mtcars Regression Analysis

Saurabh Ghadge

05/02/2022

Objective of study are:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"
  Attaching require packages:

```
data(mtcars)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Let us look at the some quick summary data-

```
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
```

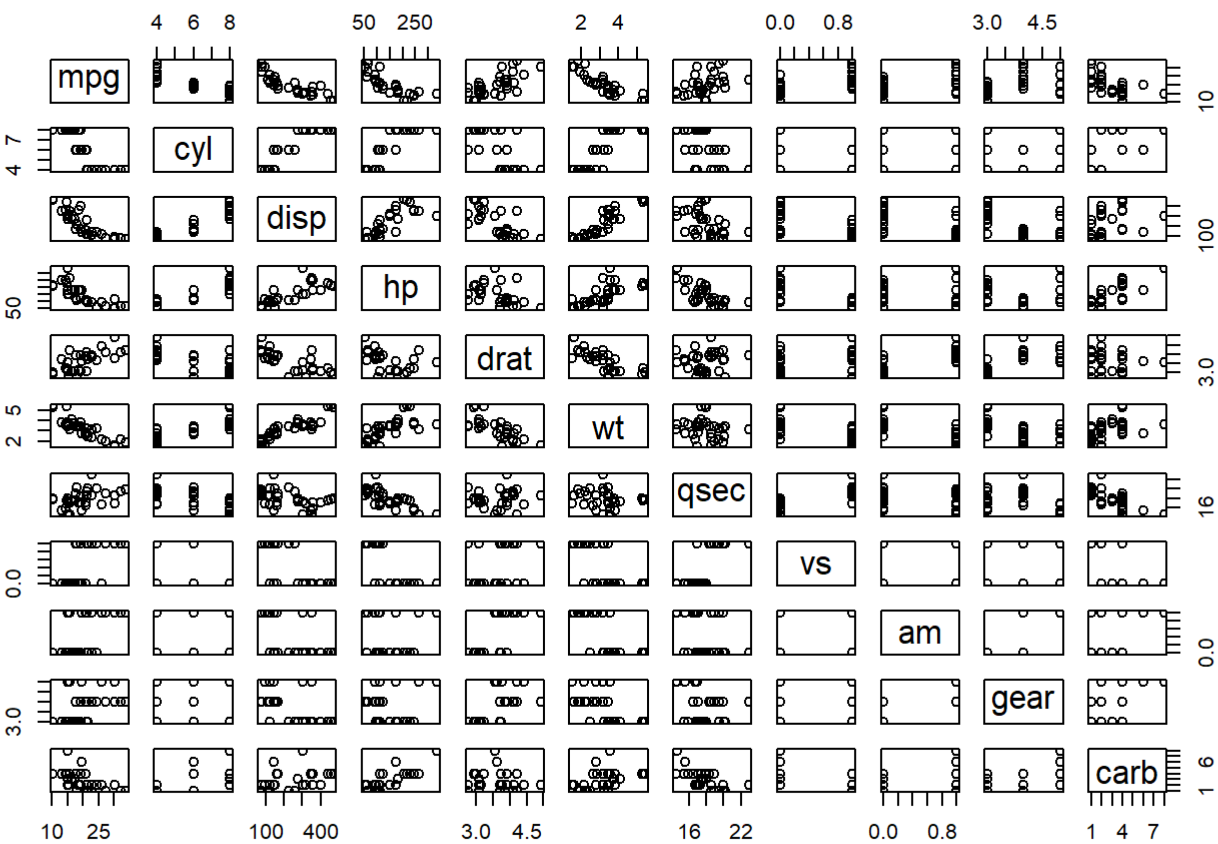
```
str(mtcars)
```

