Analysis

Yuhan Hu

2020/2/29

Contents

Research Question **1**

data management **1**

For research question 1 2

For research question 2 9

Model **11**

Research Question 1 **11**

LASSO regression 11

Backward/forward Selection 16

mixed model 22

Research Question 2 **28**

Interpretation **76**

Research Question 1 76

Research Question 2 77

Research Question

1. What characteristics of the users is associated with a higher rate of improvement for the health risk factors.
2. Is a higher frequency of usage of the platform (measured through the accumulation of points) associated with a general improvement in health as measured by the risk factors.

Higher value of scores means better.

data management

After EDA and further group discussion, I have refined the process of data cleaning

For research question 1

```
#load data
BLR_user_HRA <- read.csv('BLR_USER_HRA.csv',na.strings = c('NA','NULL'))
colnames(BLR_user_HRA)[1]<-"UserId"
BLR_user_HRA[,1] <- as.character(unlist(BLR_user_HRA[,1]))
BLR_user_HRA$FinishedDate <- as.Date(BLR_user_HRA$FinishedDate)
BLR_user_HRA$CreatedDate <- as.Date(BLR_user_HRA$CreatedDate)
length(unique(BLR_user_HRA$UserId))
```

```
## [1] 137754
```

```
################################################################################

#remove rows with N/A BLR column
rmNA<-c()
for(i in 1:nrow(BLR_user_HRA)){
  if (is.na(BLR_user_HRA[i,][4])){
    rmNA <- c(rmNA,i)
  }
}

BLR_user_HRA <- BLR_user_HRA[-rmNA,]

rm(rmNA)
rm(i)

#remove long interval
#I can't understand why some entries took the users like more than 1 day to complete, so I removed all
#
rmLongInterval1 <- c()
for(i in 1:nrow(BLR_user_HRA)){
  if (as.numeric(difftime(BLR_user_HRA$FinishedDate[i], BLR_user_HRA$CreatedDate[i],units='auto'))>1){
    rmLongInterval1 <- c(rmLongInterval1,i)
  }
}
length(rmLongInterval1)
```

```
## [1] 15370
```

```
rmLongInterval7 <- c()
for(i in 1:nrow(BLR_user_HRA)){
  if (as.numeric(difftime(BLR_user_HRA$FinishedDate[i], BLR_user_HRA$CreatedDate[i],units='auto'))>7){
    rmLongInterval7 <- c(rmLongInterval7,i)
  }
}
length(rmLongInterval7)
```

```
## [1] 13136
```

```
rmLongInterval14 <- c()
for(i in 1:nrow(BLR_user_HRA)){
  if (as.numeric(difftime(BLR_user_HRA$FinishedDate[i], BLR_user_HRA$CreatedDate[i],units='auto'))>14){
    rmLongInterval14 <- c(rmLongInterval14,i)
  }
}
length(rmLongInterval14)
```

```
## [1] 11985
```

```
#There is no big difference between

#some user with duplicated rows may become user with only one entry after remove all invalid rows
#so pick duplicated rows after remove all those invalid date point
BLR_user_HRA_rmLongInterval <- BLR_user_HRA[-rmLongInterval1,]
BLR_user_HRA_rmLongInterval_duplicated <- subset(BLR_user_HRA_rmLongInterval,
                                                 duplicated(UserId)|duplicated(UserId,fromLast = TRUE))

BLR_user_HRA_rmLongInterval_duplicated <- BLR_user_HRA_rmLongInterval_duplicated[,-c(5,14,17)]

rm(rmLongInterval1)
rm(rmLongInterval7)
rm(rmLongInterval14)
rm(i)
rm(BLR_user_HRA_rmLongInterval)
```

For the first research quesiton, if I want to build a linear model, I need to calculate difference between the first and last entry for each user.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
##
## 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
```

```
last <- BLR_user_HRA_rmLongInterval_duplicated %>%
  group_by(UserId) %>%
  filter(FinishedDate==max(FinishedDate))

first <- BLR_user_HRA_rmLongInterval_duplicated %>%
  group_by(UserId) %>%
  filter(FinishedDate==min(FinishedDate))

length(unique(BLR_user_HRA_rmLongInterval_duplicated$UserId))
```

```
## [1] 26186
```

```
nrow(last)
```

```
## [1] 27899
```

```
#there must be something wrong, there are users put in duplicated entry in the same data
#for those entries, we have no information about which entry was added in later
#so I decide to randomly pick one entry

last <- last %>% group_by(UserId) %>% sample_n(1)
first <- first %>% group_by(UserId) %>% sample_n(1)

sorted_last <- arrange(last,by_group=UserId)
sorted_first <- arrange(first,by_group=UserId)
rm(first)
rm(last)


diff <- sorted_last
diff$time_interval <- as.numeric(difftime(sorted_last$FinishedDate,sorted_first$FinishedDate,units = 'da
#Found there are users that put in data points with same finished date, i.e. time_interval = 0
#which means these data points are invalid and cant be used for analysis
#no need to go back to the begining of data management
#if a user with more than 2 data points return this problem
#it means other data points of this user also have same finished, so all data points from this user are
non_zero_interval <- diff$time_interval!=0
sorted_first <- sorted_first[non_zero_interval,]
sorted_last <- sorted_last[non_zero_interval,]
diff <- diff[non_zero_interval,]
all(diff$time_interval!=0)
```

```
## [1] TRUE
```

```
rm(non_zero_interval)

#all invalid data points removed
#now column finished date and created date are useless
diff <- diff[,-c(13,14)]
sorted_first <- sorted_first[,-c(13,14)]
sorted_last <- sorted_last[,-c(13,14)]

#check if entries in both df are in the same order, if not, next step will be erroneous
all(sorted_first$UserId == sorted_last$UserId)
```

```
## [1] TRUE
```

```
features <- colnames(sorted_first)
```

```
i=4
for (i in length(features)){
  diff[paste(features[i],'change',sep='_')] <- sorted_last[[features[i]]] - sorted_first[[features[i]]]
}
```

Don't understand why the above loop does not work, it did the same thing as code in the following chunk

```
diff[paste(features[4],'change',sep='_')] <- (sorted_last[[features[4]]] - sorted_first[[features[4]]])/
diff[paste(features[5],'change',sep='_')] <- (sorted_last[[features[5]]] - sorted_first[[features[5]]])/
diff[paste(features[6],'change',sep='_')] <- (sorted_last[[features[6]]] - sorted_first[[features[6]]])/
diff[paste(features[7],'change',sep='_')] <- (sorted_last[[features[7]]] - sorted_first[[features[7]]])/
diff[paste(features[8],'change',sep='_')] <- (sorted_last[[features[8]]] - sorted_first[[features[8]]])/
diff[paste(features[9],'change',sep='_')] <- (sorted_last[[features[9]]] - sorted_first[[features[9]]])/
diff[paste(features[10],'change',sep='_')] <- (sorted_last[[features[10]]] - sorted_first[[features[10]]
diff[paste(features[11],'change',sep='_')] <- (sorted_last[[features[11]]] - sorted_first[[features[11]]
diff[paste(features[12],'change',sep='_')] <- (sorted_last[[features[12]]] - sorted_first[[features[12]]
diff[paste(features[13],'change',sep='_')] <- (sorted_last[[features[13]]] - sorted_first[[features[13]]
diff[paste(features[14],'change',sep='_')] <- (sorted_last[[features[14]]] - sorted_first[[features[14]]
diff[paste(features[15],'change',sep='_')] <- (sorted_last[[features[15]]] - sorted_first[[features[15]]
diff[paste(features[16],'change',sep='_')] <- (sorted_last[[features[16]]] - sorted_first[[features[16]]
diff[paste(features[17],'change',sep='_')] <- (sorted_last[[features[17]]] - sorted_first[[features[17]]
diff[paste(features[18],'change',sep='_')] <- (sorted_last[[features[18]]] - sorted_first[[features[18]]
diff[paste(features[19],'change',sep='_')] <- (sorted_last[[features[19]]] - sorted_first[[features[19]]
diff[paste(features[20],'change',sep='_')] <- (sorted_last[[features[20]]] - sorted_first[[features[20]]
diff[paste(features[21],'change',sep='_')] <- (sorted_last[[features[21]]] - sorted_first[[features[21]]
#now the original value can be discarded
diff <- diff[,-seq(4,length(sorted_first))]
rm(sorted_first)
rm(sorted_last)
```

Finish data preparation for regression model

But there may be random effect between users

```
mixed_preparation <- BLR_user_HRA_rmLongInterval_duplicated %>%
  group_by(UserId) %>%
  filter(min(FinishedDate)!=max(FinishedDate))


unique_table <- table(mixed_preparation$UserId)
length(unique_table)
```

```
## [1] 24866
```

```
sum(unique_table > 2)
```

```
## [1] 9755
```

```
sum(unique_table == 3)
```

```
## [1] 5321
```

```
sum(unique_table == 4)
```

```
## [1] 2217
```

```r
sum(unique_table == 5)
```

```
## [1] 1004
```

```r
sum(unique_table > 5)
```

```
## [1] 1213
```

```r
mixed_preparation <- mixed_preparation %>%
   group_by(UserId) %>%
   mutate(time_interval = as.numeric(difftime(FinishedDate, lag(FinishedDate)), units='days'))

mixed_preparation <- mixed_preparation %>%
   group_by(UserId) %>%
   mutate(BLR_change = (BLRScore - lag(BLRScore, default = BLRScore[1]))/time_interval)

mixed_preparation <- mixed_preparation %>%
   group_by(UserId) %>%
   mutate(Heart_change = (HeartScore - lag(HeartScore, default = HeartScore[1]))/time_interval)

mixed_preparation <- mixed_preparation %>%
   group_by(UserId) %>%
   mutate(Respiratory_change = (RespiratoryScore - lag(RespiratoryScore, default = RespiratoryScore[1]))

mixed_preparation <- mixed_preparation %>%
   group_by(UserId) %>%
   mutate(Gastrointestinal_change = (GastrointestinalScore - lag(GastrointestinalScore,
                                                                 default = GastrointestinalSco

mixed_preparation <- mixed_preparation %>%
   group_by(UserId) %>%
   mutate(Diabetes_change = (DiabetesScore - lag(DiabetesScore, default = DiabetesScore[1]))/time_interv

mixed_preparation <- mixed_preparation %>%
   group_by(UserId) %>%
   mutate(Cancer_change = (CancerScore - lag(CancerScore, default = CancerScore[1]))/time_interval)


mixed_preparation <- mixed_preparation %>%
   group_by(UserId) %>%
   mutate(ArthritisPain_change = (ArthritisPainScore - lag(ArthritisPainScore, default = ArthritisPainS

mixed_preparation <- mixed_preparation %>%
   group_by(UserId) %>%
   mutate(MentalHealth_change = (MentalHealthScore - lag(MentalHealthScore, default = MentalHealthScore

mixed_preparation <- mixed_preparation %>%
   group_by(UserId) %>%
   mutate(SocioFinancial_change = (SocialFinancialRelationshipScore -
            lag(SocialFinancialRelationshipScore, default = SocialFinancialRelationshipScore[1]))/time_

mixed_preparation <- mixed_preparation %>%
```

```r
  group_by(UserId) %>%
  mutate(Diet_change = (DietScore - lag(DietScore, default = DietScore[1]))/time_interval)

mixed_preparation <- mixed_preparation %>%
  group_by(UserId) %>%
  mutate(PhysicalActivity_change = (PhysicalActivityScore - lag(PhysicalActivityScore,
                                                                default = PhysicalActivityScor

mixed_preparation <- mixed_preparation %>%
  group_by(UserId) %>%
  mutate(FinancialHealth_change = (FinancialHealthScore - lag(FinancialHealthScore,
                                                              default = FinancialHealthScore[1

mixed_preparation <- mixed_preparation %>%
  group_by(UserId) %>%
  mutate(Medication_change = (MedicationScore - lag(MedicationScore, default = MedicationScore[1]))/tim

mixed_preparation <- mixed_preparation %>%
  group_by(UserId) %>%
  mutate(Alcohol_change = (AlcoholScore - lag(AlcoholScore, default = AlcoholScore[1]))/time_interval)

mixed_preparation <- mixed_preparation %>%
  group_by(UserId) %>%
  mutate(Sleep_change = (SleepScore - lag(SleepScore, default = SleepScore[1]))/time_interval)

mixed_preparation <- mixed_preparation %>%
  group_by(UserId) %>%
  mutate(Stress_change = (StressScore - lag(StressScore, default = StressScore[1]))/time_interval)

mixed_preparation <- mixed_preparation %>%
  group_by(UserId) %>%
  mutate(Smoking_change = (SmokingScore - lag(SmokingScore, default = SmokingScore[1]))/time_interval)

mixed_preparation <- mixed_preparation %>%
  group_by(UserId) %>%
  mutate(BMI_change = (BMI - lag(BMI, default = BMI[1]))/time_interval)


mixed_preparation <- na.omit(mixed_preparation)
mixed_preparation <- subset(mixed_preparation,time_interval!=0)
unique_table <- table(mixed_preparation$UserId)
unique_table <- unique_table[unique_table >= 3]
unique_df <- as.data.frame(unique_table)
mixed_preparation <- mixed_preparation[mixed_preparation$UserId %in% unique_df$Var1,]
mixed_preparation <- mixed_preparation[,-c(4:23)]
```

```r
mixed_preparation2 <- BLR_user_HRA_rmLongInterval_duplicated %>%
  group_by(UserId) %>%
  filter(min(FinishedDate)!=max(FinishedDate))

mixed_preparation2 <- mixed_preparation2 %>%
  group_by(UserId) %>%
  mutate(BLR_change = BLRScore - lag(BLRScore, default = BLRScore[1]))
```

```r
mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Heart_change = HeartScore - lag(HeartScore, default = HeartScore[1]))

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Respiratory_change = RespiratoryScore - lag(RespiratoryScore, default = RespiratoryScore[1]))

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Gastrointestinal_change = GastrointestinalScore - lag(GastrointestinalScore,
                                                                default = GastrointestinalSco

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Diabetes_change = DiabetesScore - lag(DiabetesScore, default = DiabetesScore[1]))

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Cancer_change = CancerScore - lag(CancerScore, default = CancerScore[1]))


mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(ArthritisPain_change = ArthritisPainScore - lag(ArthritisPainScore, default = ArthritisPainSc

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(MentalHealth_change = MentalHealthScore - lag(MentalHealthScore, default = MentalHealthScore[

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(SocioFinancial_change = SocialFinancialRelationshipScore -
            lag(SocialFinancialRelationshipScore, default = SocialFinancialRelationshipScore[1]))

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Diet_change = DietScore - lag(DietScore, default = DietScore[1]))

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(PhysicalActivity_change = PhysicalActivityScore - lag(PhysicalActivityScore,
                                                                default = PhysicalActivitySco

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(FinancialHealth_change = FinancialHealthScore - lag(FinancialHealthScore,
                                                              default = FinancialHealthScore[

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Medication_change = MedicationScore - lag(MedicationScore, default = MedicationScore[1]))
```

```
mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Alcohol_change = AlcoholScore - lag(AlcoholScore, default = AlcoholScore[1]))

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Sleep_change = SleepScore - lag(SleepScore, default = SleepScore[1]))

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Stress_change = StressScore - lag(StressScore, default = StressScore[1]))

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(Smoking_change = SmokingScore - lag(SmokingScore, default = SmokingScore[1]))

mixed_preparation2 <- mixed_preparation2 %>%
   group_by(UserId) %>%
   mutate(BMI_change = BMI - lag(BMI, default = BMI[1]))

mixed_preparation2 <- na.omit(mixed_preparation2)
unique_table <- table(mixed_preparation2$UserId)
unique_table <- unique_table[unique_table >= 3]
unique_df <- as.data.frame(unique_table)
mixed_preparation2 <- mixed_preparation2[mixed_preparation2$UserId %in% unique_df$Var1,]
mixed_preparation2 <- mixed_preparation2[,-c(4:23)]
```

For research question 2

```
BLR_user_HRA <- read.csv('BLR_USER_HRA.csv',na.strings = c('NA','NULL'))
colnames(BLR_user_HRA)[1]<-"UserId"
BLR_user_HRA[,1] <- as.character(unlist(BLR_user_HRA[,1]))

p1_points_M <- read.csv("part1- points M.csv")
colnames(p1_points_M)[1]<-"UserId"

p1_points_R <- read.csv("part1- points R.csv")
colnames(p1_points_R)[1]<-"UserId"

p1_points_Z <- read.csv("part1- points Z copy.csv")
colnames(p1_points_Z)[1]<-"UserId"

p2_points_M <- read.csv("part2- points M copy.csv")
colnames(p2_points_M)[1]<-"UserId"

p2_points_R <- read.csv("part2- points R copy.csv")
colnames(p2_points_R)[1]<-"UserId"

p3_points_M <- read.csv("part3- points M.csv")
colnames(p3_points_M)[1]<-"UserId"
```

```
p3_points_R <- read.csv("part3- points R copy.csv")
colnames(p3_points_R)[1]<-"UserId"

colnames(p1_points_M) <- c('UserId','points','CreatedDate')
colnames(p2_points_M) <- c('UserId','points','CreatedDate')
colnames(p3_points_M) <- c('UserId','points','CreatedDate')
points_M <- rbind(rbind(p1_points_M,p2_points_M),p3_points_M)
rm(p1_points_M,p2_points_M,p3_points_M)

colnames(p1_points_R) <- c('UserId','points','CreatedDate')
colnames(p2_points_R) <- c('UserId','points','CreatedDate')
colnames(p3_points_R) <- c('UserId','points','CreatedDate')
points_R <- rbind(rbind(p1_points_R,p2_points_R),p3_points_R)
rm(p1_points_R,p2_points_R,p3_points_R)

colnames(p1_points_Z) <- c('UserId','points','CreatedDate')
points_Z <- p1_points_Z
rm(p1_points_Z)


points_all <- rbind(points_M,points_R,points_Z)
length(unique(points_all$UserId))
```

```
## [1] 218002
```

```
length(unique(BLR_user_HRA$UserId))
```

```
## [1] 137754
```

```
rm(points_M,points_R,points_Z)

#not all users of the platform tried BLR score
#remove those users does not try BLR system
new_points <- subset(points_all, UserId%in%unique(diff$UserId))
length(unique(new_points$UserId))
```

```
## [1] 20662
```

```
#remove those users used BLR but haven't tried points system
BLR_data_points <- subset(diff, UserId %in% unique(new_points$UserId))
length(unique(BLR_data_points$UserId))
```

```
## [1] 20662
```

```
points_arranged <- data.frame(unique(new_points$UserId))
colnames(points_arranged) <- c('UserId')
absolute_points_change <- new_points %>%
  group_by(UserId) %>%
  summarise(sum_points = sum(abs(points)))
```

```r
points_earned <- new_points %>%
  group_by(UserId) %>%
  summarise(points_earned = sum(points[points > 0]))

points_used <- new_points %>%
  group_by(UserId) %>%
  summarise(points_used = sum(points[points < 0]))

frequency <- new_points %>%
  group_by(UserId) %>%
  summarise(frequency = n())

points_arranged <- merge(points_arranged,points_earned, by = 'UserId')
points_arranged <- merge(points_arranged,points_used, by = 'UserId')
points_arranged <- merge(points_arranged,absolute_points_change, by = 'UserId')
points_arranged <- merge(points_arranged,frequency, by = 'UserId')

points_ready <- points_arranged

rm(points_earned,points_used,frequency,absolute_points_change,points_all,new_points,points_arranged)
```

Model

Research Question 1

LASSO regression

```r
#do not need user Id at this point
library(glmnet)
```

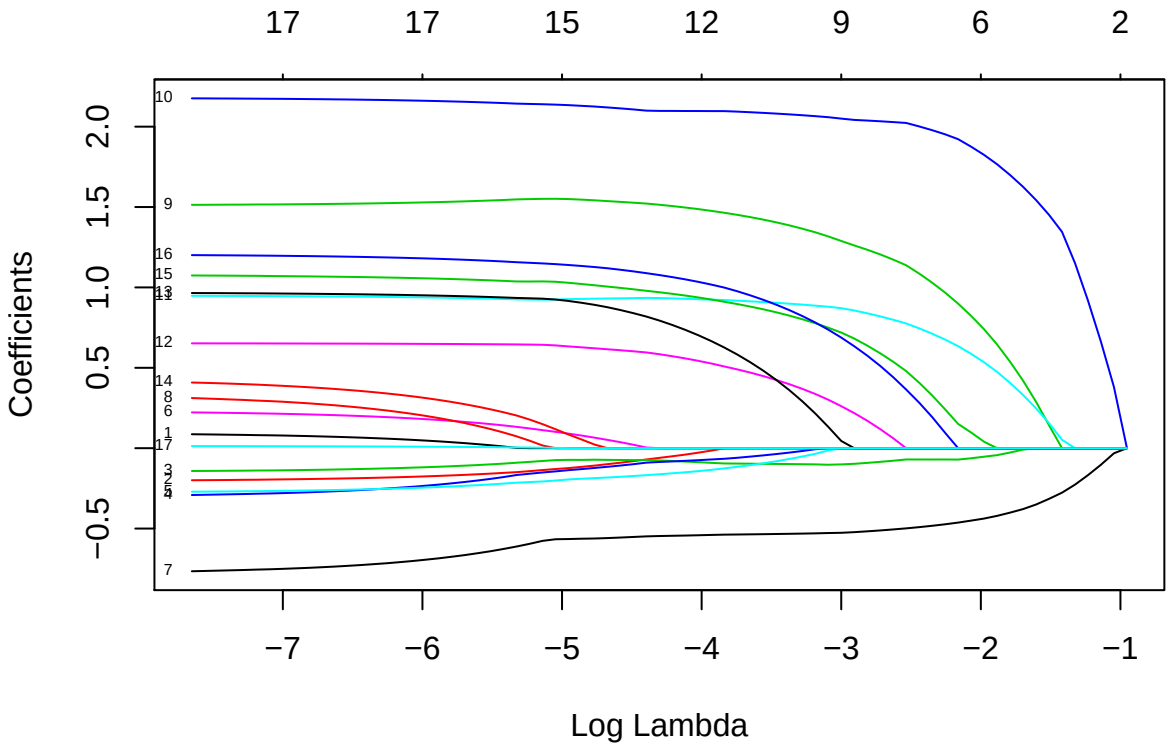
```
## Warning: package 'glmnet' was built under R version 3.6.2
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 3.6.2
```

```
## Loaded glmnet 3.0-2
```

```r
BLR_data <- na.omit(diff[-c(1,2,3,4)])

BLR_data <- na.omit(BLR_data)
BLR_response <- as.matrix(BLR_data$BLRScore_change)
BLR_predictor <- as.matrix(BLR_data[,-1])

LASSO <- glmnet(BLR_predictor,BLR_response,family = 'gaussian')
plot(LASSO,xvar='lambda',label =  TRUE)
```




