Regression Models (Motor Trend Project)

Nutan Sahoo

11 September 2017

Executive Summary

The purpose of this project is to explore the impact of a set of variables like horsepower, transmission configuration, engine cylinder configuration, etc. on `mpg`(Miles per Gallon). In particular we have two objectives:—

1. To find which one of Automatic or Manual Transmission is better for `mpg`.
2. To Quantify the `mpg` difference b/w auto and manual transmissions.

Firstly, I plotted boxplots of `mpg` against various categorical variables to visualise how `mpg` changes with changing levels and also to examine the spread and mean of the different levels of the factor. We visualise a correlation matrix using `chart.Correlation` function in `PerformanceAnalytics` Package. The variables which had high cor. with the mpg were selected as regressor variables for our regression model. Out of all possible regression models which can be made from the combination of those regressors, we select the best model using `stepAIC` function from `MASS` package on the basis on Akaike's Information Criteria, Other criteria such as adjusted R^2, Mean Square of Residuals, Generalised Variance Inflation factors were also taken into consideration.

Exploratory Analysis

We plot various box plots to visualise the distribution of `mpg` by various groups of the categorical variable.[1]

```
head(mtcars)
dim(mtcars)
```

In plot-1 **(see appendix for all plots)** we see that the mean `mpg` of cars with automatic transmission is much lower than manual. We can check if the difference b/w their mean values are statistically significant through a two-samples independent t-test.

```
aggregate(mpg~am, data = mtcars, mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

```
auto<- mtcars[mtcars$am=="0", 1]
manual<- mtcars[mtcars$am=="1",1]
t.test(auto, manual)
## t = -3.7671, df = 18.332, p-value = 0.001374
```

p value <0.05, we reject the null hypothesis that the true mean difference is equal to zero. Hence, the difference is significant and the cars with automatic transmission have a lower `mpg` on an average. Plot-2 shows that generally, `mpg` of cars with `S` engine configuration is much higher. We can confirm this using a t-test as shown above.

[1] **All plots given in appendix**

```
t.test(mpg~vs, data= mtcars)
##t = -4.6671, df = 22.716, p-value = 0.0001098
```

By looking at plot-3 it is safe to assume, higher th number of cylinders in the car, lower is the `mpg`. Such definite conclusions cannot be drawn by looking at plot-4 & plot-5 but we can perform t-tests for it.

Regression Analysis

We have graphically seen that Manual is better for `mpg`. Now we will try to qantify the difference between them.

```
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$am   <- factor(mtcars$am)
model1<- lm(mpg~am, data=mtcars)
summary(model1)
## Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385
## F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285
```

The R^2 value is less, the variable am only explains 36 percent of the variability in mpg; so we include more variables. We will now plot (Plot-6) a correlation chart with the signifcant values.

Selecting the Best Model

There is a high correlation of `mpg` with `disp, hp, drat, wt`. We exclude `disp` as disp has high correlation with all the other regressors, this creates the problem of multicollinearity and affects the accuracy of the model. So we remove `disp` from the model. We should also include categorical var. in the model as they can also very well explain the variability in `mpg`. So, we choose `am, cyl, vs` to be included in the model. We want a parsimonious model hence we only include a few important qualitative regressors.

```
model2<- lm(mpg~hp+drat+wt+am+vs+cyl, data= mtcars)
#install.packages(MASS)
library(MASS)
stepAIC(model2, direction="both")
final_model<- lm(mpg~ hp+wt+am+cyl, data= mtcars)
summary(final_model)
## F-statistic: 33.57 on 5 and 26 DF, p-value: 1.506e-10
```

stepAIC fn. will form models with all possible combinations of regressors of model 2. Then model is selected on the basis on AIC (Akaike's Information Criteria); which in this case is the one with the following regressors: `hp, wt, am, cyl`. We take this to be the best model as it has the lowest AIC (61.635). We see that the overall model is significant with a p-value of 1.5e-10 and Adjusted R^2 value of 0.8401 i.e. our model accounts for 84% of the variabilty in outcome variable. We use `vif` func. from the `car` package to check for possible multicollinearity.

