Previsão das notas finais de alunos

Silva, Guilherme Aquino

26/10/2021

Criando um modelo para previsão das notas finais de alunos através dos dados disponíveis no dataset Student Performance Dataset (https://archive.ics.uci.edu/ml/datasets/Student+Performance).

1. Carregando o dataset

```
df <- read.csv2('estudantes.csv')
```

2. Pacotes utilizados

```
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(corrplot)
```

```
## corrplot 0.90 loaded
```

```
library(ggplot2)
library(ggthemes)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

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

```
library(caTools)
```

3. Explorando os dados

```
View(df)
```

```
summary(df)
```

```
##  school   sex          age       address famsize   Pstatus      Medu      
##  GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   Min.   :0.000  
##  MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1st Qu.:2.000  
##                   Median :17.0                             Median :3.000  
##                   Mean   :16.7                             Mean   :2.749  
##                   3rd Qu.:18.0                             3rd Qu.:4.000  
##                   Max.   :22.0                             Max.   :4.000  
##       Fedu             Mjob           Fjob            reason      guardian  
##  Min.   :0.000   at_home : 59   at_home : 20   course    :145   father: 90  
##  1st Qu.:2.000   health  : 34   health  : 18   home      :109   mother:273  
##  Median :2.000   other   :141   other   :217   other     : 36   other : 32  
##  Mean   :2.522   services:103   services:111   reputation:105               
##  3rd Qu.:3.000   teacher : 58   teacher : 29                                
##  Max.   :4.000                                                              
##    traveltime      studytime        failures      schoolsup famsup     paid    
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344   no :153   no :214  
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51   yes:242   yes:181  
##  Median :1.000   Median :2.000   Median :0.0000                                
##  Mean   :1.448   Mean   :2.035   Mean   :0.3342                                
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000                                
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000                                
##  activities nursery   higher    internet  romantic      famrel     
##  no :194    no : 81   no : 20   no : 66   no :263   Min.   :1.000  
##  yes:201    yes:314   yes:375   yes:329   yes:132   1st Qu.:4.000  
##                                                     Median :4.000  
##                                                     Mean   :3.944  
##                                                     3rd Qu.:5.000  
##                                                     Max.   :5.000  
##     freetime         goout            Dalc            Walc      
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000  
##  Median :3.000   Median :3.000   Median :1.000   Median :2.000  
##  Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291  
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000  
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
##      health         absences            G1              G2       
##  Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00  
##  1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00  
##  Median :4.000   Median : 4.000   Median :11.00   Median :11.00  
##  Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71  
##  3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00  
##  Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00  
##        G3       
##  Min.   : 0.00  
##  1st Qu.: 8.00  
##  Median :11.00  
##  Mean   :10.42  
##  3rd Qu.:14.00  
##  Max.   :20.00
```