```
colnames(BLR_predictor)[c(9,10)]
```

```
## [1] "DietScore_change"             "PhysicalActivityScore_change"
```

```
knitr::kable(as.matrix(coef(LASSO,s=exp(-3))),digits = 3,caption = 'coefficient estimate when lambda =
```

Table 1: coefficient estimate when lambda = 1

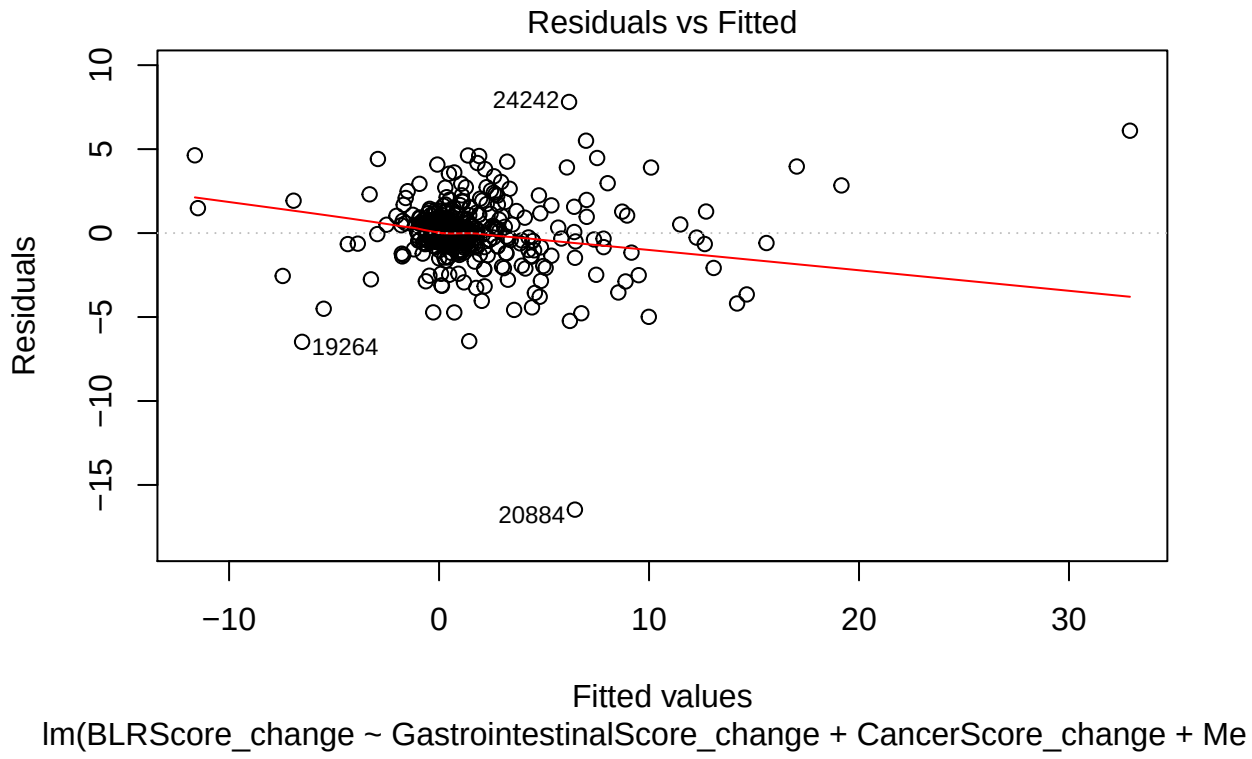
	1
(Intercept)	0.009
HeartScore_change	0.000
RespiratoryScore_change	0.000
GastrointestinalScore_change	-0.101
DiabetesScore_change	0.000
CancerScore_change	0.000
ArthritisPainScore_change	0.000
MentalHealthScore_change	-0.526
SocialFinancialRelationshipScore_change	0.000
DietScore_change	1.289
PhysicalActivityScore_change	2.050
FinancialHealthScore_change	0.871
MedicationScore_change	0.261
AlcoholScore_change	0.045
SleepScore_change	0.000

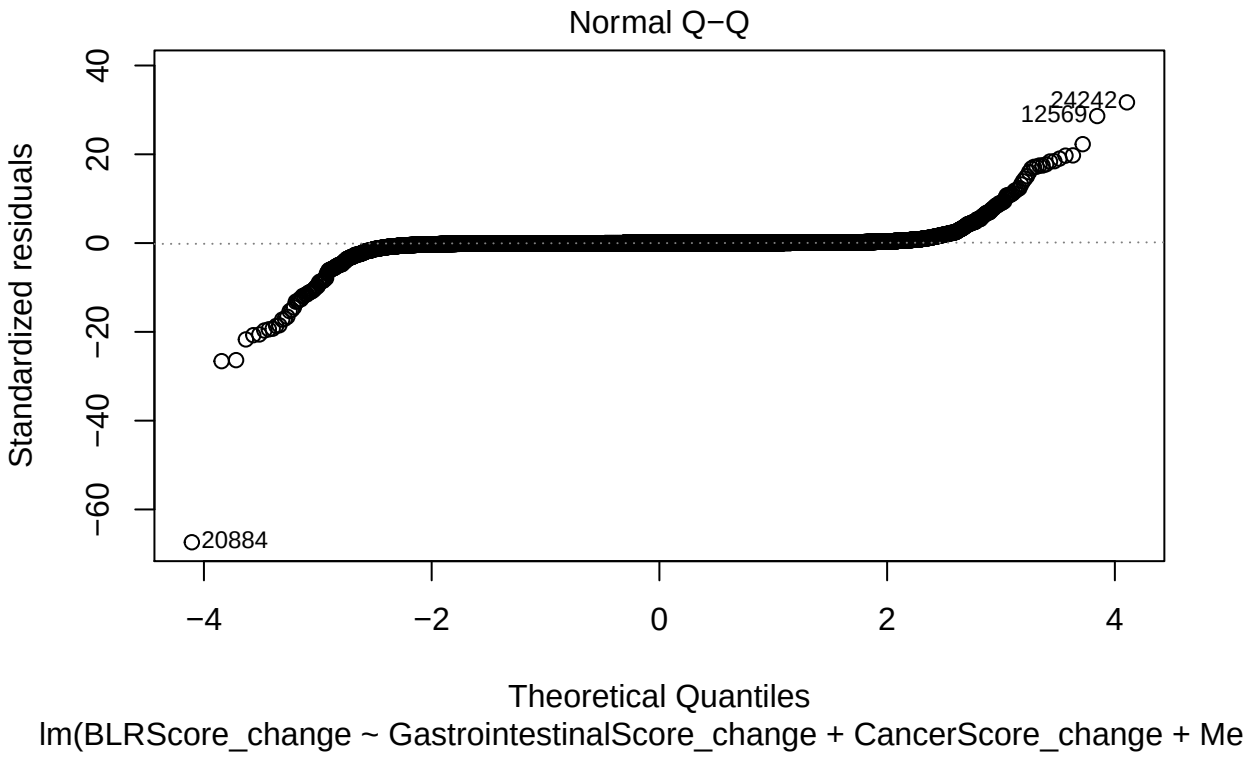
	1
StressScore_change	0.719
SmokingScore_change	0.688
BMI_change	0.000

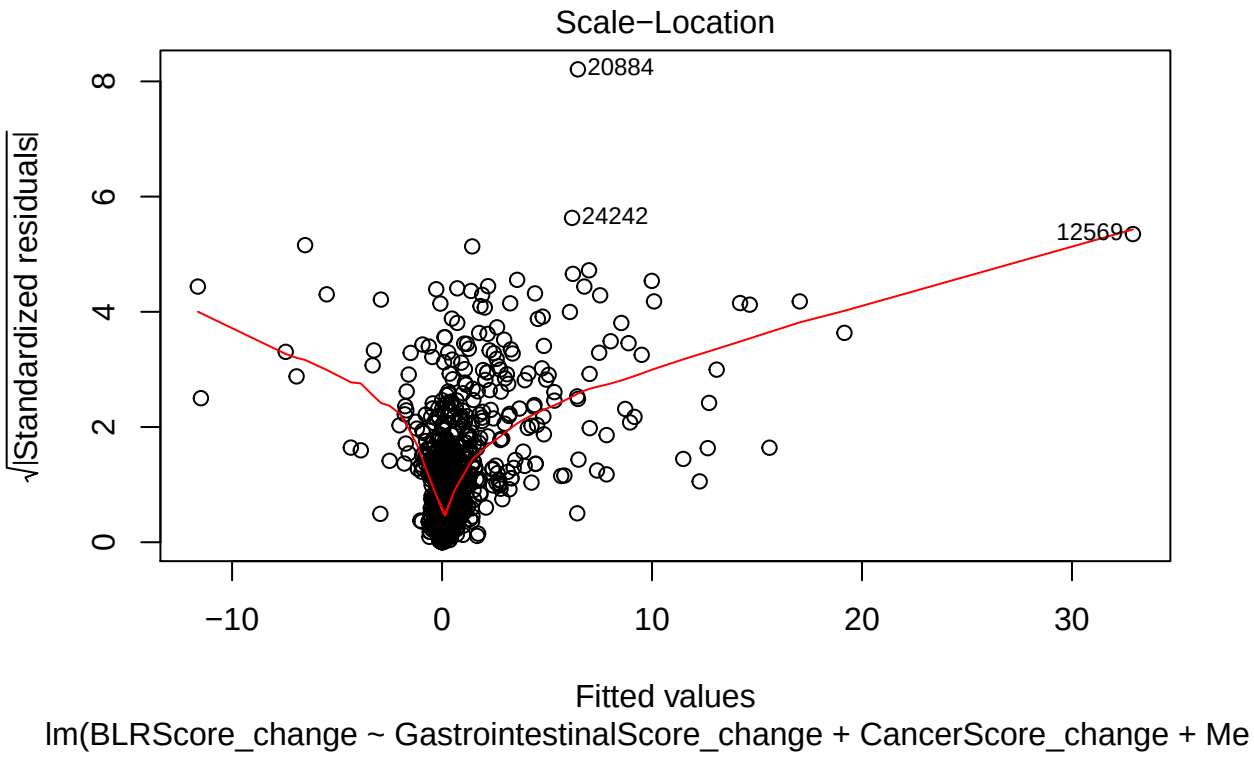
```
residual_fit<-lm(BLRScore_change ~ GastrointestinalScore_change+CancerScore_change+MentalHealthScore_cha
                 +DietScore_change+PhysicalActivityScore_change+FinancialHealthScore_change+MedicationS
                 +AlcoholScore_change+StressScore_change+SmokingScore_change,data=BLR_data)
coef(residual_fit)
```

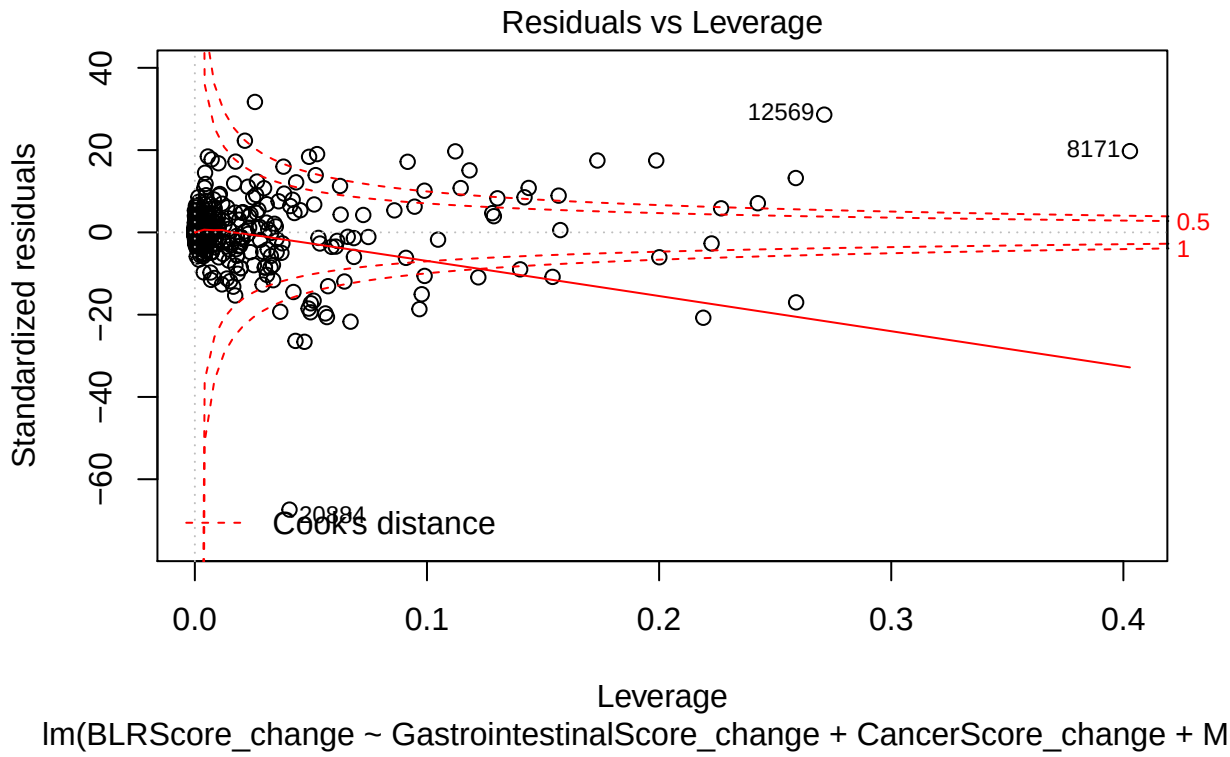
```
##                   (Intercept) GastrointestinalScore_change
##                   0.002916714                  -0.097404394
##            CancerScore_change     MentalHealthScore_change
##                  -0.293101631                  -0.543581164
##              DietScore_change PhysicalActivityScore_change
##                   1.606246290                   2.127719843
##  FinancialHealthScore_change       MedicationScore_change
##                   0.952631501                   0.727067727
##           AlcoholScore_change           StressScore_change
##                   1.059192182                   1.055111212
##           SmokingScore_change
##                   1.240058132
```

```
plot(residual_fit)
```




Normal Q−Q
Standardized residuals
40
20
0
−20
−40
−60
24242
12569
20884
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ GastrointestinalScore_change + CancerScore_change + Me


Scale−Location
20884
24242
12569
√|Standardized residuals|
0
2
4
6
8
−10
0
10
20
30
Fitted values
lm(BLRScore_change ~ GastrointestinalScore_change + CancerScore_change + Me



Backward/forward Selection

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.6.2
```

```
##
## Attaching package: 'MASS'
```

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

```
fit <- lm(BLRScore_change ~ .,data = BLR_data)
step <- stepAIC(fit, direction = 'both')
```

```
## Start:  AIC=-69946.6
## BLRScore_change ~ HeartScore_change + RespiratoryScore_change +
##     GastrointestinalScore_change + DiabetesScore_change + CancerScore_change +
##     ArthritisPainScore_change + MentalHealthScore_change + SocialFinancialRelationshipScore_change +
##     DietScore_change + PhysicalActivityScore_change + FinancialHealthScore_change +
##     MedicationScore_change + AlcoholScore_change + SleepScore_change +
```

```
##     StressScore_change + SmokingScore_change + BMI_change
##
##                                           Df Sum of Sq    RSS    AIC
## <none>                                                  1490.0 -69947
## - BMI_change                               1      2.83 1492.9 -69901
## - SleepScore_change                        1      3.57 1493.6 -69889
## - HeartScore_change                        1      3.96 1494.0 -69883
## - GastrointestinalScore_change             1      4.25 1494.3 -69878
## - RespiratoryScore_change                  1      8.26 1498.3 -69811
## - DiabetesScore_change                     1     11.52 1501.6 -69757
## - CancerScore_change                       1     11.60 1501.6 -69756
## - SocialFinancialRelationshipScore_change  1     19.97 1510.0 -69618
## - ArthritisPainScore_change                1     23.99 1514.0 -69551
## - AlcoholScore_change                      1     55.33 1545.4 -69042
## - StressScore_change                       1     65.26 1555.3 -68883
## - MedicationScore_change                   1     72.76 1562.8 -68763
## - FinancialHealthScore_change              1    143.37 1633.4 -67664
## - MentalHealthScore_change                 1    192.59 1682.6 -66926
## - DietScore_change                         1    233.91 1724.0 -66323
## - SmokingScore_change                      1    320.13 1810.2 -65110
## - PhysicalActivityScore_change             1    461.24 1951.3 -63243
```

```
summary(step)
```

```
##
## Call:
## lm(formula = BLRScore_change ~ HeartScore_change + RespiratoryScore_change +
##     GastrointestinalScore_change + DiabetesScore_change + CancerScore_change +
##     ArthritisPainScore_change + MentalHealthScore_change + SocialFinancialRelationshipScore_change +
##     DietScore_change + PhysicalActivityScore_change + FinancialHealthScore_change +
##     MedicationScore_change + AlcoholScore_change + SleepScore_change +
##     StressScore_change + SmokingScore_change + BMI_change, data = BLR_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2714  -0.0085  -0.0021   0.0050   7.3314
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               0.002964   0.001559   1.901   0.0573
## HeartScore_change                         0.097079   0.011953   8.122 4.81e-16
## RespiratoryScore_change                  -0.205271   0.017490 -11.737  < 2e-16
## GastrointestinalScore_change             -0.146131   0.017360  -8.418  < 2e-16
## DiabetesScore_change                     -0.305779   0.022067 -13.857  < 2e-16
## CancerScore_change                       -0.276893   0.019910 -13.907  < 2e-16
## ArthritisPainScore_change                 0.232880   0.011643  20.002  < 2e-16
## MentalHealthScore_change                 -0.785656   0.013864 -56.669  < 2e-16
## SocialFinancialRelationshipScore_change  0.341149   0.018693  18.250  < 2e-16
## DietScore_change                          1.509238   0.024166  62.453  < 2e-16
## PhysicalActivityScore_change              2.179764   0.024855  87.698  < 2e-16
## FinancialHealthScore_change               0.950757   0.019445  48.895  < 2e-16
## MedicationScore_change                    0.651932   0.018717  34.831  < 2e-16
## AlcoholScore_change                       0.967532   0.031852  30.376  < 2e-16
## SleepScore_change                         0.428925   0.055616   7.712 1.28e-14
```

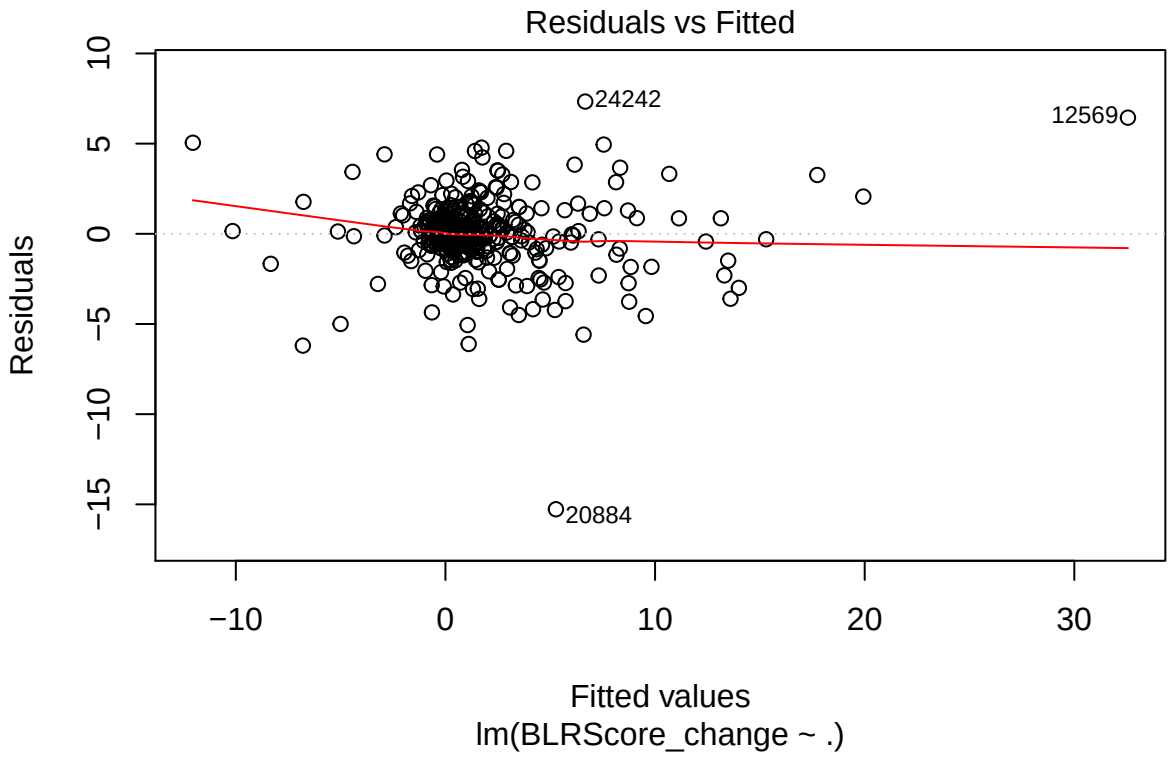
```
## StressScore_change                          1.077883   0.032674  32.989  < 2e-16
## SmokingScore_change                         1.205836   0.016504  73.063  < 2e-16
## BMI_change                                  0.013734   0.001998   6.874 6.41e-12
##
## (Intercept)                                .
## HeartScore_change                          ***
## RespiratoryScore_change                    ***
## GastrointestinalScore_change               ***
## DiabetesScore_change                       ***
## CancerScore_change                         ***
## ArthritisPainScore_change                  ***
## MentalHealthScore_change                   ***
## SocialFinancialRelationshipScore_change    ***
## DietScore_change                           ***
## PhysicalActivityScore_change               ***
## FinancialHealthScore_change                ***
## MedicationScore_change                     ***
## AlcoholScore_change                        ***
## SleepScore_change                          ***
## StressScore_change                         ***
## SmokingScore_change                        ***
## BMI_change                                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2449 on 24846 degrees of freedom
## Multiple R-squared:  0.8146, Adjusted R-squared:  0.8145
## F-statistic:  6422 on 17 and 24846 DF,  p-value: < 2.2e-16
```

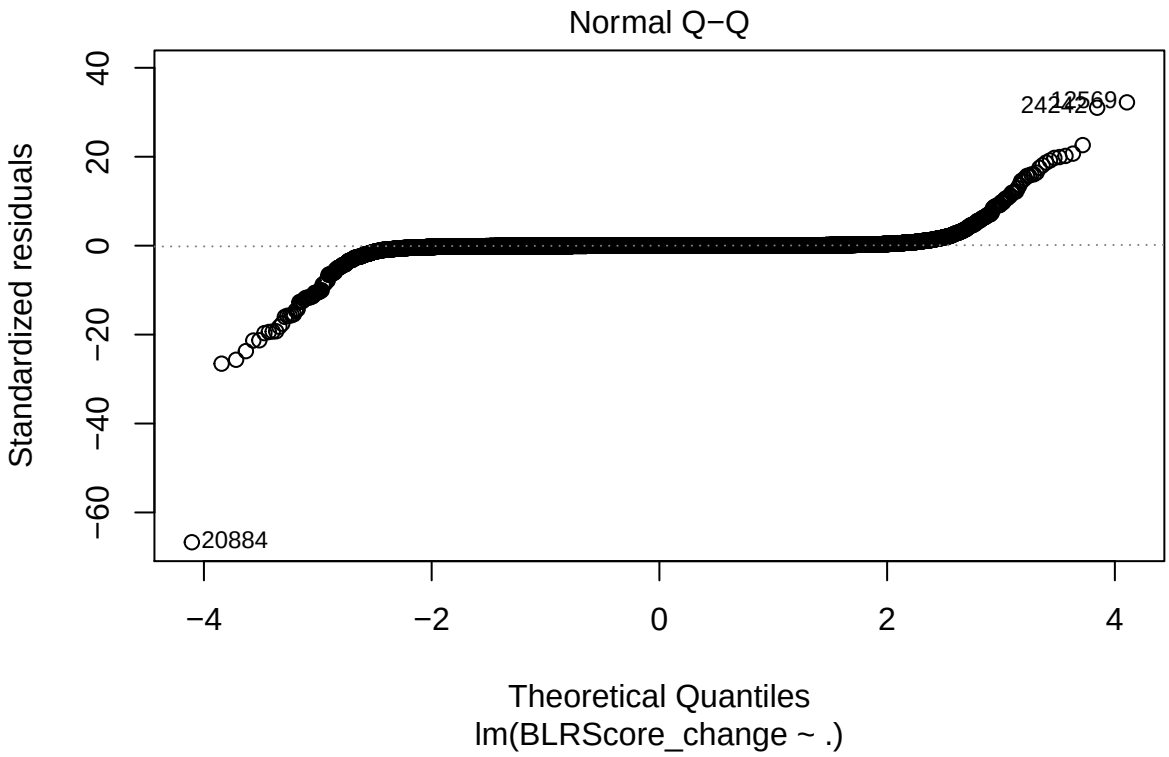
```
knitr::kable(coef(fit),caption = 'coefficeints estimate obtained by Backward/forward Selection')
```

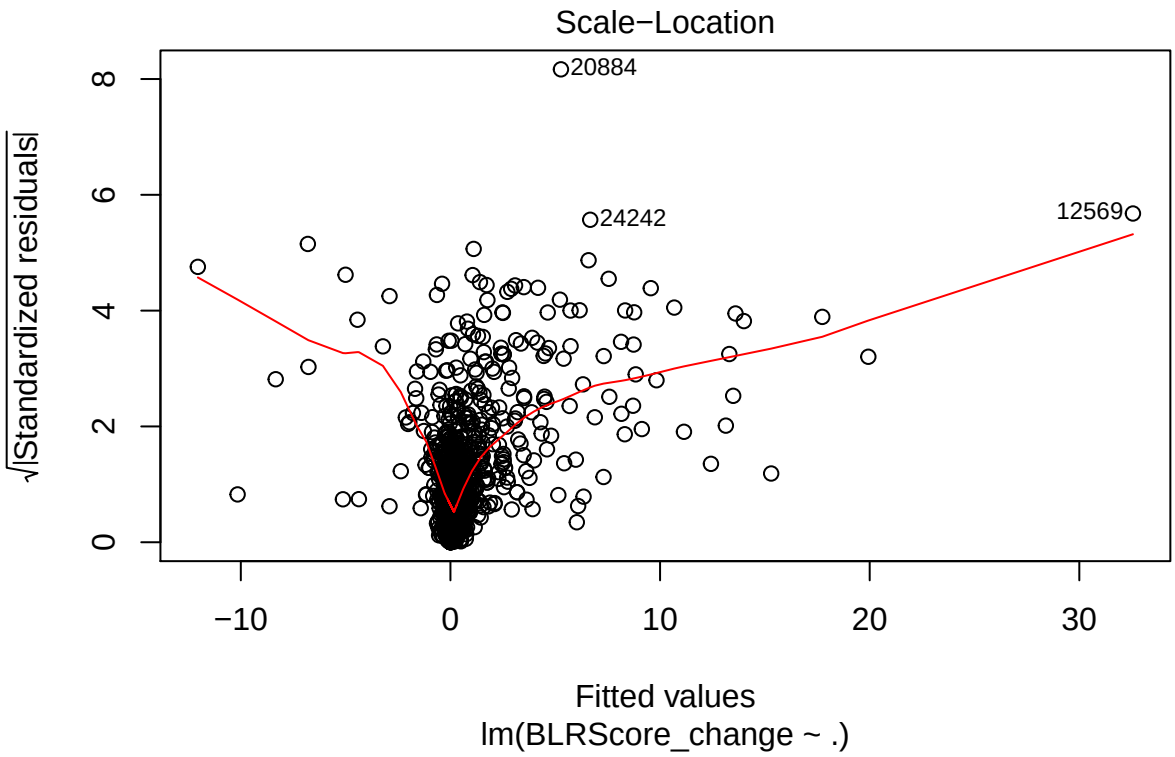
Table 2: coefficeints estimate obtained by Backward/forward Selection

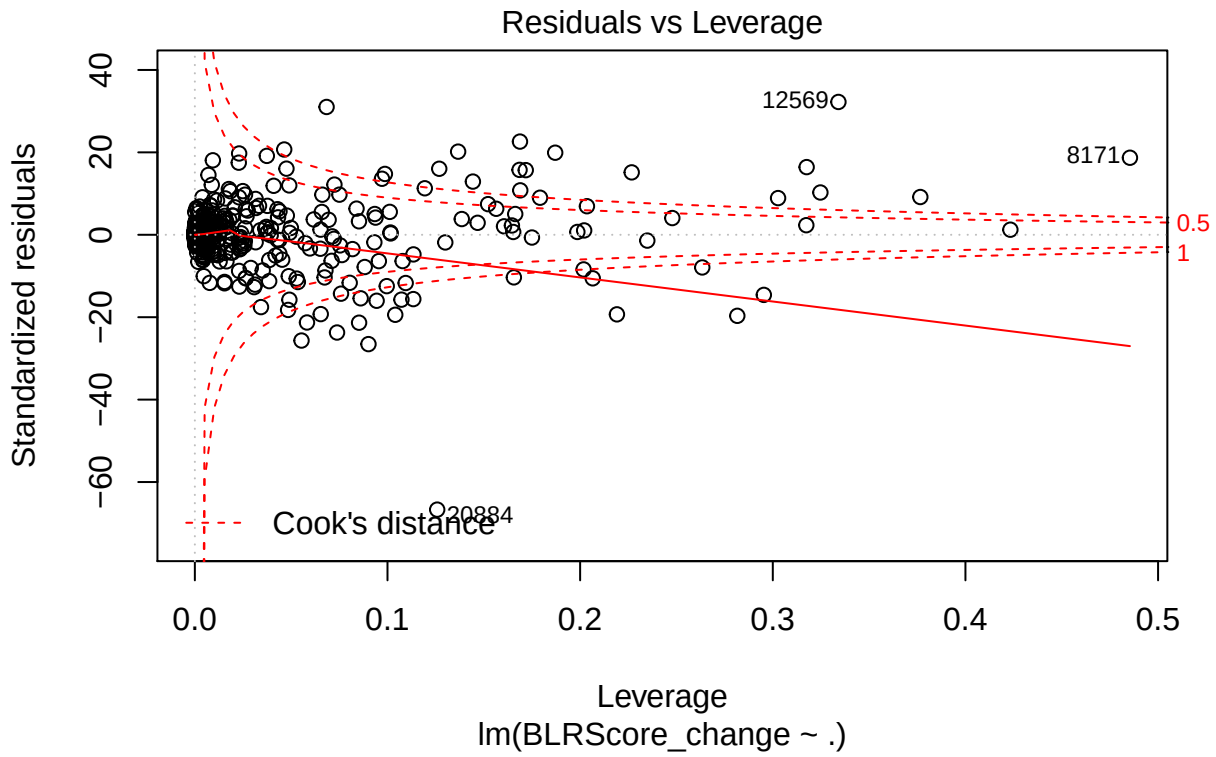
	x
(Intercept)	0.0029640
HeartScore_change	0.0970790
RespiratoryScore_change	-0.2052711
GastrointestinalScore_change	-0.1461309
DiabetesScore_change	-0.3057787
CancerScore_change	-0.2768932
ArthritisPainScore_change	0.2328797
MentalHealthScore_change	-0.7856560
SocialFinancialRelationshipScore_change	0.3411494
DietScore_change	1.5092383
PhysicalActivityScore_change	2.1797642
FinancialHealthScore_change	0.9507569
MedicationScore_change	0.6519318
AlcoholScore_change	0.9675317
SleepScore_change	0.4289246
StressScore_change	1.0778826
SmokingScore_change	1.2058356
BMI_change	0.0137343

```
plot(fit)
```


Residuals vs Fitted
24242
12569
20884
Residuals
10
5
0
−5
−10
−15
−10
0
10
20
30
Fitted values
lm(BLRScore_change ~ .)


Normal Q−Q
Standardized residuals
40
20
0
−20
−40
−60
20884
24242
12569
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ .)


Scale−Location
20884
24242
12569
√|Standardized residuals|
0
2
4
6
8
−10
0
10
20
30
Fitted values
lm(BLRScore_change ~ .)

