R project

2024-08-03

```
df1<- read.csv("E:/3RD 1ST SEM/R project/online_course_engagement_data.csv")
head(df1)
```

```
##   TimeSpentOnCourse NumberOfVideosWatched NumberOfQuizzesTaken QuizScores
## 1          29.97972                    17                    3   50.36566
## 2          27.80264                     1                    5   62.61597
## 3          86.82048                    14                    2   78.45896
## 4          35.03843                    17                   10   59.19885
## 5          92.49065                    16                    0   98.42829
## 6          79.46613                    12                    7   70.23333
##   CompletionRate DeviceType M1 M2 M3 M4
## 1       20.86077          1  1  0  0  0
## 2       65.63242          1  0  1  0  0
## 3       63.81201          1  0  1  0  0
## 4       95.43316          0  0  0  1  0
## 5       18.10248          0  0  0  0  1
## 6       76.48402          0  1  0  0  0
```

```
#fitting linear regression model
model1<- lm(CompletionRate~.,data=df1)
summary(model1)
```

```
## 
## Call:
## lm(formula = CompletionRate ~ ., data = df1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -52.387 -24.760   0.038  25.163  51.518 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           50.64447    1.97877  25.594   <2e-16 ***
## TimeSpentOnCourse      0.02083    0.01071   1.944   0.0519 .  
## NumberOfVideosWatched  0.07618    0.05066   1.504   0.1327    
## NumberOfQuizzesTaken   0.06678    0.09670   0.691   0.4898    
## QuizScores            -0.02445    0.02124  -1.151   0.2496    
## DeviceType            -0.28887    0.61049  -0.473   0.6361    
## M1                     0.08593    0.95744   0.090   0.9285    
## M2                    -0.69233    0.97167  -0.713   0.4762    
## M3                    -1.24486    0.95839  -1.299   0.1940    
## M4                    -0.57760    0.95923  -0.602   0.5471    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 28.95 on 8990 degrees of freedom
## Multiple R-squared:  0.001196,   Adjusted R-squared:  0.0001965
## F-statistic: 1.197 on 9 and 8990 DF,  p-value: 0.2921
```

