Manual or Automatic

Marco Venturi

2022-09-09

Abstract

In the Motor Trend magazine, we looked for a answer about which trasmission is better for ratio mile per gallon (mpg). This report go throught the mtcar dataset supplied by R, with some Exploratory data analisys and differents models that helped to discover wich trasmission has a better impact to mpg.

Getting and Cleaning Data

```
##       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
```

The summary show that all features are numeric, let's investigate further.

```
str(df)
```

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

```
#am unique values
table(df$am)
```

```
##
##  0  1
## 19 13
```

```
#cyl unique values
table(df$cyl)
```

```
##
##  4  6  8
## 11  7 14
```

```
#vs unique values
table(df$vs)
```

```
##
##  0  1
## 18 14
```

```
#gear unique values
table(df$gear)
```

```
##
##  3  4  5
## 15 12  5
```

```
#carb unique values
table(df$carb)
```

```
##
##  1  2  3  4  6  8
##  7 10  3 10  1  1
```

There are differrentfactor variable, they are going to be factorize.

```
## '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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

```
#sum of N/A values
sum(is.na(df))
```

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

```
head(df,5)
```

```
##                   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
```

Now df is clean and there are any missing values.

Exploratory data Analisys

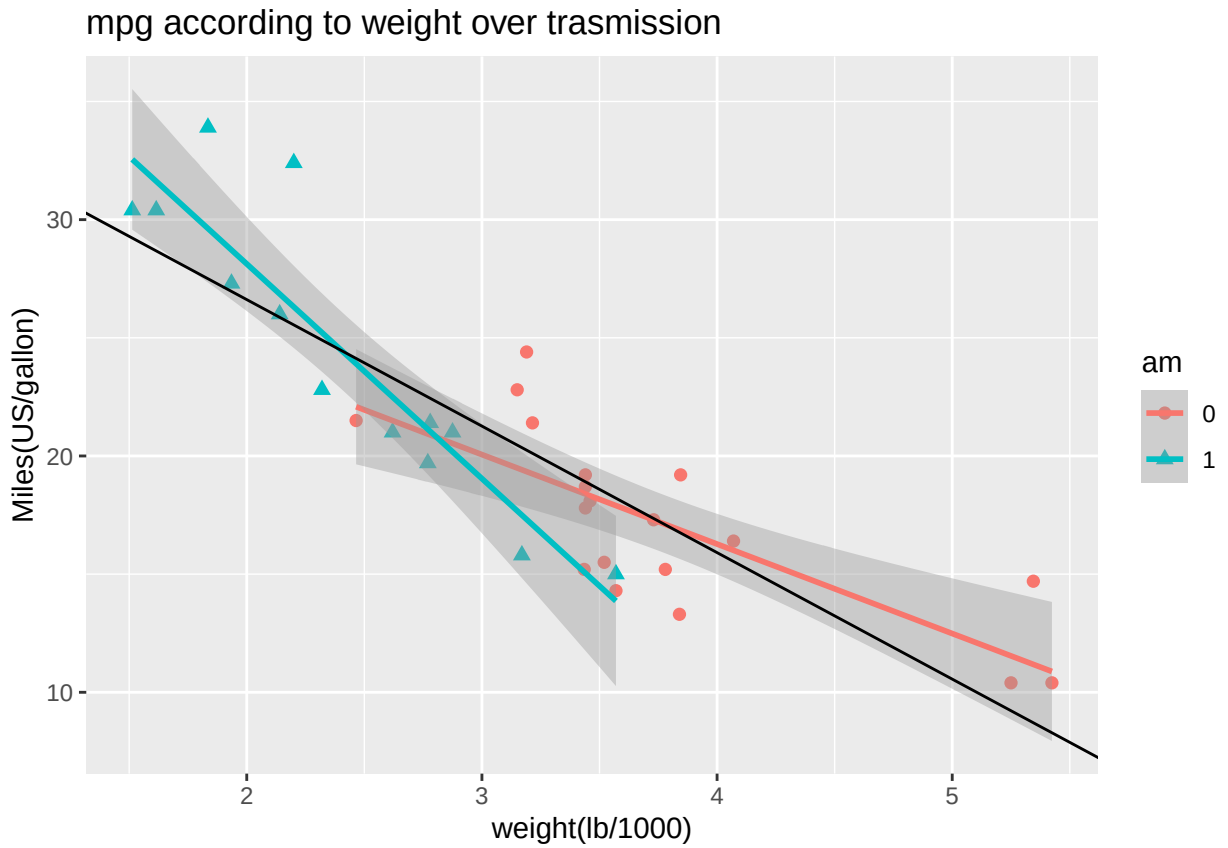
```
##     Estimate Std. Error  t value     Pr(>|t|)
## am0 17.14737   1.124603 15.24749 1.133983e-15
## am1 24.39231   1.359578 17.94109 1.376283e-17
```

Comparing *mpg* to *automatic* and *manual* , it seems that automatic cars consume less than 18 per mile, while manual it's around 24/25 mpg. Let's see how horse-power and weight influence those statistics.

Horse power and weight investigation

Weight:

