Analysis

Thiloshon Nagarajah

4/25/2017

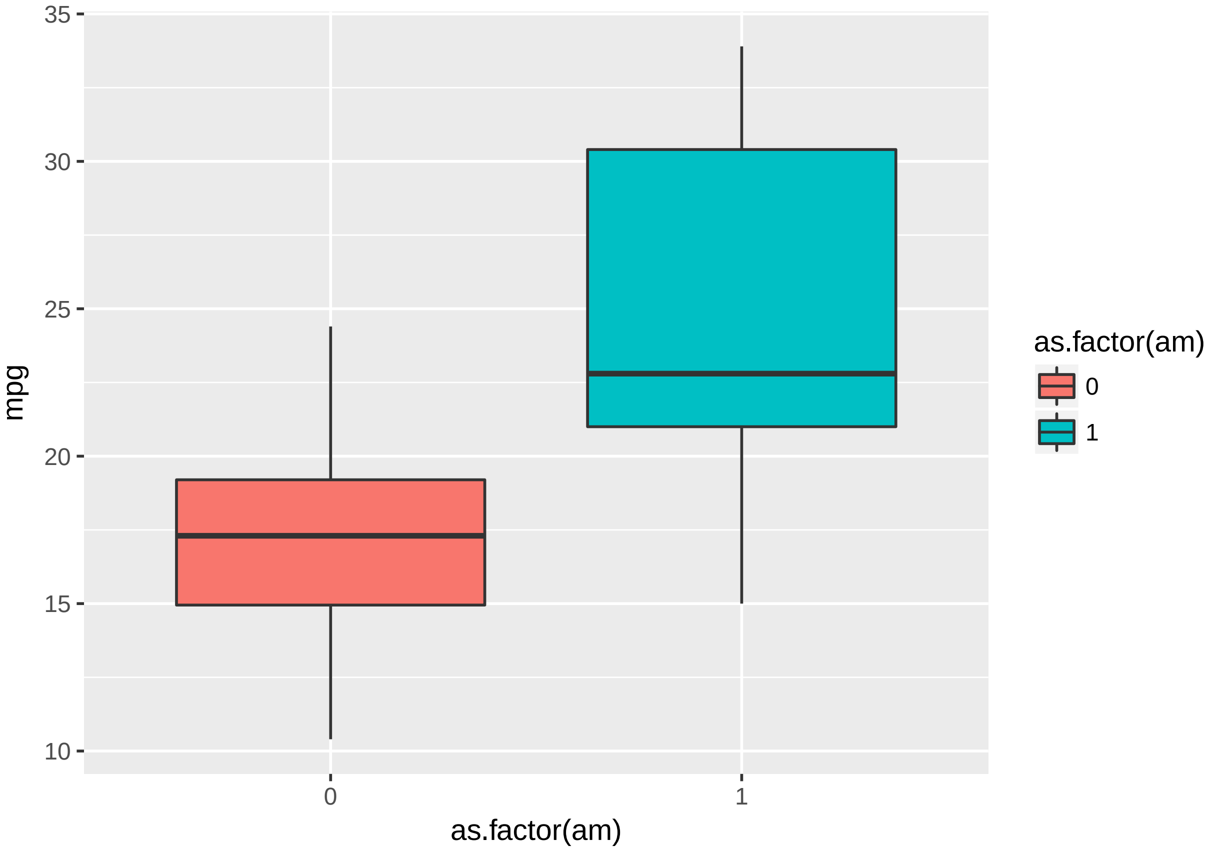
Synopsis

This analysis explores the relationship between a set of variables and miles per gallon (MPG) (outcome) of motor cars. We are particularly interested in the following two questions:

1. "Is an automatic or manual transmission better for MPG?"
2. "Quantify the MPG difference between automatic and manual transmissions"

Exploratory Analysis

```
cars<-mtcars
library(ggplot2)
ggplot(cars, aes(as.factor(am), mpg)) + geom_boxplot(aes(fill = as.factor(am)))
```



From this plot we can get a rough idea the Transmission mode has correlation with MPG. But this is not enough to quantify or conclude this is the only correlation.

Analysis

In order to quantify how correlated Transmission mode is, we need to first find what are the other variants that are correlated. To find that we can use correlation function.

```
correlation <-cor(cars$mpg, cars)
order <- correlation[,order(abs(correlation), decreasing = T)]
order
```

```
##        mpg         wt        cyl       disp         hp       drat
##  1.0000000 -0.8676594 -0.8521620 -0.8475514 -0.7761684  0.6811719
##         vs         am       carb       gear       qsec
##  0.6640389  0.5998324 -0.5509251  0.4802848  0.4186840
```

Now, lets select only the variables that are as or more correlated than Transmission mode.

```
variates <- names(order)[1:8]
relavantData<-cars[,names(cars) %in% variates]
head(relavantData)
```

```
##                    mpg cyl disp  hp drat    wt vs am
## Mazda RX4         21.0   6  160 110 3.90 2.620  0  1
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875  0  1
## Datsun 710        22.8   4  108  93 3.85 2.320  1  1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215  1  0
## Hornet Sportabout 18.7   8  360 175 3.15 3.440  0  0
## Valiant           18.1   6  225 105 2.76 3.460  1  0
```

Model Selection

Now we have subsetted the data, lets fit a linear regression model.

```
basicFit <- lm(mpg ~ am, relavantData)
summary(basicFit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = relavantData)
##
## 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 ***
## am             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
```

Here the R-squared value is 35. So our besic fit only explains 35% of the varience. Lets model a multi variate regression.