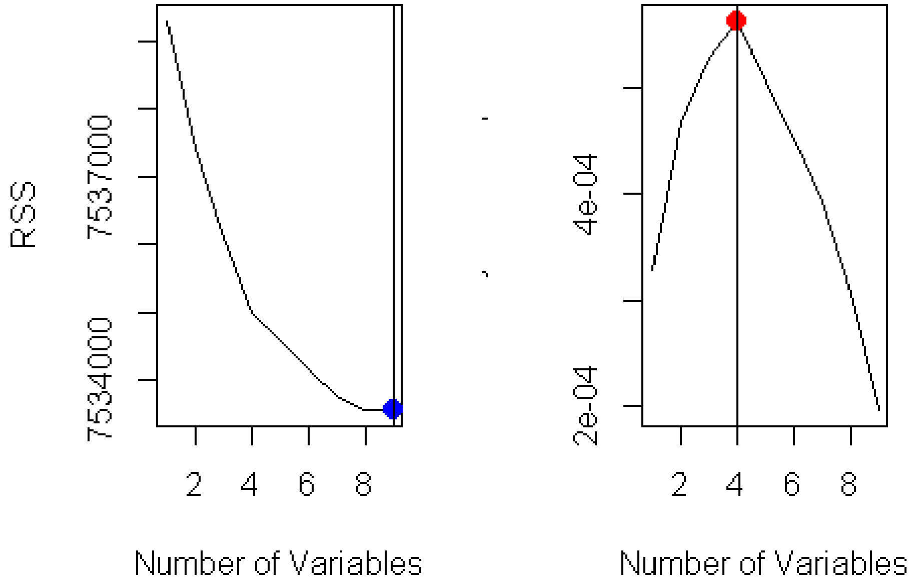
```
#perform best subset selection fitting 10 variables model using nv max argume
nt
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.3.3
```

```
model2<- regsubsets(CompletionRate~.,df1,nvmax = 9)
model2_summary<-summary(model2)
model2_summary
```

```
## Subset selection object
## Call: regsubsets.formula(CompletionRate ~ ., df1, nvmax = 9)
## 9 Variables  (and intercept)
##                         Forced in Forced out
## TimeSpentOnCourse           FALSE      FALSE
## NumberOfVideosWatched       FALSE      FALSE
## NumberOfQuizzesTaken        FALSE      FALSE
## QuizScores                  FALSE      FALSE
## DeviceType                  FALSE      FALSE
## M1                          FALSE      FALSE
## M2                          FALSE      FALSE
## M3                          FALSE      FALSE
## M4                          FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          TimeSpentOnCourse NumberOfVideosWatched NumberOfQuizzesTaken
## 1  ( 1 ) "*"               " "                   " "                 
## 2  ( 1 ) "*"               "*"                   " "                 
## 3  ( 1 ) "*"               "*"                   " "                 
## 4  ( 1 ) "*"               "*"                   " "                 
## 5  ( 1 ) "*"               "*"                   "*"                 
## 6  ( 1 ) "*"               "*"                   " "                 
## 7  ( 1 ) "*"               "*"                   "*"                 
## 8  ( 1 ) "*"               "*"                   "*"                 
## 9  ( 1 ) "*"               "*"                   "*"                 
##          QuizScores DeviceType M1  M2  M3  M4 
## 1  ( 1 ) " "        " "        " " " " " " " "
## 2  ( 1 ) " "        " "        " " " " " " " "
## 3  ( 1 ) " "        " "        " " " " "*" " "
## 4  ( 1 ) "*"        " "        " " " " "*" " "
## 5  ( 1 ) "*"        " "        " " " " "*" " "
## 6  ( 1 ) "*"        " "        " " "*" "*" "*"
## 7  ( 1 ) "*"        " "        " " "*" "*" "*"
## 8  ( 1 ) "*"        "*"        " " "*" "*" "*"
## 9  ( 1 ) "*"        "*"        "*" "*" "*" "*"
```

```
#Plotting the RSS and adjusted R2 and add a point where R2 is at its maximum
using the
#which.max() function

par(mfrow=c(1,2))
plot(model2_summary$rss, xlab = "Number of Variables", ylab = "RSS", type = "
l")
RSs.min<-which.min(model2_summary$rss)
points(RSs.min,model2_summary$rss[RSs.min],col="blue",cex = 2, pch = 20)
abline(v=RSs.min)

plot(model2_summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq
", type = "l")
adjr2.max <- which.max(model2_summary$adjr2)
points(adjr2.max, model2_summary$adjr2[adjr2.max], col = "red", cex = 2, pch
= 20)
abline(v=adjr2.max)
```



```
#Plotting the the (C_p) statistic and BIC and identify the minimum points
par(mfrow = c(1, 2))
plot(model2_summary$cp, xlab = "Number of Variables", ylab = "Cp", type = "l"
)
cp.min <- which.min(model2_summary$cp)
points(cp.min, model2_summary$cp[cp.min], col = "red", cex = 2, pch = 20)
bic.min <- which.min(model2_summary$bic)
plot(model2_summary$bic, xlab = "Number of Variables", ylab = "BIC", type = "
l")
```

```
points(bic.min, model2_summary$bic[bic.min], col = "red", cex = 2, pch = 20)


#subset selection by forward method

Model_fward <- regsubsets(CompletionRate~.,df1,nvmax = 9,method = "forward")
summary(Model_fward)
```