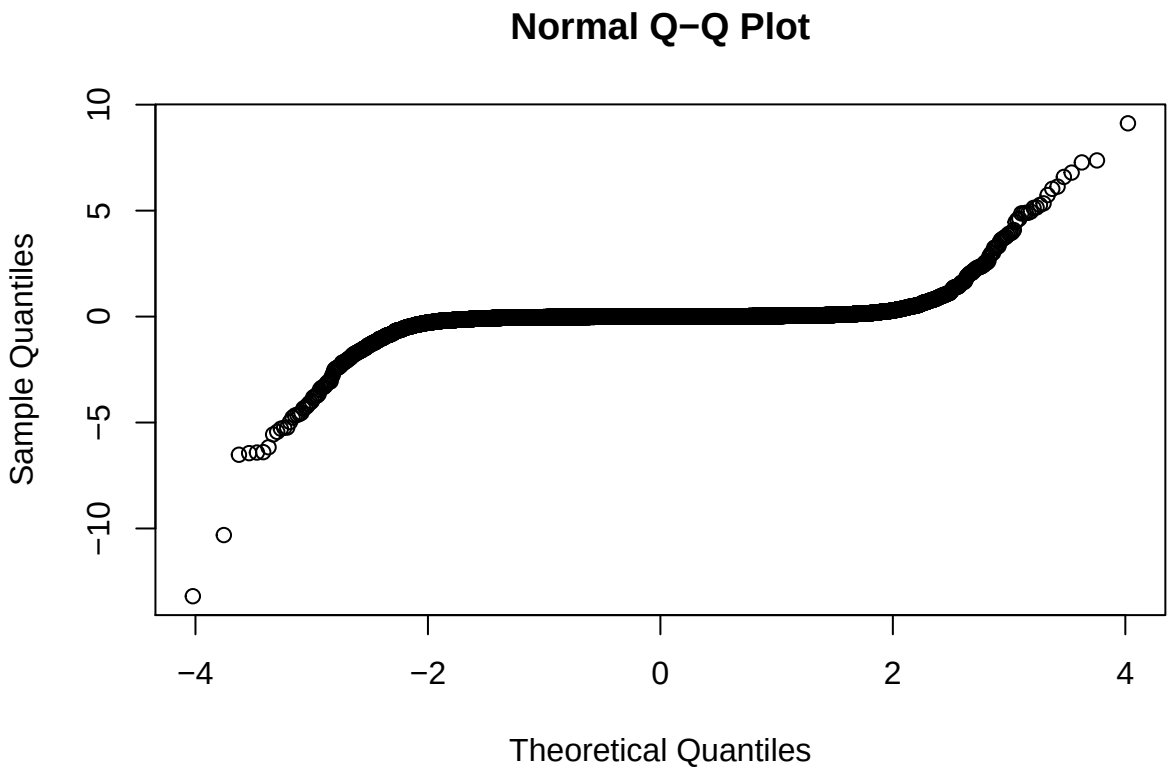



```
lmtest::lrtest(step,residual_fit)
```

```
## Likelihood ratio test
##
## Model 1: BLRScore_change ~ HeartScore_change + RespiratoryScore_change +
##     GastrointestinalScore_change + DiabetesScore_change + CancerScore_change +
##     ArthritisPainScore_change + MentalHealthScore_change + SocialFinancialRelationshipScore_change +
##     DietScore_change + PhysicalActivityScore_change + FinancialHealthScore_change +
##     MedicationScore_change + AlcoholScore_change + SleepScore_change +
##     StressScore_change + SmokingScore_change + BMI_change
## Model 2: BLRScore_change ~ GastrointestinalScore_change + CancerScore_change +
##     MentalHealthScore_change + DietScore_change + PhysicalActivityScore_change +
##     FinancialHealthScore_change + MedicationScore_change + AlcoholScore_change +
##     StressScore_change + SmokingScore_change
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1  19 -289.19
## 2  12 -760.28 -7 942.2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

mixed model

```
mixed_fit <- nlme::lme(BLR_change ~ Heart_change+Respiratory_change+Gastrointestinal_change+
                         Diabetes_change+Cancer_change+ArthritisPain_change+MentalHealth_change+SocioFi
                         Diet_change+PhysicalActivity_change+FinancialHealth_change+Medication_change+A
                         Sleep_change+Stress_change+Smoking_change+BMI_change,random = ~1|UserId, data =

qqnorm(resid(mixed_fit))
```

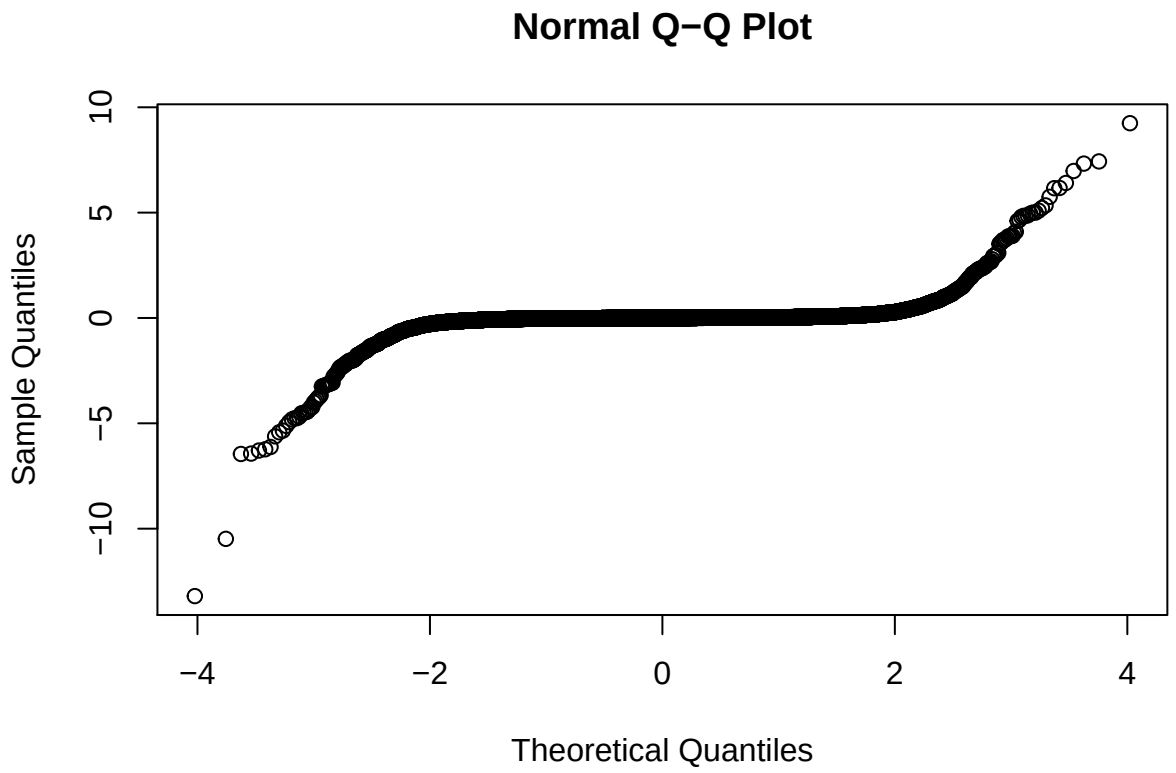



```
knitr::kable(Pmisc::lmeTable(mixed_fit),digits = 3)
```

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	-0.001	0.003	13234	-0.273	0.785
Heart_change	-0.349	0.012	13234	-30.183	0.000
Respiratory_change	-0.140	0.016	13234	-8.659	0.000
Gastrointestinal_change	-0.133	0.017	13234	-7.610	0.000
Diabetes_change	-0.184	0.027	13234	-6.704	0.000
Cancer_change	-0.546	0.025	13234	-21.456	0.000
ArthritisPain_change	0.129	0.012	13234	10.600	0.000
MentalHealth_change	-0.226	0.017	13234	-13.505	0.000
SocioFinancial_change	-0.218	0.020	13234	-11.040	0.000
Diet_change	1.134	0.029	13234	39.626	0.000
PhysicalActivity_change	1.488	0.034	13234	44.096	0.000
FinancialHealth_change	1.554	0.023	13234	66.349	0.000
Medication_change	0.550	0.020	13234	27.221	0.000

	MLE	Std.Error	DF	t-value	p-value
Alcohol_change	1.905	0.035	13234	54.240	0.000
Sleep_change	1.705	0.051	13234	33.689	0.000
Stress_change	2.061	0.039	13234	52.442	0.000
Smoking_change	0.807	0.011	13234	75.519	0.000
BMI_change	-0.003	0.000	13234	-21.261	0.000
σ	0.000	NA	NA	NA	NA
τ	0.399	NA	NA	NA	NA

```
mixed_fit <- nlme::lme(BLR_change ~ Heart_change+Respiratory_change+Gastrointestinal_change+
                         Diabetes_change+Cancer_change+MentalHealth_change+SocioFinancial_change+
                         Diet_change+PhysicalActivity_change+FinancialHealth_change+Medication_change+Al
                         Sleep_change+Stress_change+Smoking_change+BMI_change,random = ~1|UserId, data =

qqnorm(resid(mixed_fit))
```

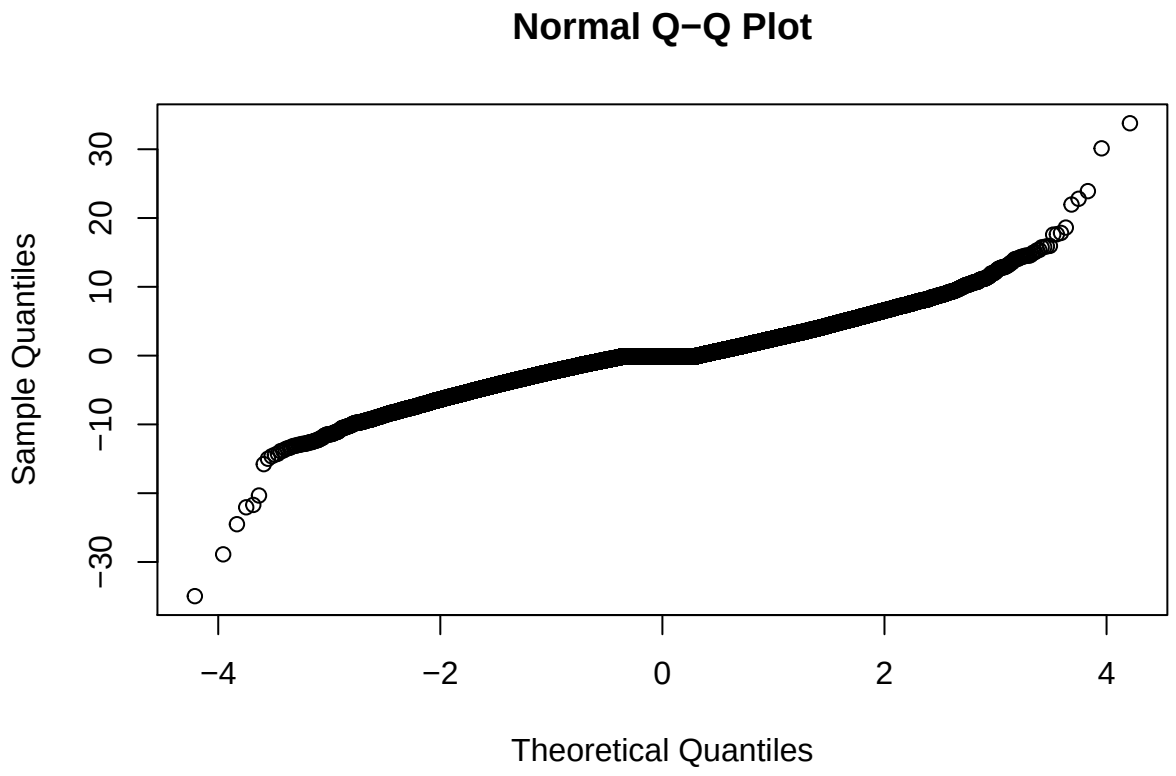



```
knitr::kable(Pmisc::lmeTable(mixed_fit),digits = 3)
```

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	-0.001	0.003	13235	-0.318	0.751
Heart_change	-0.321	0.011	13235	-28.432	0.000
Respiratory_change	-0.125	0.016	13235	-7.729	0.000
Gastrointestinal_change	-0.125	0.018	13235	-7.149	0.000

	MLE	Std.Error	DF	t-value	p-value
Diabetes_change	-0.169	0.027	13235	-6.160	0.000
Cancer_change	-0.489	0.025	13235	-19.599	0.000
MentalHealth_change	-0.208	0.017	13235	-12.472	0.000
SocioFinancial_change	-0.224	0.020	13235	-11.326	0.000
Diet_change	1.132	0.029	13235	39.410	0.000
PhysicalActivity_change	1.503	0.034	13235	44.432	0.000
FinancialHealth_change	1.546	0.023	13235	65.847	0.000
Medication_change	0.531	0.020	13235	26.302	0.000
Alcohol_change	1.923	0.035	13235	54.639	0.000
Sleep_change	1.588	0.050	13235	32.046	0.000
Stress_change	2.064	0.039	13235	52.372	0.000
Smoking_change	0.831	0.010	13235	79.313	0.000
BMI_change	-0.003	0.000	13235	-21.102	0.000
σ	0.000	NA	NA	NA	NA
τ	0.400	NA	NA	NA	NA

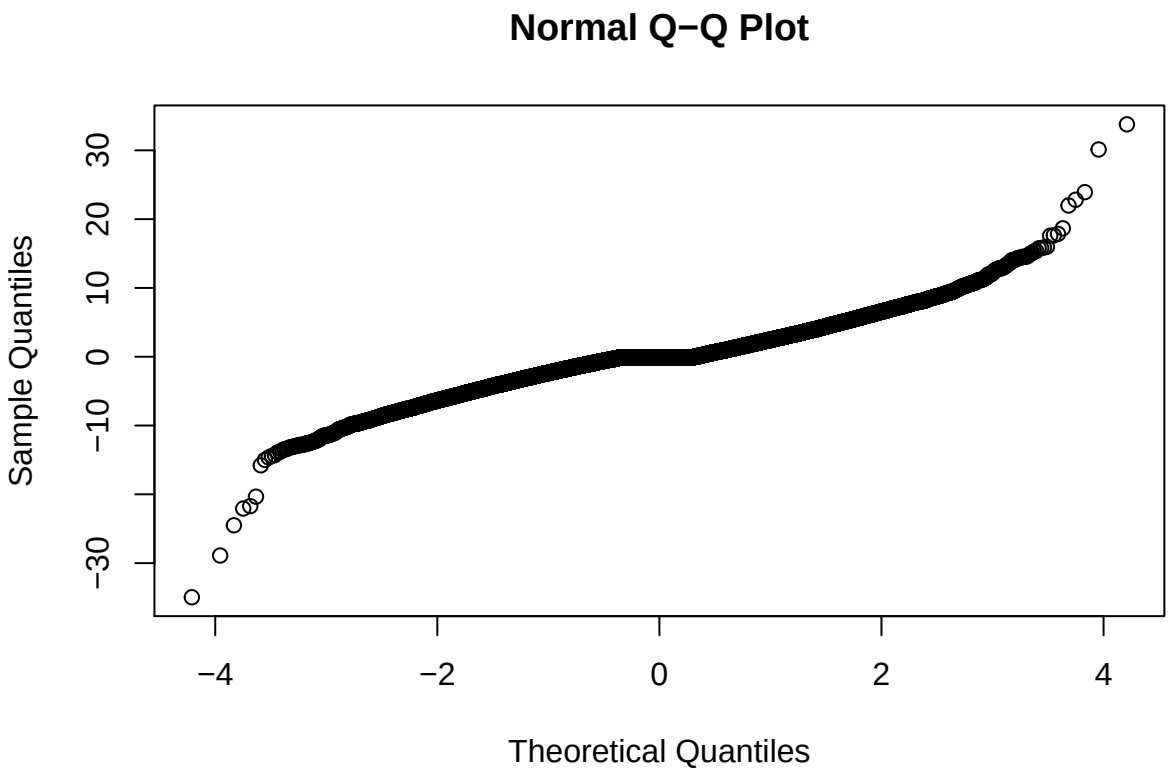
```
mixed_fit <- nlme::lme(BLR_change ~ Heart_change+Respiratory_change+Gastrointestinal_change+
                         Diabetes_change+Cancer_change+ArthritisPain_change+MentalHealth_change+SocioFi
                         Diet_change+PhysicalActivity_change+FinancialHealth_change+Medication_change+A
                         Sleep_change+Stress_change+Smoking_change+BMI_change,random = ~1|UserId, data =

qqnorm(resid(mixed_fit))
```



```
knitr::kable(Pmisc::lmeTable(mixed_fit),digits = 3)
```

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	0.142	0.015	29486	9.459	0.000
Heart_change	-0.085	0.010	29486	-8.074	0.000
Respiratory_change	-0.113	0.014	29486	-7.982	0.000
Gastrointestinal_change	-0.203	0.014	29486	-14.673	0.000
Diabetes_change	-0.171	0.018	29486	-9.273	0.000
Cancer_change	-0.429	0.016	29486	-27.625	0.000
ArthritisPain_change	0.004	0.010	29486	0.407	0.684
MentalHealth_change	-0.276	0.012	29486	-22.355	0.000
SocioFinancial_change	-0.282	0.017	29486	-16.556	0.000
Diet_change	1.419	0.022	29486	65.841	0.000
PhysicalActivity_change	1.417	0.019	29486	73.004	0.000
FinancialHealth_change	1.267	0.017	29486	76.309	0.000
Medication_change	0.325	0.018	29486	18.550	0.000
Alcohol_change	1.226	0.031	29486	39.367	0.000
Sleep_change	1.184	0.040	29486	29.415	0.000
Stress_change	1.198	0.030	29486	40.595	0.000
Smoking_change	0.812	0.013	29486	64.332	0.000
BMI_change	0.000	0.000	29486	-0.934	0.350
σ	0.000	NA	NA	NA	NA
τ	2.926	NA	NA	NA	NA

```
mixed_fit <- nlme::lme(BLR_change ~ Heart_change+Respiratory_change+Gastrointestinal_change+
                         Diabetes_change+Cancer_change+MentalHealth_change+SocioFinancial_change+
                         Diet_change+PhysicalActivity_change+FinancialHealth_change+Medication_change+Al
                         Sleep_change+Stress_change+Smoking_change,random = ~1|UserId, data = mixed_prep

qqnorm(resid(mixed_fit))
```

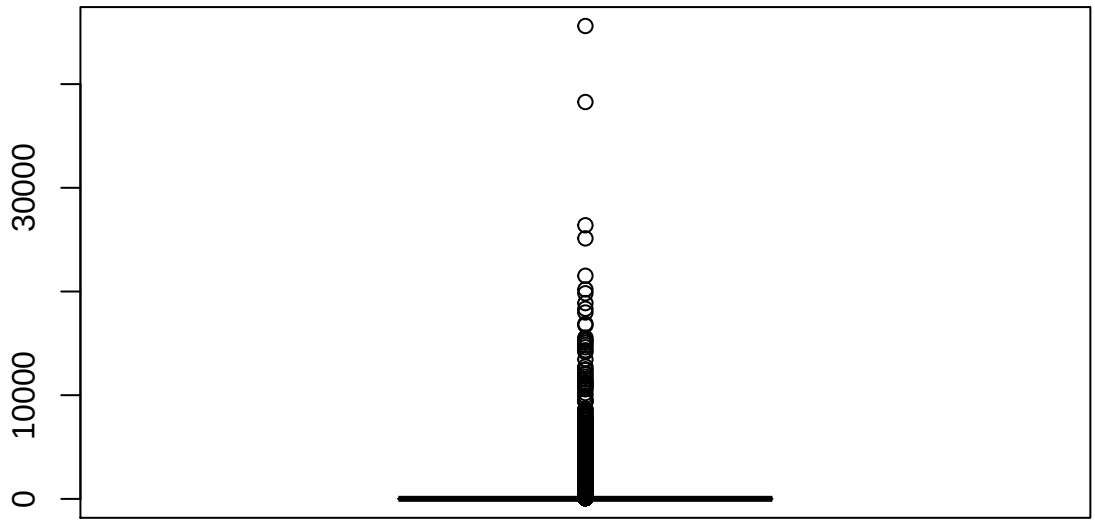



```
knitr::kable(Pmisc::lmeTable(mixed_fit),digits = 3)
```

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	0.142	0.015	29488	9.457	0
Heart_change	-0.084	0.010	29488	-8.067	0
Respiratory_change	-0.113	0.014	29488	-7.970	0
Gastrointestinal_change	-0.203	0.014	29488	-14.692	0
Diabetes_change	-0.170	0.018	29488	-9.428	0
Cancer_change	-0.428	0.015	29488	-27.889	0
MentalHealth_change	-0.275	0.012	29488	-22.813	0
SocioFinancial_change	-0.283	0.017	29488	-16.640	0
Diet_change	1.419	0.022	29488	65.840	0
PhysicalActivity_change	1.416	0.019	29488	73.208	0
FinancialHealth_change	1.267	0.017	29488	76.318	0
Medication_change	0.325	0.018	29488	18.578	0
Alcohol_change	1.226	0.031	29488	39.369	0
Sleep_change	1.182	0.040	29488	29.879	0
Stress_change	1.199	0.030	29488	40.625	0
Smoking_change	0.812	0.013	29488	64.353	0
σ	0.000	NA	NA	NA	NA
τ	2.926	NA	NA	NA	NA

Research Question 2

```
boxplot(points_ready$frequency)
```

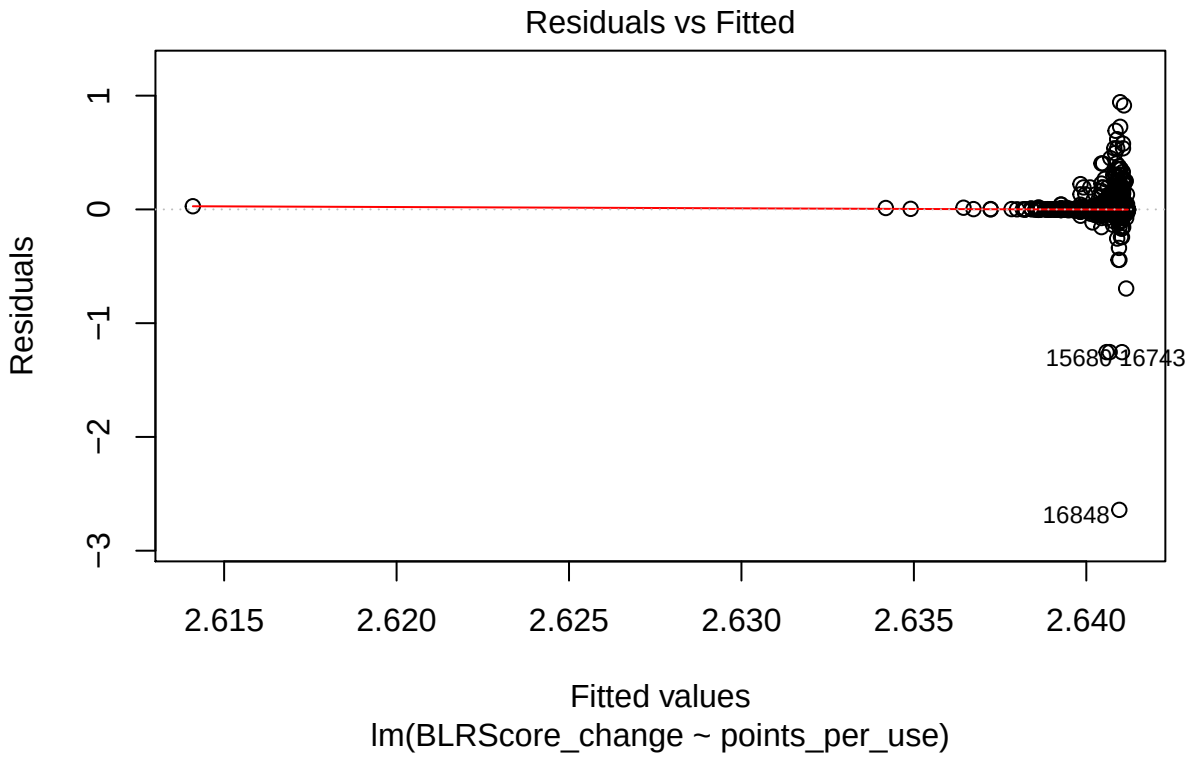


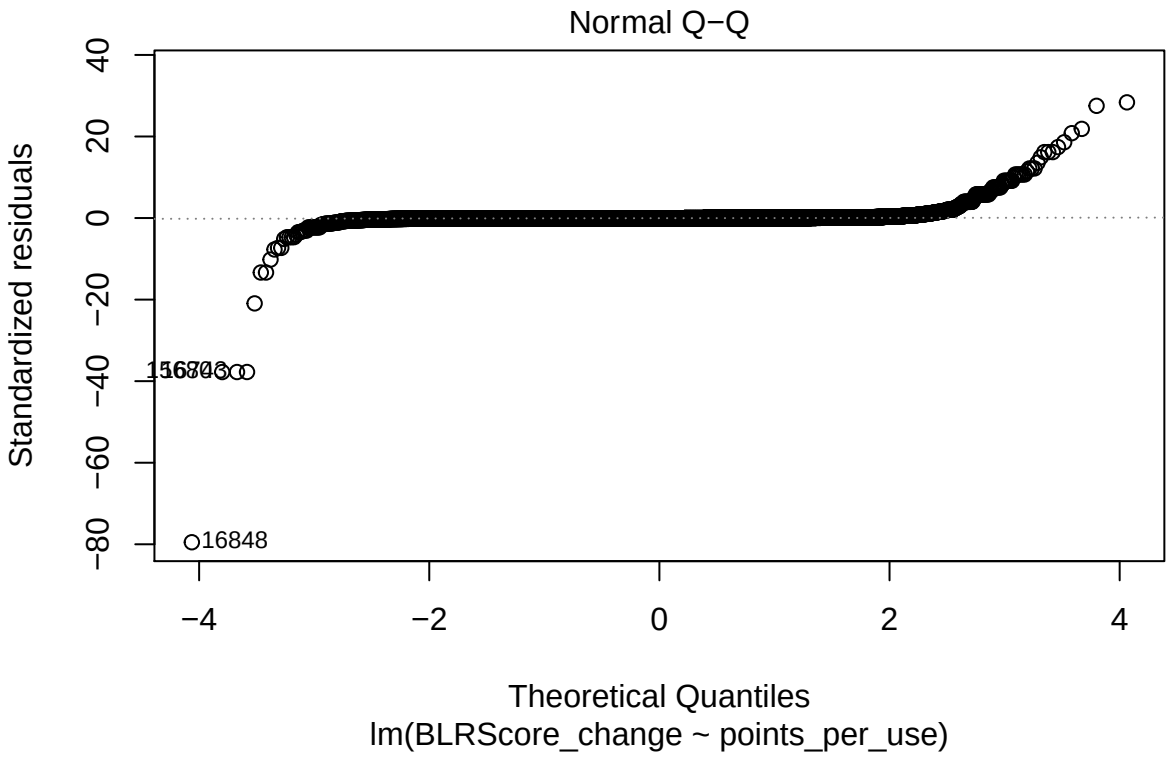

```
points_BLR <- merge(points_ready, diff, by = 'UserId')
points_BLR$BLRScore_change <- log(points_BLR$BLRScore_change+abs(min(points_BLR$BLRScore_change))+1)
points_BLR$points_per_use <- points_BLR$sum_points/points_BLR$frequency
outlier <- boxplot(points_BLR$frequency,plot = FALSE)$out
points_BLR_rmOutlier <- points_BLR[-which(points_BLR$frequency %in% outlier),]


model1 <- lm(BLRScore_change ~ points_per_use, data = points_BLR)
summary(model1)
```

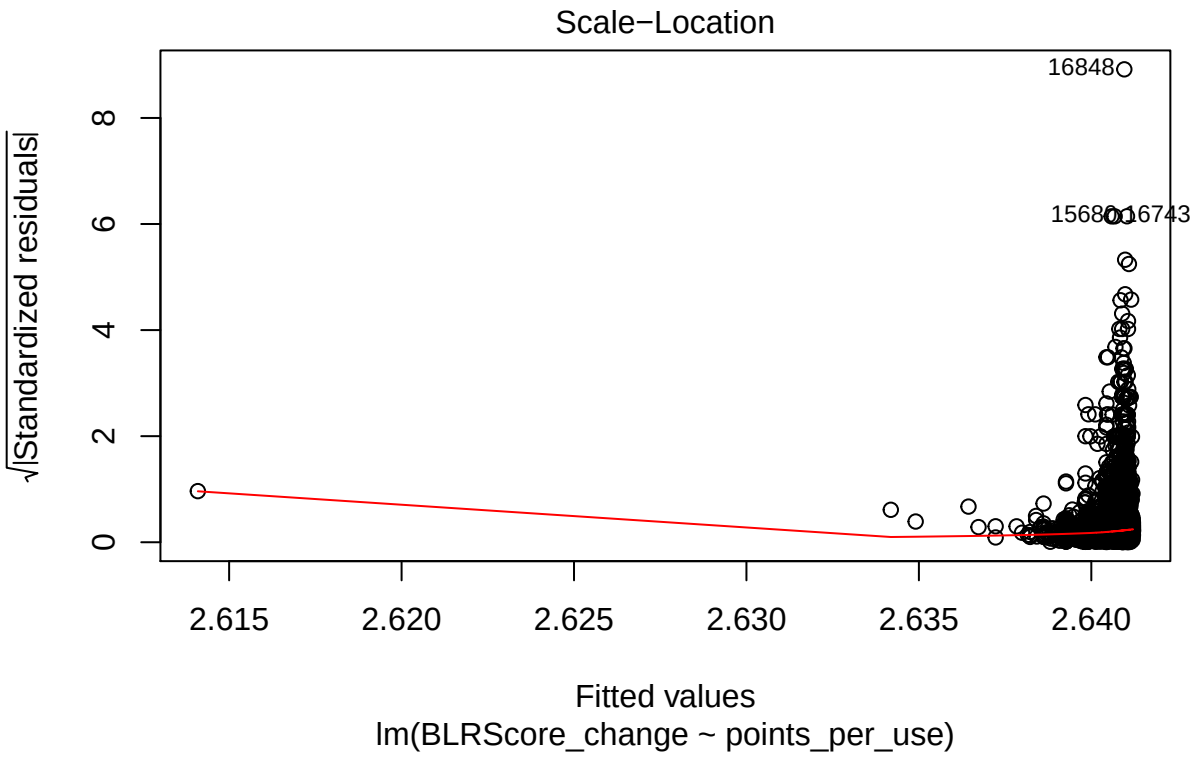
```
##
## Call:
## lm(formula = BLRScore_change ~ points_per_use, data = points_BLR)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64096 -0.00198 -0.00136 -0.00042  0.94254
##
## Coefficients:
##                  Estimate Std. Error  t value Pr(>|t|)
```

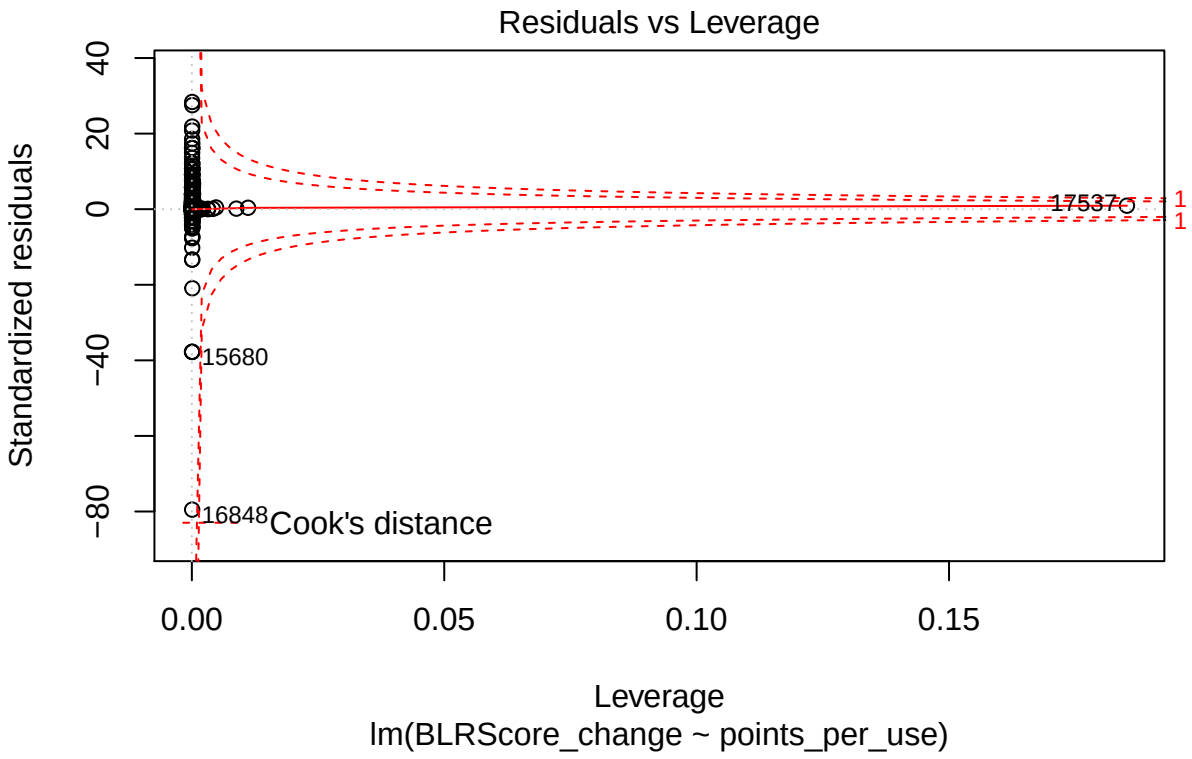
```
## (Intercept)     2.641e+00  3.645e-04 7247.123   <2e-16 ***
## points_per_use -2.605e-06  1.400e-06   -1.861   0.0628 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03322 on 20660 degrees of freedom
## Multiple R-squared:  0.0001676,  Adjusted R-squared:  0.0001192
## F-statistic: 3.463 on 1 and 20660 DF,  p-value: 0.06278
```

