Regression Models Final Project

Jungsun Lee

2021/7/19

Introduction

This report is answer the following project questions.
1.Is an automatic or manual transmission better for MPG
2.Quantify the MPG difference between automatic and manual transmissions

```
library(ggplot2)
library(magrittr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(reshape2)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
data("mtcars")
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 ...
```

As we can see above data type, some factor variables are numeric data type. The following is change those variables to factor.

```
cols <- c("cyl", "vs", "gear", "carb")
mtcars %<>% mutate_at(cols, funs(factor(.)))
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

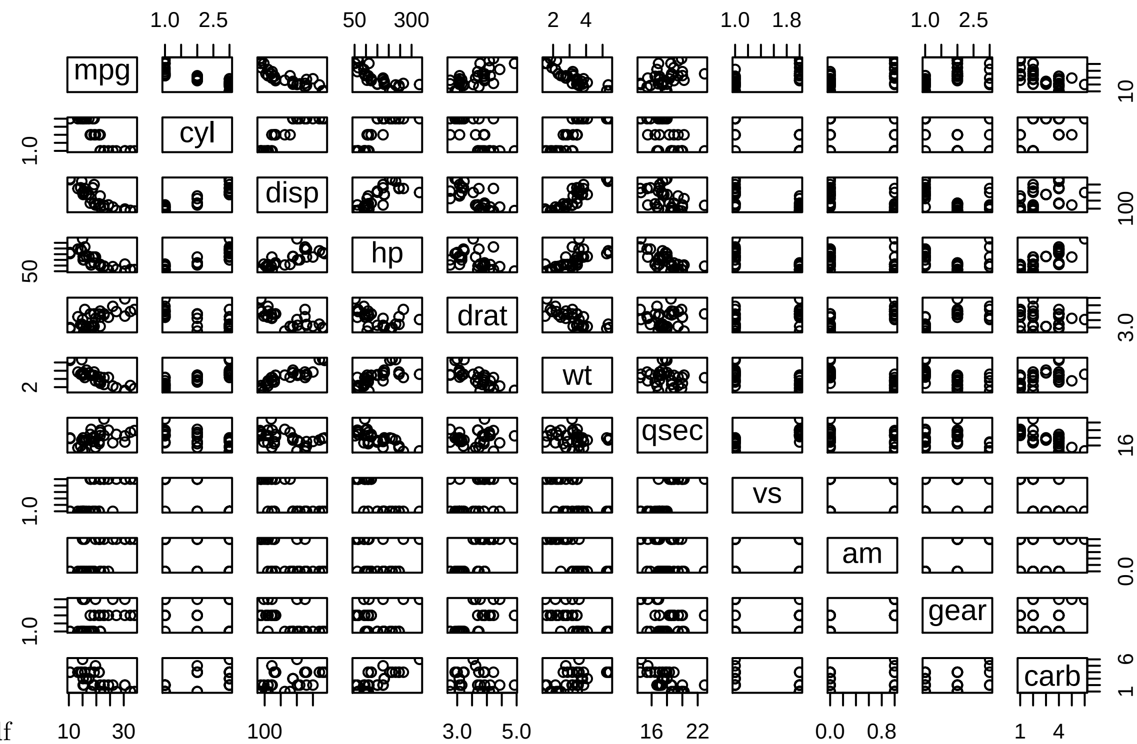
```
str(mtcars)
```

```
## '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 : 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  : num 1 1 1 0 0 0 0 0 0 0 ...
## $ 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 ...
```

Exploratory Data Analysis

Here, I use pairs function to analyze relation between all variables and perform correlation analysis