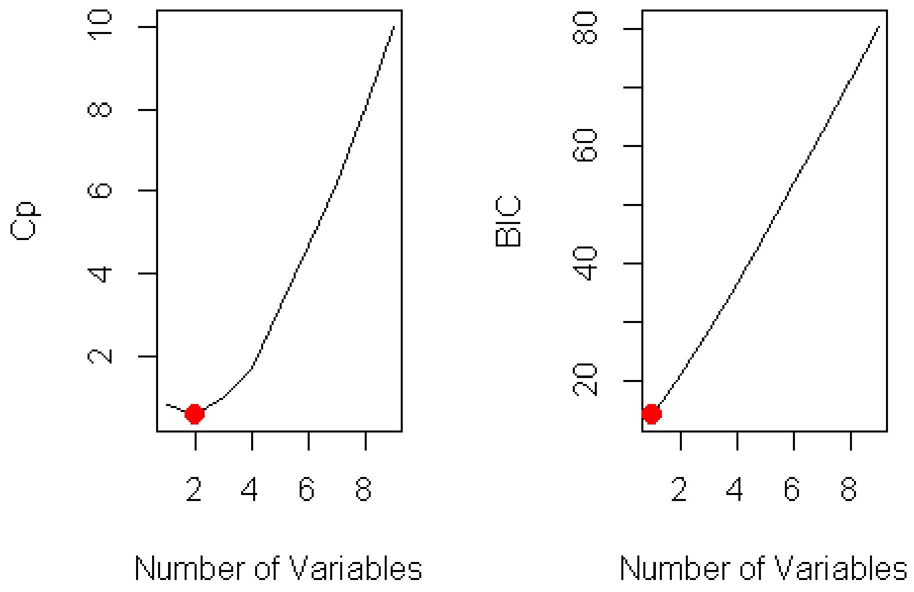
```
## Subset selection object
## Call: regsubsets.formula(CompletionRate ~ ., df1, nvmax = 9, method = "for
ward")
## 9 Variables  (and intercept)
##                       Forced in Forced out
## TimeSpentOnCourse         FALSE      FALSE
## NumberOfVideosWatched     FALSE      FALSE
## NumberOfQuizzesTaken      FALSE      FALSE
## QuizScores                FALSE      FALSE
## DeviceType                FALSE      FALSE
## M1                        FALSE      FALSE
## M2                        FALSE      FALSE
## M3                        FALSE      FALSE
## M4                        FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: forward
##          TimeSpentOnCourse NumberOfVideosWatched NumberOfQuizzesTaken
## 1  ( 1 ) "*"               " "                   " "                 
## 2  ( 1 ) "*"               "*"                   " "                 
## 3  ( 1 ) "*"               "*"                   " "                 
## 4  ( 1 ) "*"               "*"                   " "                 
## 5  ( 1 ) "*"               "*"                   "*"                 
## 6  ( 1 ) "*"               "*"                   "*"                 
## 7  ( 1 ) "*"               "*"                   "*"                 
## 8  ( 1 ) "*"               "*"                   "*"                 
## 9  ( 1 ) "*"               "*"                   "*"                 
##          QuizScores DeviceType M1  M2  M3  M4 
## 1  ( 1 ) " "        " "        " " " " " " " "
## 2  ( 1 ) " "        " "        " " " " " " " "
## 3  ( 1 ) " "        " "        " " " " "*" " "
## 4  ( 1 ) "*"        " "        " " " " "*" " "
## 5  ( 1 ) "*"        " "        " " " " "*" " "
## 6  ( 1 ) "*"        " "        " " "*" "*" " "
## 7  ( 1 ) "*"        " "        " " "*" "*" "*"
## 8  ( 1 ) "*"        "*"        " " "*" "*" "*"
## 9  ( 1 ) "*"        "*"        "*" "*" "*" "*"
```

```
#Create a data frame including all the crieterion values for all the models

res.sum <- summary(Model_fward)
criterion<-data.frame(
model=1:9,
Adj.R2 = (res.sum$adjr2),
CP = (res.sum$cp),
BIC = (res.sum$bic),
RSS=res.sum$rss
)
head(criterion)
```