```
plot(model1)
```




Normal Q−Q
Standardized residuals
40
20
0
−20
−40
−60
−80
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ points_per_use)

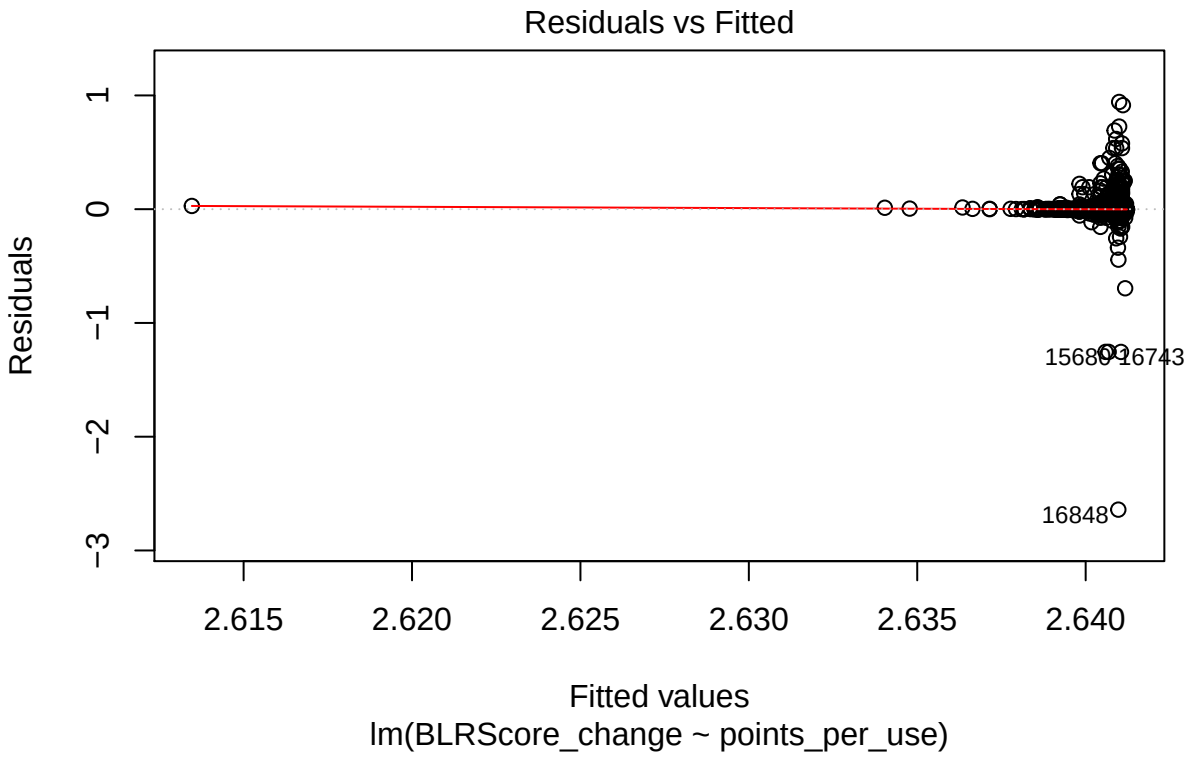

Scale−Location
16848
15680
16743
√|Standardized residuals|
0
2
4
6
8
2.615
2.620
2.625
2.630
2.635
2.640
Fitted values
lm(BLRScore_change ~ points_per_use)

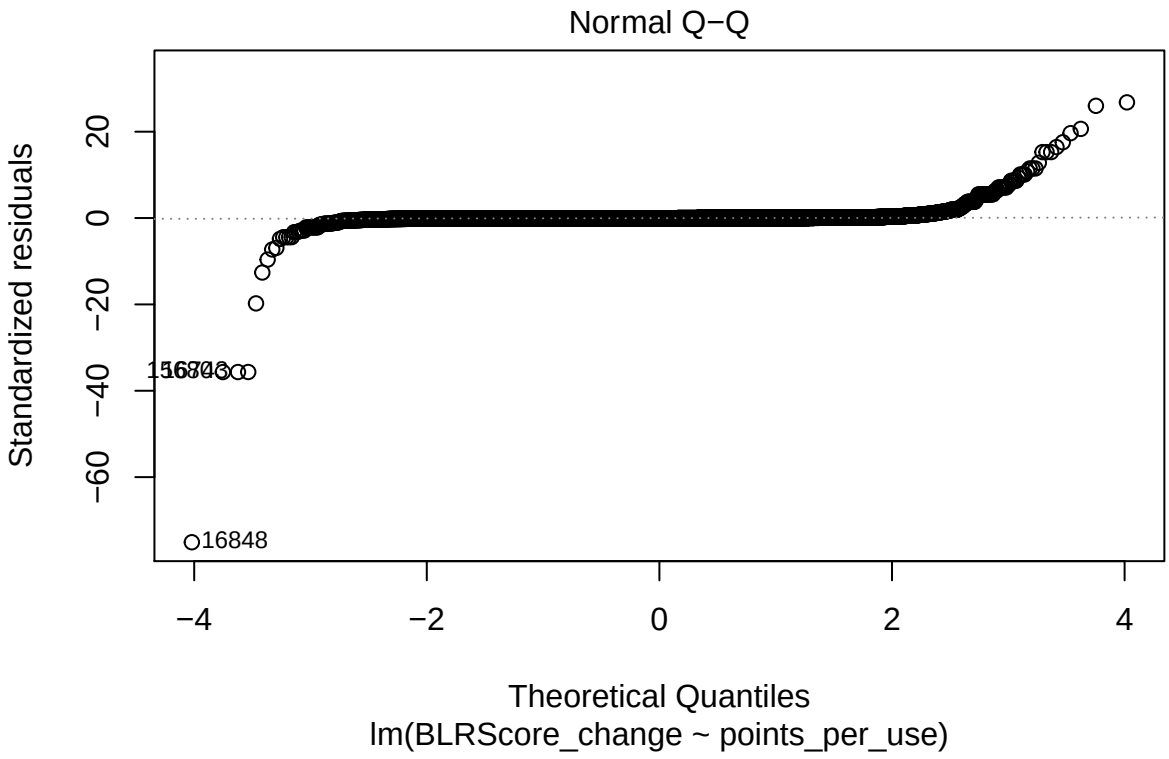


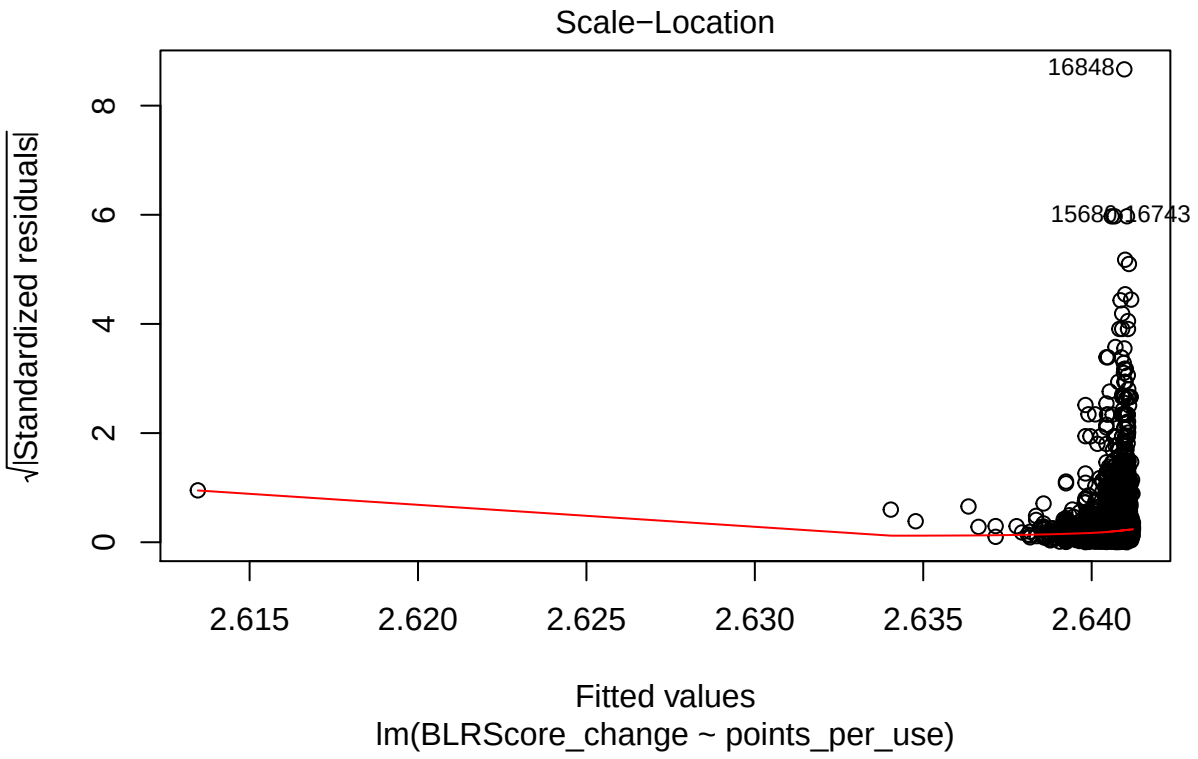

```
model1_rmOutlier <- lm(BLRScore_change ~ points_per_use, data = points_BLR_rmOutlier)
summary(model1_rmOutlier)
```

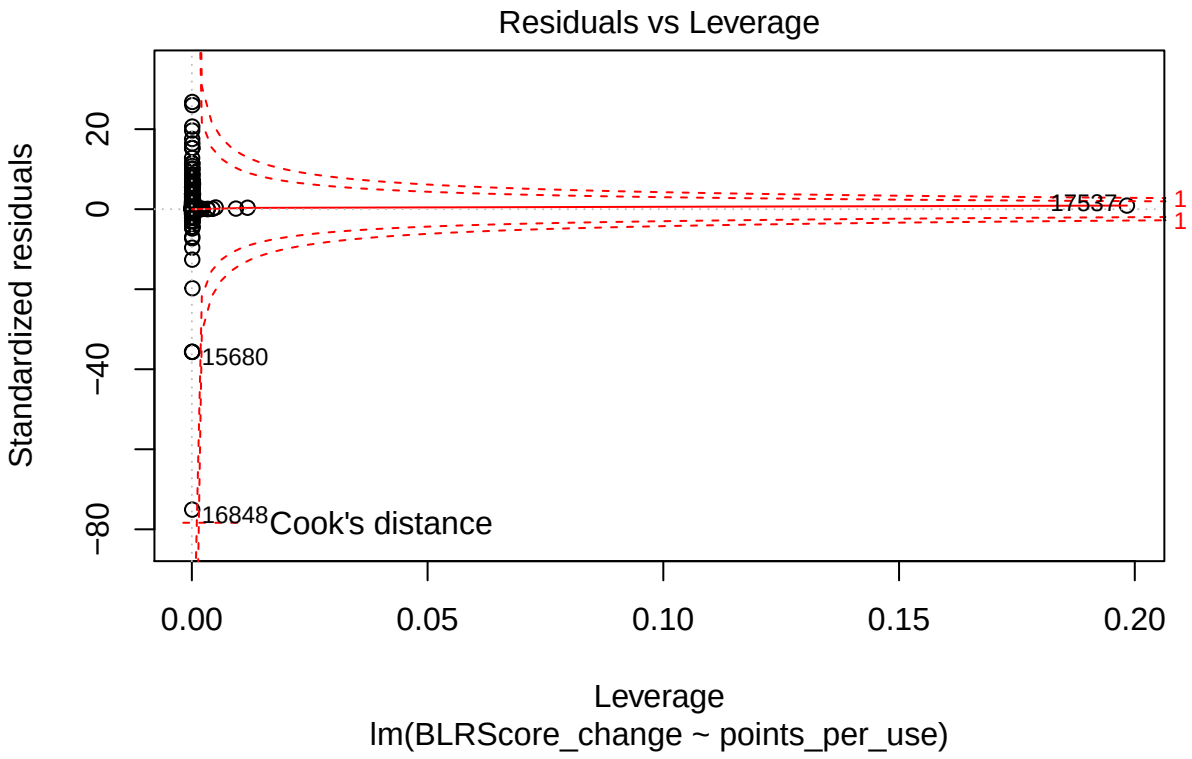
```
##
## Call:
## lm(formula = BLRScore_change ~ points_per_use, data = points_BLR_rmOutlier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64097 -0.00197 -0.00132 -0.00033  0.94252
##
## Coefficients:
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     2.641e+00  4.276e-04 6177.165   <2e-16 ***
## points_per_use -2.668e-06  1.536e-06   -1.736   0.0825 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03518 on 17199 degrees of freedom
## Multiple R-squared:  0.0001753,  Adjusted R-squared:  0.0001172
## F-statistic: 3.015 on 1 and 17199 DF,  p-value: 0.0825
```

```
plot(model1_rmOutlier)
```

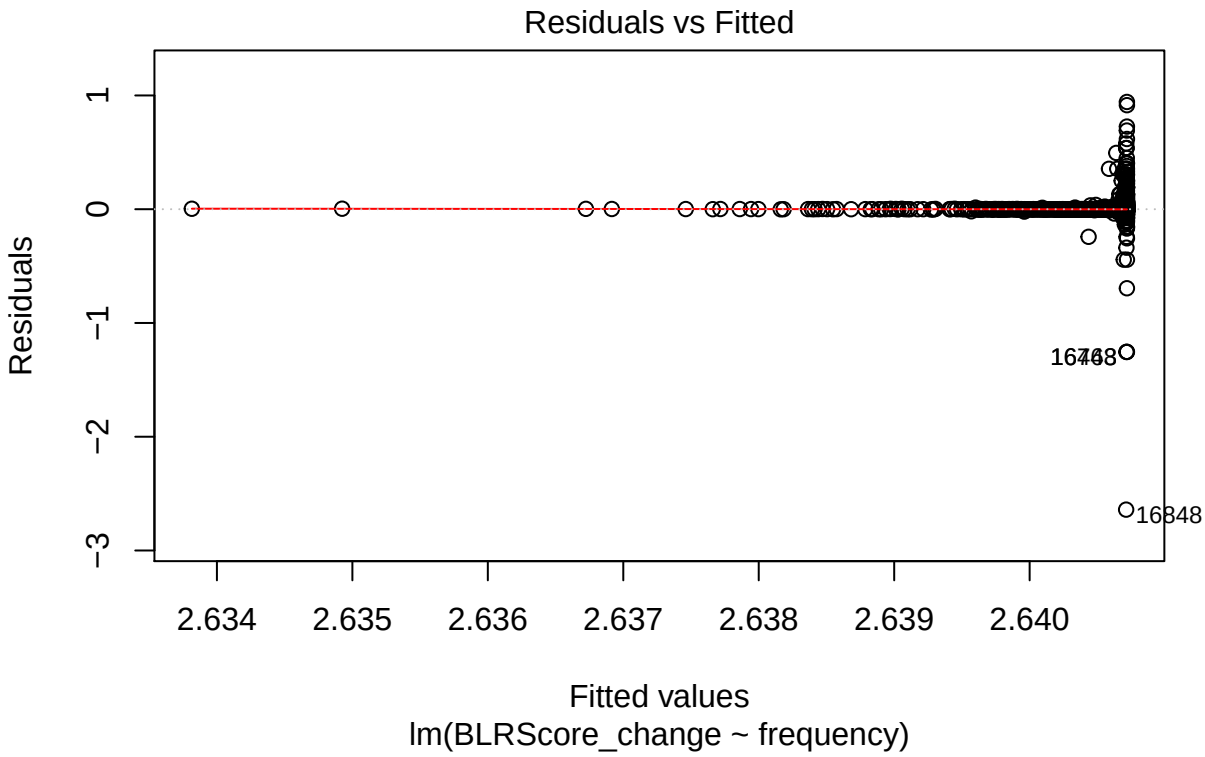

Residuals vs Fitted
Residuals
1
0
−1
−2
−3
15680
16743
16848
2.615
2.620
2.625
2.630
2.635
2.640
Fitted values
lm(BLRScore_change ~ points_per_use)

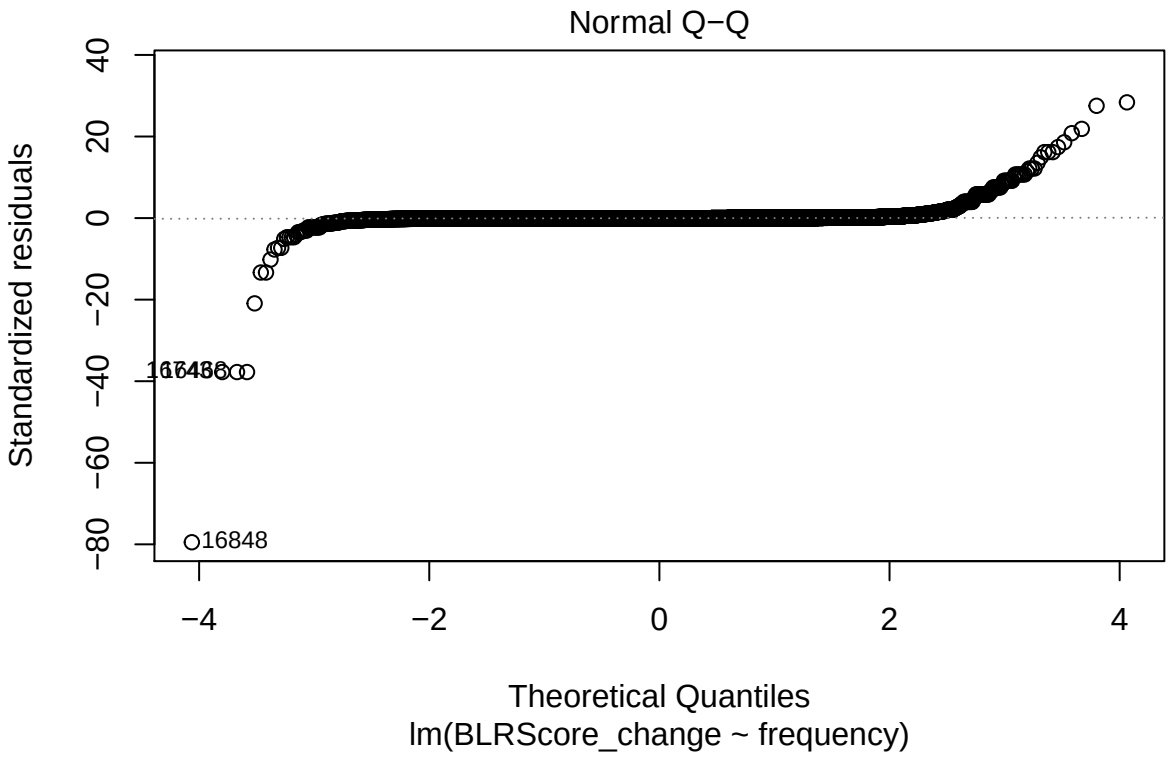

Normal Q−Q
Standardized residuals
20
0
−20
−40
−60
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ points_per_use)

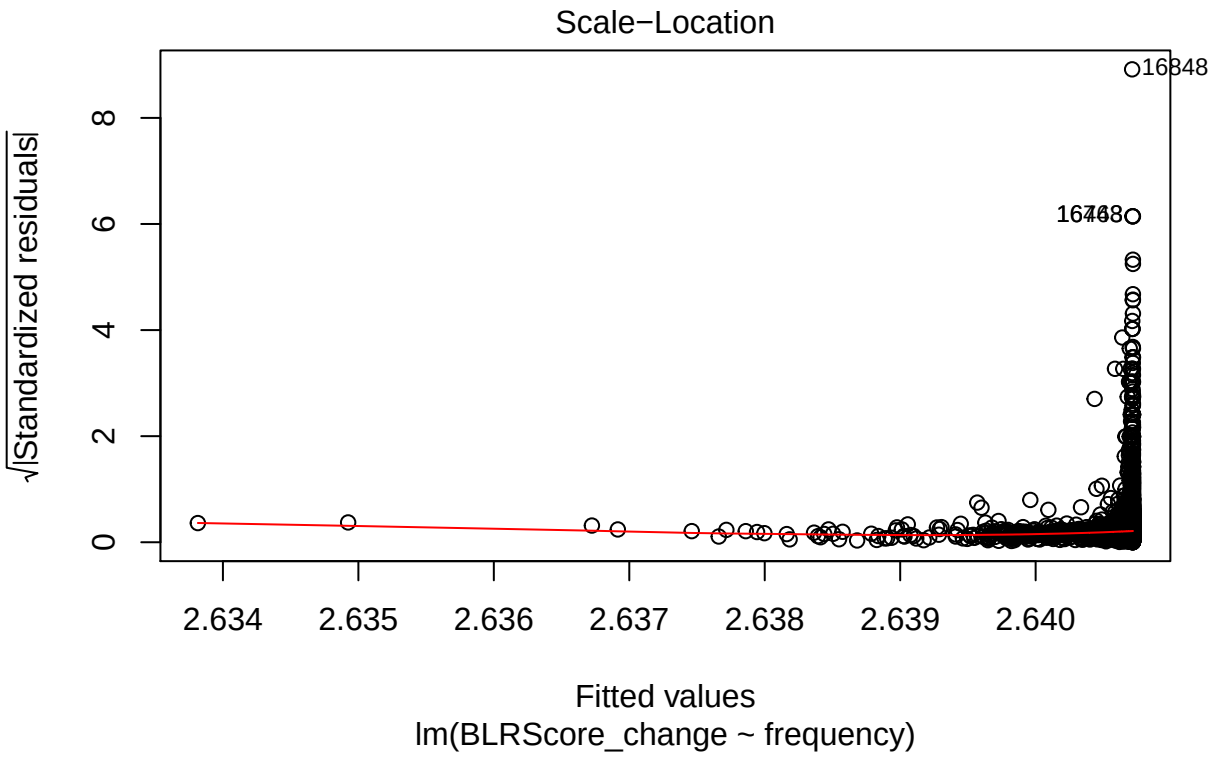

Scale−Location
16848
15680
16743
√|Standardized residuals|
0
2
4
6
8
2.615
2.620
2.625
2.630
2.635
2.640
Fitted values
lm(BLRScore_change ~ points_per_use)

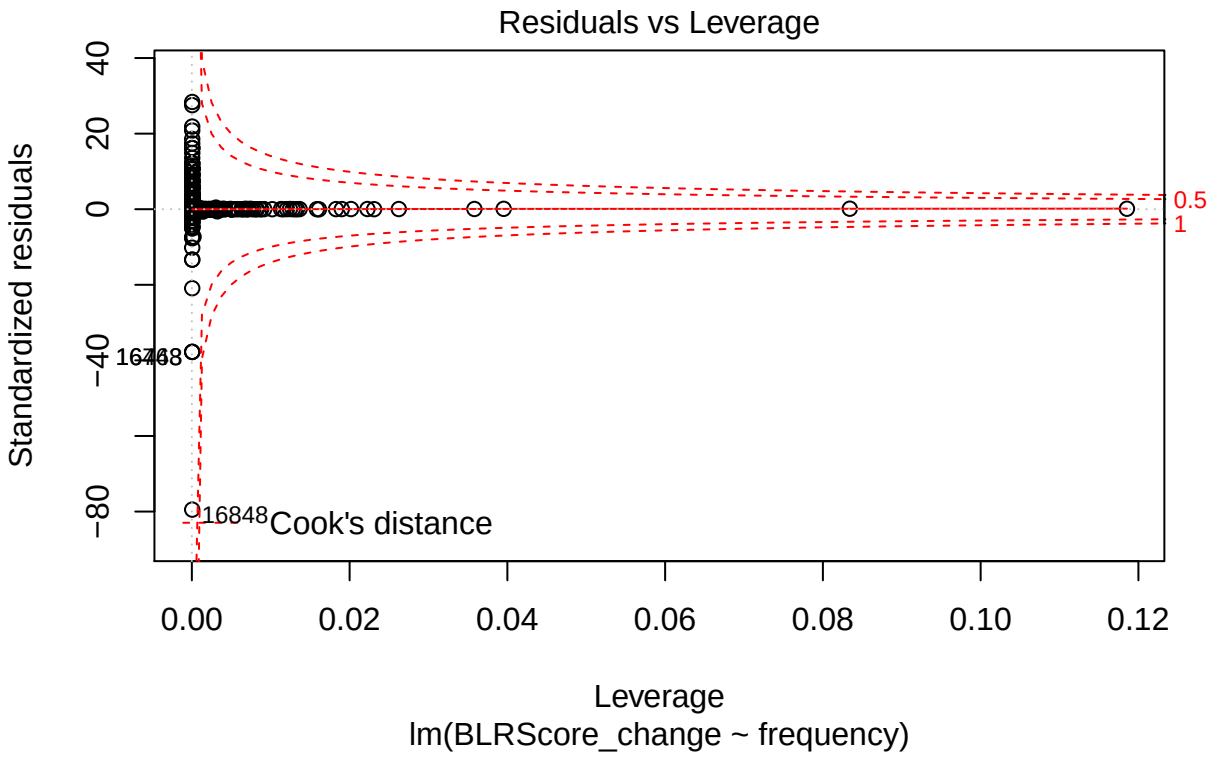



