Regression Models Course Project

Brief Introduction

In this project we are going to analyse the data related to cars. mtcars is a dataset in which the data collected on a collection of cars is saved. The primary objective of this analysis is to answer following questions.

1) Is the automatic or manual gear transmission effects the Miles per Gallon(MPG) or not?
2) Quanitfying the effect of gear shift type on MPG?

```
## [1] 32 11
```

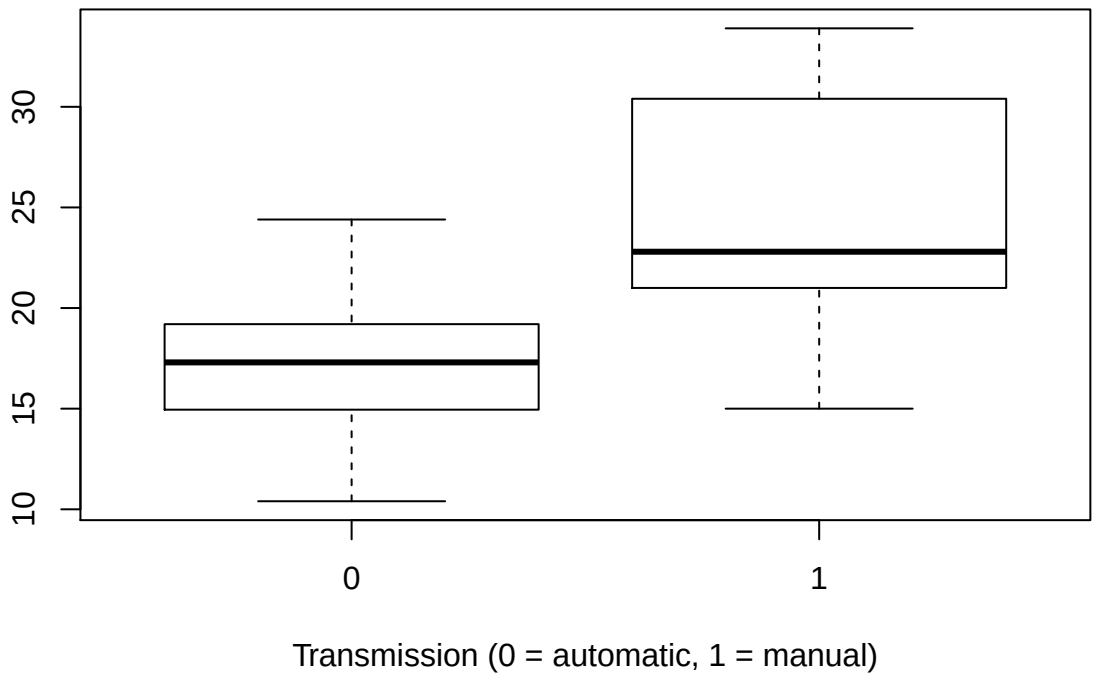
```
##       mpg       cyl      disp        hp      drat        wt      qsec
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        vs        am      gear      carb
## "numeric" "numeric" "numeric" "numeric"
```

The mtcars data set have 32 observations and 11 variables. all the variables are numeric in nature and the variles "mpg" is the target or dependent variable for our analysis.

subtitle: Test whether there is difference in MPG by great shfit type

Since our analysis is by variables vm,am conver these two variabels as factor variables.

We mainly focus on the relationship between variables mpg (Miles/(US) gallon) and am (Transmission). Box plot shows that there's a good separation of groups based on gas mileage.



Since there are 10 predictor variables in the data set. Some may play bigger role to determination of mpg. To identify the best variables out of these 10 predictor variales let's do an analysis of variance model as below.

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          1  817.7   817.7 116.425 5.03e-10 ***
## disp         1   37.6    37.6   5.353  0.03091 *
## hp           1    9.4     9.4   1.334  0.26103
## drat         1   16.5    16.5   2.345  0.14064
## wt           1   77.5    77.5  11.031  0.00324 **
## qsec         1    3.9     3.9   0.562  0.46166
## vs           1    0.1     0.1   0.018  0.89317
## am           1   14.5    14.5   2.061  0.16586
## gear         1    1.0     1.0   0.138  0.71365
## carb         1    0.4     0.4   0.058  0.81218
## Residuals   21  147.5     7.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above output gives the relative imporatnce of each variables with respect to MPG and obviously, variables with p-value below 0.05 are more important. Based on this, we choose cyl, disp, wt, drat, am as predictor variables for first model.