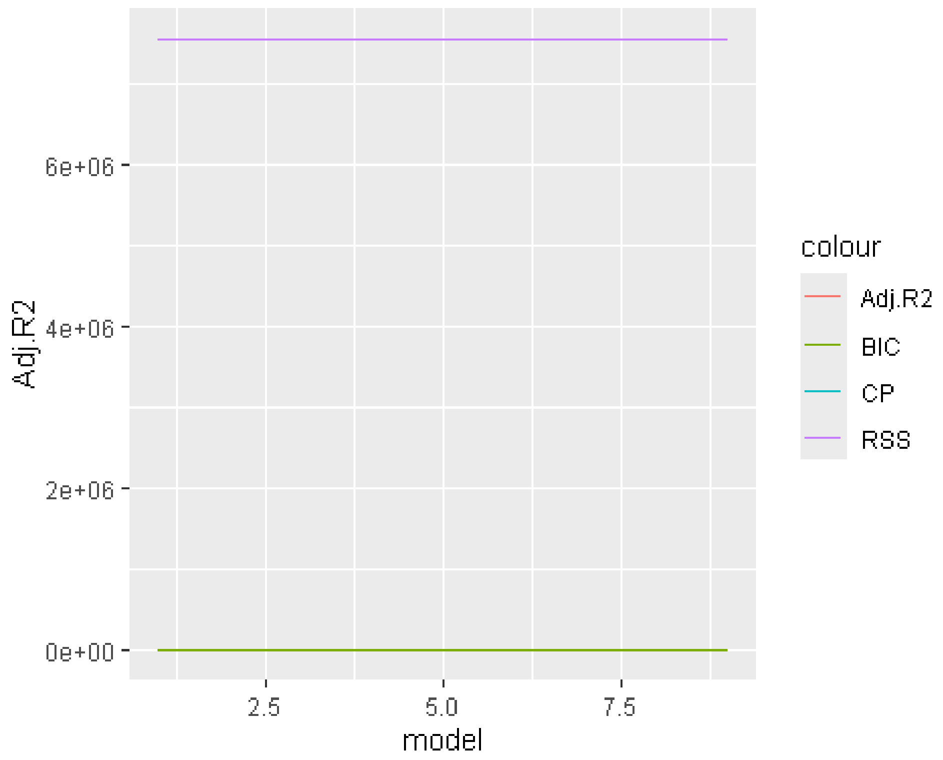
```
##   model       Adj.R2        CP      BIC     RSS
## 1     1 0.0003271584 0.8240873 14.26489 7539288
## 2     2 0.0004671522 0.5644462 21.10914 7537394
## 3     3 0.0005275739 1.0210571 28.66967 7536101
## 4     4 0.0005628189 1.7042965 36.45677 7534998
## 5     5 0.0005052920 3.2221618 45.07916 7534594
## 6     6 0.0004424050 4.7881251 53.74967 7534230
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

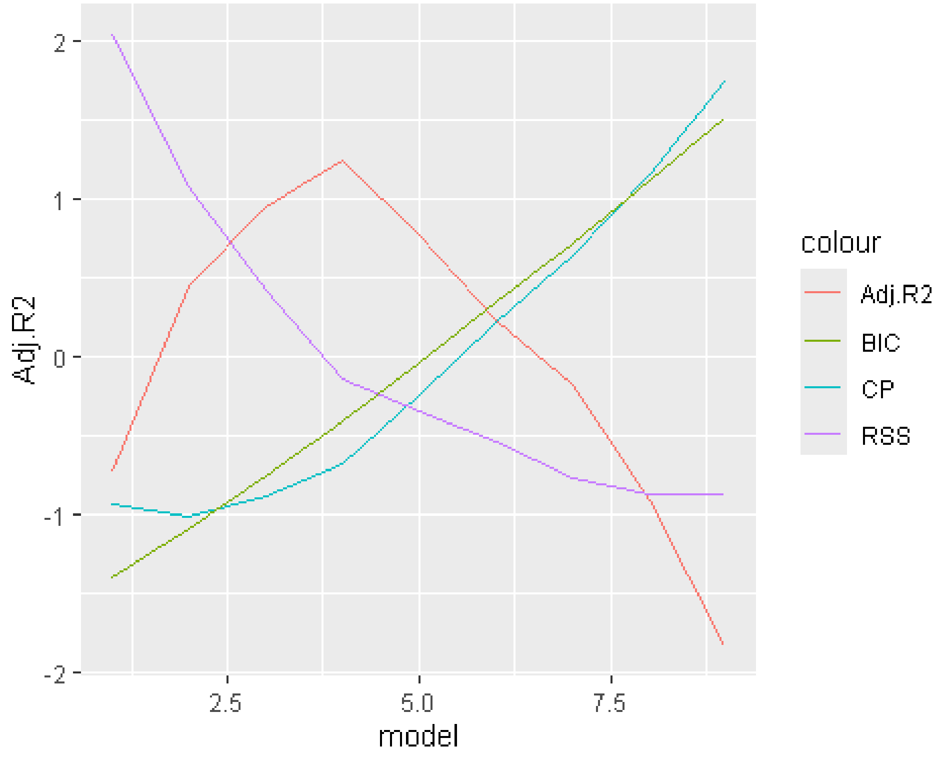


