Regression Models Course Project

Sriraman Krishnamurthy

February 11, 2017

EXECUTIVE SUMMARY

This paper explores the relationship between miles-per-gallon (MPG) and other variables in the mtcars data set. In particular, the analysis attempts to determine whether an automatic or manual transmission is better for MPG, and quantifies the MPG difference.. Particularly, following question will be answered by the data set:

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

Analysis Conclusion : Briefly, cars with a manual transmission have a slightly better than automatic for MPG, but this different is statistical insignificant. According to our best predictive model the impact of having a manual distribution system only enhance by 1.80921 Miles per Gallon the efficiency of a car in comparison to automatic distribution system

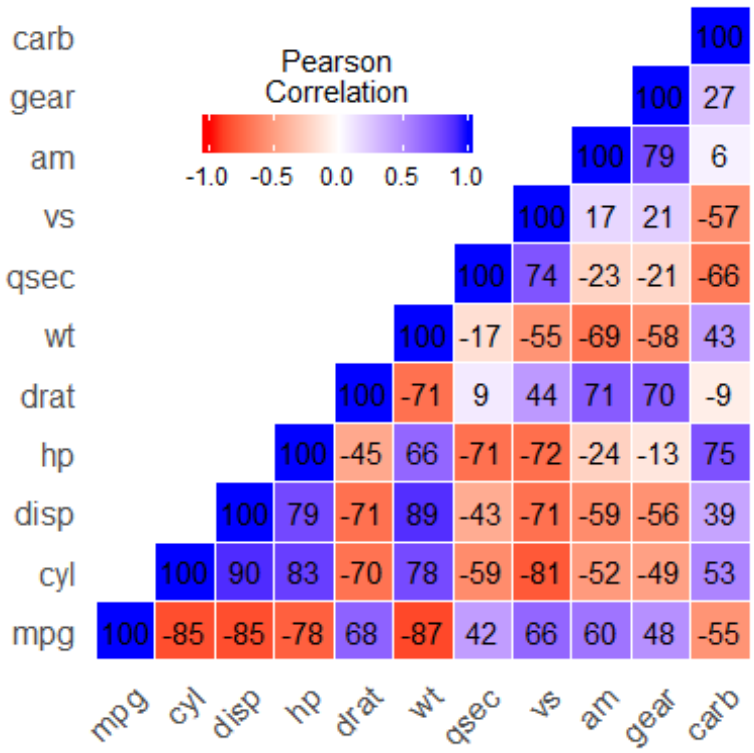
Visualizations

We are interested by exploring the relationship between variables. For this purpose, we will construct a correlation matrix

```
#Take a look at what the datasets consists of
#Regression model course Project
library(knitr)

## Warning: package 'knitr' was built under R version 3.3.2

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.2

library(datasets)
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

# Keeping the original data
mtcars.orig <- mtcars

#converting Variables to factor
```

```r
mtcars$cyl <- factor(mtcars$cyl)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)


# exploring the relationship between variables

cor_mtcars <- round(cor(mtcars.orig), 2)

# Get lower triangle of the correlation matrix
get_lower_tri<-function(cormat){
  cormat[upper.tri(cormat)] <- NA
  return(cormat)
}

# Get upper triangle of the correlation matrix
get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)
}

upper_tri <- get_upper_tri(cor_mtcars)

# Melt the correlation matrix
library(reshape2)

## Warning: package 'reshape2' was built under R version 3.3.2

melted_cormat <- melt(upper_tri) # melting the upper triangle
melted_cormat <- na.omit(melted_cormat)

# Creating ggheatmap
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "red", high = "blue", mid = "white",
                       midpoint = 0, limit = c(-1,1),
name="Pearson\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1),
        axis.text.y = element_text(size = 12))+
  coord_fixed()

ggheatmap +
  geom_text(aes(Var2, Var1, label = value*100), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
```

```
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))
#printing the created heatmap
```



We are especially interested in exploring the relationship between miles per gallon (MPG) and the other parameters. For this purpose we compute the absolute correlation to highlight parameters that best explain the MPG parameter.