```
str(df)
```

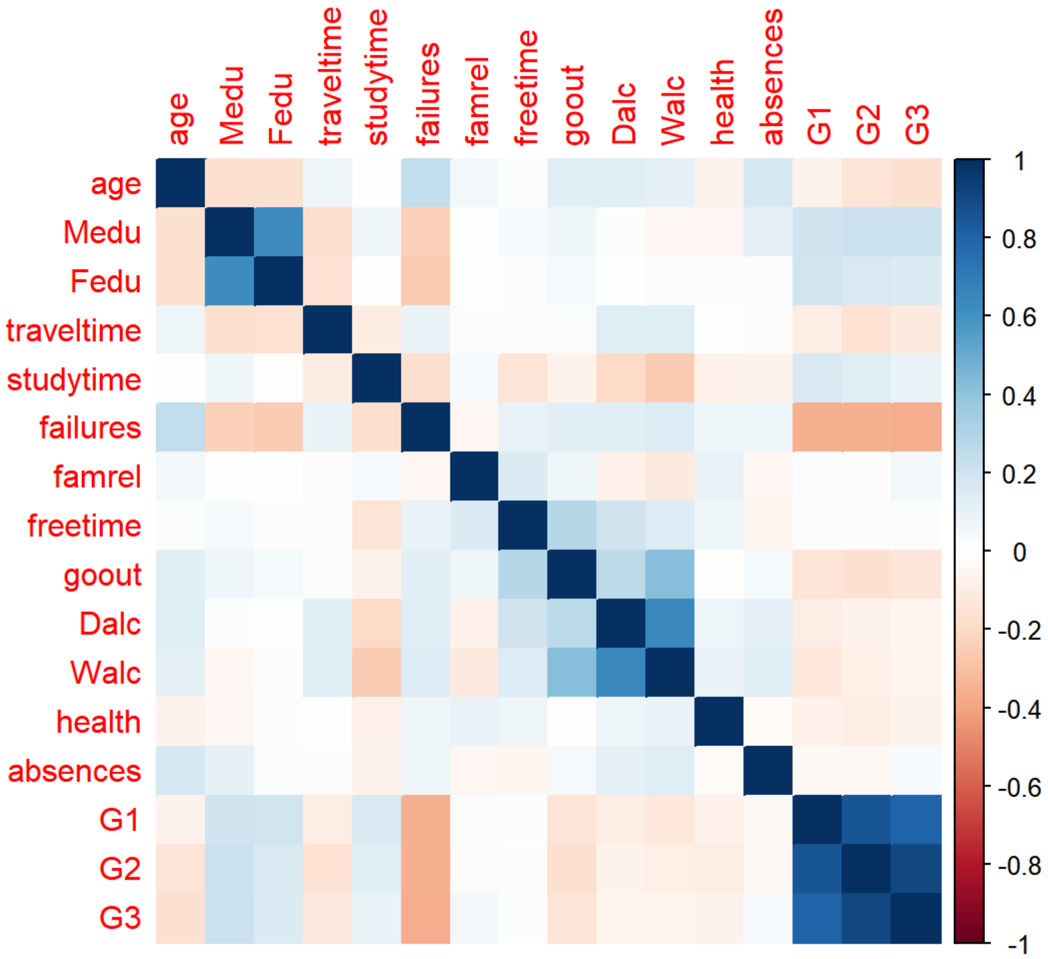
```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
##  $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
##  $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
##  $ Fjob      : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
##  $ reason    : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
##  $ guardian  : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
##  $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
##  $ paid      : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 2 2 ...
##  $ activities: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
##  $ nursery   : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
##  $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ internet  : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
##  $ romantic  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

```
any(is.na(df)) # verificação de valores NA no dataset
```

```
## [1] FALSE
```