```
## `geom_smooth()` using formula 'y ~ x'
```



The graph show a relation between *weight* and *mpg* over *transmission*, less a car weight more mpg does the car, if it's applied transmission information too, it seems that lower weight cars that have manual transmission, they consume in general less respect to automatic cars, while this assumption is not true if we talk about medium size car, but any light cars data with automatic transmission are given for prove the first assumption.

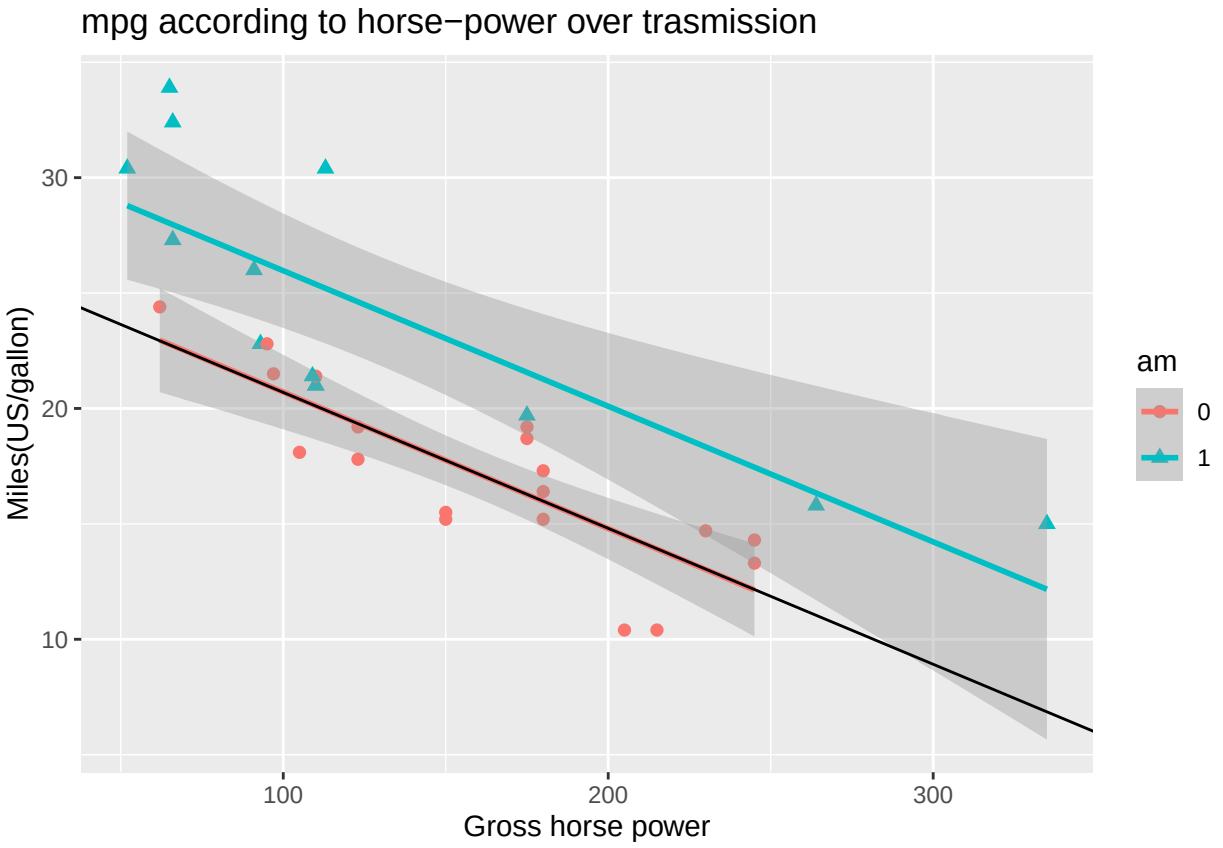
let's check this assumption:

```
##          Estimate Std. Error   t value     Pr(>|t|)
## wt      -3.785908  0.7856478 -4.818836 4.551182e-05
## am0     31.416055  3.0201093 10.402291 4.001043e-11
## am1     46.294478  3.0101489 15.379465 3.488923e-15
## wt:am1 -5.298360  1.4446993 -3.667449 1.017148e-03
```

the automatic trasmission seems to consume more than manual adjusting with weight

Horse_power:

```
## `geom_smooth()` using formula 'y ~ x'
```



The graph show a relation between *gross horse-power* and *mpg* over *transmission*, it seems that more horse power has the car less mpg are. Transmission to influence the mpg, and it show that automatic cars use less mpg when horse power grow then manual ones.

let's check this assumption:

```
summary(lm(mpg ~ hp*am -1, data = df))
```

```
##
## Call:
## lm(formula = mpg ~ hp * am - 1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3818 -2.2696  0.1344  1.7058  5.8752
##
## Coefficients:
```

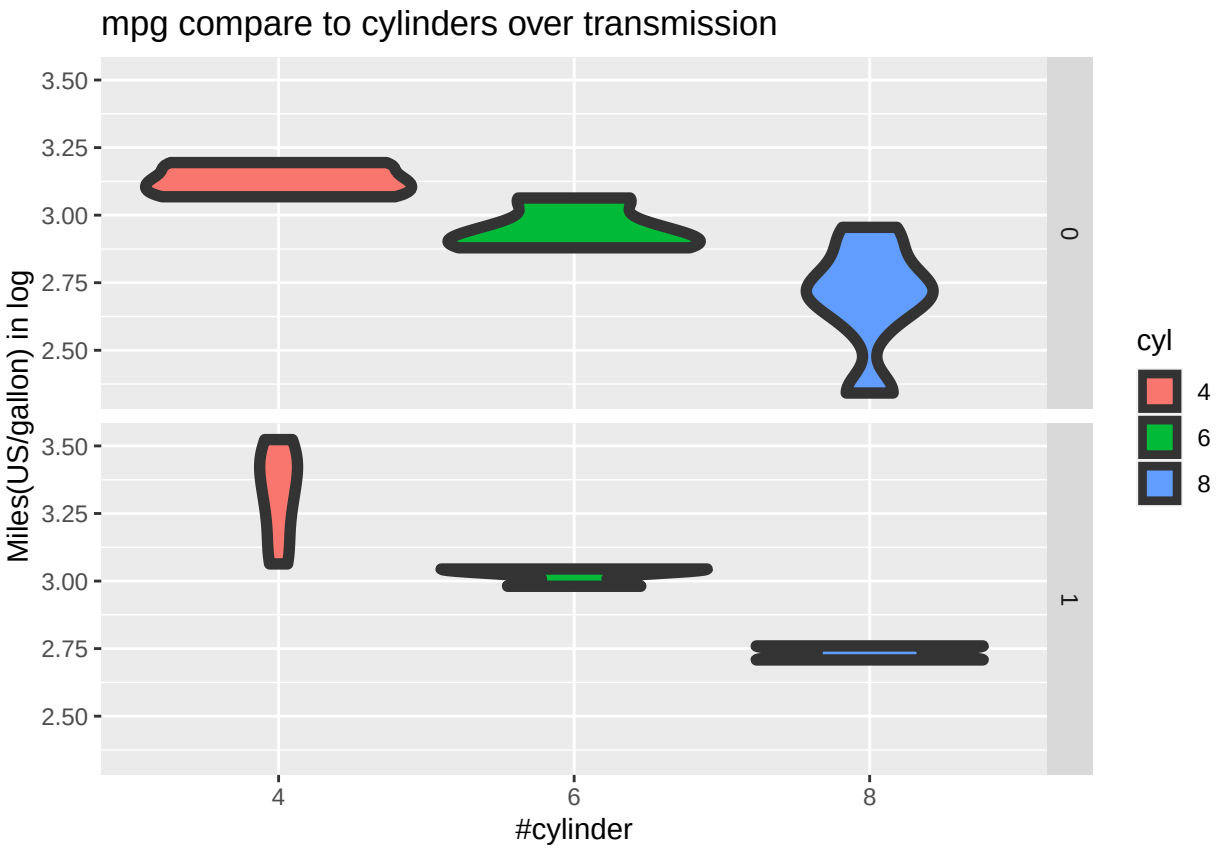
```
##          Estimate Std. Error t value Pr(>|t|)
## hp     -0.0591370  0.0129449  -4.568 9.02e-05 ***
## am0    26.6248479  2.1829432  12.197 1.01e-12 ***
## am1    31.8425012  1.5288820  20.827  < 2e-16 ***
## hp:am1  0.0004029  0.0164602   0.024    0.981
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 28 degrees of freedom
## Multiple R-squared:  0.9825, Adjusted R-squared:   0.98
## F-statistic: 393.5 on 4 and 28 DF,  p-value: < 2.2e-16
```

Comparing cylinderrs, C-shape engine, gear and carb to mpg over trasmission before start let's check which features seems to influence the mpg prediction and them coefficients