```
## '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 ...
```

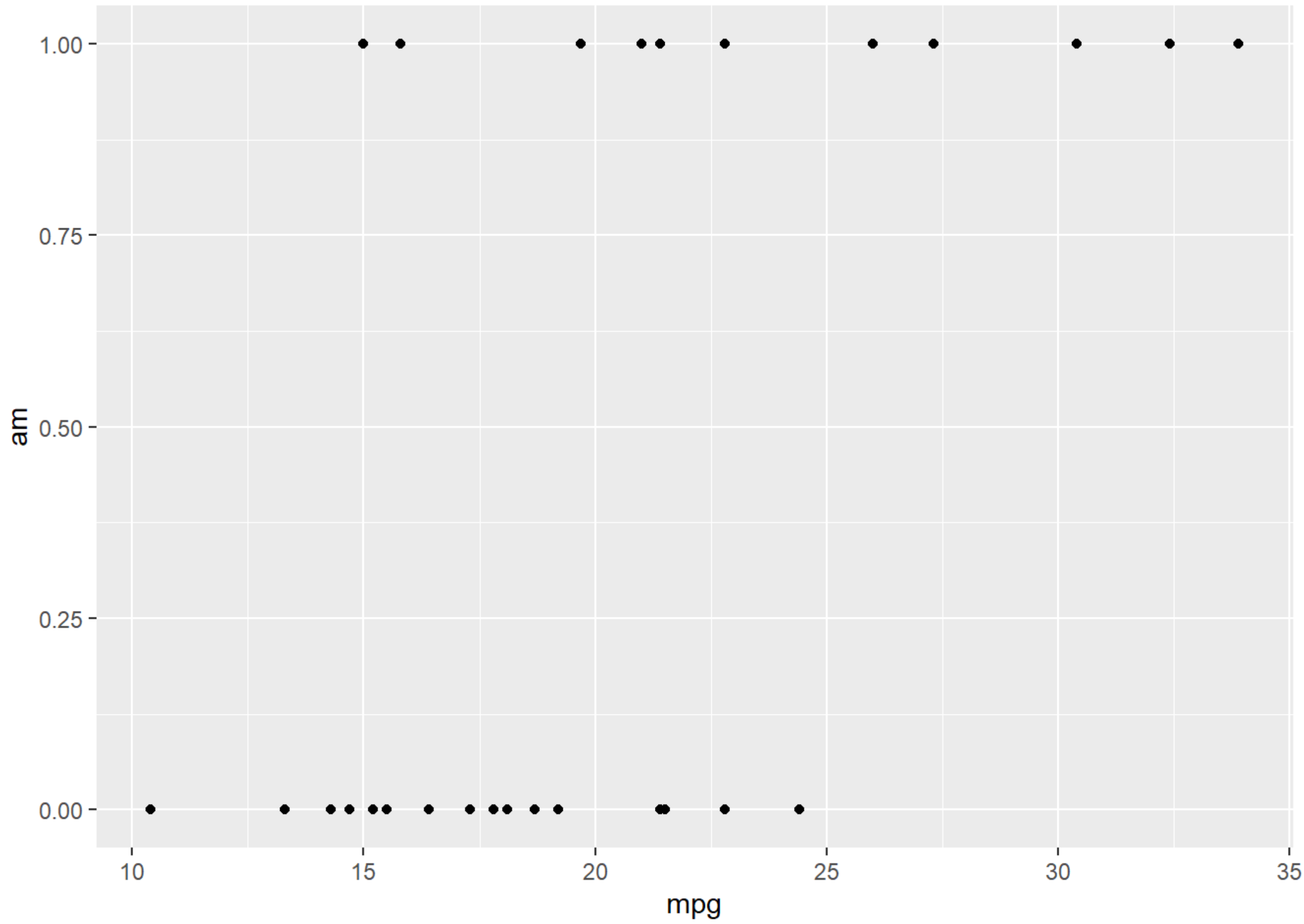
EDA

first we plot pair plot:

