Project

Praveen S

8 February 2018

Regression Models - Course Project

Question

Take the mtcars data set and write up an analysis to answer the below two questions using regression models and exploratory data analyses.

1. Is an automatic or manual transmission better for MPG

2. Quantify the MPG difference between automatic and manual transmissions

Loading mtcars data

```
data(mtcars)
```

Convert required mtcars variables into factors

```
mtcars$cyl <- factor(mtcars$cyl)

mtcars$vs <- factor(mtcars$vs)

mtcars$am <- factor(mtcars$am, labels = c("Automatc", "Manual"))

mtcars$gear  <- factor(mtcars$gear)

mtcars$carb <- factor(mtcars$carb)
```

Just analyzing data based on question - To check the impact on mpg with transmission mode, we can plot a basic box-polt as in plot 1 given in the Appendix, we can observe that there is an impact on MPG with automatic transmission which is having a lower MPG.

Model building using Regression

Mathematically check the impact on MPG with transmission type

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

```
##          am      mpg
## 1 Automatc 17.14737
## 2   Manual 24.39231
```

Checking the difference in MPG for automatic and manual transmission type

```
aggregate(mpg ~ am, data = mtcars, mean)[2,2] - aggregate(mpg ~ am, data = mtcars,
mean)[1,2]
```

```
## [1] 7.244939
```

So we can hypothesize that automatic cars have MPG 7.2449 lower than manual cars.

Test using t-test to check whether this differnece is significant

```
T_automatic <- mtcars[mtcars$am=="Automatc",]

T_manual <- mtcars[mtcars$am=="Manual",]

t.test(T_automatic$mpg,T_manual$mpg)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  T_automatic$mpg and T_manual$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
```

The p-value obtained is 0.001374. So this is a significant differenec.

Build a model using linear regression

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

summary(first_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
```

Summary shows that the average MPG for automatic is 17.15 MPG and manual is 7.2 MPG higher.The R^2 value is 0.36 means this model explains only 36 % of the variance. So we have to build a multivariate linear regression.

We can include other variables to make it more accurate. Selection of the new vraiables to be included is based on the pairs plot which shows the variable correlation with mpg.The pairs plot is given as plot - 3 in the Appendix.

From the pairs plot we can observe that cyl, disp, hp, wt have the strongest correlation with mpg.So we can build a new model based on multivariate linear regression and compare it to the initial model.

Build a second model and use anova fuction to compare it to first model

```
second_model <- lm(mpg ~ am + cyl + disp + hp + wt, data = mtcars)


anova(first_model, second_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     25 150.41  5    570.49 18.965 8.637e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The new p-value is 8.637e-08. Which is far better than the initial model

We check again the residuals for non-normality and can observe that all normally distributed and homoskedastic. Please refer plot -3 given in the Appendix.

Check the parameters of second model

```
summary(second_model)
```

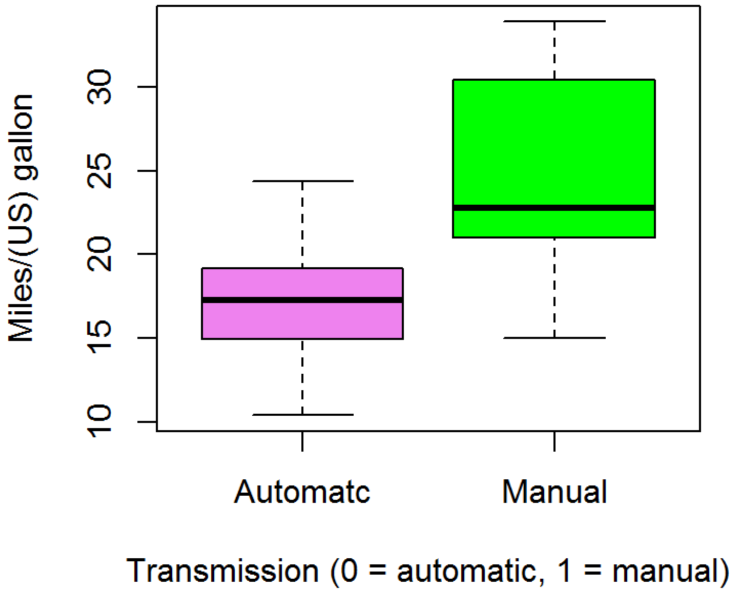
```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## amManual     1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## hp          -0.032480   0.013983  -2.323   0.0286 *
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

So it is clear from the summary that model explains 86.64 % of the variance and as a result variables cyl, hp, disp, wt affected the correlation between mpg and am.

So we can say that the difference between automatic and manual transmissions is 1.81 MPG. So manual transmission is better in terms of MPG.

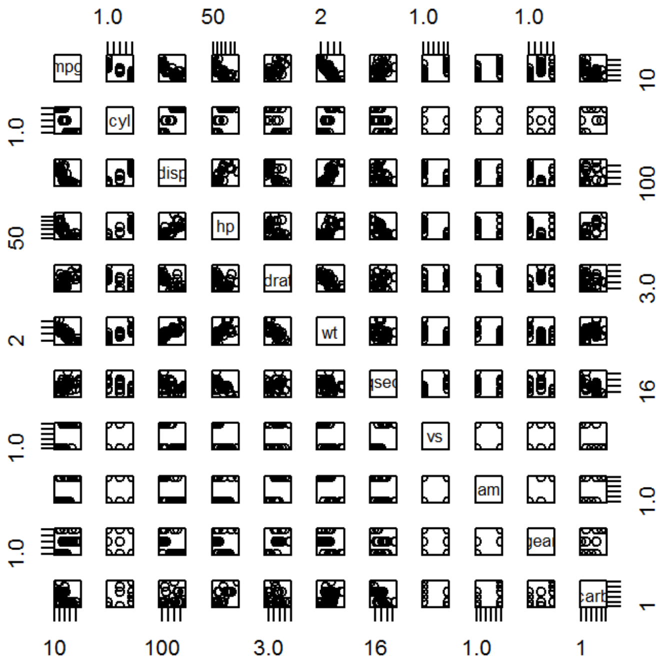
Appendix : Plot 1 Box-plot

```
boxplot(mpg ~ am, data = mtcars, col = (c("violet","green")), ylab = "Miles/(US) gallon", xlab = "Transmission (0 = automatic, 1 = manual)")
```



Plot 2 - Pairs plot

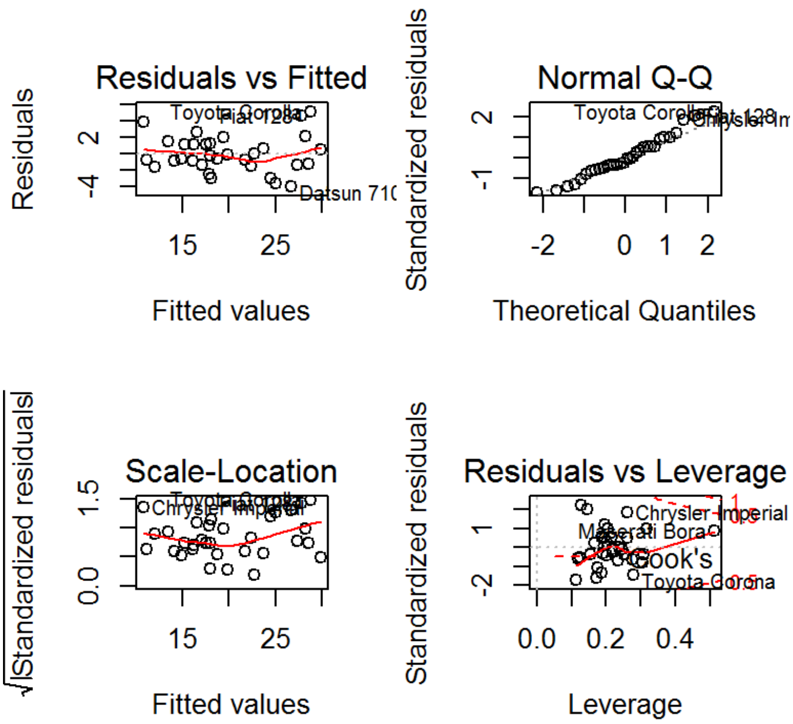
```
pairs(mpg ~ ., data = mtcars)
```



Plot 3 - Residuals

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

plot(second_model)
```


Residuals vs Fitted
Residuals
2
-4
15
25
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q-Q
Standardized residuals
2
-1
-2
0
1
2
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Im
Scale-Location
√|Standardized residuals|
1.5
0.0
15
25
Fitted values
Toyota Corolla
Fiat 128
Chrysler Imperial
Residuals vs Leverage
Standardized residuals
1
-2
0.0
0.2
0.4
Leverage
Chrysler Imperial
Maserati Bora
Cook's
Toyota Corona
1
0.5
0.5