```
##          Estimate  Std. Error    t value   Pr(>|t|)
## cyl4  23.87913244 20.06582026  1.1900402 0.25252548
## cyl6  21.23043717 18.33416483  1.1579713 0.26498157
## cyl8  23.54296946 18.22249667  1.2919728 0.21591810
## disp   0.03554632  0.03189920  1.1143329 0.28267339
## hp    -0.07050683  0.03942556 -1.7883534 0.09393155
## drat   1.18283018  2.48348458  0.4762784 0.64073922
## wt    -4.52977584  2.53874584 -1.7842573 0.09461859
## qsec   0.36784482  0.93539569  0.3932505 0.69966720
## vs1    1.93085054  2.87125777  0.6724755 0.51150791
## am1    1.21211570  3.21354514  0.3771896 0.71131573
## gear4  1.11435494  3.79951726  0.2932886 0.77332027
## gear5  2.52839599  3.73635801  0.6767007 0.50889747
## carb2 -0.97935432  2.31797446 -0.4225044 0.67865093
## carb3  2.99963875  4.29354611  0.6986390 0.49546781
## carb4  1.09142288  4.44961992  0.2452845 0.80956031
## carb6  4.47756921  6.38406242  0.7013668 0.49381268
## carb8  7.25041126  8.36056638  0.8672153 0.39948495
```

Cylinders seems to have more impact respect to other features.