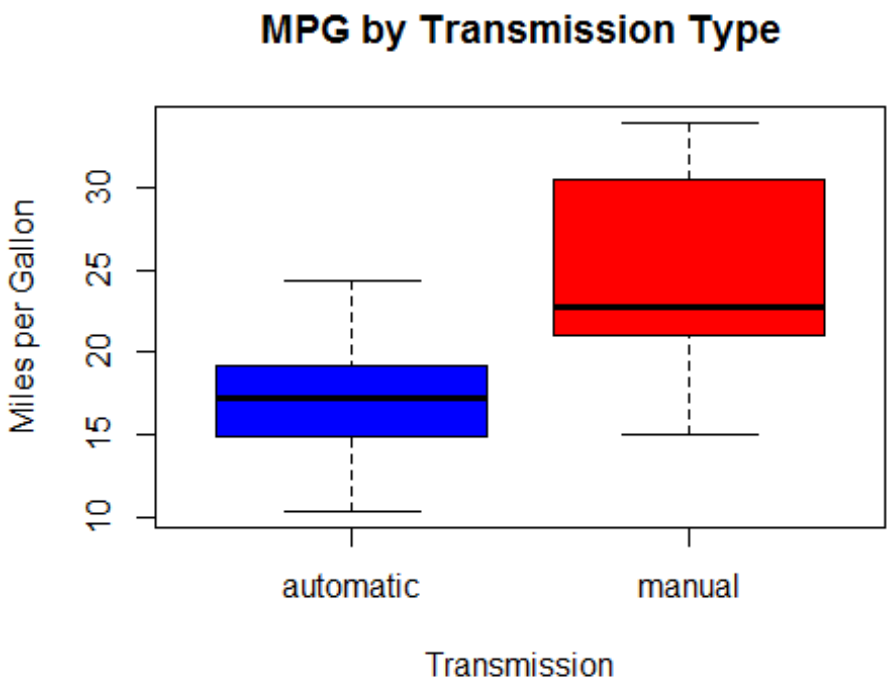
```
# relationship between by exploring miles per gallon (MPG)
abs_cor_MPG <- abs(cor_mtcars)[1,2:11]
abs_cor_MPG <- abs_cor_MPG[order(abs_cor_MPG,decreasing = TRUE)]
abs_cor_MPG

##   wt  cyl disp   hp drat   vs   am carb gear qsec
## 0.87 0.85 0.85 0.78 0.68 0.66 0.60 0.55 0.48 0.42
```

wt, cyl, disp parameters best explain the MPG parameter. Those variables are quite negatively correlated with MPG meaning that the miles driven for a given quantity of fuel, tend to decrease when the weight, the number of cylinders, the displacement size of an engine increase.

Question 1: Is an automatic or manual transmission better for MPG?

Let us Visualize the MPG distribution with respect to type of transmission

```r
par(mfrow = c(1, 1))
boxplot(mpg~am, data = mtcars,
col = c("blue", "red"),
xlab = "Transmission",
ylab = "Miles per Gallon",
main = "MPG by Transmission Type",
names= c("automatic","manual"))
```



We clearly see a difference between automatic and manual distribution systems. Manual distribution systems are mostly less fuel consuming. But we want assert that statistically. For this purpose we implement a t?test in the second part of our analysis

Question 2 : Quantify the MPG difference between automatic and manual transmissions.

Lets do a two samples t-test on the MPG parameter distinguishing auto vs manual systems *Null hypothesis*: There is no difference between MPG means for automatic and manual transmissions

```r
# Two samples t-test on the MPG parameter distinguishing auto vs manual
systems.
auto <-mtcars.orig[mtcars.orig$am==0,c("mpg")]
manual <- mtcars.orig[mtcars.orig$am==1,c("mpg")]
t.test(mtcars$mpg ~ mtcars$am)

##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$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 0 mean in group 1
##        17.14737        24.39231
```

We get a p value of 0.00137, so we reject the null hypothesis. We are 99.86% confident that the mean of both transmissions are significantly different. And the average miles per gallon for Manual Transmission is 24.39 which is 7.24 higher than the average miles of Automatic Transmission. The boxplot with respect to the Transmission show significant overlap, indicating that it may not be the best predictor of the MPG. To determine the best predictor, further analysis needs to be done.

First of all, lets take a look to what extent a model only based on the am parameter can explain in regard to MPG consumption.

```
# model only based on the am parameter
mpg_am_model <- lm(mpg~am, mtcars)
sum_mpg_am_model <- summary(mpg_am_model)

# R-squared (mpg~am)
round(100*sum_mpg_am_model$r.squared,2)

## [1] 35.98

# R-squared adjusted (mpg~am)
round(100*sum_mpg_am_model$adj.r.squared,2)

## [1] 33.85
```

