Normal Q-Q plot** consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- The **Scale-Location plot** consists of points scattered in a constant band pattern, indicating constant variance.
- The **Residuals vs Leverage** some distinct points of interest (outliers or leverage points) in the top right of the plots.

Conclusions

Based on observations on our best fit model we can conclude that

- Cars with Manual transmission get more miles per gallon mpg compared to cars with Automatic transmission. **(1.8 adjusted by hp, cyl, and wt).**
- mpg will decrease by **2.5**(adjusted by hp, cyl, and am) for every **1000 lb** increase in wt. , mpg decreases negligibly with increase of hp.
- If number of cylinders, cyl increases from **4 to 6 and 8**, mpg will decrease by a factor of **3 and 2.2** respectively (adjusted by hp, wt, and am).