```
## Warning: Ignoring unknown parameters: coloir
```



Checking variance of mpg compare to #cylinders over transmision it's seems that automatic transmision as flatter variance then manual transmission does on 4 cyl while the opposite occur for 6 and 8 cyl.

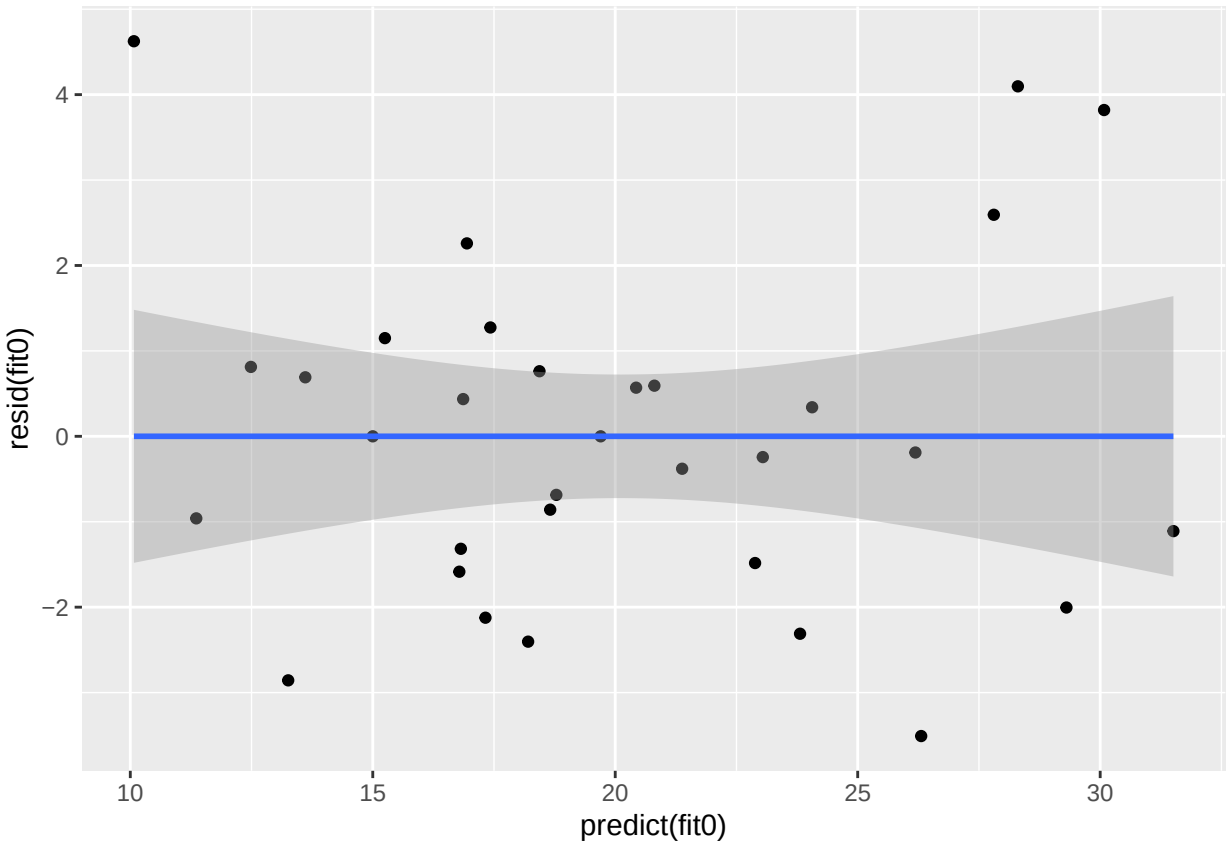
Diagnostic

```
##          Estimate  Std. Error    t value   Pr(>|t|)
## cyl4  23.87913244 20.06582026  1.1900402 0.25252548
## cyl6  21.23043717 18.33416483  1.1579713 0.26498157
## cyl8  23.54296946 18.22249667  1.2919728 0.21591810
## disp   0.03554632  0.03189920  1.1143329 0.28267339
## hp    -0.07050683  0.03942556 -1.7883534 0.09393155
## drat   1.18283018  2.48348458  0.4762784 0.64073922
## wt    -4.52977584  2.53874584 -1.7842573 0.09461859
## qsec   0.36784482  0.93539569  0.3932505 0.69966720
## vs1    1.93085054  2.87125777  0.6724755 0.51150791
## am1    1.21211570  3.21354514  0.3771896 0.71131573
## gear4  1.11435494  3.79951726  0.2932886 0.77332027
## gear5  2.52839599  3.73635801  0.6767007 0.50889747
## carb2 -0.97935432  2.31797446 -0.4225044 0.67865093
## carb3  2.99963875  4.29354611  0.6986390 0.49546781
## carb4  1.09142288  4.44961992  0.2452845 0.80956031
## carb6  4.47756921  6.38406242  0.7013668 0.49381268
## carb8  7.25041126  8.36056638  0.8672153 0.39948495
```

Looking the coefficients it's seems that all features are necessary, so we need to first the *residual plot* for look throught the data pattern and then check over *vif* for look variance on features.

Residual plot Now that we know the main features that influence mpg overam lets look the residual plot

```
## `geom_smooth()` using formula 'y ~ x'
```



residual plot have a pattern that suggest a linear regresion should fit the pattern, but we still need to investigate over outliers for check if some data can be adjust for fit in a better way the model.