```
model2 <- lm(BLRScore_change ~ frequency,data = points_BLR)
summary(model2)
plot(model2)
```

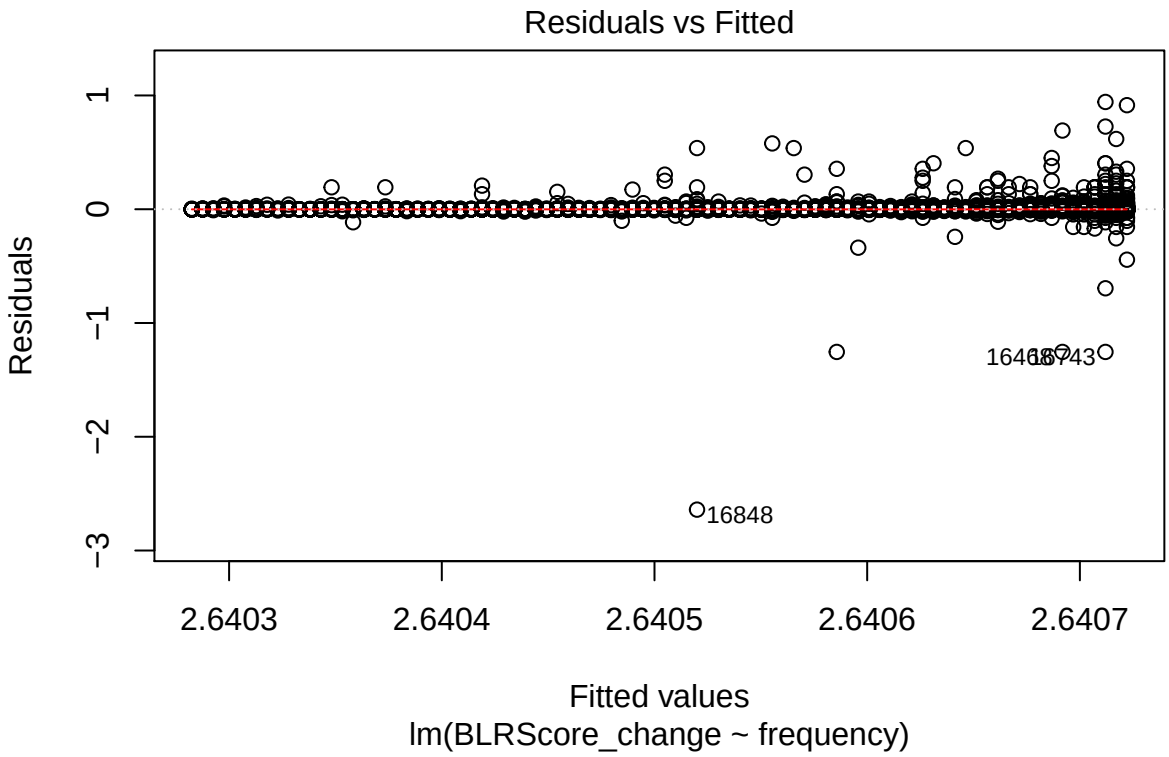

Residuals vs Fitted
Residuals
1
0
−1
−2
−3
2.634
2.635
2.636
2.637
2.638
2.639
2.640
Fitted values
lm(BLRScore_change ~ frequency)

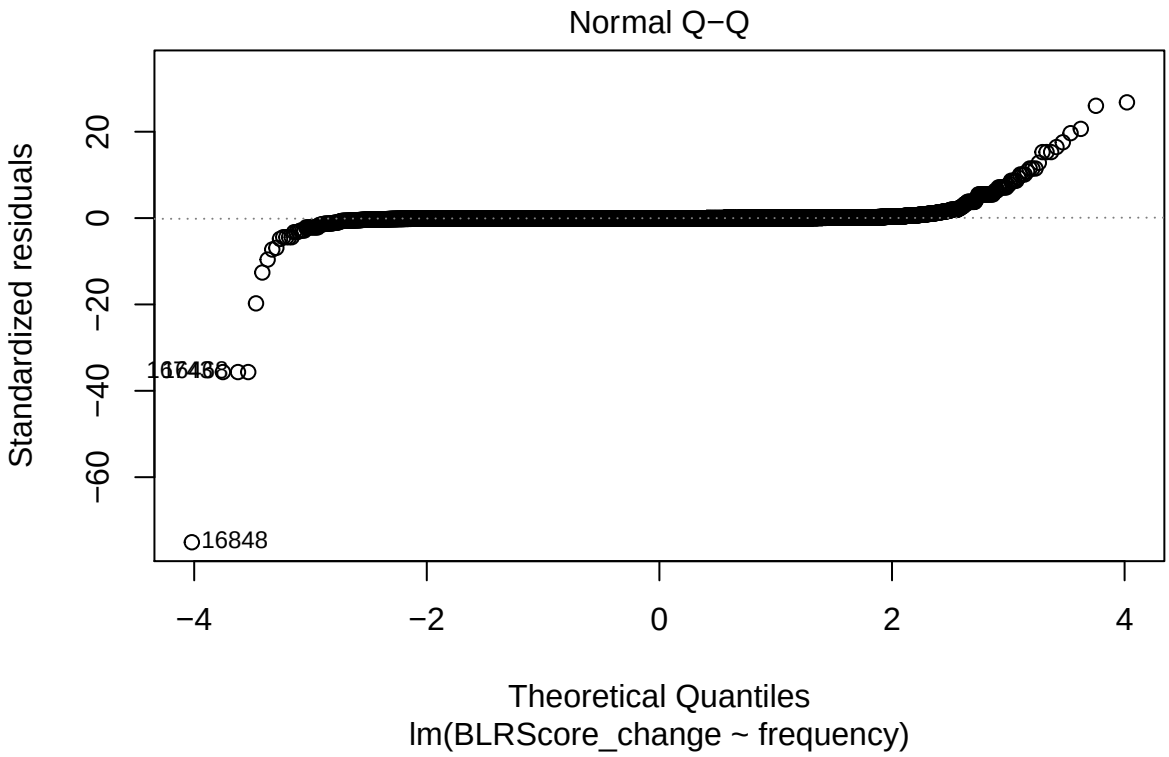

Normal Q−Q
Standardized residuals
40
20
0
−20
−40
−60
−80
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ frequency)

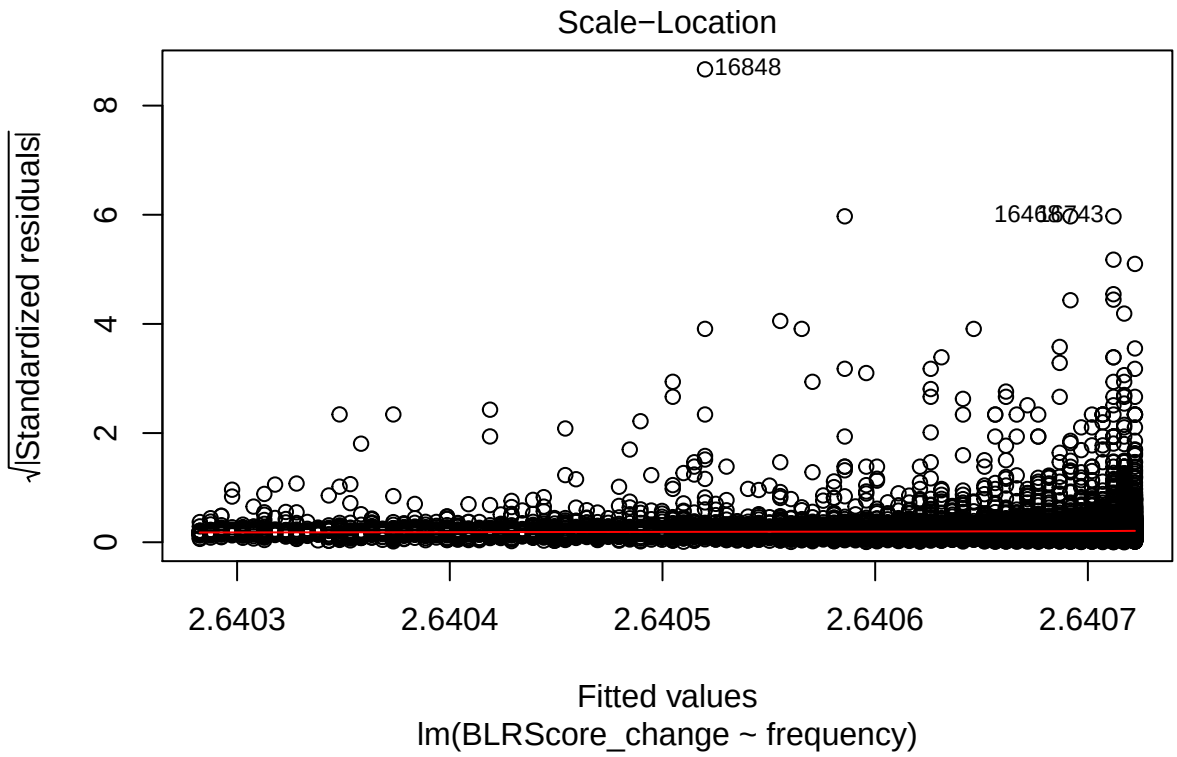

Scale−Location
√|Standardized residuals|
0
2
4
6
8
16848
16468
2.634
2.635
2.636
2.637
2.638
2.639
2.640
Fitted values
lm(BLRScore_change ~ frequency)

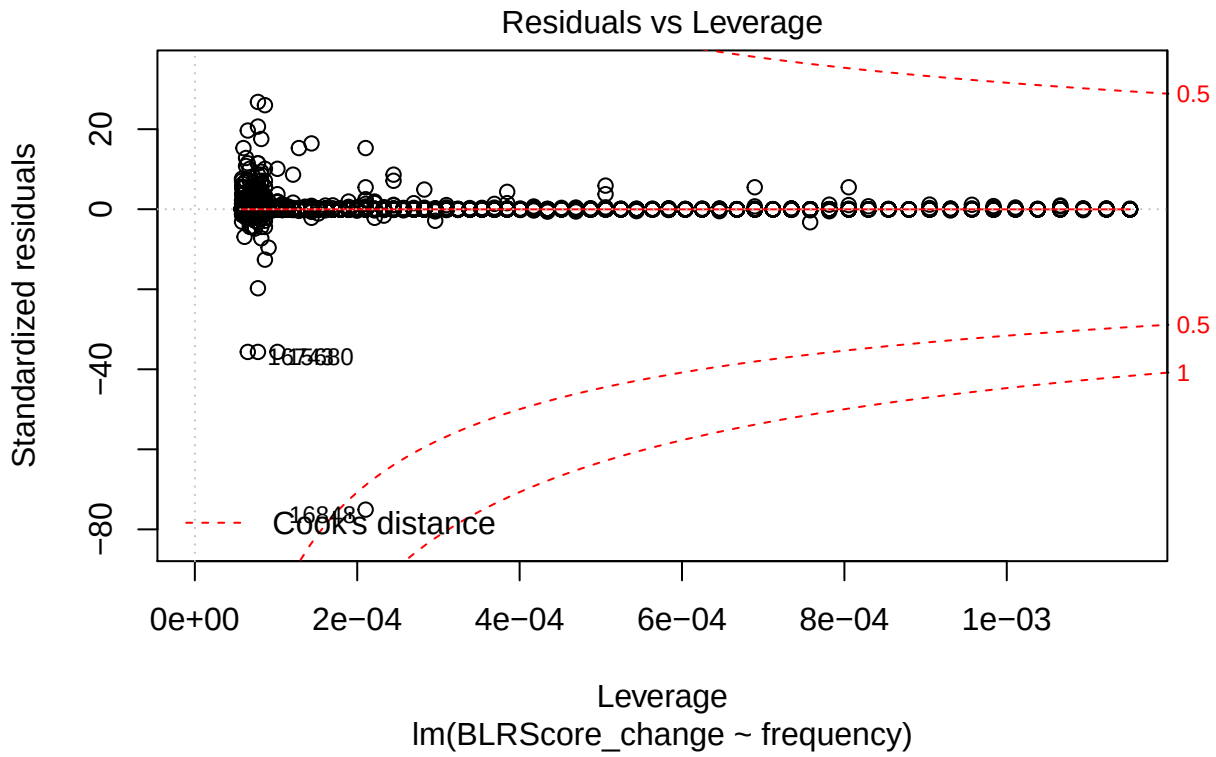



```
#qqnorm(sort(residuals(model1))[-c(1:400,seq(length(residuals(model1))-400,length(residuals(model1))))]

model2_rmOutlier <- lm(BLRScore_change ~ frequency,data = points_BLR_rmOutlier)
summary(model2_rmOutlier)
plot(model2_rmOutlier)
```

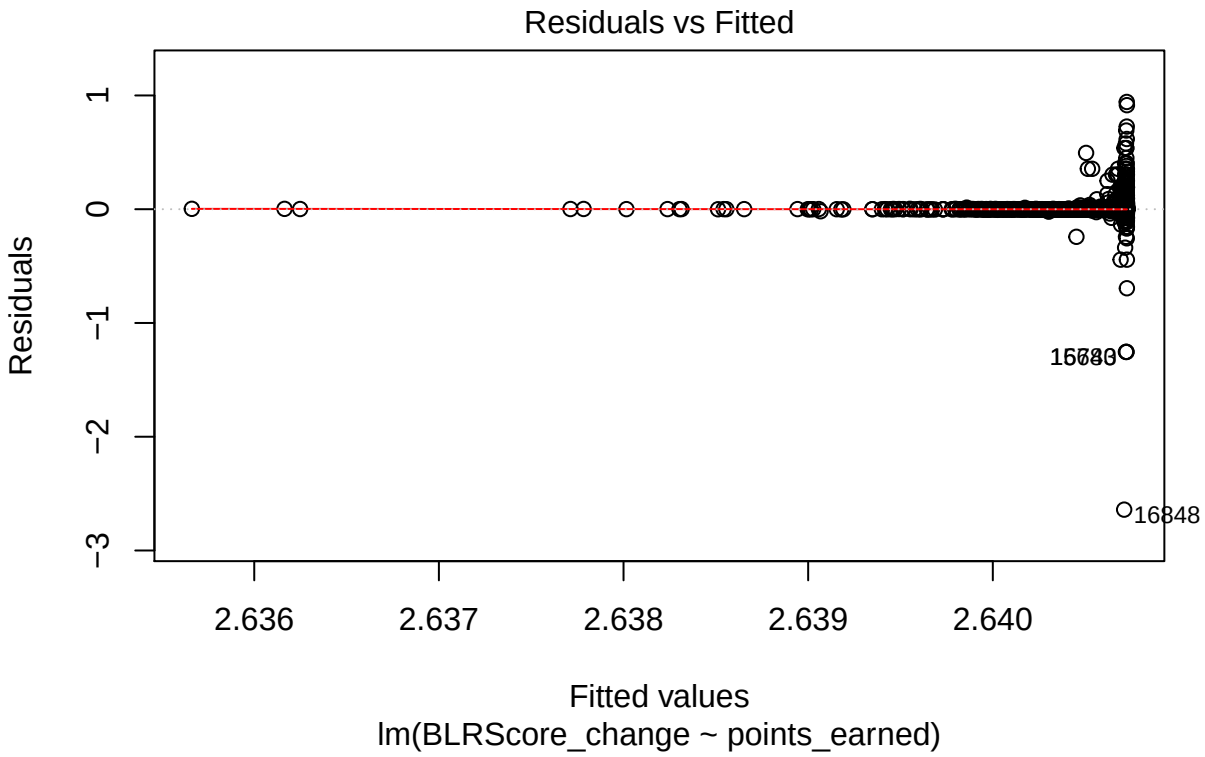

Residuals vs Fitted
Residuals
1
0
−1
−2
−3
16848
2.6403
2.6404
2.6405
2.6406
2.6407
Fitted values
lm(BLRScore_change ~ frequency)

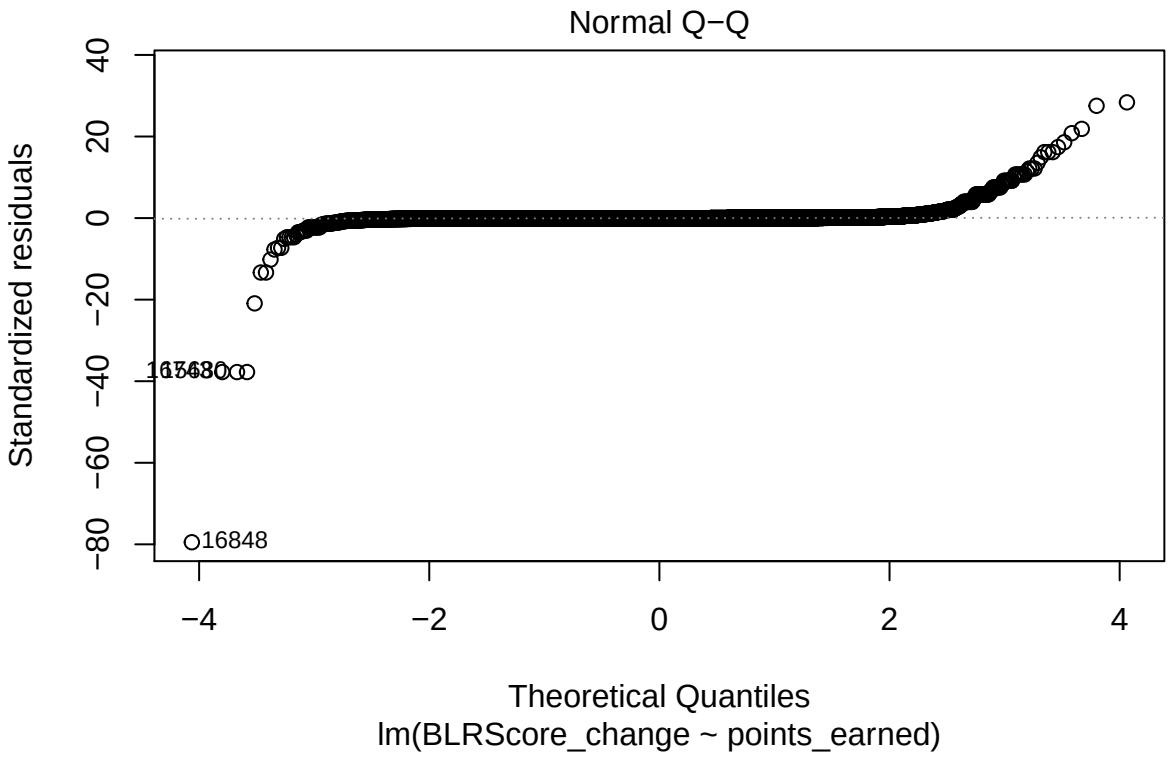

Normal Q−Q
Standardized residuals
20
0
−20
−40
−60
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ frequency)

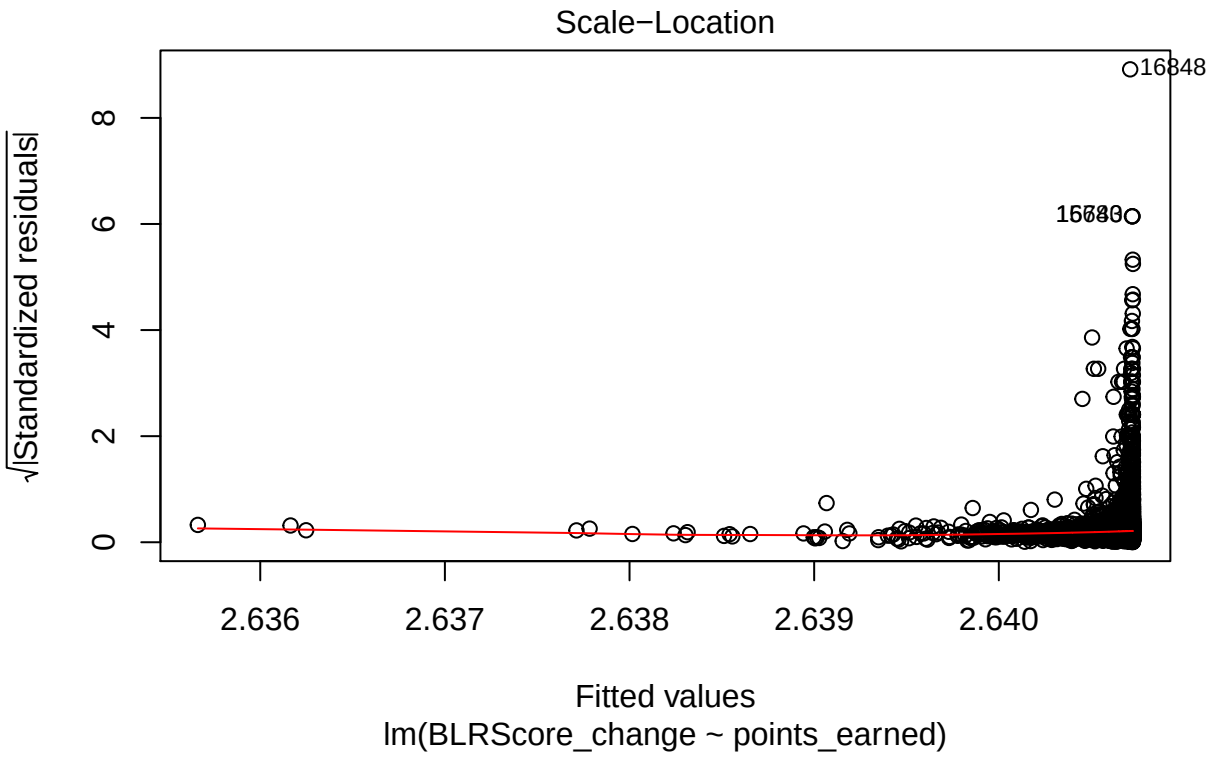

Scale−Location
16848
√|Standardized residuals|
0
2
4
6
8
2.6403
2.6404
2.6405
2.6406
2.6407
Fitted values
lm(BLRScore_change ~ frequency)

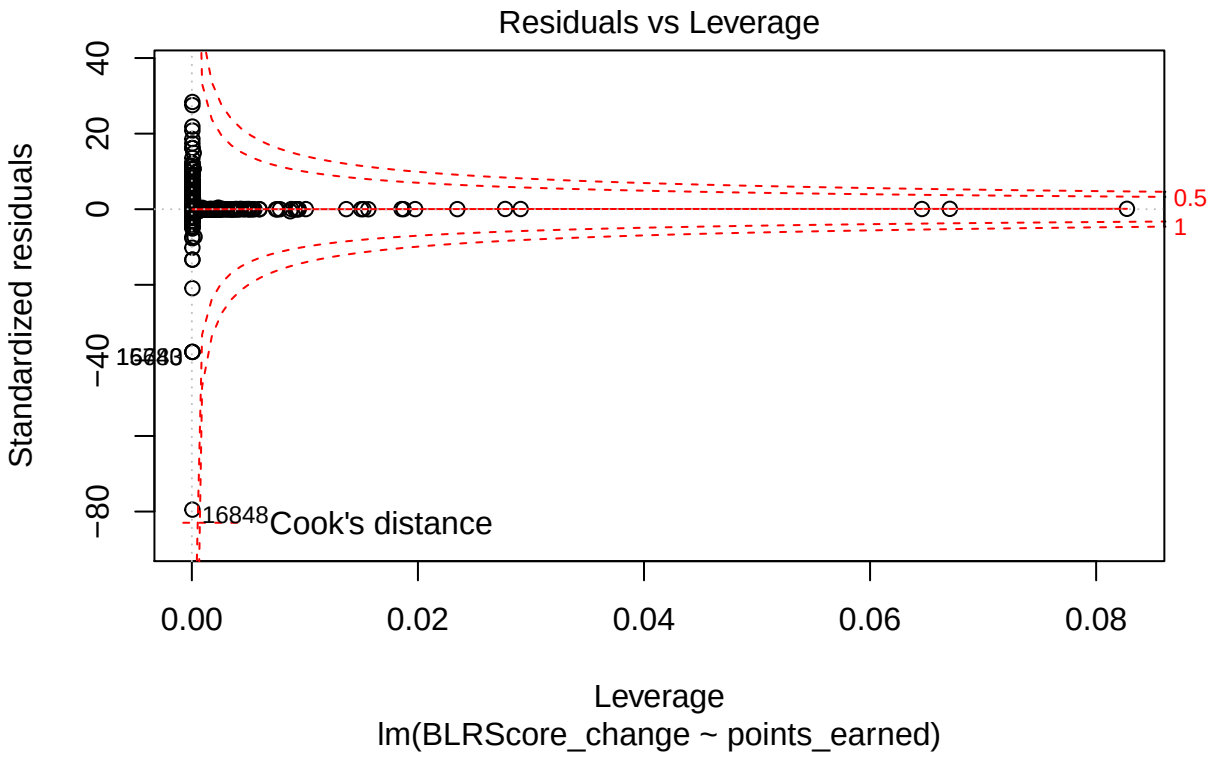