```
pairs(mtcars)
```



plot-1.pdf

```
mydf <- mtcars[, c(1,3,4,5,6,7)]
cormat <- round(cor(mydf),2)
cormat
```

```
##         mpg  disp    hp  drat    wt  qsec
## mpg   1.00 -0.85 -0.78  0.68 -0.87  0.42
## disp -0.85  1.00  0.79 -0.71  0.89 -0.43
## hp   -0.78  0.79  1.00 -0.45  0.66 -0.71
## drat  0.68 -0.71 -0.45  1.00 -0.71  0.09
## wt   -0.87  0.89  0.66 -0.71  1.00 -0.17
## qsec  0.42 -0.43 -0.71  0.09 -0.17  1.00
```

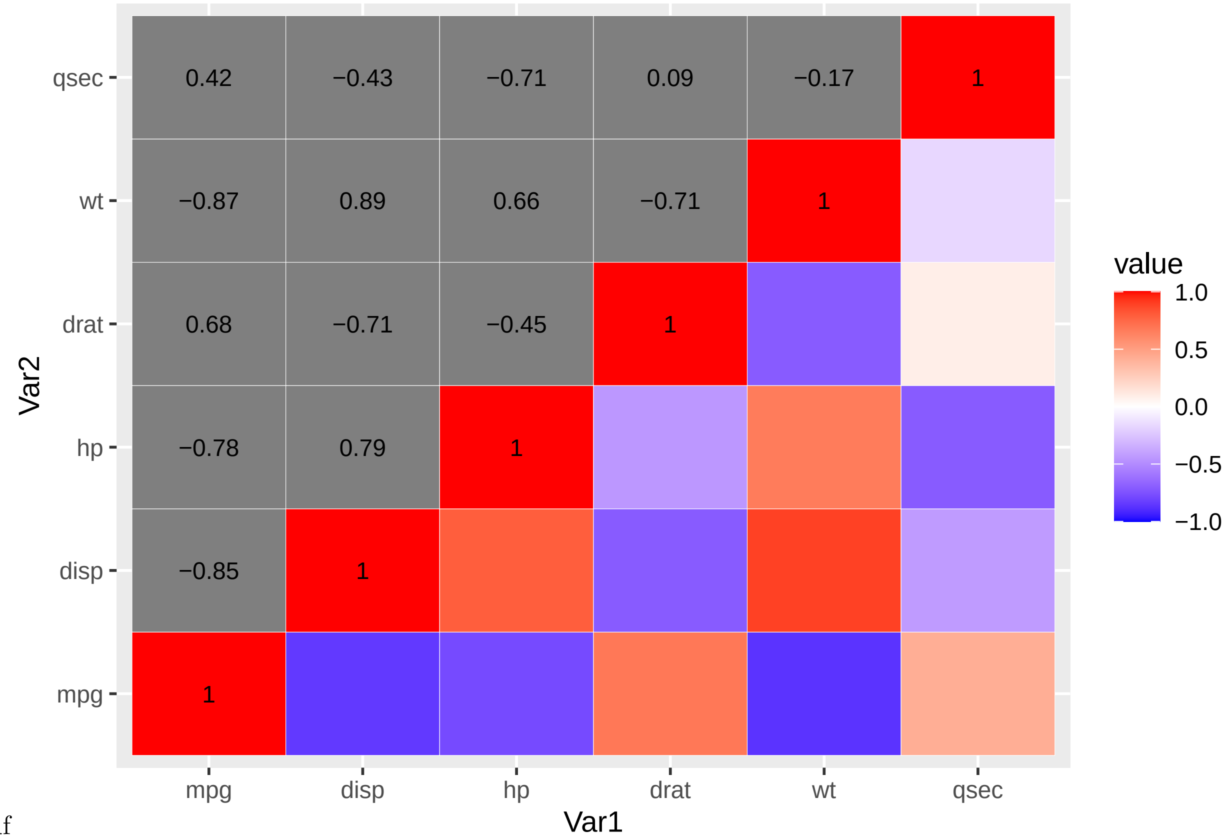
```
get_lower_tri<-function(cormat){
  cormat[upper.tri(cormat)] <- NA
  return(cormat)
}
lower_tri <- get_lower_tri(cormat)
melted_cormat <- melt(lower_tri)
head(melted_cormat)
```

```
##   Var1 Var2 value
## 1  mpg  mpg  1.00
## 2 disp  mpg -0.85
## 3   hp  mpg -0.78
## 4 drat  mpg  0.68
## 5   wt  mpg -0.87
## 6 qsec  mpg  0.42
```

```
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
```

```
                     midpoint = 0, limit = c(-1,1), space = "Lab")+
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 3)
```