```
ggplot(criterion, aes(model)) +
 geom_line(aes(y = Adj.R2, colour = "Adj.R2")) +
 geom_line(aes(y = CP, colour = "CP"))+
 geom_line(aes(y = BIC, colour = "BIC"))+
 geom_line(aes(y = RSS, colour = "RSS"))
```

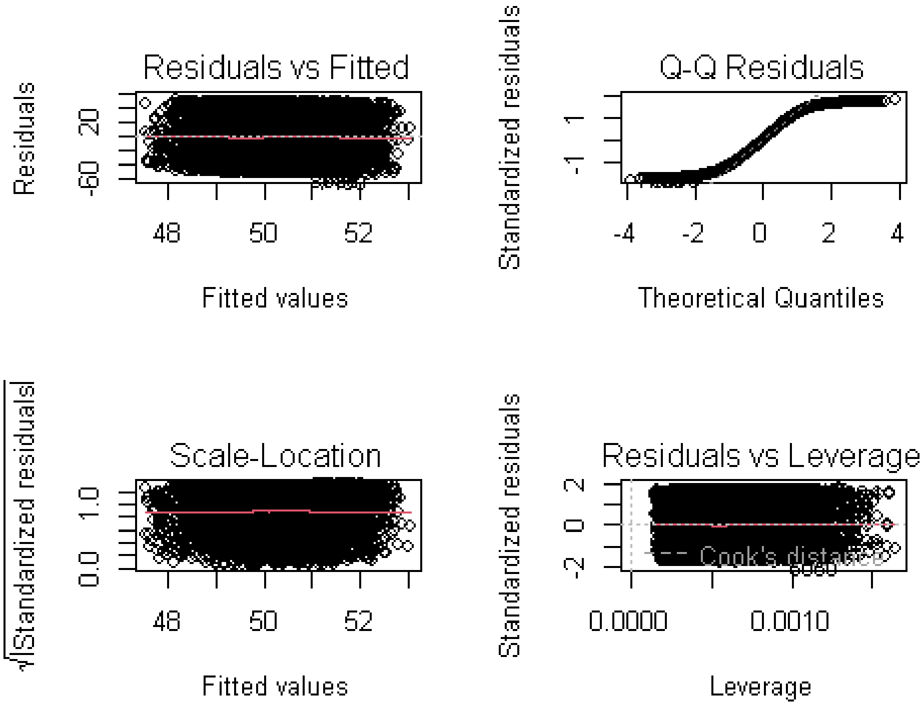


```
#standarizing
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.3.3

##
## 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

criterion_std<-cbind(model=criterion$model, scale(criterion[,-1]))
criterion_std<-as.data.frame(criterion_std)
head(criterion_std)

##   model     Adj.R2         CP        BIC        RSS
## 1     1 -0.7309015 -0.9368492 -1.3883198  2.0366679
## 2     2  0.4429570 -1.0124795 -1.0879360  1.0746353
## 3     3  0.9495970 -0.8794744 -0.7561162  0.4175440
## 4     4  1.2451295 -0.6804552 -0.4143524 -0.1430612
## 5     5  0.7627619 -0.2383198 -0.0359292 -0.3483280
## 6     6  0.2354492  0.2178259  0.3446060 -0.5331173
```

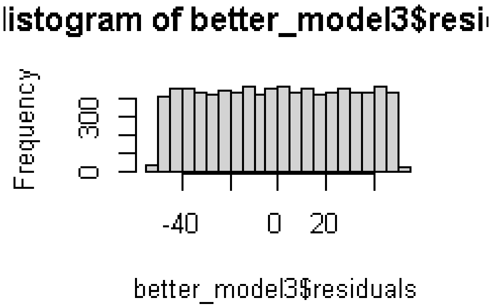
```
#after standarizing
ggplot(criterion_std, aes(model)) +
 geom_line(aes(y = Adj.R2, colour = "Adj.R2")) +
 geom_line(aes(y = CP, colour = "CP"))+
 geom_line(aes(y = BIC, colour = "BIC"))+
 geom_line(aes(y = RSS, colour = "RSS"))
```

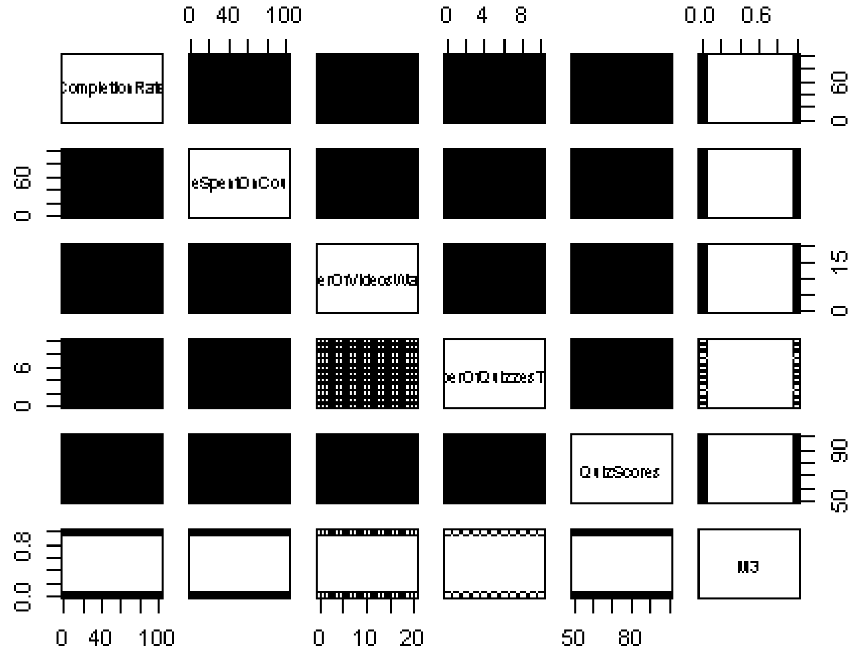