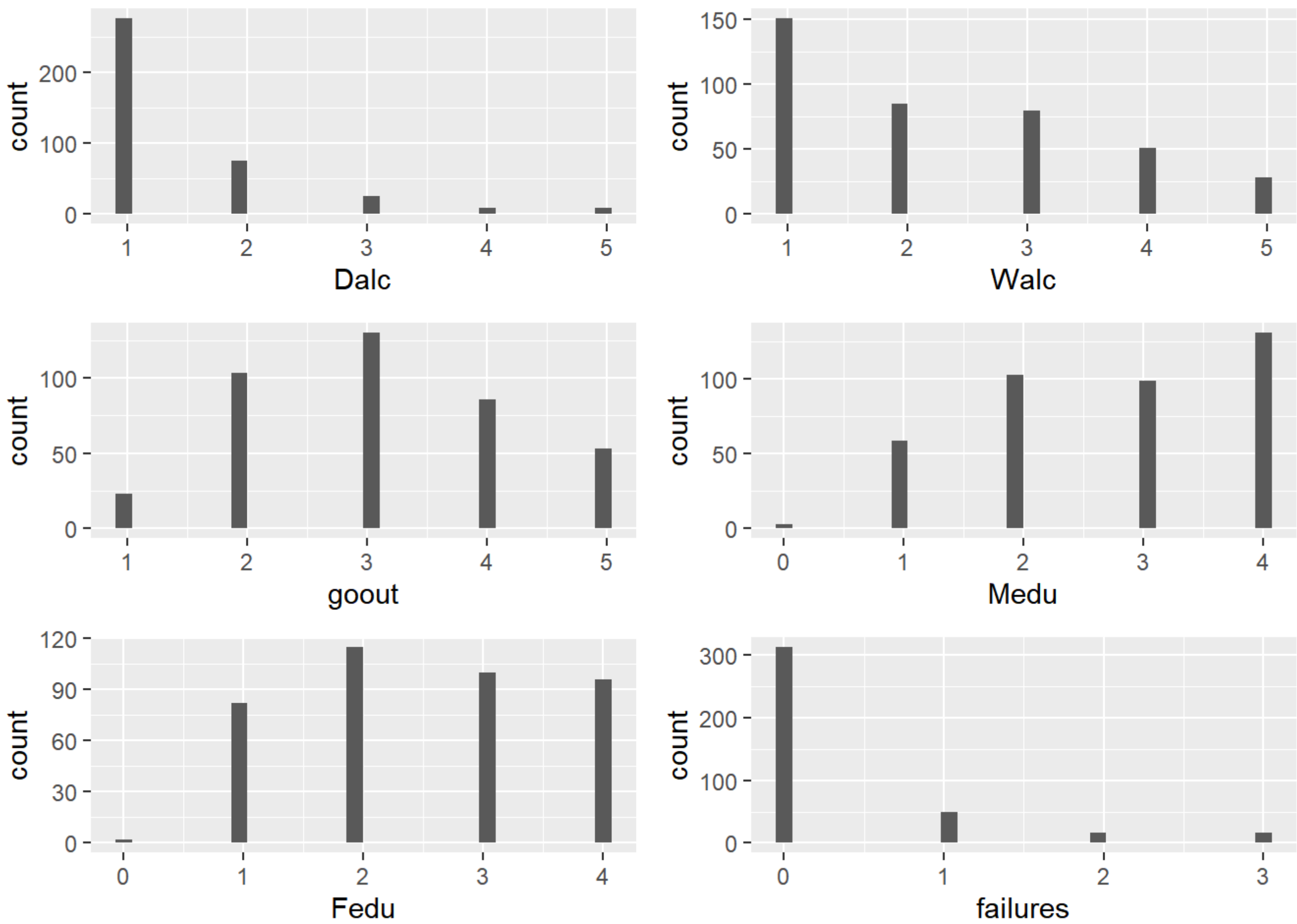
4. Verificando a correlação entre as colunas numéricas

```
col_numericas <- sapply(df, is.numeric) # extraindo as colunas numéricas

corrplot(cor(df[, col_numericas]), method = 'color') # plotando a correlação
```