Best Linear Model

To determine which variables to include in our model and to avoid multi?colinearity issue, we used an R stepwise regression function. This function adds and removes independent variables to the model until it finds the combination of independent variables minimizing the Akaike Information Criterion

```
##
## 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 **
```

```
## am           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 summary of the best model is seen above

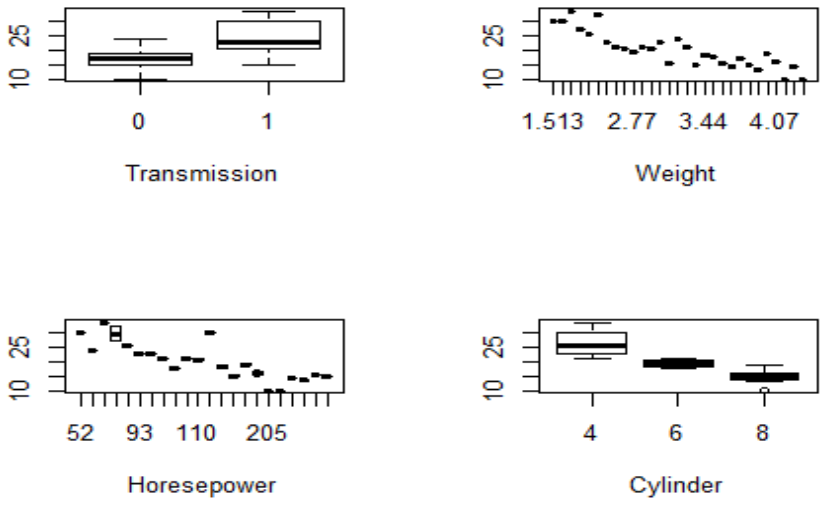
After computation our best model get an R?squared adjusted of 84.01%. wt, hp and cyl are the variables that best explain miles per gallon consumption if we look at the asterix marks.

According to our best predictive model the impact of having a manual distribution system only enhance by 1.80921 Miles per Gallon the efficiency of a car in comparison to automatic distribution system

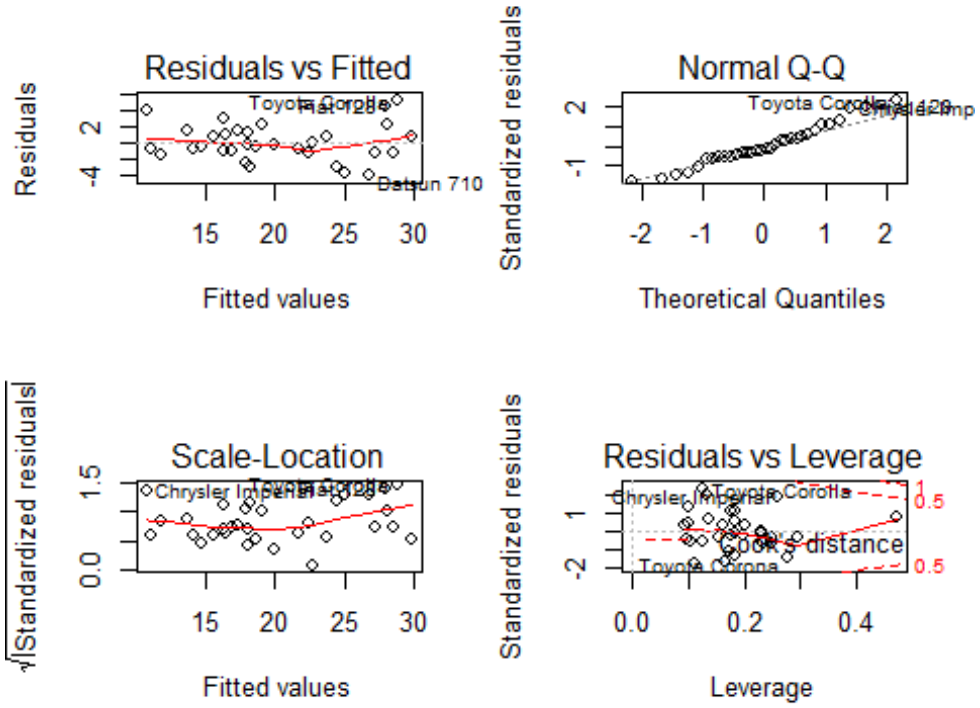
Residual analysis

Lets look at residual Plots to see if we can improve the model

```
# Plotting the box-plot into 2 * 2 matrix
par(mfrow=c(2, 2))
# Impact of Transmission on the Fuel consumption
boxplot(mtcars$mpg ~ mtcars$am, xlab="Transmission")
# Impact of weight on the Fuel consumption
boxplot(mtcars$mpg ~ mtcars$wt, xlab="Weight")
# Impact of Horsepower on the Fuel consumption
boxplot(mtcars$mpg ~ mtcars$hp, xlab = "Horesepower")
# Impact of Cylinder on the Fuel consumption
boxplot(mtcars$mpg ~ mtcars$cyl, xlab="Cylinder")
```



```r
# Residual analysis
par(mfrow=c(2,2))
plot(best_model)
```



In the residual plot, we don't see any pattern that causes us to believe that the Fuel consumption could be explained more by any other predictors available in the dataset.