```
pairs(mtcars)
```



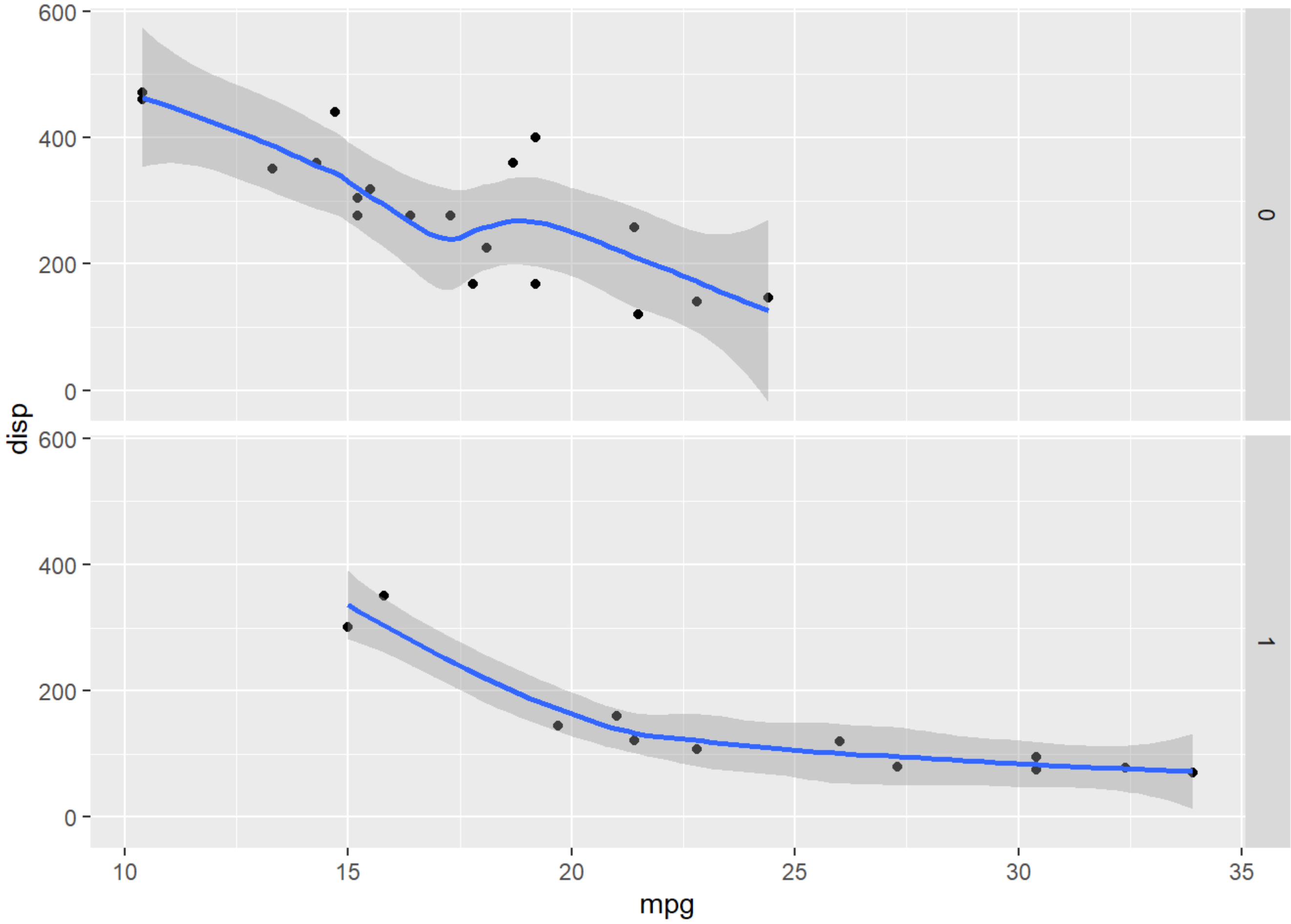

```
g <- ggplot(data = mtcars, mapping = aes(x = mpg,y = am))
g + geom_point()
```



Plot gives us idea about manual transmission are tends to give more mileage.

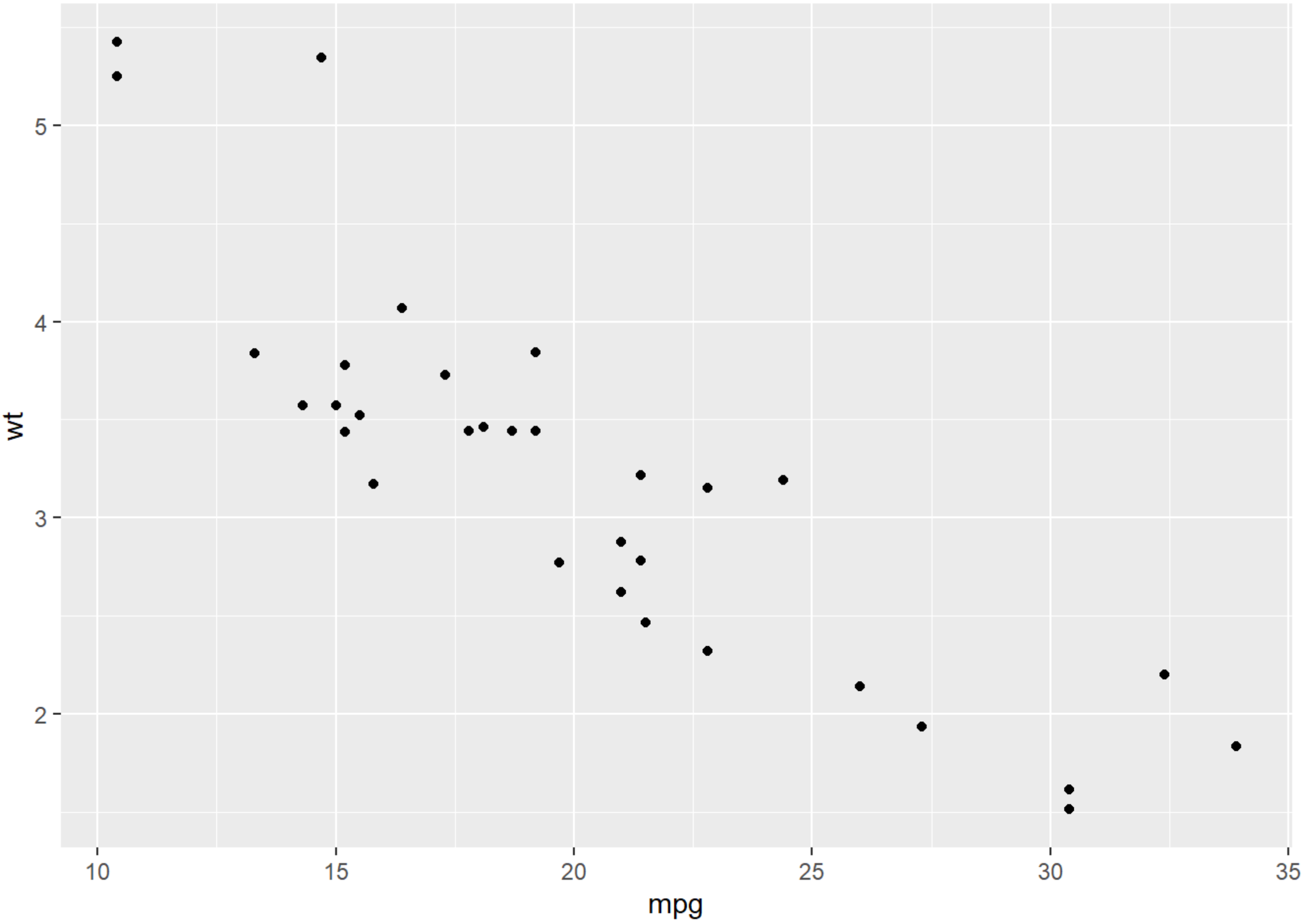
```
g <- ggplot(data = mtcars, mapping = aes(x = mpg,y = disp))
g + geom_point()+facet_grid(am~.)+geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Plot shows that Miles/(US) gallon is negatively correlated Displacement (cu.in.) which is seen deeply negatively correlated with cars having manual transmission (1).

```
g <- ggplot(data = mtcars, mapping = aes(x = mpg,y = wt))
g + geom_point()
```