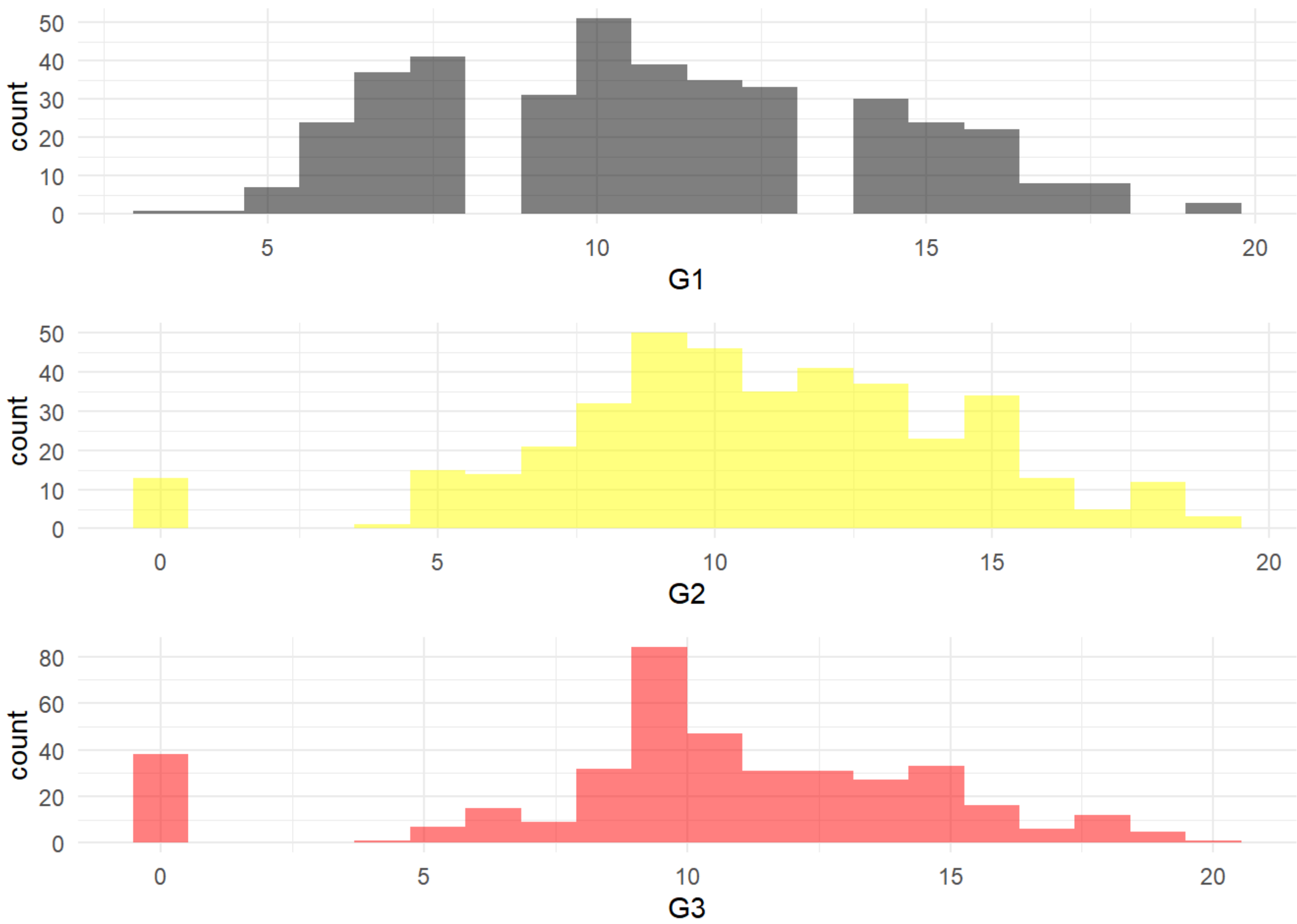
5. Analisando as variáveis:

```
hist1 <- ggplot(df, aes(Dalc)) +
  geom_histogram(bins = 30) # Consumação de Álcool durante de trabalho

hist2 <- ggplot(df, aes(Walc)) +
  geom_histogram(bins = 30) # Consumação de Álcool no final de semana

hist3 <- ggplot(df, aes(x = goout)) +
  geom_histogram(bins = 30) # Frequências de saídas com os amigos

hist4 <- ggplot(df, aes(x = Medu)) +
  geom_histogram(bins = 30) # Escolaridade da mãe

hist5 <- ggplot(df, aes(x = Fedu)) +
  geom_histogram(bins = 30) # Escolaridade do pai

hist6 <- ggplot(df, aes(x = failures)) +
  geom_histogram(bins = 30) # Frequência de reprovações

grid.arrange(hist1, hist2, hist3, hist4, hist5, hist6)
```



6. Analisando as variáveis G1, G2 e G3

```
plot1 <- ggplot(df, aes(G1)) +
  geom_histogram(bins = 20,
                 alpha = 0.5,
                 fill = 'black') +
  theme_minimal()

plot2 <- ggplot(df, aes(G2)) +
  geom_histogram(bins = 20,
                 alpha = 0.5,
                 fill = 'yellow') +
  theme_minimal()

plot3 <- ggplot(df, aes(G3)) +
  geom_histogram(bins = 20,
                 alpha = 0.5,
                 fill = 'red') +
  theme_minimal()

grid.arrange(plot1, plot2, plot3, ncol = 1)
```