```r
model3 <- lm(BLRScore_change ~ points_earned,data = points_BLR)
summary(model3)
plot(model3)
```

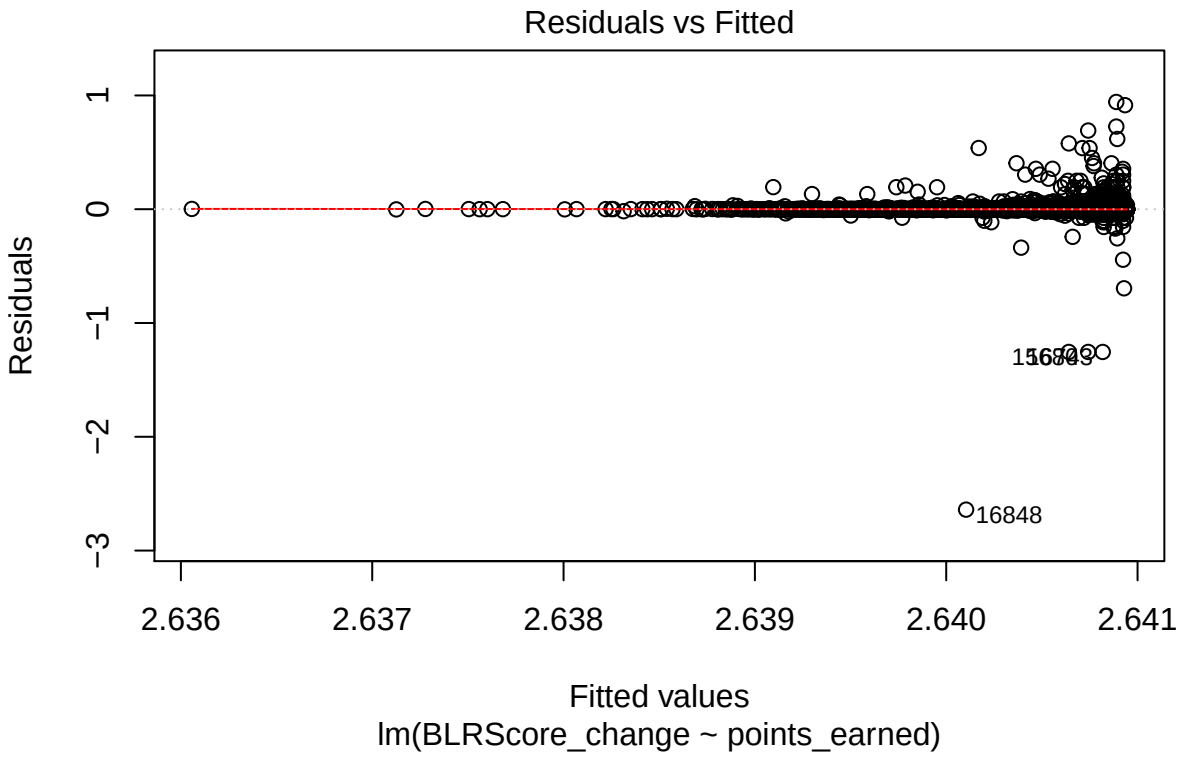

Residuals vs Fitted
Residuals
1
0
−1
−2
−3
16848
2.636
2.637
2.638
2.639
2.640
Fitted values
lm(BLRScore_change ~ points_earned)

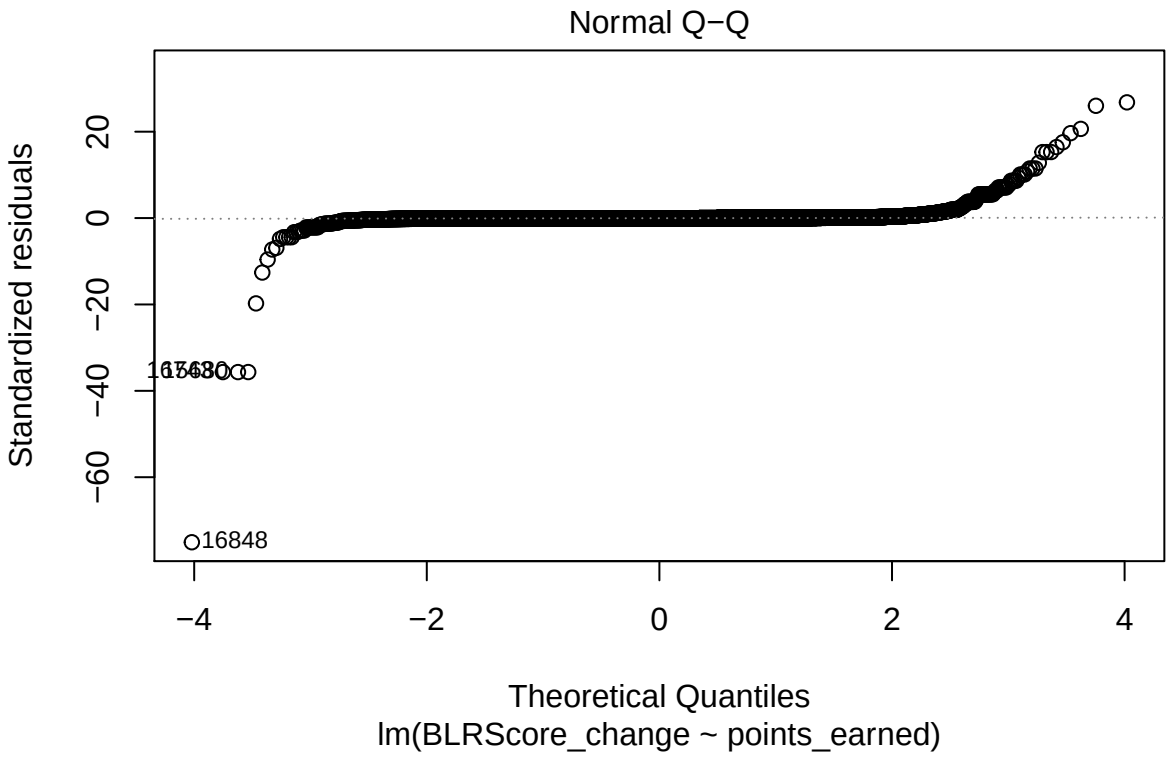

Normal Q−Q
Standardized residuals
40
20
0
−20
−40
−60
−80
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ points_earned)

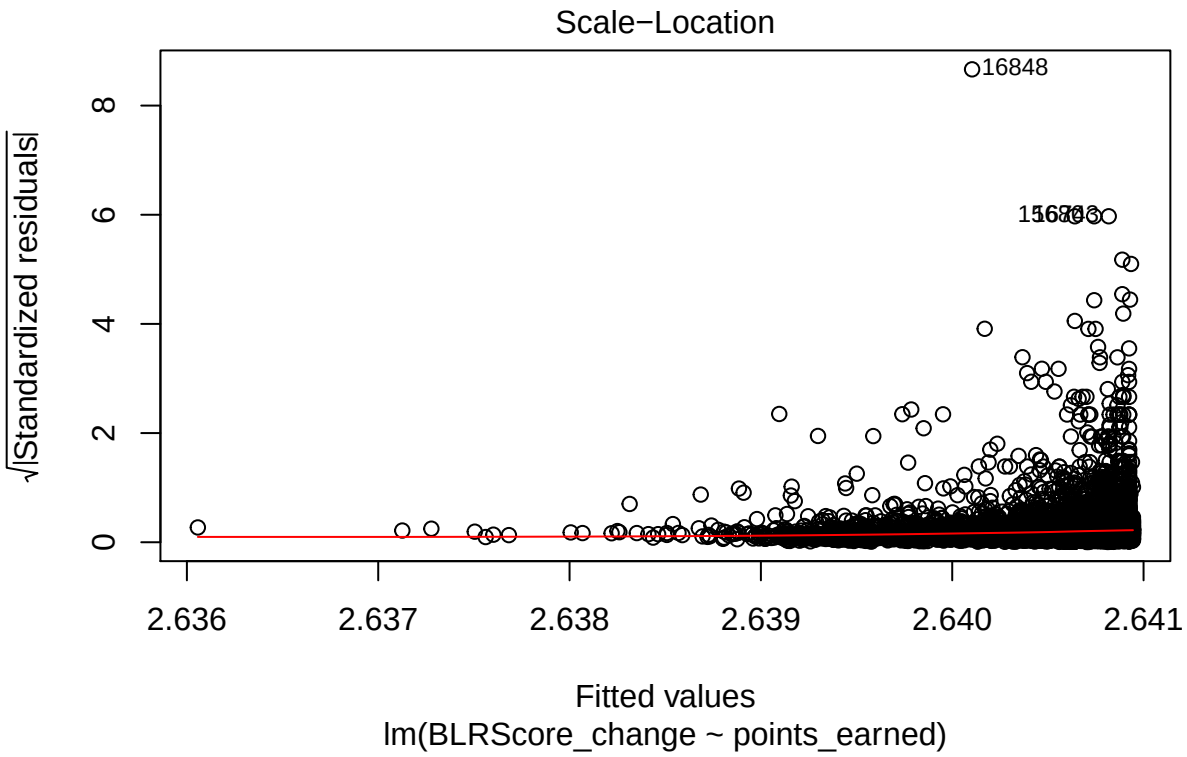

Scale−Location
16848
√|Standardized residuals|
0
2
4
6
8
2.636
2.637
2.638
2.639
2.640
Fitted values
lm(BLRScore_change ~ points_earned)

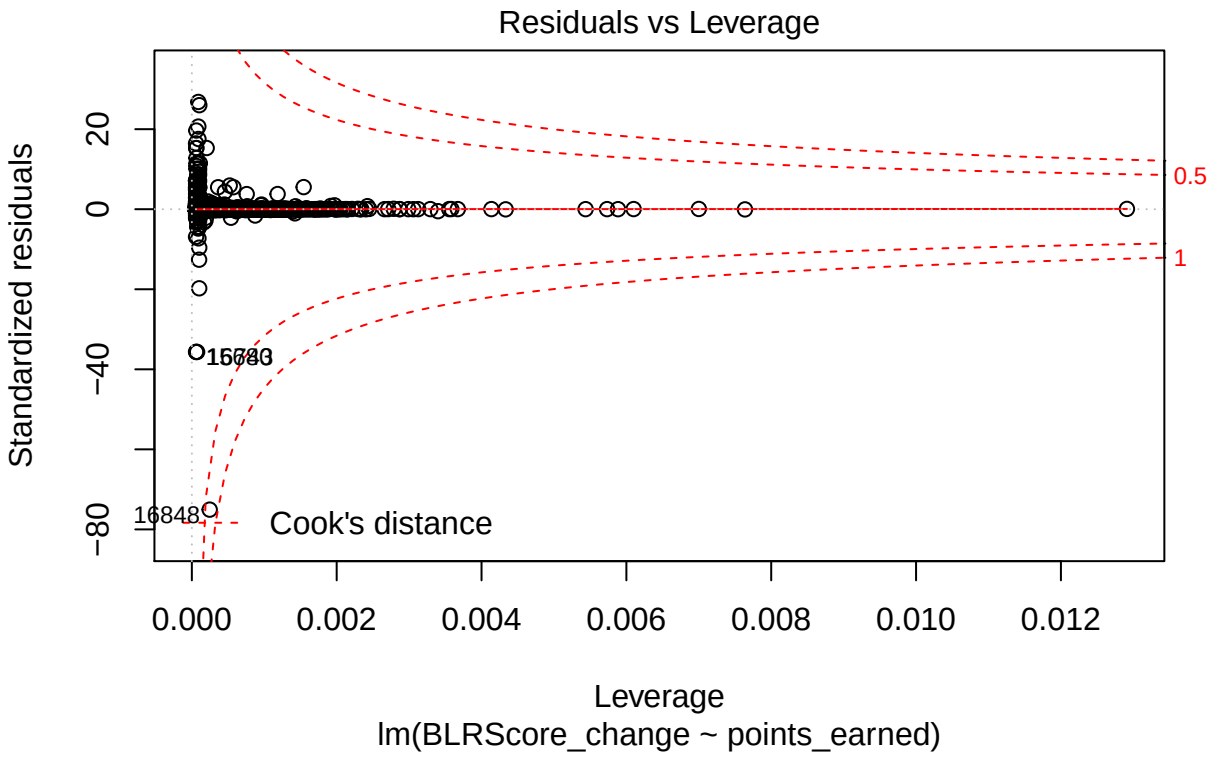



```
model3_rmOutlier <- lm(BLRScore_change ~ points_earned,data = points_BLR_rmOutlier)
summary(model3_rmOutlier)
plot(model3_rmOutlier)
```

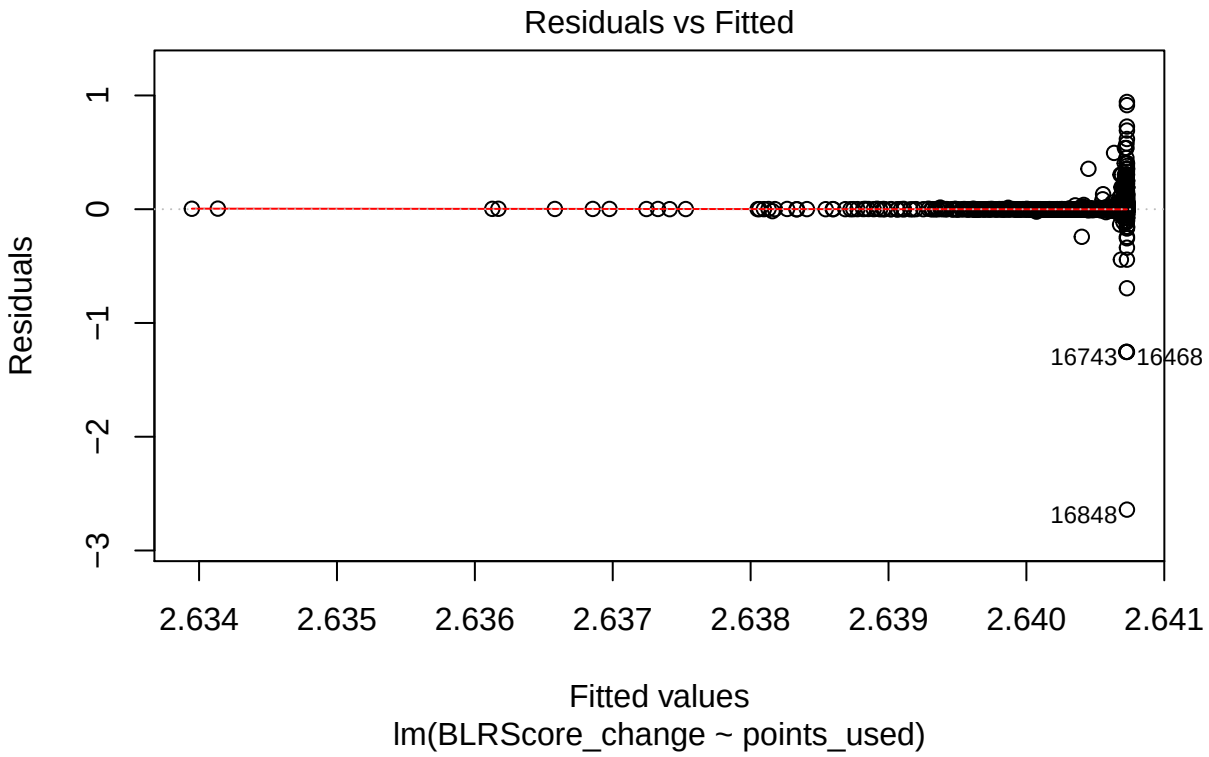

Residuals vs Fitted
Residuals
1
0
−1
−2
−3
2.636
2.637
2.638
2.639
2.640
2.641
16848
Fitted values
lm(BLRScore_change ~ points_earned)

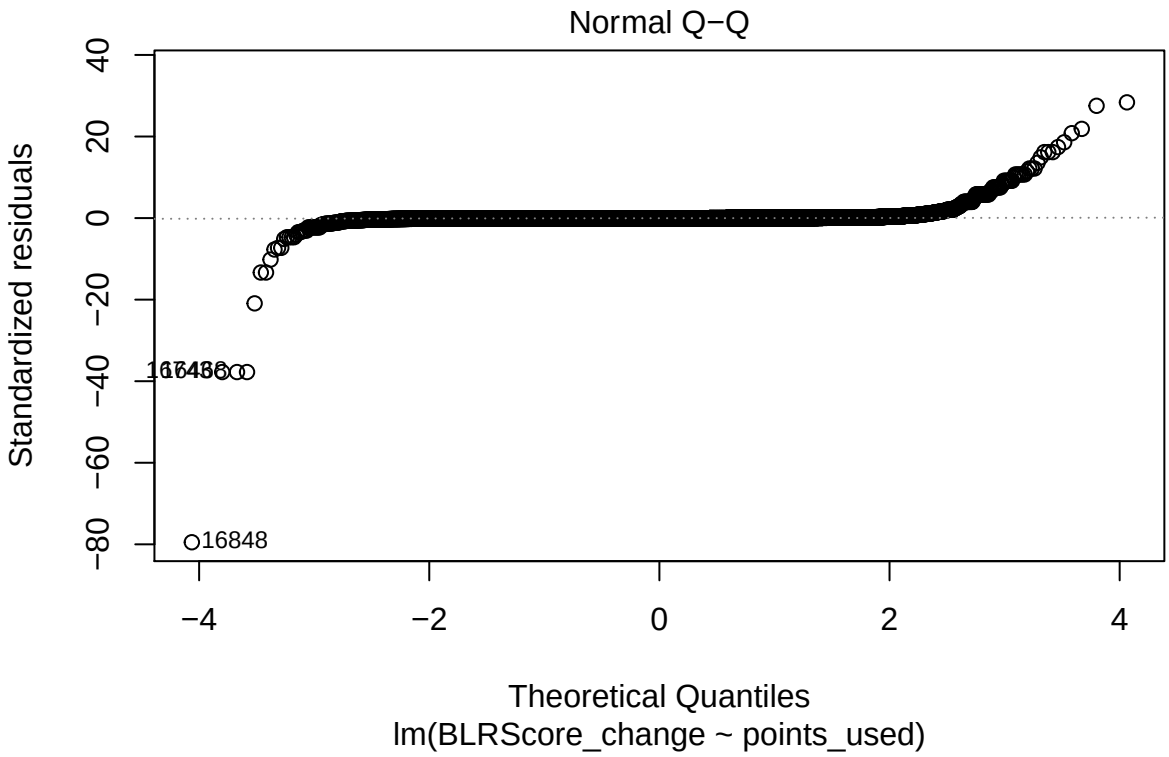

Normal Q−Q
Standardized residuals
20
0
−20
−40
−60
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ points_earned)

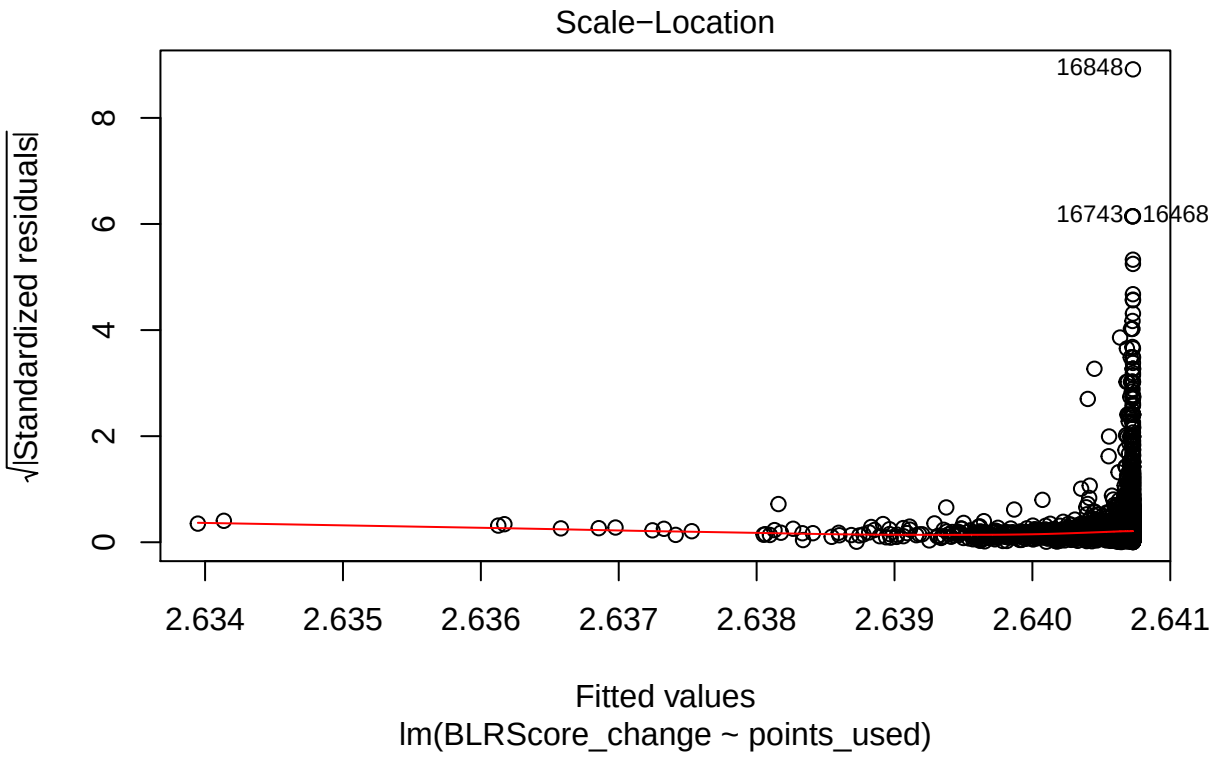

Scale−Location
16848
√|Standardized residuals|
0
2
4
6
8
2.636
2.637
2.638
2.639
2.640
2.641
Fitted values
lm(BLRScore_change ~ points_earned)

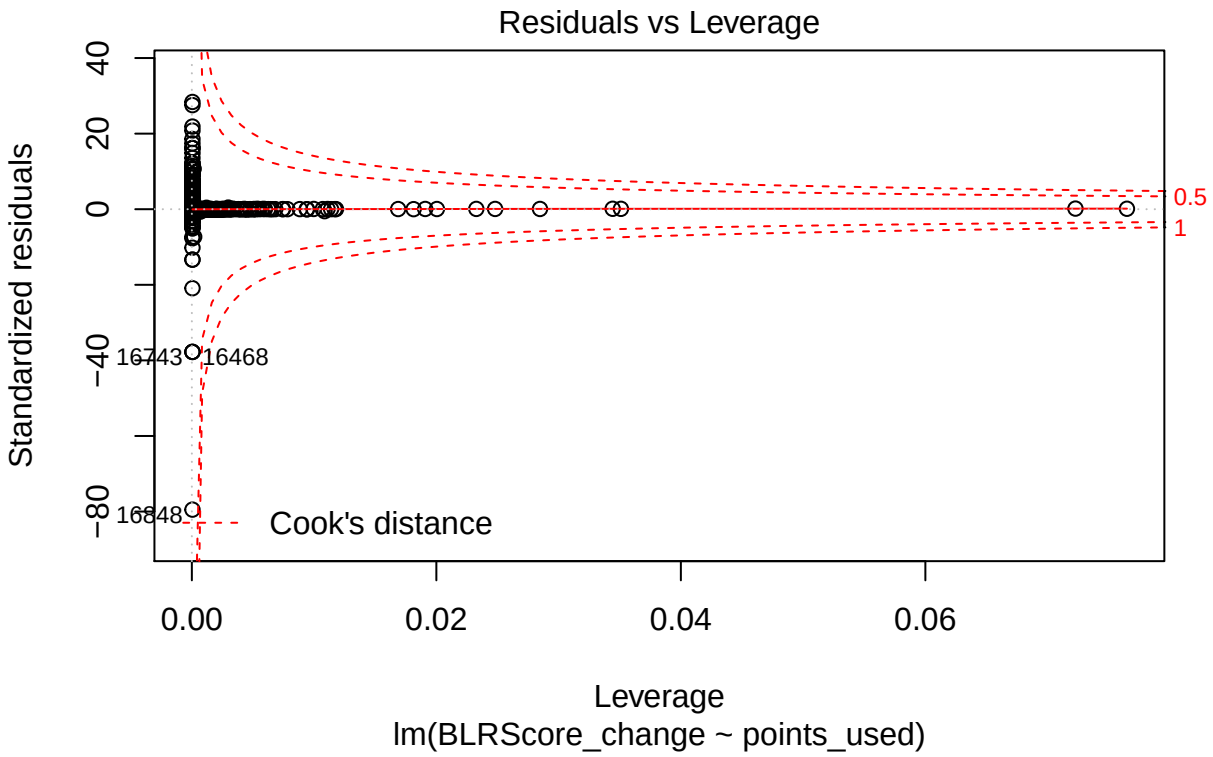



```
model4 <- lm(BLRScore_change ~ points_used,data = points_BLR)
summary(model4)
plot(model4)
```

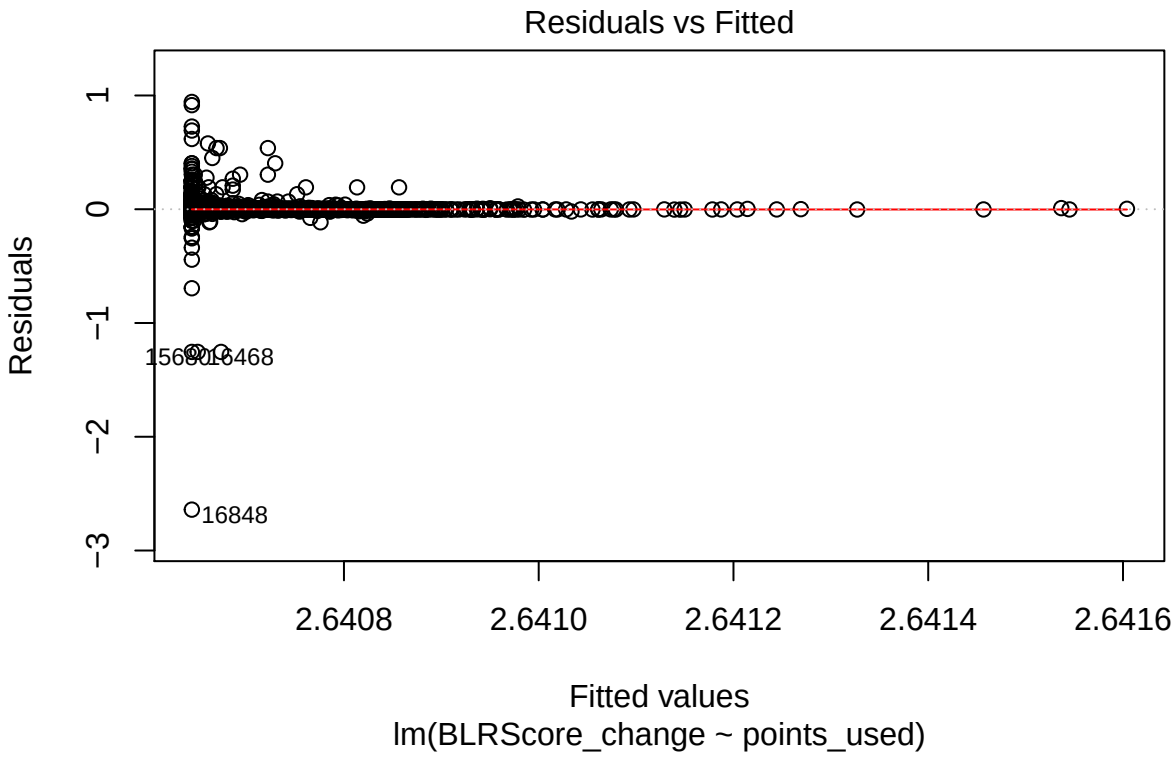

Residuals vs Fitted
Residuals
1
0
−1
−2
−3
16743
16468
16848
2.634
2.635
2.636
2.637
2.638
2.639
2.640
2.641
Fitted values
lm(BLRScore_change ~ points_used)

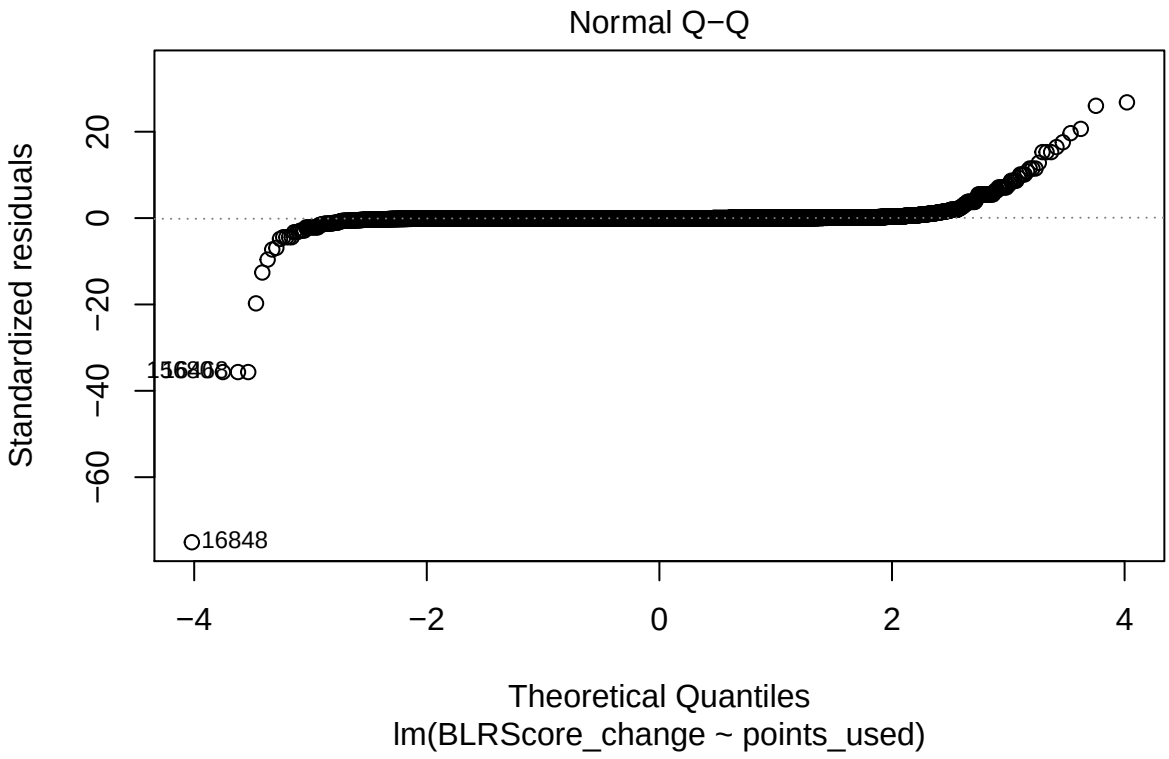

Normal Q−Q
Standardized residuals
40
20
0
−20
−40
−60
−80
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ points_used)

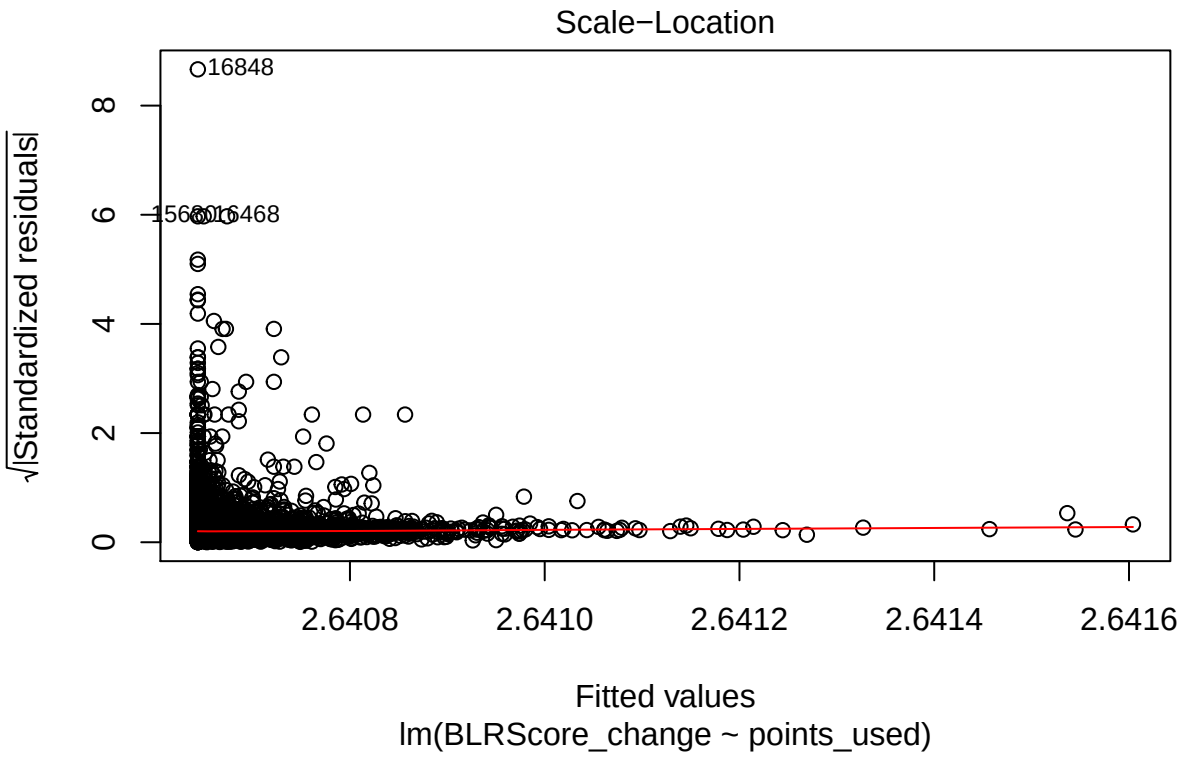

Scale−Location
16848
16743
16468
√|Standardized residuals|
0
2
4
6
8
2.634
2.635
2.636
2.637
2.638
2.639
2.640
2.641
Fitted values
lm(BLRScore_change ~ points_used)

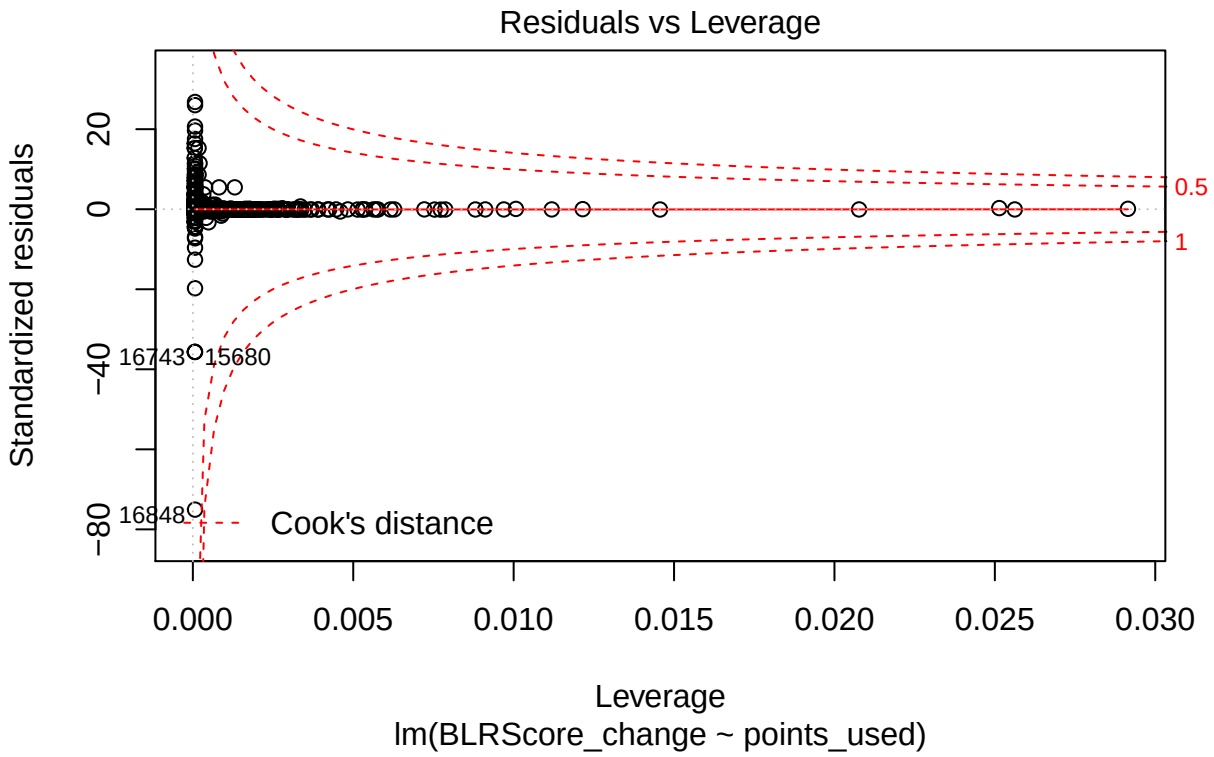