```
#5 model is better one
#getting coefficients of 5th
coef(Model_fward,5)

##          (Intercept)     TimeSpentOnCourse NumberOfVideosWatched 
##          50.19920521            0.02086613            0.07495978 
##  NumberOfQuizzesTaken            QuizScores                    M3 
##           0.06712478           -0.02417303           -0.95786033

better_model3 <- lm(CompletionRate~TimeSpentOnCourse+NumberOfVideosWatched+Nu
mberOfQuizzesTaken+QuizScores+M3,data=df1 )
summary(better_model3)

## 
## Call:
## lm(formula = CompletionRate ~ TimeSpentOnCourse + NumberOfVideosWatched + 
##     NumberOfQuizzesTaken + QuizScores + M3, data = df1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.944 -24.850   0.106  25.183  51.322
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            50.19921    1.86957  26.851   <2e-16 ***
## TimeSpentOnCourse       0.02087    0.01071   1.948   0.0515 .
## NumberOfVideosWatched   0.07496    0.05063   1.480   0.1388
## NumberOfQuizzesTaken    0.06712    0.09666   0.694   0.4874
## QuizScores             -0.02417    0.02123  -1.139   0.2548
## M3                     -0.95786    0.76069  -1.259   0.2080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.94 on 8994 degrees of freedom
## Multiple R-squared:  0.001061,   Adjusted R-squared:  0.0005053
## F-statistic:  1.91 on 5 and 8994 DF,  p-value: 0.08917

par(mfrow=c(2,2))
plot(better_model3)
```



```
hist(better_model3$residuals)
#checking Multicolinearity


df_subset <- subset(df1, select = c("CompletionRate", "TimeSpentOnCourse","Nu
mberOfVideosWatched","NumberOfQuizzesTaken","QuizScores",
"M3"))
head(df_subset)

##   CompletionRate TimeSpentOnCourse NumberOfVideosWatched NumberOfQuizzesTa
ken
## 1       20.86077          29.97972                    17
3
## 2       65.63242          27.80264                     1
5
## 3       63.81201          86.82048                    14
2
## 4       95.43316          35.03843                    17
10
## 5       18.10248          92.49065                    16
0
## 6       76.48402          79.46613                    12
7
##   QuizScores M3
## 1   50.36566  0
## 2   62.61597  0
## 3   78.45896  0
## 4   59.19885  1
## 5   98.42829  0
## 6   70.23333  0

pairs(df_subset)
```





```
#install.packages('car')
library('car')

## Warning: package 'car' was built under R version 4.3.3
## Loading required package: carData
```

```
## 
## Attaching package: 'car'
```

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

```
vif(better_model3)
```

```
##     TimeSpentOnCourse NumberOfVideosWatched  NumberOfQuizzesTaken 
##              1.000640              1.001271              1.000713 
##            QuizScores                    M3 
##              1.000719              1.000450
```