Above plot shows that weight and mpg are strongly negatively correlated.

here,R is treating some of the column as numeric instead of that they should be treated as factor, And then we will find correlation of mpg with other continuous variable.

```
mtcars <- tibble(mtcars)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
mtcars$cyl <- as.factor(mtcars$cyl)
head(mtcars)
```

```
## # A tibble: 6 x 11
##     mpg cyl   disp    hp  drat    wt  qsec vs    am    gear  carb
##   <dbl> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct> <fct>
## 1  21   6       160   110  3.9   2.62  16.5 0     1     4     4
## 2  21   6       160   110  3.9   2.88  17.0 0     1     4     4
## 3  22.8 4       108    93  3.85  2.32  18.6 1     1     4     1
## 4  21.4 6       258   110  3.08  3.22  19.4 1     0     3     1
## 5  18.7 8       360   175  3.15  3.44  17.0 0     0     3     2
## 6  18.1 6       225   105  2.76  3.46  20.2 1     0     3     1
```

```
cont_mtcars <- mtcars %>% select(mpg,disp,hp,drat,wt,qsec)
cor(cont_mtcars)
```

```
##            mpg       disp         hp        drat         wt        qsec
## mpg   1.0000000 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403
## disp -0.8475514  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788
## hp   -0.7761684  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339
## drat  0.6811719 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476
## wt   -0.8676594  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588
## qsec  0.4186840 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000
```

Here, we see that mpg and qsec are moderately correlated so we are including in our study.
Now we will fit the model(multiple linear regression) model with all other continuous variable and with including only one categorical variable *am*(Transmission (0 = automatic, 1 = manual)).