```
library(car)
vif(final_model)
##          hp    wt   am    cyl
## GVIF     4.66  4.0  2.56  2.4
```

It gives generalised inflation factors (GVIF) and GVIF^(1/2(df)), we can simply apply the same standard rules of thumb as for GVIF for detecting possible multicollinearity. If GIV is greater than 4, it indicates multicollinearity. We see that it is smaller than 4 or a little greater than 4. Hence, we conclude that there is no serious problem of multicollinearity in `final_model`.

Checking basic assumptions of regression

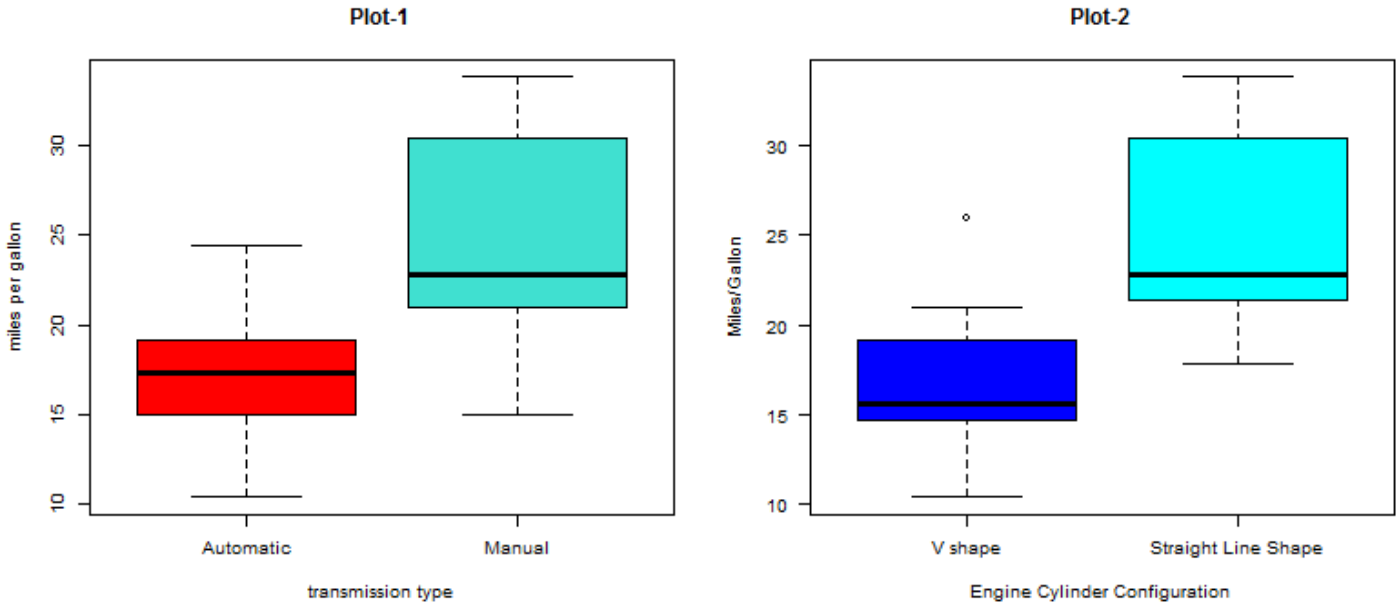
```
res<- final_model$residuals
shapiro.test(res)
## p-value = 0.4479
```

p>0.05 hence we fail to reject the null hypothesis that residuals are normally distributed. By looking at Plot-7, we can say that the residuals seem to be homoscedastic. The scatter points in QQ Plot also seem to lie on the line, thereby