```
##            GVIF Df GVIF^(1/(2*Df))
## cyl  128.120962  2        3.364380
## disp  60.365687  1        7.769536
## hp    28.219577  1        5.312210
## drat   6.809663  1        2.609533
## wt    23.830830  1        4.881683
## qsec  10.790189  1        3.284842
## vs     8.088166  1        2.843970
## am     9.930495  1        3.151269
## gear  50.852311  2        2.670408
## carb 503.211851  5        1.862838
```

```
##           GVIF       Df GVIF^(1/(2*Df))
## cyl  11.319053 1.414214        1.834225
## disp  7.769536 1.000000        2.787389
## hp    5.312210 1.000000        2.304823
## drat  2.609533 1.000000        1.615405
## wt    4.881683 1.000000        2.209453
## qsec  3.284842 1.000000        1.812413
## vs    2.843970 1.000000        1.686407
## am    3.151269 1.000000        1.775181
```

```
## gear  7.131081 1.414214        1.634138
## carb 22.432384 2.236068        1.364858
```

Model selection

In this section we are interest in find the best model for *mpg* and then check which *am* is better.

Now i will check which feature is better to exclude by the use of *anova* function and the pre-knoledge obtain with *vif* test.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + vs + qsec
## Model 3: mpg ~ am + vs + qsec + hp + wt
## Model 4: mpg ~ am + vs + qsec + hp + wt + drat
## Model 5: mpg ~ am + vs + qsec + hp + wt + drat + disp
## Model 6: mpg ~ am + vs + qsec + hp + wt + drat + disp + gear
## Model 7: mpg ~ am + vs + qsec + hp + wt + drat + disp + carb
## Model 8: mpg ~ am + vs + qsec + hp + wt + drat + disp + cyl
## Model 9: mpg ~ (cyl + disp + hp + drat + wt + qsec + vs + am + gear +
##     carb) - 1
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 315.89  2    405.01 25.2286 1.589e-05 ***
## 3     26 159.82  2    156.07  9.7216  0.001961 **
## 4     25 158.56  1      1.26  0.1569  0.697630
## 5     24 149.45  1      9.11  1.1351  0.303548
## 6     22 146.57  2      2.88  0.1795  0.837437
## 7     19 134.15  3     12.42  0.5157  0.677744
## 8     22 139.02 -3     -4.87  0.2024  0.893093
## 9     15 120.40  7     18.62  0.3314  0.927349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

the best suitable model with linear regression seems to be m5 the best.

Aswer and conclusion

```
b_m <- lm(mpg ~ am + vs + qsec + hp + wt + drat + disp -1 , data = df)
summary(b_m)$coefficients
```

```
##         Estimate  Std. Error    t value    Pr(>|t|)
## am0  12.49804962 12.48038774  1.0014152 0.326616519
## am1  15.52206608 12.21052872  1.2712034 0.215839083
## vs1   0.59016578  1.83303033  0.3219618 0.750269228
## qsec  0.87148931  0.61331436  1.4209504 0.168194583
## hp   -0.02282191  0.01525893 -1.4956426 0.147781592
## wt   -3.94973602  1.26261038 -3.1282303 0.004567014
## drat  0.95532814  1.40737217  0.6788028 0.503756614
## disp  0.01373930  0.01135852  1.2096028 0.238211942
```

Answer Automatic transmission seems to be the best transmission respect to mpg because consume around 12 mpg with a variance of 12, respect to manual transmission that consume 15 mpg with 12 of mpg variance.