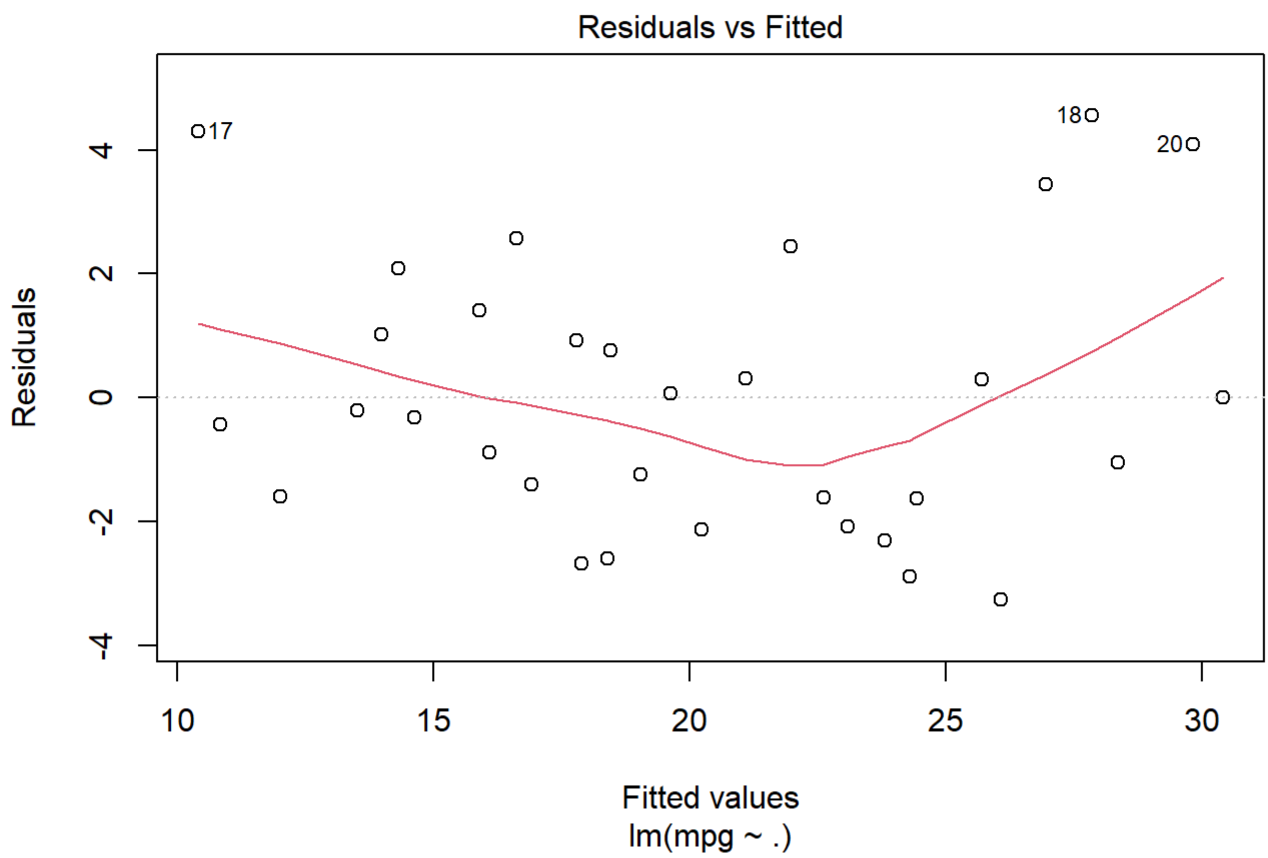
```
new_mtcars <- mtcars %>% select(mpg,disp,hp,drat,wt,qsec,am)
fit1 <- lm(mpg~.,data = new_mtcars)
summary(fit1)
```

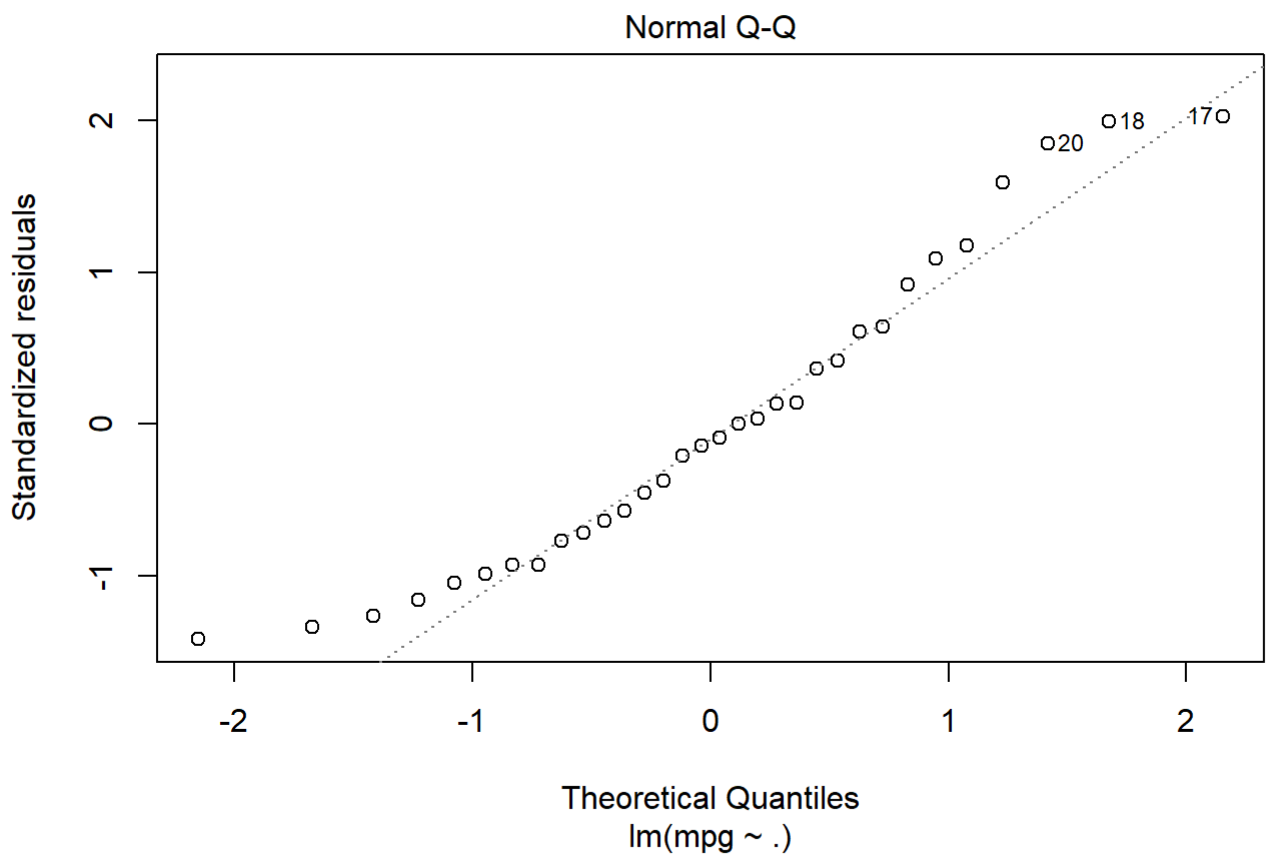
```
##
## Call:
## lm(formula = mpg ~ ., data = new_mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2669 -1.6148 -0.2585  1.1220  4.5564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.71062   10.97539   0.976  0.33848
## disp         0.01310    0.01098   1.193  0.24405
## hp          -0.02180    0.01465  -1.488  0.14938
## drat         1.02065    1.36748   0.746  0.46240
## wt          -4.04454    1.20558  -3.355  0.00254 **
## qsec         0.99073    0.48002   2.064  0.04955 *
## am1          2.98469    1.63382   1.827  0.07969 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.45 on 25 degrees of freedom
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8347
## F-statistic: 27.09 on 6 and 25 DF,  p-value: 8.637e-10
```