confirming their normality. Basic assumptions of regression are met. We can finally, say that **compared to cars which had automatic transmission(0) we would expect mileage of cars with manual transmission(1) 1.8 miles per gallon more on average given values of other regressor variables remain same**

Appendix

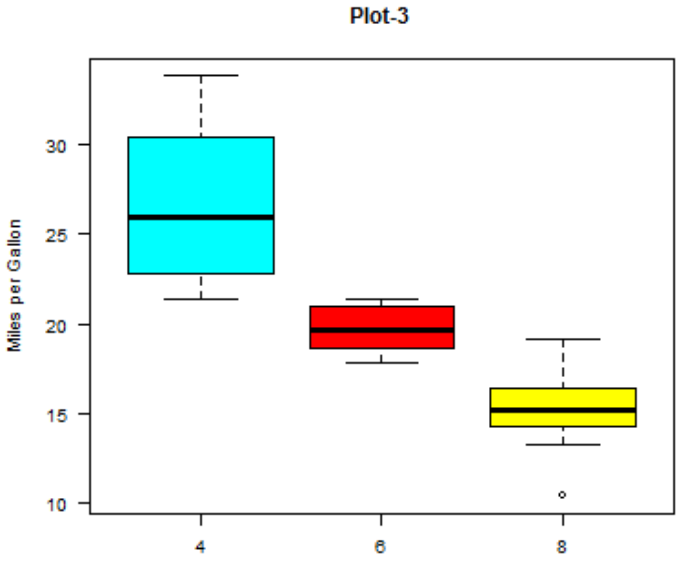
This section includes all the above mentioned Plots.

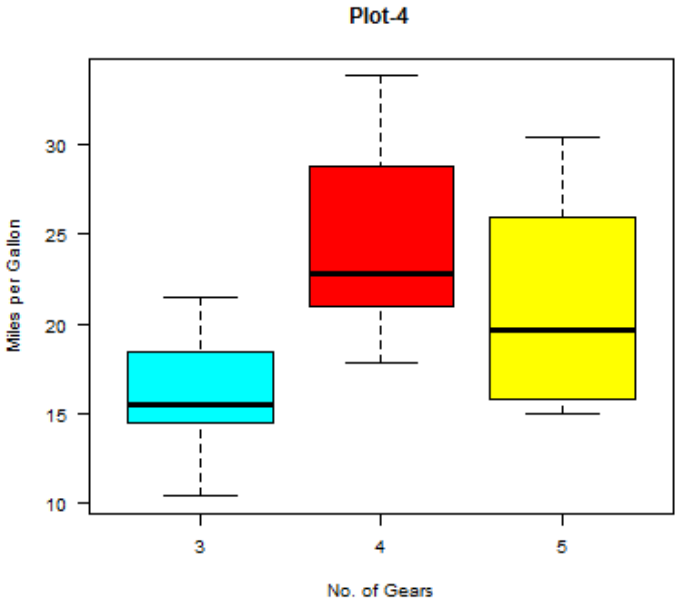
```
layout(matrix(c(1,2,1,2),2,2, byrow = TRUE))
boxplot(mpg~am,data=mtcars,col=c("red", "turquoise"),xlab="transmission type",ylab="miles per gallon",
        names= c("Automatic","Manual"),main="Plot-1")
boxplot(mpg~vs, data= mtcars, col=c(4,"cyan"), xlab="Engine Cylinder Configuration",
        ylab="Miles/Gallon", las=TRUE, names= c("V shape", "Straight Line Shape"), main="Plot-2")
```

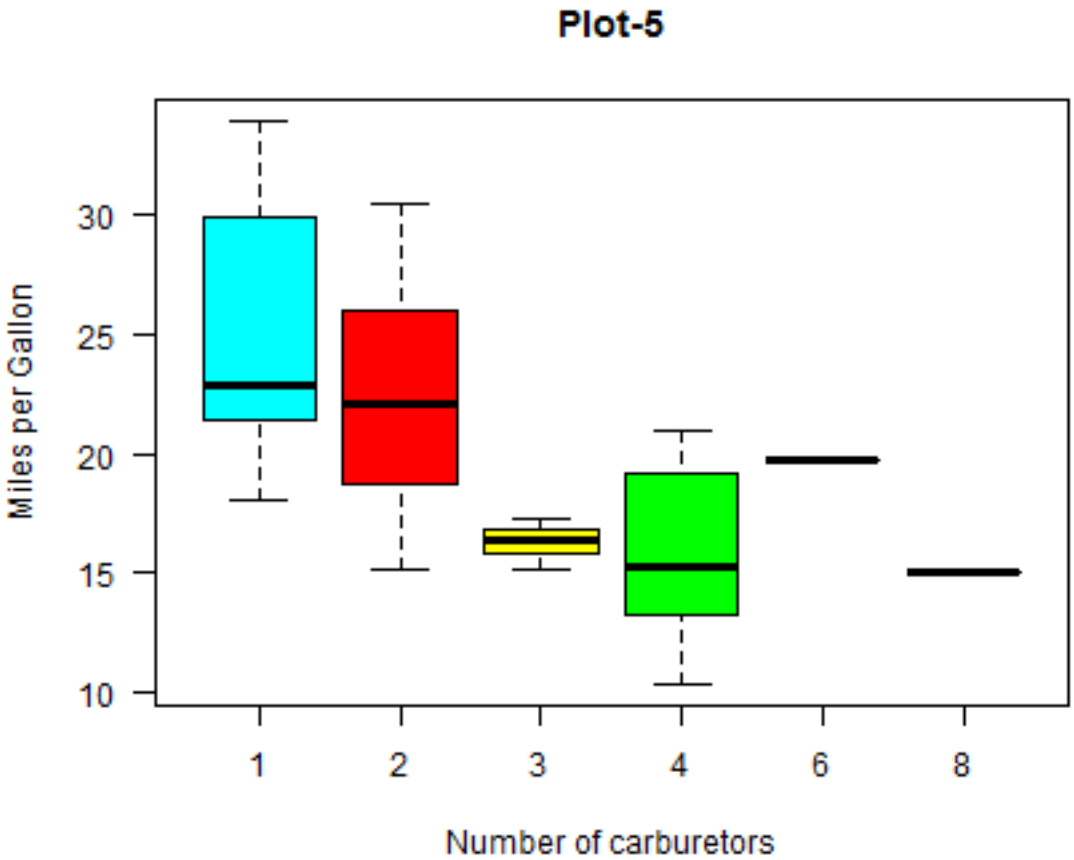