```
## Warning: Removed 15 rows containing missing values (geom_text).
```

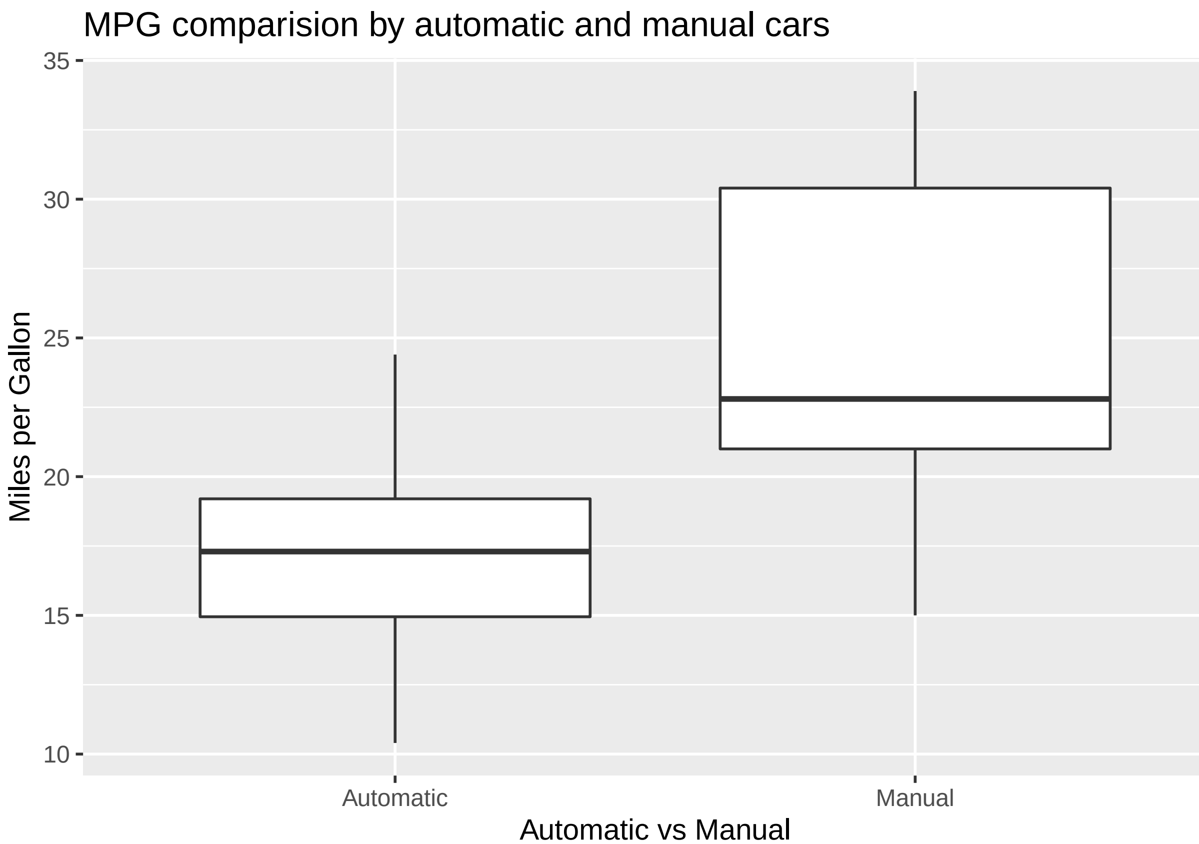


matrix plot-1.pdf

To determine relationship between variables,perform correlation matrix and plot pair graphs. As we can see from the correlation matrix,weight, rear axle ratio, Gross horsepower,and displacement have highly correlated with dependent variable ,Miles per gallon(wt(-0.87), drat(0.68),hp(-0.78),disp(-0.85)). We have to bear in mind those variables are higly correlated so need to adjust to avoid false conclusions.

Question 1. Is an automatic or manual transmission better for MPG?

```
mtcars$amlabel<-factor(mtcars$am, labels=c("Automatic","Manual"))
ggplot(data=mtcars,aes(x=amlabel,y=mpg))+
  geom_boxplot()+
  xlab('Automatic vs Manual')+
  ylab('Miles per Gallon')+
  ggtitle('MPG comparision by automatic and manual cars')
```



or manual plot-1.pdf

```
aggregate(mtcars$mpg,by=list(mtcars$amlabel), FUN = "mean")
```

```
##     Group.1        x
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

The above result shown we can conclude there is difference of mean value between automatic and manual transmission for MPG. Automatic has far lower mean than manual transmission and IQR of automatic is narrower than manual.