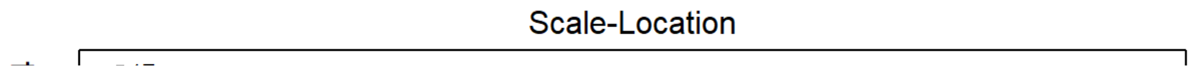
When transmission of car is automatic (am = 0),we can interpret intercept as expected Miles/(US) gallon when all other regressors are held constant or all regressor haves value equal to zero for automatic transmission(beta_0),and when transmission of car is manual the intercept becomes Intercept + am1 = 10.71 + 2.98 = 13.69 which is expected Miles/(US) gallon for manual transmission car when all other regressors are held constant or all regressor haves value equal to zero.
Also from summary of fitted model we see that p value for variable *disp*, *hp*, *drat* is not significant, i.e. it is try to say that slope of that variable is nearly close to zero, So excluding those variable from model will not affect the model that much.
p_value for *qsec*, *wt* is highly significant suggesting that they are playing important role in this model.
The Value of *R square* is 0.8867, which tells that about 87% variability in target variable/output data is explained by our model.

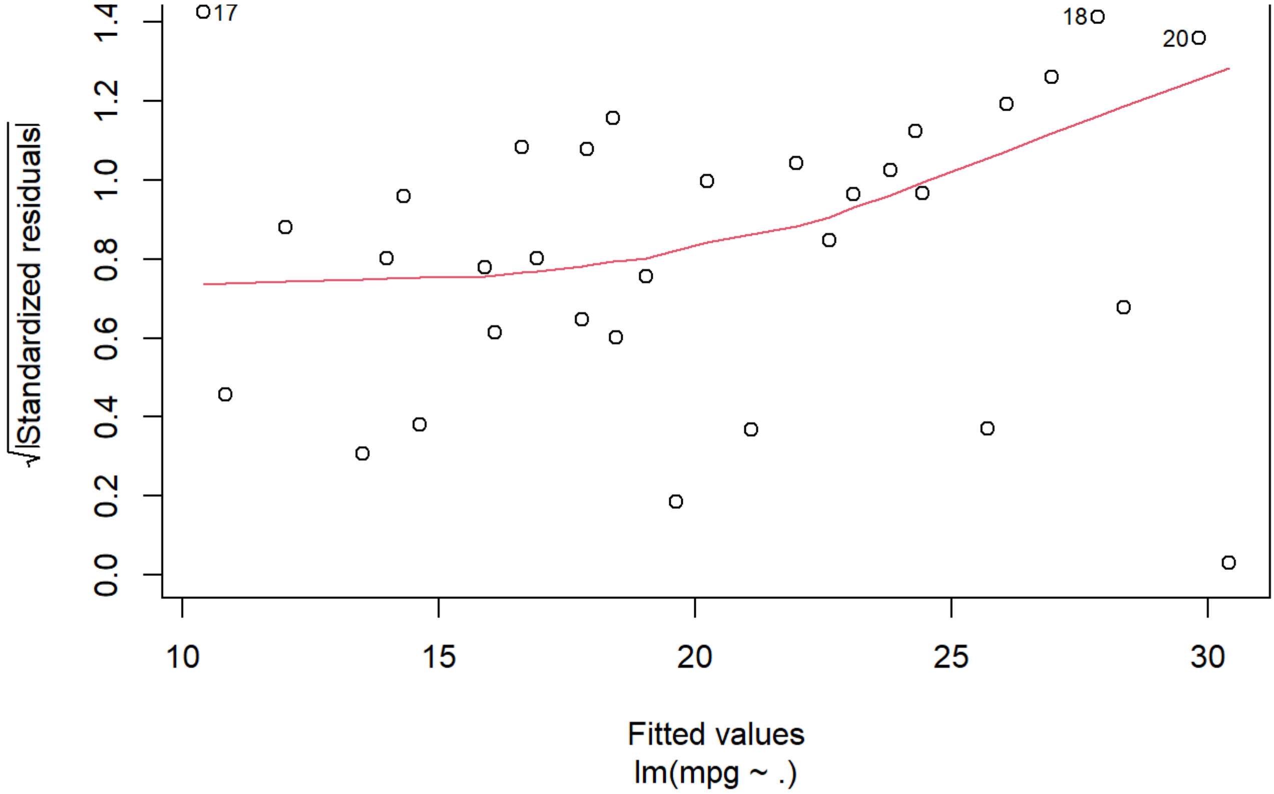
Following are residual diagnostic plot:

```
plot(fit1)
```


Residuals vs Fitted
Residuals
4
2
0
-2
-4
10
15
20
25
30
17
18
20
Fitted values
lm(mpg ~ .)


Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-1
0
1
2
20
18
17
Theoretical Quantiles
lm(mpg ~ .)


Scale-Location



From Residual Vs Fitted plot we see that, as time goes, spread of data is somewhat seems to be increasing.. which tells that residuals are increasing function of target, i.e assumption of constant variance is violated here.Which can be made stabilized using log transformation on output/target variable before fitting. Quantile plot of residuals is showing our residuals are normally distributed.Plot can we made more interpretative as residuals are following Gaussian(Normal) distribution if we use log transformation to target at starting of model building.
Residual Vs leverage plot shows that there are no outlier in our data.

Leaverage and Outlier :

```
hatvalues(fit1)
```

```
##          1          2          3          4          5          6          7
## 0.16148727 0.16225320 0.11806146 0.14351038 0.18462126 0.23448333 0.19932582
##          8          9         10         11         12         13         14
## 0.16340542 0.48867565 0.26642121 0.21880667 0.14339389 0.09410647 0.09210156
##         15         16         17         18         19         20         21
## 0.28930770 0.28440803 0.25173079 0.13156185 0.32927212 0.18295746 0.20310492
##         22         23         24         25         26         27         28
## 0.20484136 0.11024720 0.26362755 0.19582828 0.11465363 0.19194769 0.21532367
##         29         30         31         32
## 0.37700712 0.27202320 0.57662468 0.13487918
```

Only observation at 31 time point haves a high value for Hat. Lets try to fit a model without a including Transmission of car,and copare with previous model.

```
new_mtcars <- mtcars %>% select(mpg,disp,hp,drat,wt,qsec)
fit2 <- lm(mpg~.,data = new_mtcars)
summary(fit2)
```

```
## 
## Call:
## lm(formula = mpg ~ ., data = new_mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.5404 -1.6701 -0.4264  1.1320  5.4996 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.53357   10.96423   1.508  0.14362   
## disp         0.00872    0.01119   0.779  0.44281   
## hp          -0.02060    0.01528  -1.348  0.18936   
## drat         2.01578    1.30946   1.539  0.13579   
## wt          -4.38546    1.24343  -3.527  0.00158 **
## qsec         0.64015    0.45934   1.394  0.17523   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.558 on 26 degrees of freedom
## Multiple R-squared:  0.8489, Adjusted R-squared:  0.8199 
## F-statistic: 29.22 on 5 and 26 DF,  p-value: 6.892e-10
```

The value of both *R_square* and *adj R_sqaure* decreases in this model if we compared them with previous model, which indicate that Transmission plays significant role in estimating Miles/(US) gallon of car.
For more validation Let us compare both model by AIC..

```
AIC(fit1,fit2)
```

```
##      df      AIC
## fit1  8 156.2687
## fit2  7 158.2784
```

Criteria is that we use model with higher AIC, here AIC for model 2 high,which shows Transmission adds significant linear prediction beyond the other variable.