```
model4_rmOutlier <- lm(BLRScore_change ~ points_used,data = points_BLR_rmOutlier)
summary(model4_rmOutlier)
plot(model4_rmOutlier)
```

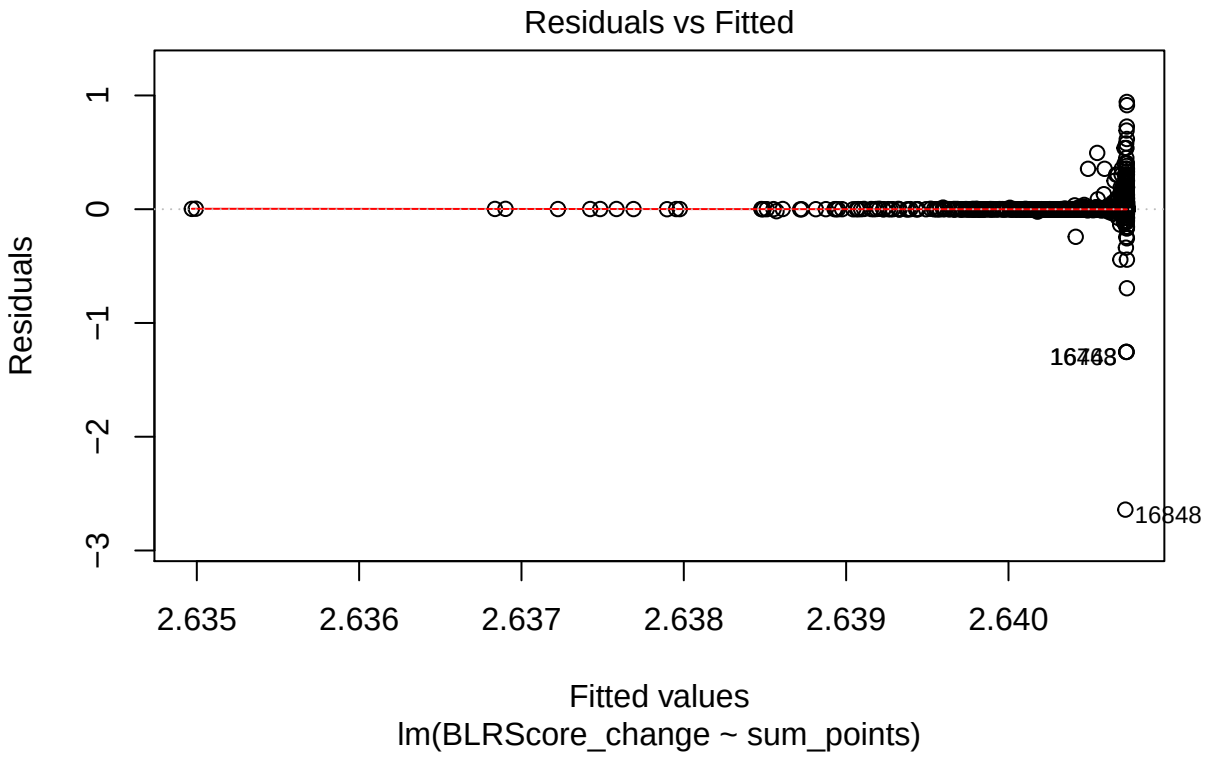

Residuals vs Fitted
Residuals
1
0
−1
−2
−3
15680
16468
16848
2.6408
2.6410
2.6412
2.6414
2.6416
Fitted values
lm(BLRScore_change ~ points_used)

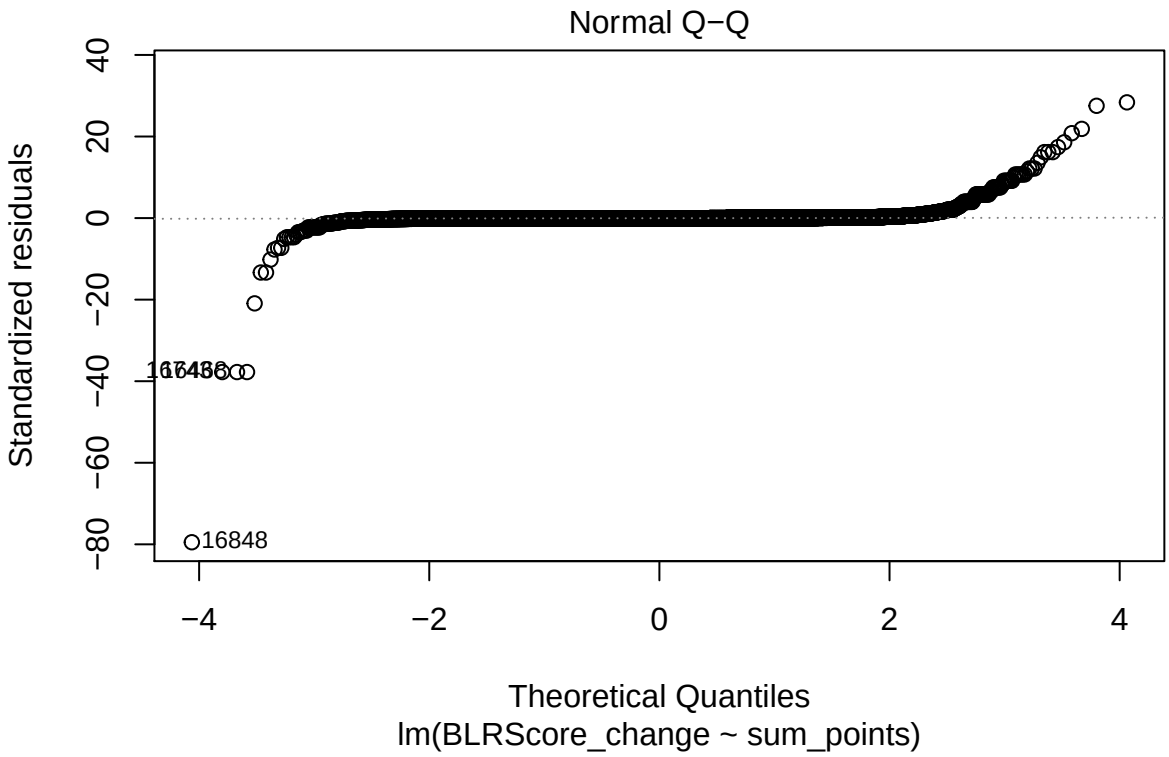

Normal Q−Q
Standardized residuals
20
0
−20
−40
−60
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ points_used)

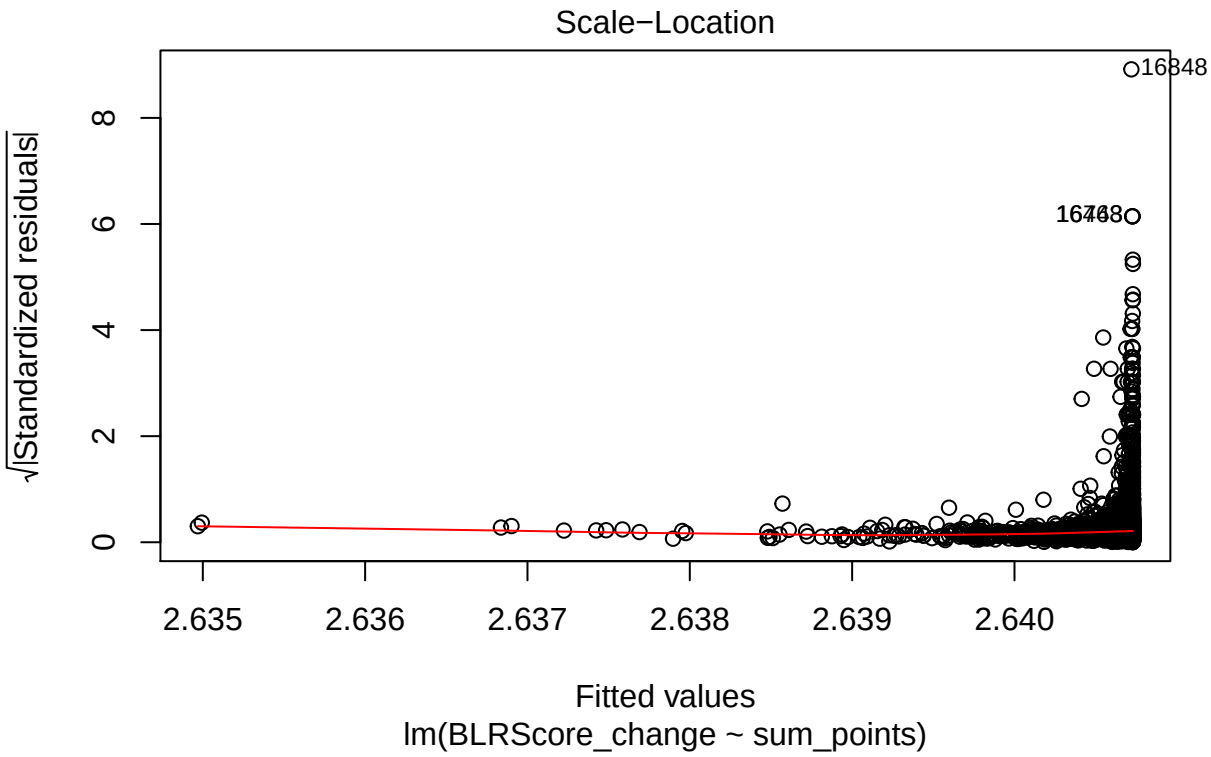

Scale−Location
16848
√|Standardized residuals|
0
2
4
6
8
2.6408
2.6410
2.6412
2.6414
2.6416
Fitted values
lm(BLRScore_change ~ points_used)

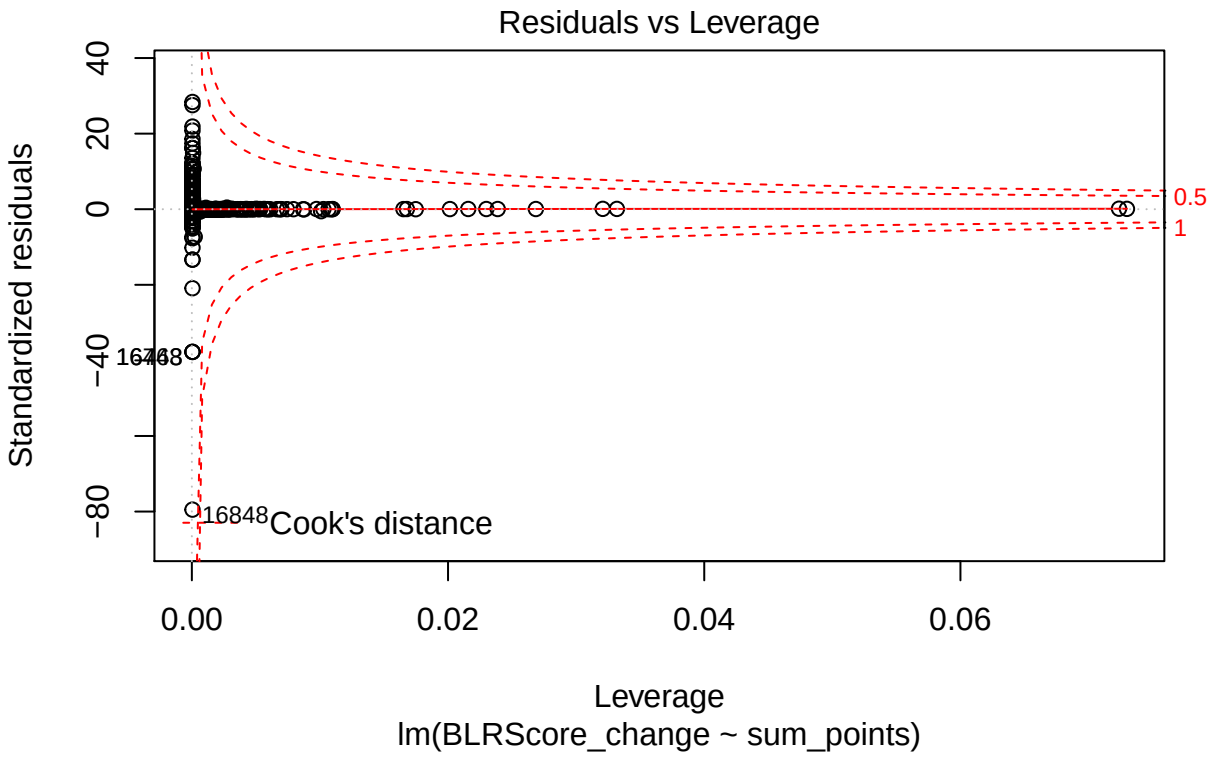



```
model5 <- lm(BLRScore_change ~ sum_points,data = points_BLR)
summary(model5)
plot(model5)
```

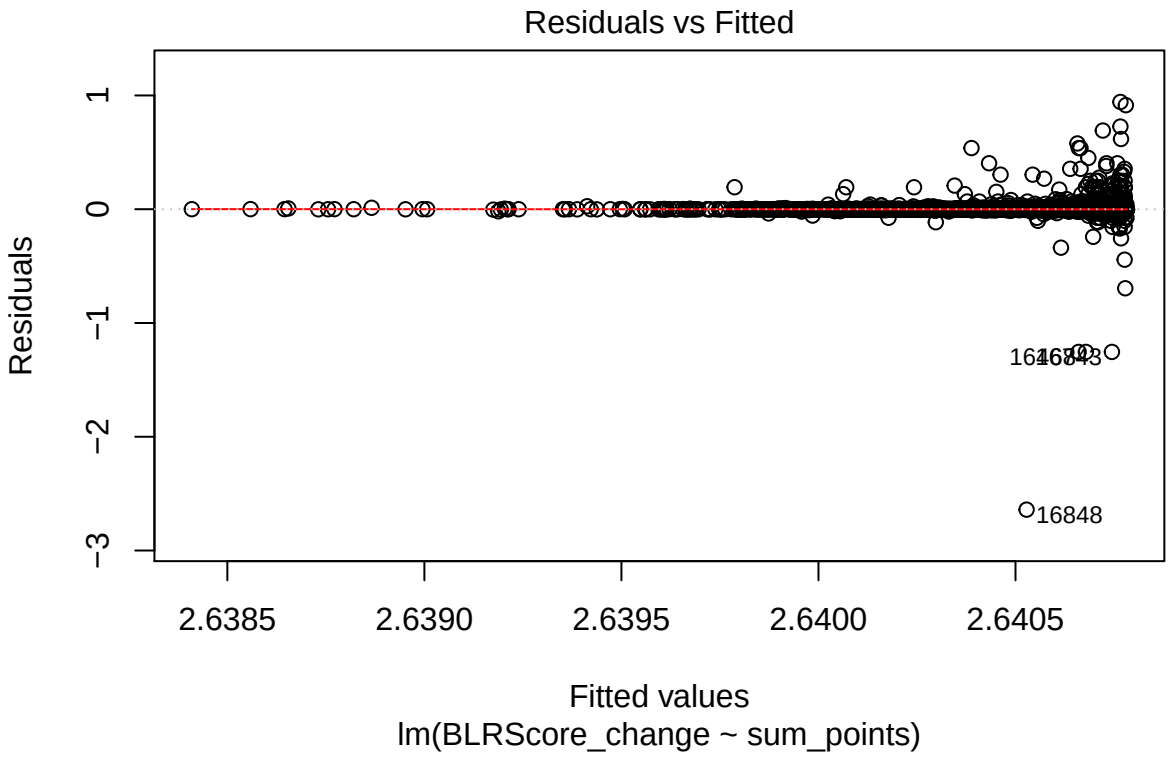

Residuals vs Fitted
Residuals
1
0
−1
−2
−3
16748
16848
2.635
2.636
2.637
2.638
2.639
2.640
Fitted values
lm(BLRScore_change ~ sum_points)

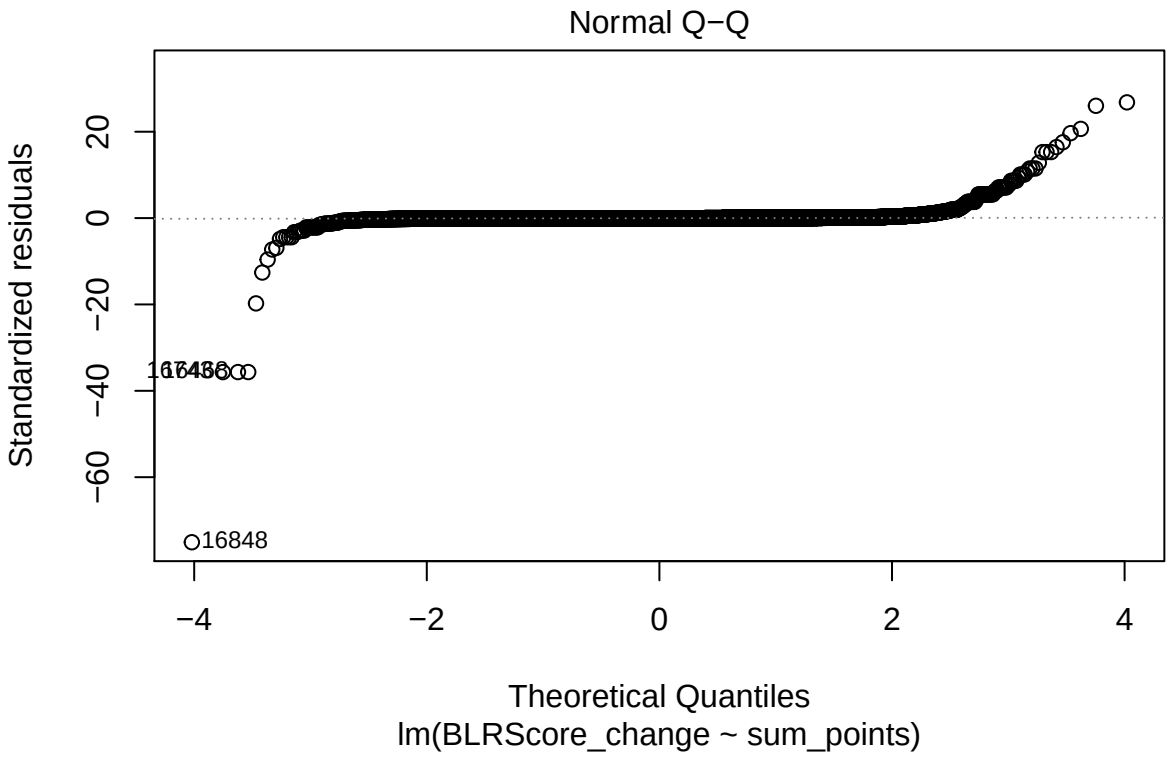

Normal Q−Q
Standardized residuals
40
20
0
−20
−40
−60
−80
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ sum_points)

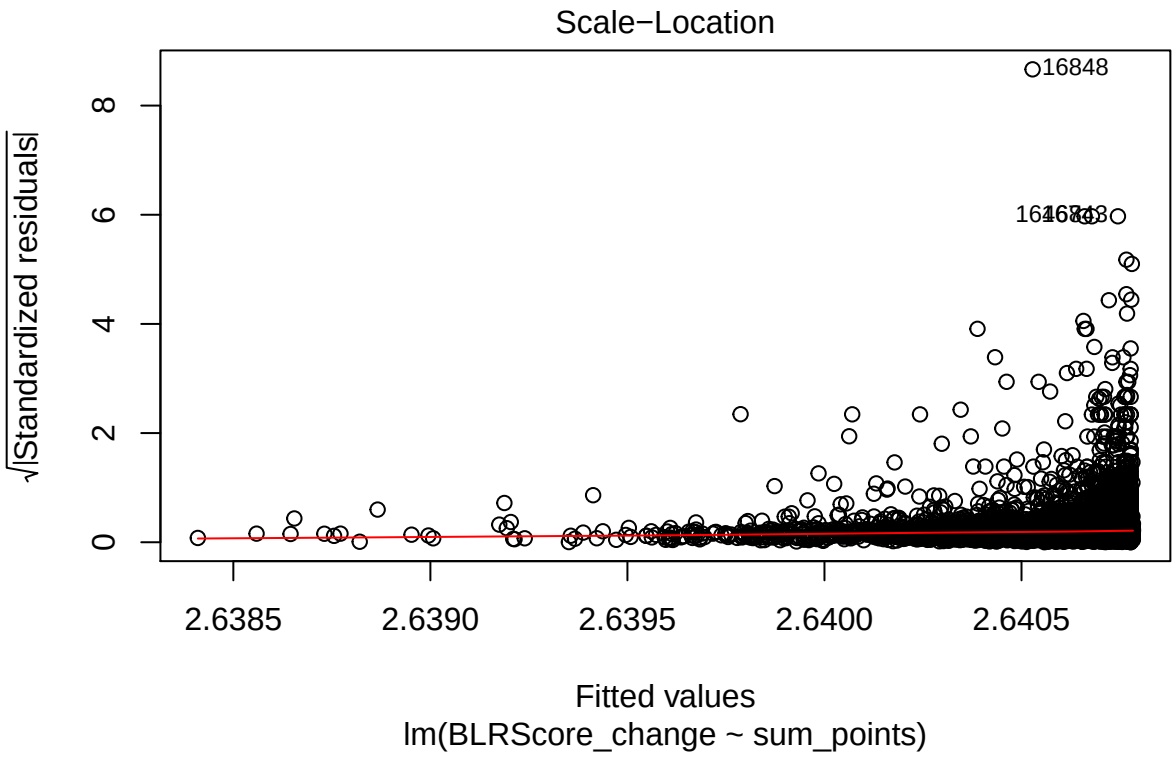

Scale−Location
16348
√|Standardized residuals|
0
2
4
6
8
2.635
2.636
2.637
2.638
2.639
2.640
Fitted values
lm(BLRScore_change ~ sum_points)

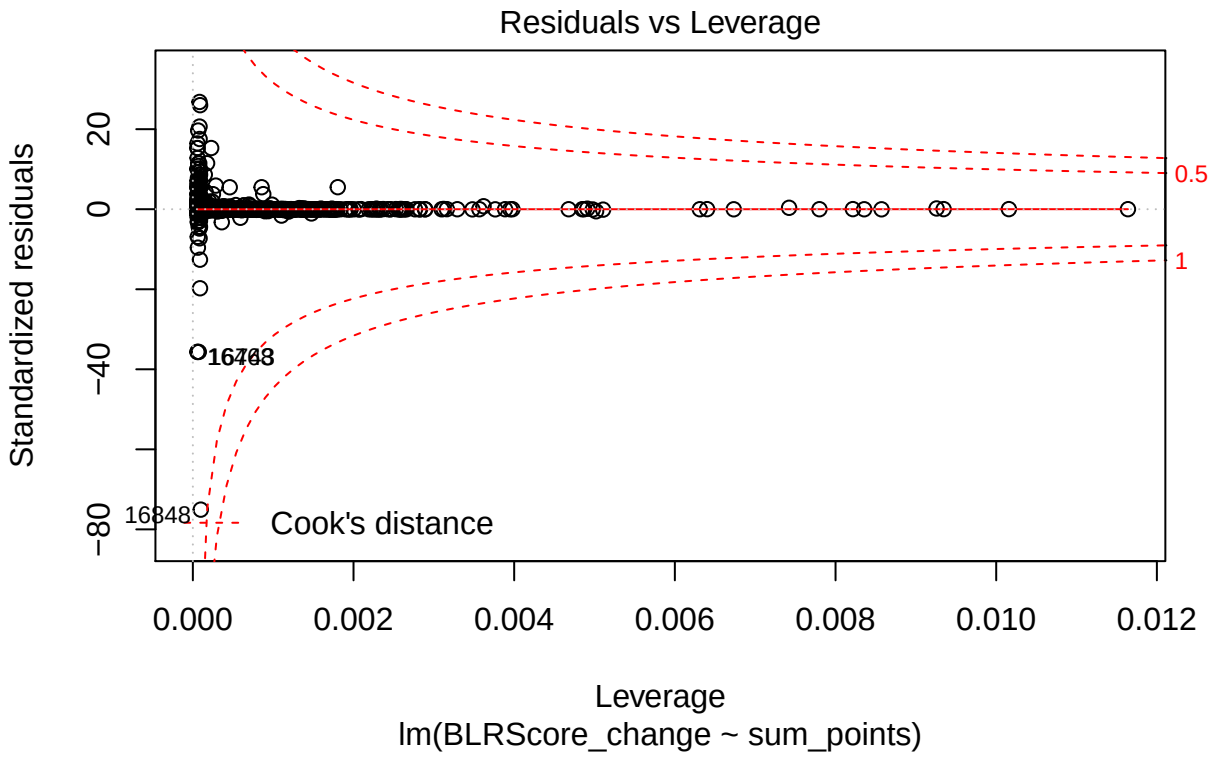