Question 2. Quantify the MPG difference between automatic and manual transmissions

Simple linear regression

```
fit1<-lm(mpg ~ amlabel,data=mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ amlabel, 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 ***
## amlabelManual   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
```

The linear regression model expected 7.245 increase in manual transmission for mpg compared to automatic one. Automatic transmission 17.147 is expected. Based on p-value, we can accept those estimations are not 0. R-squared of this model is 0.3598 which means it explains only 36% of variance.

Variance Inflation Factors

```
vif(lm(mpg~amlabel+cyl+disp+hp+drat+wt+qsec+vs+gear+carb,data = mtcars))
```

```
##               GVIF Df GVIF^(1/(2*Df))
## amlabel   9.930495  1        3.151269
## 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
## gear     50.852311  2        2.670408
## carb    503.211851  5        1.862838
```

This measure how much variance inflation among the variables causes the regressors. When VIF is higher than 10 , it is considered to that variable is highly correlated with other independent variables. However, in this data set, VIF is less than 10. This data does not need to fix multicollineaity.

Multivariate regression with ANOVA

```
fit2<-update(fit1,mpg~amlabel+cyl,data = mtcars)
fit3<-update(fit1,mpg~amlabel+cyl+disp)
fit4<-update(fit1,mpg~amlabel+cyl+disp+hp)
fit5<-update(fit1,mpg~amlabel+cyl+disp+hp+drat)
fit6<-update(fit1,mpg~amlabel+cyl+disp+hp+drat+wt)
fit7<-update(fit1,mpg~amlabel+cyl+disp+hp+drat+wt+qsec)
fit8<-update(fit1,mpg~amlabel+cyl+disp+hp+drat+wt+qsec+vs)
fit9<-update(fit1,mpg~amlabel+cyl+disp+hp+drat+wt+qsec+vs+gear)
fit10<-update(fit1,mpg~amlabel+cyl+disp+hp+drat+wt+qsec+vs+gear+carb)
anova(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8,fit9,fit10)
```

```
## Analysis of Variance Table
##
## Model  1: mpg ~ amlabel
## Model  2: mpg ~ amlabel + cyl
## Model  3: mpg ~ amlabel + cyl + disp
```

```
## Model  4: mpg ~ amlabel + cyl + disp + hp
## Model  5: mpg ~ amlabel + cyl + disp + hp + drat
## Model  6: mpg ~ amlabel + cyl + disp + hp + drat + wt
## Model  7: mpg ~ amlabel + cyl + disp + hp + drat + wt + qsec
## Model  8: mpg ~ amlabel + cyl + disp + hp + drat + wt + qsec + vs
## Model  9: mpg ~ amlabel + cyl + disp + hp + drat + wt + qsec + vs + gear
## Model 10: mpg ~ amlabel + cyl + disp + hp + drat + wt + qsec + vs + gear +
##     carb
##    Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1      30 720.90
## 2      28 264.50  2    456.40 28.4297 7.89e-06 ***
## 3      27 230.46  1     34.04  4.2402  0.05728 .
## 4      26 183.04  1     47.42  5.9078  0.02809 *
## 5      25 182.38  1      0.66  0.0820  0.77855
## 6      24 150.10  1     32.28  4.0216  0.06331 .
## 7      23 141.21  1      8.89  1.1081  0.30916
## 8      22 139.02  1      2.18  0.2719  0.60964
## 9      20 134.00  2      5.02  0.3128  0.73606
## 10     15 120.40  5     13.60  0.3388  0.88144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fit2 model contains 1 variable (cyl) than fit1 model. P-value shows fit2 model is necessary over fit1 However, other models except fit 4 does not give reasons adding more variables are necessary over the previous model.

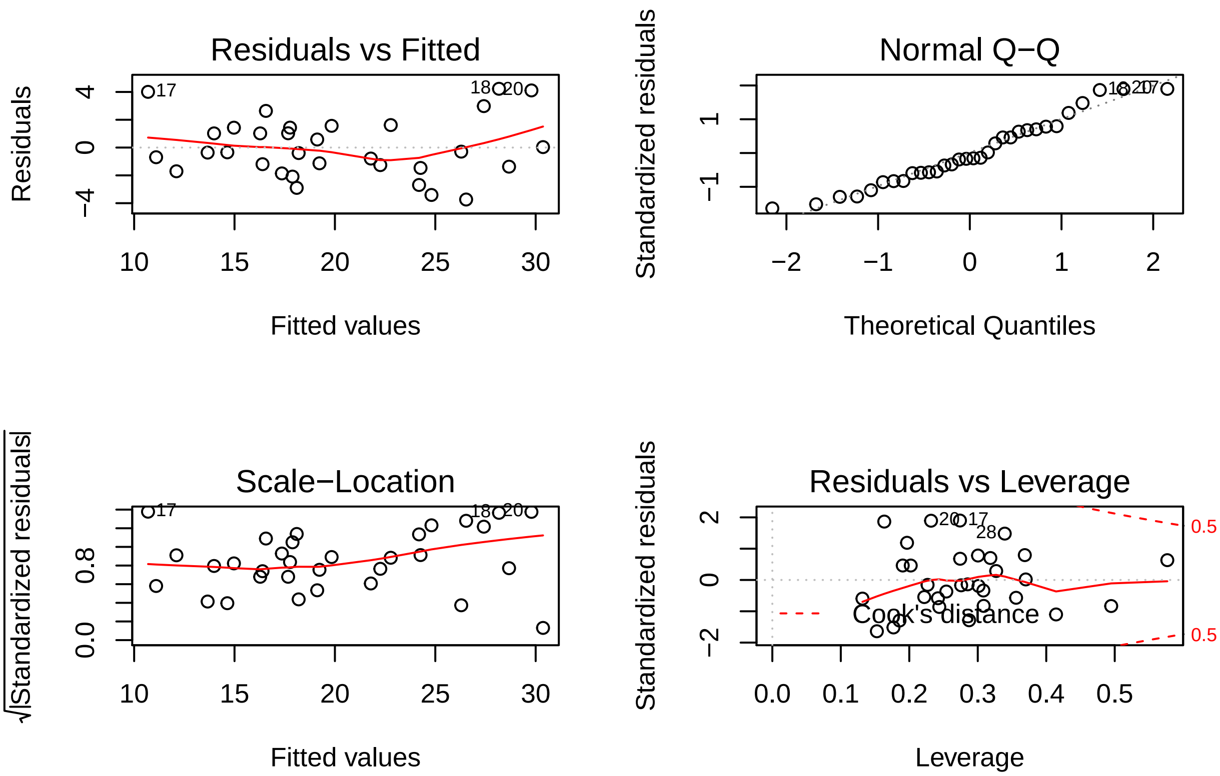
```
model<-c('fit1','fit2','fit3','fit4','fit5','fit6','fit7','fit8','fit9','fit10')