As in both model we observe that variable disp, hp, drat, are not playing that much of significant role in model. So next we are going to fit the model by dropping them. And also we use log transformation to target variable to make variability of target constant over time.

```
cars <- mtcars %>% select(mpg,wt,qsec,am)
fit3 <- lm(log(mpg)~.,data = cars)
summary(fit3)
```

```
##
## Call:
## lm(formula = log(mpg) ~ ., data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13879 -0.08114 -0.03466  0.07030  0.26575
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.69410    0.31326   8.600 2.40e-09 ***
## wt          -0.22456    0.03201  -7.015 1.25e-07 ***
## qsec         0.05329    0.01299   4.101  0.00032 ***
## am1          0.08558    0.06351   1.347  0.18863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1107 on 28 degrees of freedom
## Multiple R-squared:  0.8752, Adjusted R-squared:  0.8619
## F-statistic: 65.47 on 3 and 28 DF,  p-value: 9.036e-13
```

All regression coefficients have their respective meaning as explained earlier in this case we have to exp(coef) as we use a log transformation at start of model. We see that now all variables are playing significant role in model building, which can seen by their respective p_values.Also this model is explaining about 88% variability from the output variable.
To verify validity we can see residual diagnosis which in case are looking satisfying all criteria i.e. constant variance of residual,normality of residual.

For choosing best model with best set of parameters we can use stepwise selection statistical procedure to do so..
* Best Subset Regression Select the subset of predictors that do the best at meeting some well-defined objective criterion, such as having the largest R2 value or the smallest MSE, Mallow’s Cp or AIC.

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
best_fit <- ols_step_best_subset(lm(mpg~.,data = mtcars))
best_fit
```

```
##              Best Subsets Regression
## ------------------------------------------------------
## Model Index    Predictors
## ------------------------------------------------------
##     1         wt
##     2         cyl wt
##     3         cyl hp wt
##     4         cyl hp wt am
##     5         cyl hp wt gear carb
##     6         cyl disp hp wt gear carb
##     7         cyl disp hp wt vs gear carb
##     8         cyl disp hp drat wt vs gear carb
##     9         cyl disp hp drat wt qsec vs gear carb
##    10         cyl disp hp drat wt qsec vs am gear carb
## ------------------------------------------------------
##
##                                       Subsets Regression Summary
## -----------------------------------------------------------------------------------------------------------------
------------
##                    Adj.       Pred
## Model    R-Square    R-Square    R-Square    C(p)      AIC       SBIC      SBC       MSEP      FPE      H
SP      APC
## -----------------------------------------------------------------------------------------------------------------
------------
##  1      0.7528     0.7446     0.7087    6.6739    166.0294    74.8970    170.4266    296.9167    9.8572
0.3199    0.2801
##  2      0.8374     0.8200     0.793    -3.1942    156.6223    65.9041    163.9510    202.2635    7.1507
0.2335    0.1962
##  3      0.8572     0.8361     0.8041    -3.9700    154.4692    65.7237    163.2636    184.2244    6.6991
0.2205    0.1836
##  4      0.8659     0.8401     0.8015    -3.1849    154.4669    67.6803    164.7270    179.7059    6.7163
0.2234    0.1838
##  5      0.8754     0.8069      -Inf    -2.5267    164.0994    69.8814    183.1540    173.5660    8.3277
0.2805    0.1820
##  6      0.8862     0.8144      -Inf    -2.0419    163.1968    72.4372    183.7172    165.1205    8.2165    0.
2809    0.1775
##  7      0.8898     0.8102      -Inf    -0.5371    164.1880    76.2000    186.1741    166.9526    8.6193    0.
2998    0.1837
##  8      0.8917     0.8024      -Inf     1.1997    165.6386    80.2787    189.0904    171.5699    9.1953    0.
3262    0.1931
##  9      0.8921     0.7909      -Inf     3.1423    167.5175    84.4924    192.4350    179.0613    9.9705    0.
3617    0.2061
## 10      0.8931     0.7790      -Inf     5.0000    169.2155    88.7678    195.5987    186.2479    10.7861
0.4013    0.2189
## -----------------------------------------------------------------------------------------------------------------
------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

which shows that model 4 having regressor cyl hp wt am is doing well.

summary :-

- Model with Regressor cyl,hp,wt,am are best set of regressor as it generates better adjusted R_square and also other evaluation metrics values as compared other subsets of regressor involving in model.

Conclussions :-

- From above we see that including Transmission is play significant role.
- Manual transmission tends to give a better mileage as compared to automatic transmission.