```
model5_rmOutlier <- lm(BLRScore_change ~ sum_points,data = points_BLR_rmOutlier)
summary(model5_rmOutlier)
plot(model5_rmOutlier)
```

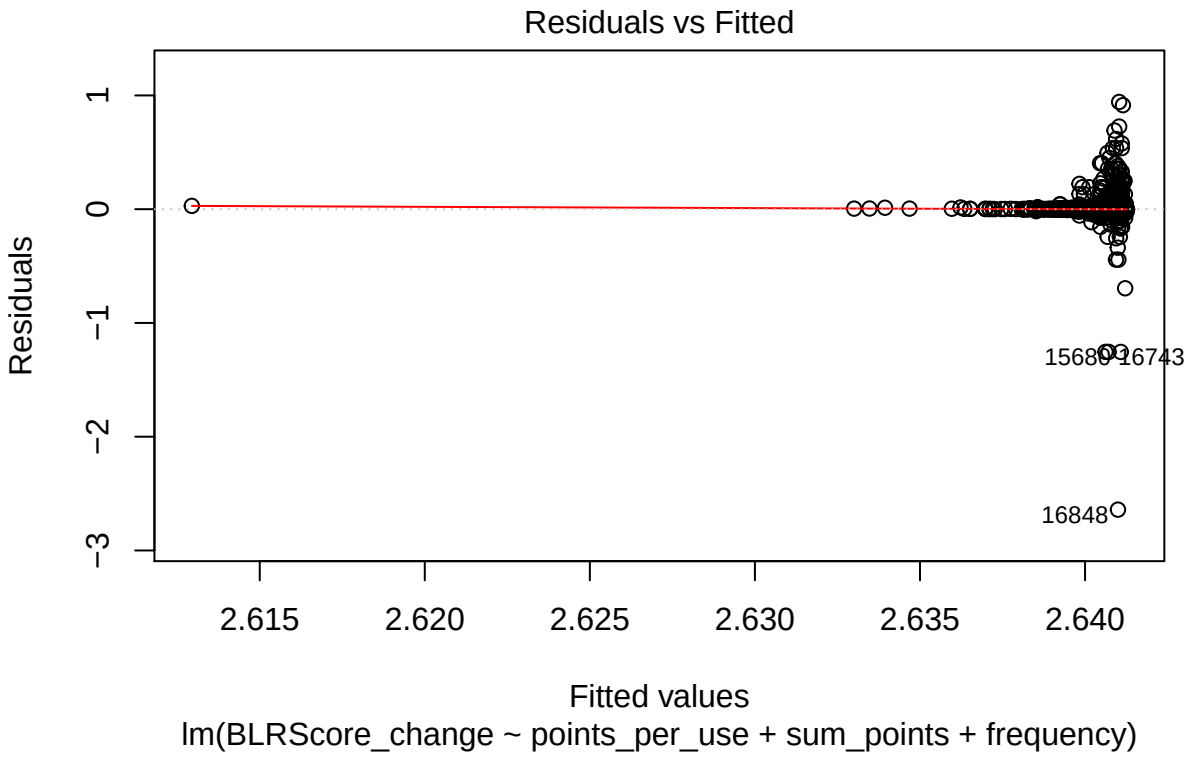

Residuals vs Fitted
Residuals
1
0
−1
−2
−3
16848
2.6385
2.6390
2.6395
2.6400
2.6405
Fitted values
lm(BLRScore_change ~ sum_points)

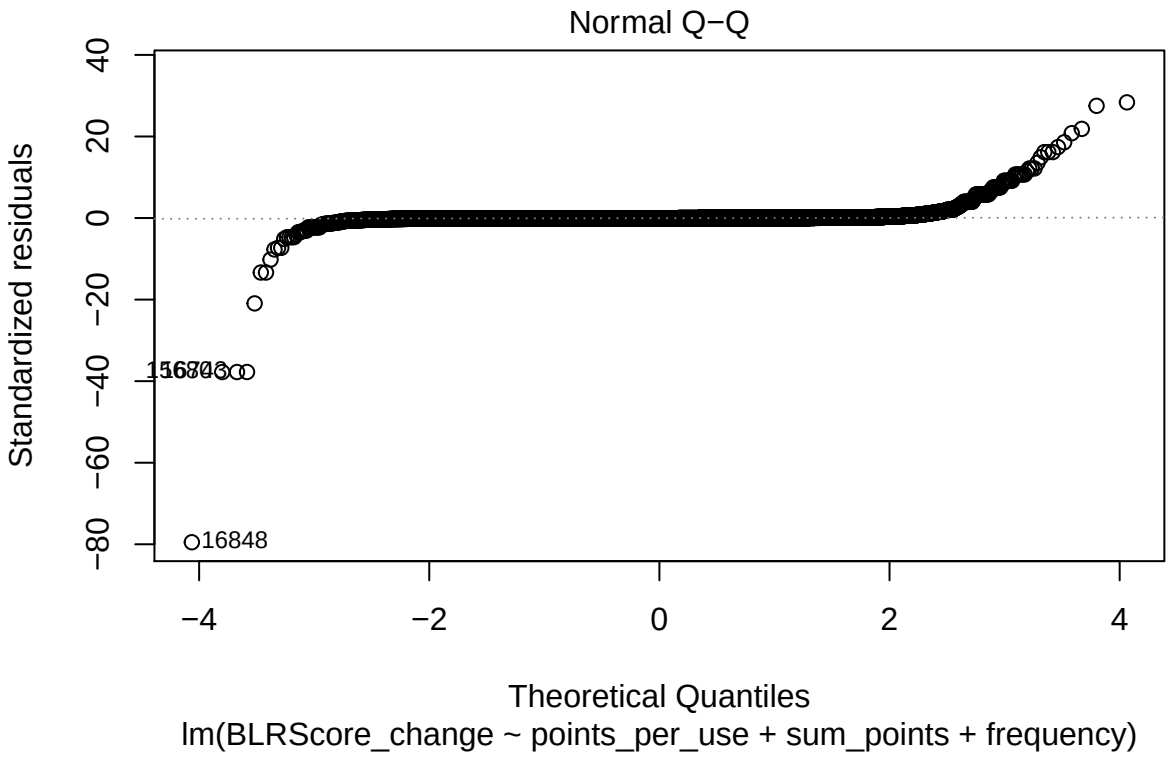

Normal Q−Q
Standardized residuals
20
0
−20
−40
−60
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ sum_points)

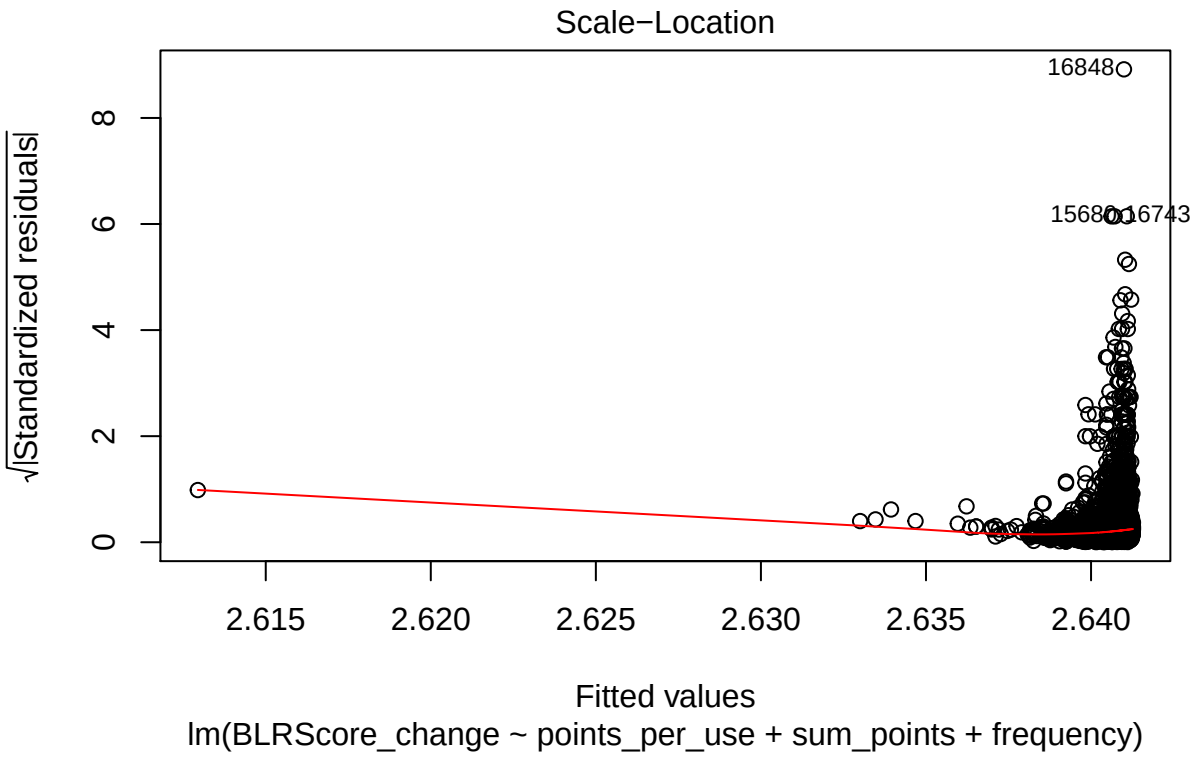

Scale−Location
16848
√|Standardized residuals|
0
2
4
6
8
2.6385
2.6390
2.6395
2.6400
2.6405
Fitted values
lm(BLRScore_change ~ sum_points)

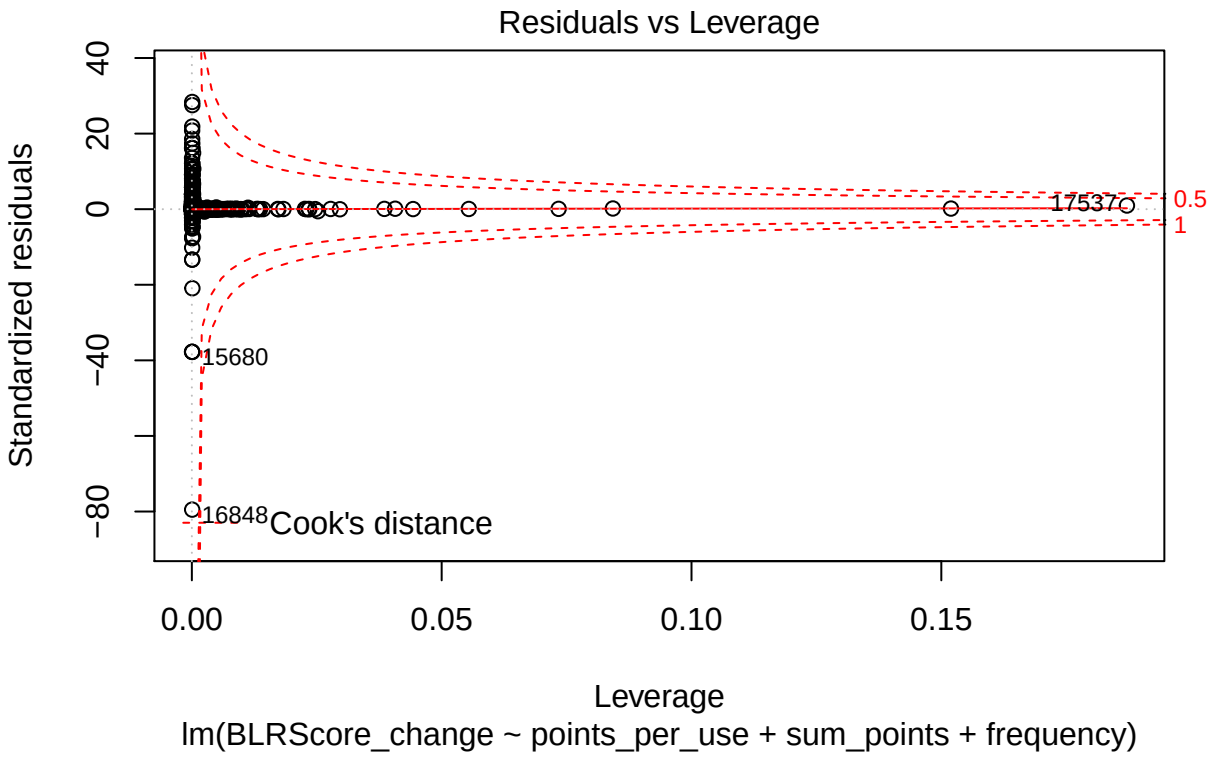



```
model6 <- lm(BLRScore_change ~ points_per_use+sum_points+frequency, data = points_BLR)
summary(model6)
plot(model6)
```


Residuals vs Fitted
Residuals
1
0
−1
−2
−3
15680
16743
16848
2.615
2.620
2.625
2.630
2.635
2.640
Fitted values
lm(BLRScore_change ~ points_per_use + sum_points + frequency)


Normal Q−Q
Standardized residuals
40
20
0
−20
−40
−60
−80
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ points_per_use + sum_points + frequency)

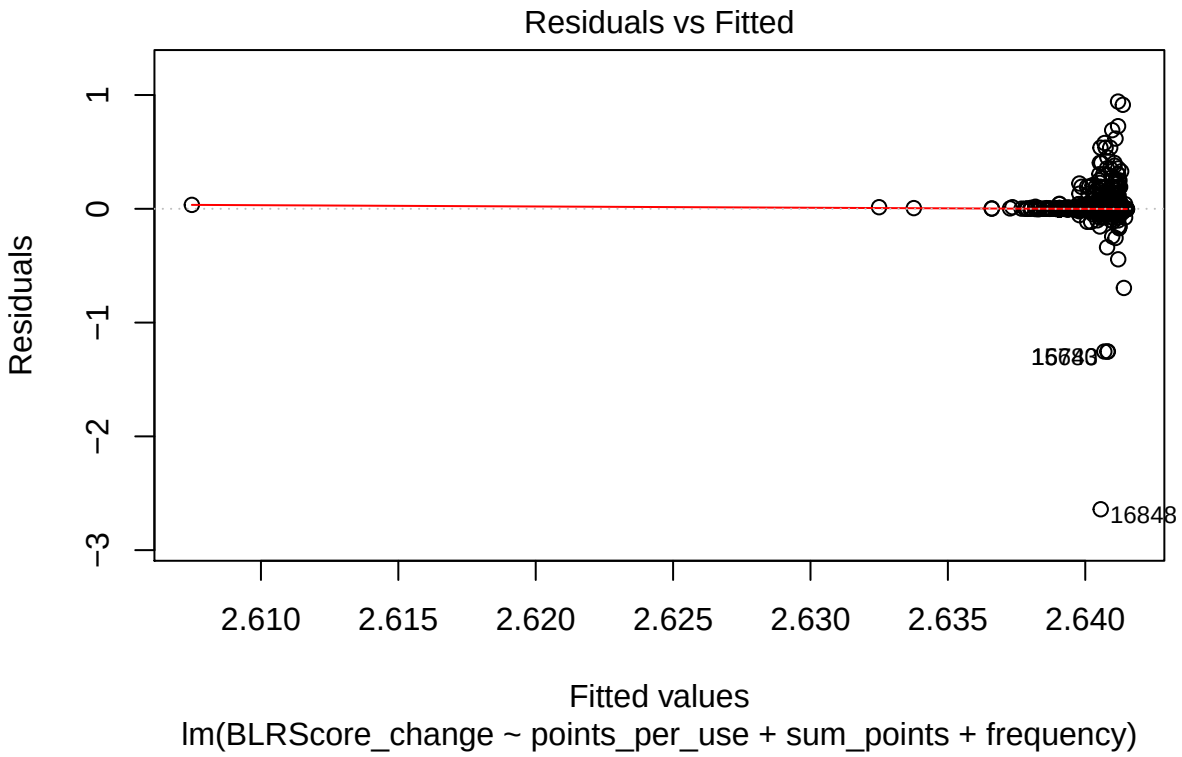

Scale−Location
16848
15680
16743
√|Standardized residuals|
0
2
4
6
8
2.615
2.620
2.625
2.630
2.635
2.640
Fitted values
lm(BLRScore_change ~ points_per_use + sum_points + frequency)

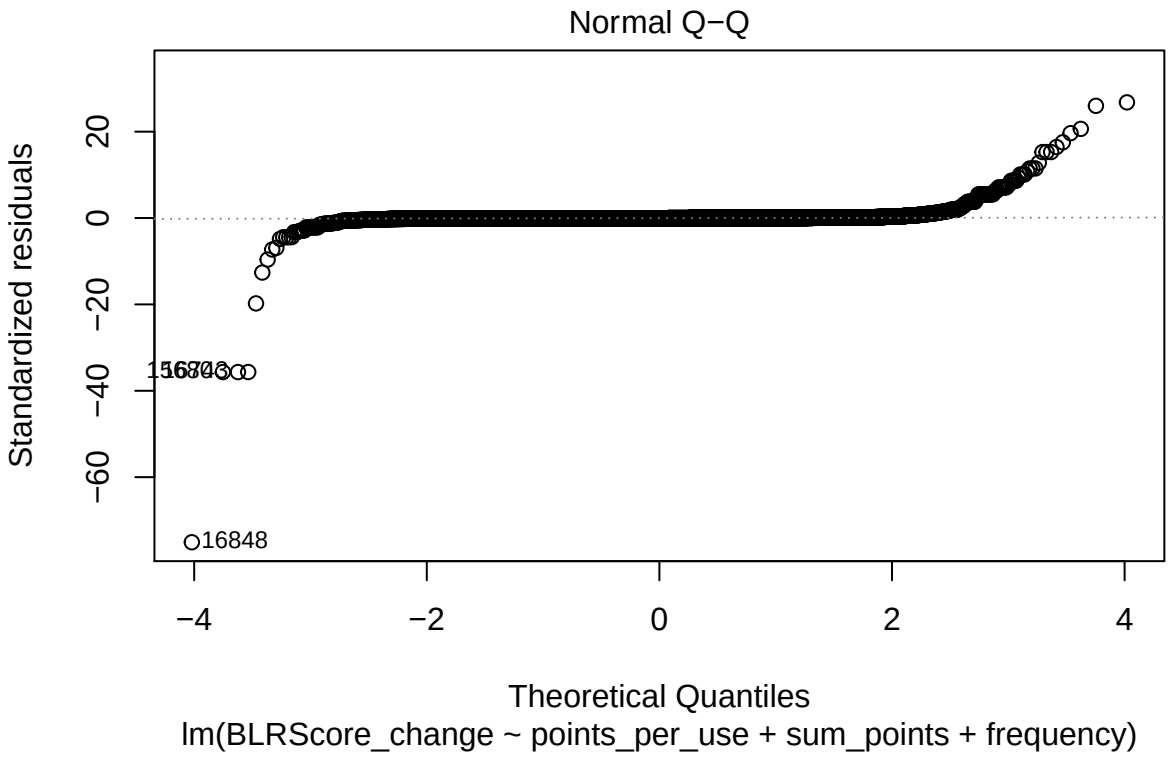


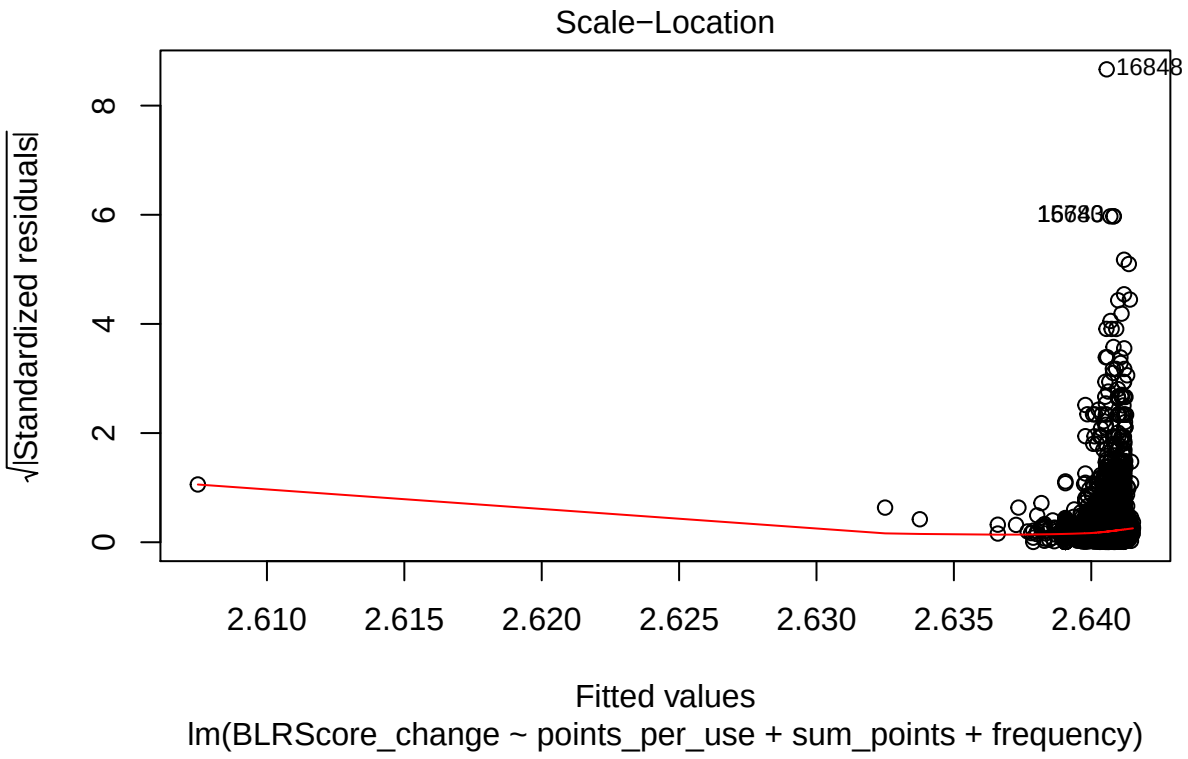

```
model6_rmOutlier <- lm(BLRScore_change ~ points_per_use+sum_points+frequency, data = points_BLR_rmOutlie
summary(model6_rmOutlier)
```

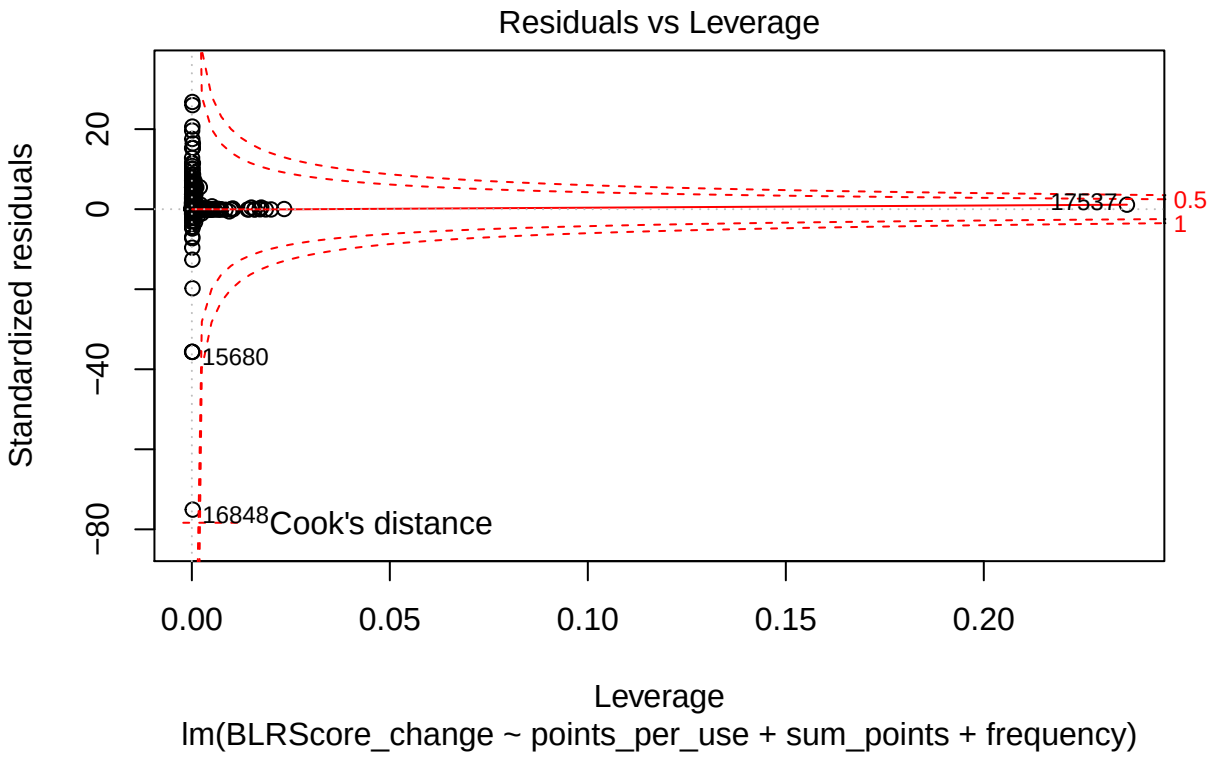
```
##
## Call:
## lm(formula = BLRScore_change ~ points_per_use + sum_points +
##     frequency, data = points_BLR_rmOutlier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64056 -0.00200 -0.00129 -0.00032  0.94233
##
## Coefficients:
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     2.642e+00  5.437e-04 4858.584   <2e-16 ***
## points_per_use -3.427e-06  1.826e-06   -1.876   0.0606 .
## sum_points      7.975e-08  1.838e-07    0.434   0.6643
## frequency      -2.391e-05  2.865e-05   -0.835   0.4040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03518 on 17197 degrees of freedom
## Multiple R-squared:  0.0002277,  Adjusted R-squared:  5.327e-05
## F-statistic: 1.305 on 3 and 17197 DF,  p-value: 0.2707
```

```r
plot(model6_rmOutlier)
```




Normal Q−Q
Standardized residuals
20
0
−20
−40
−60
16848
−4
−2
0
2
4
Theoretical Quantiles
lm(BLRScore_change ~ points_per_use + sum_points + frequency)


Scale−Location
16848
√|Standardized residuals|
0
2
4
6
8
2.610
2.615
2.620
2.625
2.630
2.635
2.640
Fitted values
lm(BLRScore_change ~ points_per_use + sum_points + frequency)




```
lmtest::lrtest(model1,model6)
```

```
## Likelihood ratio test
##
## Model 1: BLRScore_change ~ points_per_use
## Model 2: BLRScore_change ~ points_per_use + sum_points + frequency
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   3  41029
## 2   5  41029  2 0.6706     0.7151
```

Interpretation

Research Question 1

Without sufficient information, it's hard to decide what covariates are supposed to be removed. So I decided to let algorithm do the variable selection. Two models are obtained by using LASSO and backward/forward selection. By LASSO, a simpler model are obtained. Choice of lambda are determined with respect to the graph. Since backward/forward selectio returns full model, I want a simpler model from LASSO and compre these two models, so I picked a value of lambda that may elinminate several covariates, then use likelihod ratio test to compare these two model. Based on the p-value of likelihood ratio test, null hypothesis rejected;thus we prefer the simpler model. However, recall the research question **What characteristics of the users is associated with a higher rate of improvement for the health risk factors.**, we concern more about which covariate has greatest coefficient rather than the exact value of coefficient. In terms of this question, both model returns same result, change in physical activity has the greatest impact on change

in BLR score. In addition, mixed model was built, the mixed model using data normalized with respect to time_interval shows a different result. While the mixed model without such normalization shows similar result. However, both model shows that $\sigma^2/\ (\sigma^2 + \tau^2) = 0$, which means the random effect almost explain no variability in models, thus the mixed model can be ignored.

Research Question 2

Though all covariates has a negative coefficients, the value of coefficients are very close to zero and the p-value is not significant, which suggest that the coefficient are actually zero and the corresponding covariate has no effect on response variable BLRScore_change. Then added a new covariates, points_per_use, stands for average changes in account points(either use or earn) with respect to frequency of using the platform. The coefficient are negative and significant but close to zero. This means increase in average points change will lead to BLR score decrease. A reasonble interpretation of this coefficient is given here. High averge points change means more time spends on electronical device thus less physical activity, therefore BLR decrease. Alternatively, spending more time on reward from BLR platform may imply a bad financial status, so a lower financial score, which has a postive coefficient refer to research question 1; thus decrease BLR score. Nonetheless, since the value is extremely close to zero, impact of platform usage on BLR scores is not so significant that requires attention.

In conclusion, change in physical activity has the greatest impact on change in BLR, and frequency of usage of platform has no effect on BLR change.

I set most output to echo = FALSE, it can be viewed by run the code.