adjr=list(summary(fit1)$adj.r.squared,
        summary(fit2)$adj.r.squared,
        summary(fit3)$adj.r.squared,
        summary(fit4)$adj.r.squared,
        summary(fit5)$adj.r.squared,
        summary(fit6)$adj.r.squared,
        summary(fit7)$adj.r.squared,
        summary(fit8)$adj.r.squared,
        summary(fit9)$adj.r.squared,
        summary(fit10)$adj.r.squared)

comp=data.frame(unlist(model),unlist(adjr))
colnames(comp)<-c('Model','adjusted R')
comp<-comp[rev(order(comp$'adjusted R')),]
comp
```

```
##    Model adjusted R
## 7   fit7  0.8309831
## 6   fit6  0.8278226
## 8   fit8  0.8260321
## 9   fit9  0.8155474
## 4   fit4  0.8061901
## 5   fit5  0.7991624
## 10 fit10  0.7790215
## 3   fit3  0.7650169
## 2   fit2  0.7399447
## 1   fit1  0.3384589
```

If we simply compared the models based on adjusted R squares, model fit7 gives us the highest adjusted R squared and better fit models than others.

Normality & Residual plot

```
par(mfrow = c(2,2))
plot(fit7)
```



The diagnostic plots shows fit7 model have randomly distributed residuals and lies on normality QQ plot