```
layout(matrix(c(1,2,1,2),2,2, byrow = T))
boxplot(mpg~cyl, data= mtcars, col=c("cyan",42,23),
        ylab="Miles per Gallon", las=TRUE, main="Plot-3")
boxplot(mpg~gear, data= mtcars, col=c("cyan",42,23), xlab="No. of Gears",
        ylab="Miles per Gallon", las=T, main= "Plot-4")


boxplot(mpg~carb, data= mtcars, col=c("cyan",42,23,"green",600), xlab="Number of carburetors",
        ylab="Miles per Gallon", las=T, main="Plot-5")
```


Plot-3
Miles per Gallon
30
25
20
15
10
4
6
8

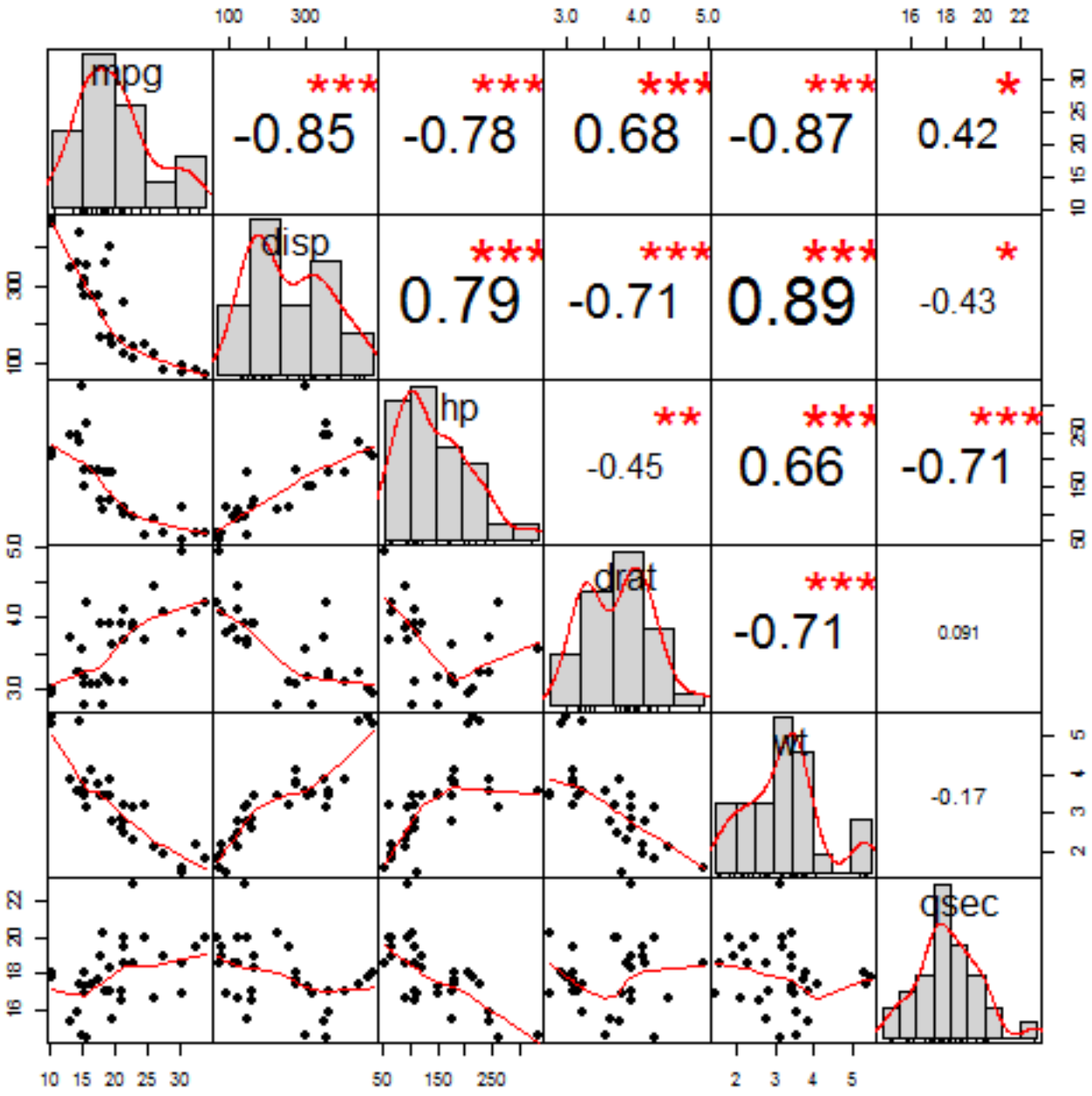

Plot-4
Miles per Gallon
30
25
20
15
10
3
4
5
No. of Gears


Plot-5
Miles per Gallon
30
25
20
15
10
1
2
3
4
6
8
Number of carburetors

Plot-6 Correlation Chart