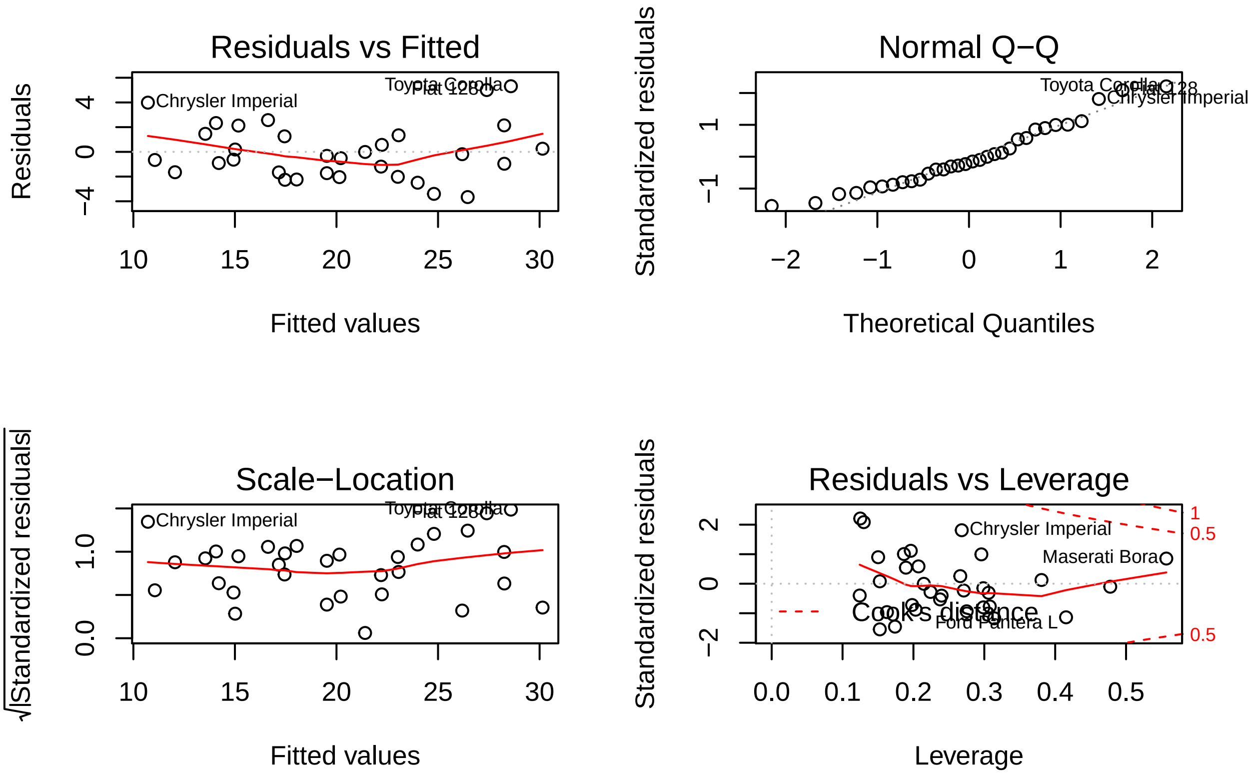
```r
multiFit <- lm(mpg ~ ., relavantData)
summary(multiFit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = relavantData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.660 -1.678 -0.417  1.371  5.312
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.45671    9.02383   3.597  0.00145 **
## cyl         -0.63992    0.89674  -0.714  0.48235
## disp         0.01348    0.01212   1.112  0.27695
## hp          -0.03032    0.01469  -2.063  0.05005 .
## drat         0.54696    1.51009   0.362  0.72037
## wt          -3.24531    1.16754  -2.780  0.01041 *
## vs           1.39761    1.84843   0.756  0.45694
## am           1.95201    1.75665   1.111  0.27749
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.571 on 24 degrees of freedom
## Multiple R-squared:  0.8591, Adjusted R-squared:  0.818
## F-statistic: 20.91 on 7 and 24 DF,  p-value: 9.089e-09
```

Now, the R-squared value has increased significantly. 86% of the varience is explained in this model which is considerably good. This model shows the Transmission mode from automatic to manual has increased 1.96 MPG.

Lets also check the Residuals plots.

```r
par(mfrow = c(2,2))
plot(multiFit)
```



The residuals show mostly homoskedastic behaviour, thus can conclude its a fairly good model.

Summary

The analysis build one by one from basic exploratory analusis to get a rough idea to a fairly complex multivariate model with 86% variance explained by the choosen variates. The analysis answers the questions we had quantitatively.