subtitle: Fitting the leaner model

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + drat + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3176 -1.3829 -0.4728  1.3229  6.0596
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.296380   7.538394   5.478 9.56e-06 ***
## cyl         -1.793995   0.650540  -2.758  0.01051 *
## disp         0.007375   0.012319   0.599  0.55462
## wt          -3.587041   1.210500  -2.963  0.00643 **
## drat        -0.093628   1.548780  -0.060  0.95226
## am1          0.172981   1.530043   0.113  0.91085
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.692 on 26 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8005
## F-statistic: 25.88 on 5 and 26 DF,  p-value: 2.528e-09
```

from the above output we can observe that the coefficient of drat has p value which shows that this variable is not siginificant in the model. let's run the model by ignoring this variable.

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## disp         0.007404   0.012081   0.613  0.54509
## wt          -3.583425   1.186504  -3.020  0.00547 **
## am1          0.129066   1.321512   0.098  0.92292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079
```

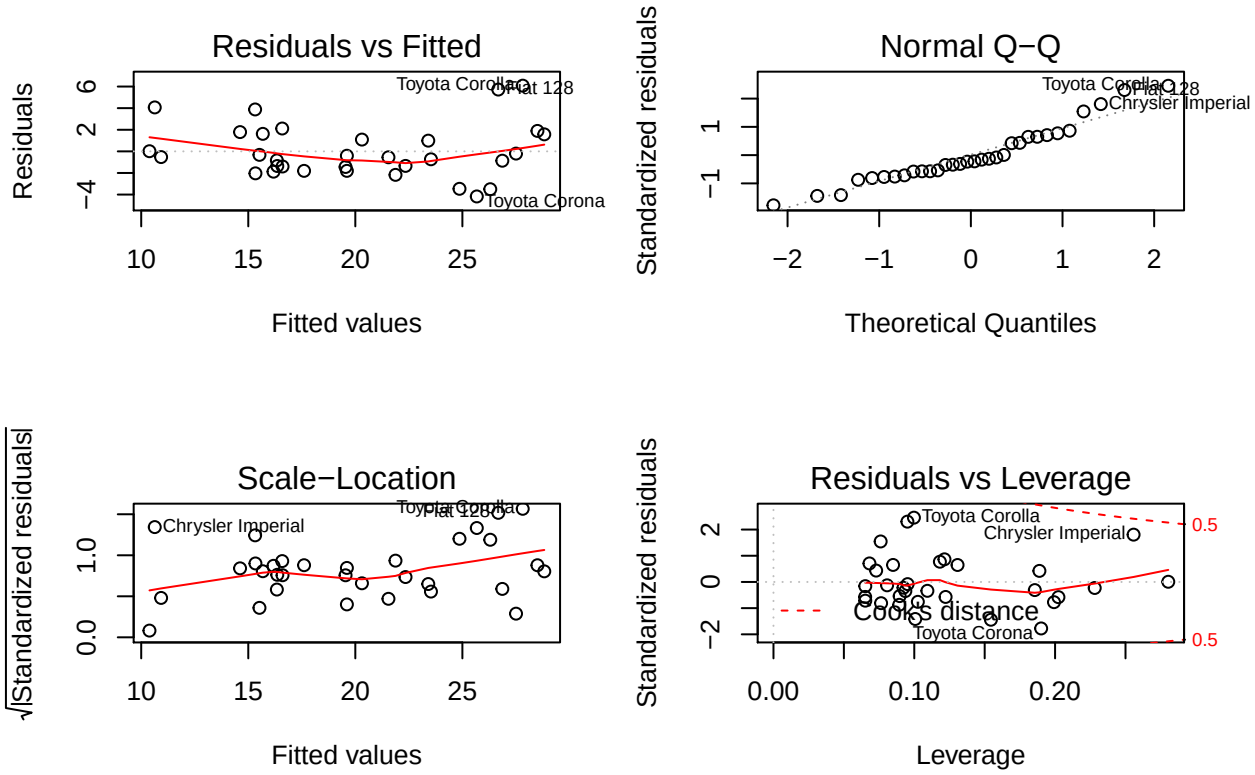
```
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

Now it is evident that disp is not significant in the model so remove it and re run the model with remaining variables.

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## wt           -3.1251     0.9109  -3.431  0.00189 **
## am1           0.1765     1.3045   0.135  0.89334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

The adjusted r-squared is 0.83 and this is our final model. Clearly, with cylinders and weight as confounding variables, the ceofficient of the am variable is small but has a large p-value. We cannot reject the hypothesis that the coefficient of am is 0.

To diagnostic the model, we apply the plot() to the object returned by the lm(). There is no discernible pattern found according to upper left graph. The normal Q-Q plot (upper right) indicates the model met the normality assumption. Scale-Location graph (bottom left) shows constant variance assumption are satisfied.



By looking at the plots we can conclude that the weight and number of cylenders paly an important role in determination of MPG.