```
#install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)
mydata<- mtcars[ , c(1,3,4,5,6,7)] #exclude the categorical variables from this
chart.Correlation(mydata, histogram = T, pch= 19)
```

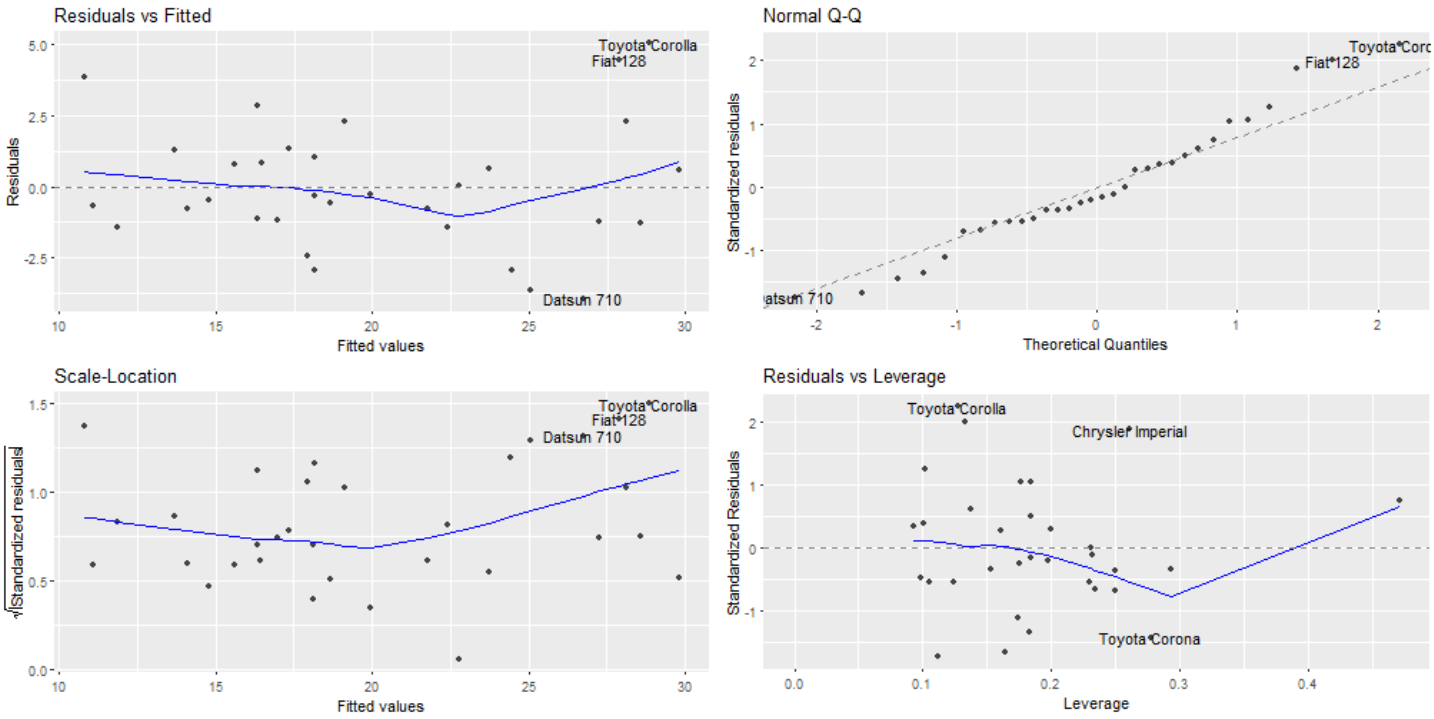
Chart-1.png



Plot-7 Diagnostic Plots

```
#install.packages("ggfortify")
library(ggfortify)
autoplot(final_model, label.size=4)
```

Plots-1.png


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Scale-Location
√|Standardized residuals|
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Residuals vs Leverage
Standardized Residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona