Regression Models Course Project

Tong Tat

December 24, 2017

Rpub Link: Click Here

1. Executive Summary

This report is for Coursera Regression Models Course's Final Project. The dataset of interest is the **mtcars** dataset. The objective is to explore the relationship between the set of variables and miles per gallon (MPG) (outcome). Two main questions will be addressed.

– "Is an automatic or manual transmission better for MPG"

– "Quantify the MPG difference between automatic and manual transmissions"

2. Data Exploration

First, let's load the mtcars data. Then we'll take a look at the summary and look for any NA. Missing dependencies check will be done to look for any missing package that require installation.

```
# Clear cache
rm(list=ls())

# Load dataset
mcar <- data.frame(mtcars)
data(mcar)

# Summary of mtcats data
summary(mcar)
```

```
##       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
```

```
str(mcar)
```

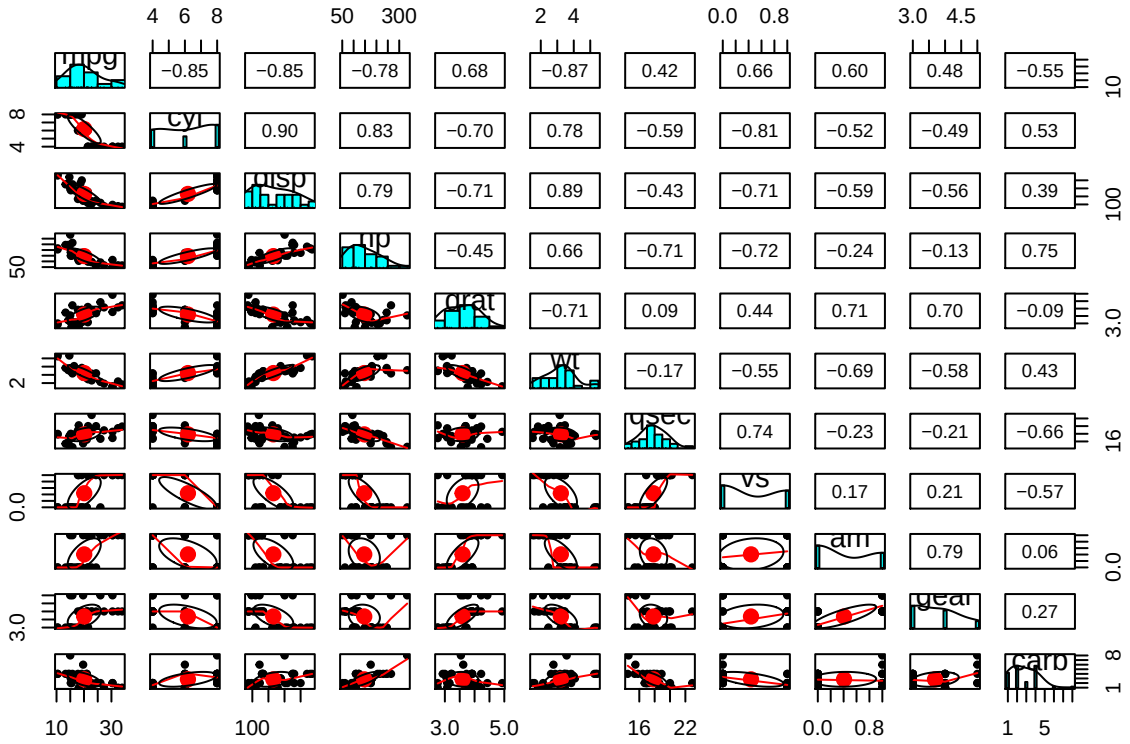
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76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

```
# Check NA
sum(is.na(mcar))
```

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

```
# Check for missing dependencies and load necessary R packages
if(!require(stats)){install.packages('stats')}; library(stats)
if(!require(psych)){install.packages('psych')}; library(psych)
if(!require(MASS)){install.packages('MASS')}; library(MASS)
if(!require(ggplot2)){install.packages('ggplot2')}; library(ggplot2)

# Check initial overview of correlation
pairs.panels(mcar)
```



Based on the Pearson correlation seen from the upper right corner of the pair.panels plot, we can see the variables which correlates with mpg are **cyl**, **disp**, **hp** and **wt**.

3. Data Cleaning

Note the variable **am** is numeric. Data transformation will be done to clean up this variable.

```
# Data Cleaning for AM variable.
mcar$am <- as.factor(mcar$am)
mcar$am <- gsub("0", "Automatic", mcar$am)
mcar$am <- gsub("1", "Manual", mcar$am)
```

4. T-Test for Automatic vs Manual Transmission

Using Welch Two Sample T-test, we investigate whether there is any significant between Automatic and Manual transmission. Since the p-value (<0.05), we reject the null hypothesis and conclude there is significant diference between Automatic and Manual transmission for MPG.

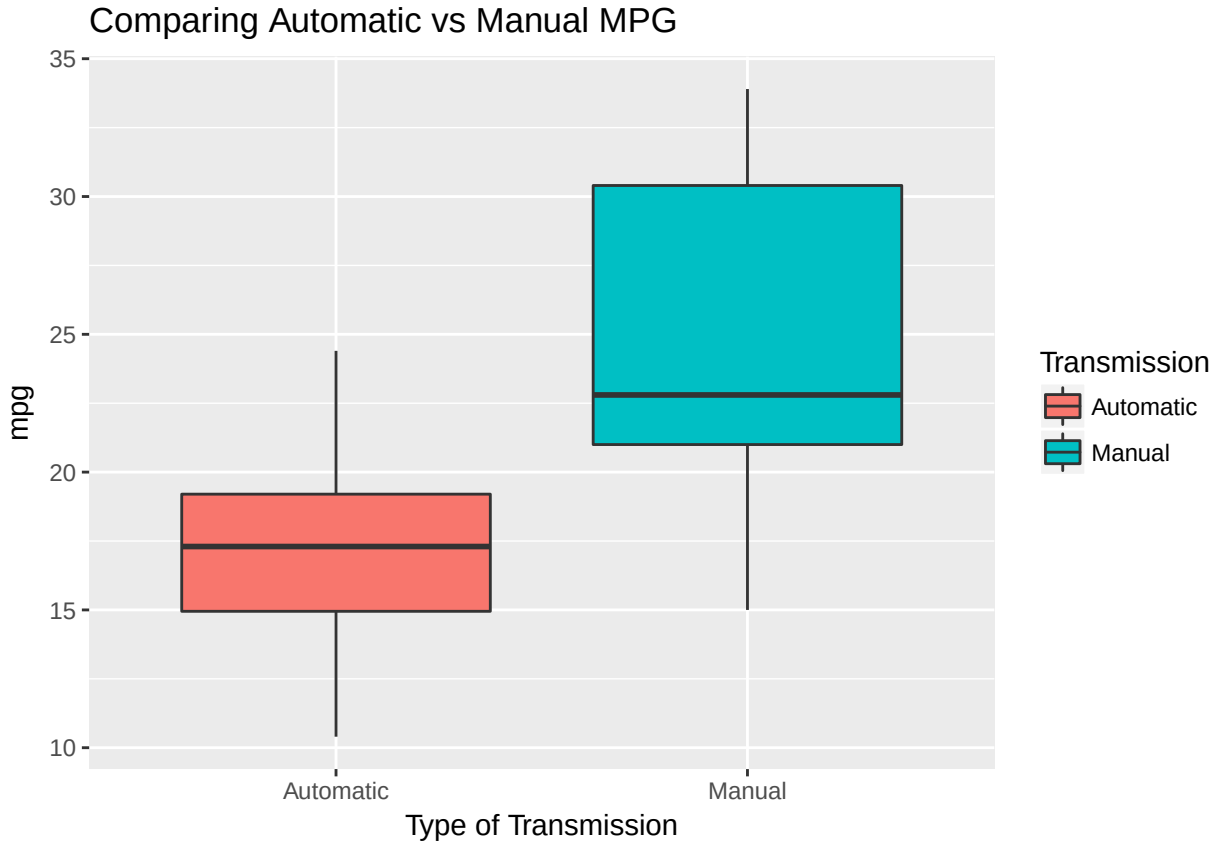
```
# Subset Automtic
auto <- subset(mcar, am=="Automatic")

# Subset Manual
man <- subset(mcar, am=="Manual", select=c(mpg, am))
```

```r
# Welch Two Sample T-test
t.test(auto$mpg, man$mpg, paired=FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  auto$mpg and man$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

```r
# Boxplot for Auto vs Manual mpg
gg1 <- ggplot(mcar, aes(x=am, y=mpg, fill=am)) +
  labs(title="Comparing Automatic vs Manual MPG", x="Type of Transmission", y="mpg") +
  scale_fill_discrete(name = "Transmission") +
  geom_boxplot()
gg1
```



As we can seen above, the mean mpg for Automatic transmission is 17.1473684 and the mean mpg for Manual transmission is 24.3923077. Manual transmission give a better mpg than Automatic transmission.

4. Model Selection

Using stepAIC from the MASS library, we will select the predictor variables by performing stepwise selection in both direction. The concept of stepAIC is to perform stepwise model selection by exact AIC (Akaike Information Criterion).

```
# Contruct lm model for stepAIC to consume later
lm.all <- lm(mpg ~., data=mtcars)

# Run stepAIC for best model selection
best.model <- stepAIC(lm.all, direction="both", trace=FALSE)
summary(best.model)
```

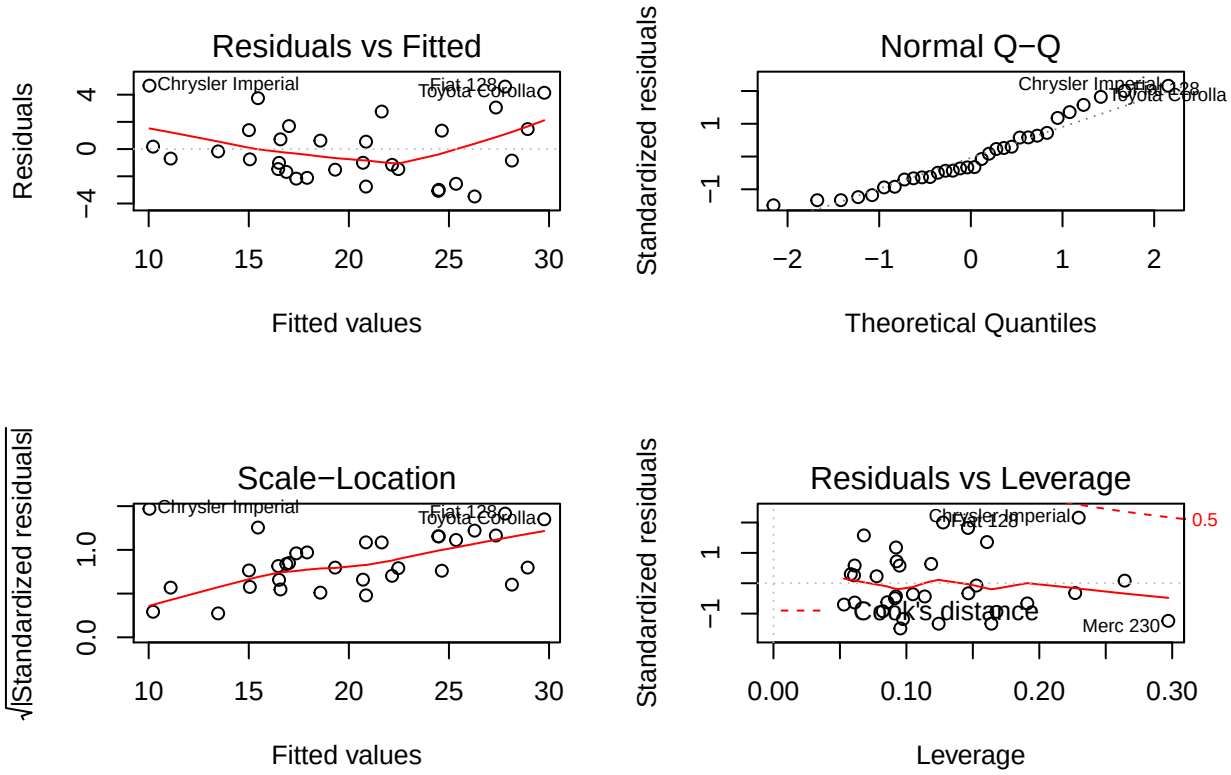
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Based on the Akaike Information Criterion, our best model is using **wt**, **qsec** and **am**. The adjusted R-squared value is 0.8336 which means that the model explains 83% of the variation in mpg indicating it is a robust and highly predictive model.

5. Residual Plot

Based on the residual plots below, we can see there is no heteroskedascity for the dataset.

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



Computing the residual term below, since the value is very close to zero, we further confirm there is no heteroskedascity.

```
y <- mcar$mpg
e <- resid(best.model)
yhat <- predict(best.model)
max(abs(e-(y-yhat)))
```

```
## [1] 6.439294e-14
```