7. Criando as amostras de forma randômica

```
amostra <- sample.split(df$age, SplitRatio = 0.70)
```

8. Criando dados de treino

```
treino <- subset(df, amostra == T)
```

9. Criando dados de teste

```
teste <- subset(df, amostra == F)
```

10. Criando os modelos

```
modelo_1 <- lm(G3 ~ ., treino)
modelo_2 <- lm(G3 ~ G1 + G2, treino)
modelo_3 <- lm(G3 ~ absences, treino)
modelo_4 <- lm(G3 ~ Medu, treino)
modelo_5 <- lm(G3 ~ Fedu, treino)
modelo_6 <- lm(G3 ~ failures, treino)
modelo_7 <- lm(G3 ~ goout, treino)
modelo_8 <- lm(G3 ~ Walc, treino)
```

11. Analisando os modelos

```
summary(modelo_1)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = treino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3705 -0.5825  0.2521  1.1126  4.5113
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.9007290  2.8157414  -0.320 0.749336
## schoolMS          1.0490878  0.5049300   2.078 0.038824 *
## sexM              0.0569180  0.3003663   0.189 0.849868
## age              -0.1622022  0.1330141  -1.219 0.223901
## addressU          0.3437237  0.3698278   0.929 0.353628
## famsizeLE3        0.2869877  0.2961524   0.969 0.333514
## PstatusT         -0.1554525  0.4356682  -0.357 0.721551
## Medu              0.0822083  0.1957504   0.420 0.674895
## Fedu             -0.0978228  0.1693973  -0.577 0.564171
## Mjobhealth        0.1559138  0.6489136   0.240 0.810331
## Mjobother         0.1211584  0.4093877   0.296 0.767529
## Mjobservices      0.3009823  0.4670577   0.644 0.519930
## Mjobteacher       0.0839619  0.6100958   0.138 0.890658
## Fjobhealth        0.5856904  0.8716706   0.672 0.502297
## Fjobother        -0.0001111  0.6757371   0.000 0.999869
## Fjobservices     -0.4697645  0.6978656  -0.673 0.501517
## Fjobteacher       0.0360505  0.8192577   0.044 0.964939
## reasonhome       -0.1016154  0.3308777  -0.307 0.759033
## reasonother       0.7355860  0.5190958   1.417 0.157792
## reasonreputation  0.2932738  0.3456473   0.848 0.397036
## guardianmother    0.2391386  0.3285550   0.728 0.467430
## guardianother    -0.2671150  0.6459659  -0.414 0.679608
## traveltime       -0.0746758  0.2072615  -0.360 0.718948
## studytime        -0.1112513  0.1736727  -0.641 0.522420
## failures         -0.1529513  0.2109223  -0.725 0.469079
## schoolsupyes      0.2832558  0.4157343   0.681 0.496328
## famsupyes         0.0719403  0.2906317   0.248 0.804713
## paidyes          -0.0160147  0.2929979  -0.055 0.956457
## activitiesyes    -0.4916750  0.2645147  -1.859 0.064308 .
## nurseryyes       -0.5713035  0.3370151  -1.695 0.091365 .
## higheryes         0.2518655  0.6341159   0.397 0.691587
## internetyes      -0.1330295  0.3915749  -0.340 0.734364
## romanticyes      -0.5823536  0.2893517  -2.013 0.045296 *
## famrel            0.3846184  0.1486294   2.588 0.010262 *
## freetime          0.0727401  0.1434165   0.507 0.612494
## goout             0.0119249  0.1371396   0.087 0.930782
## Dalc             -0.3986011  0.2128513  -1.873 0.062355 .
## Walc              0.2133330  0.1501522   1.421 0.156707
## health            0.1002375  0.0969652   1.034 0.302317
## absences          0.0573059  0.0166550   3.441 0.000686 ***
## G1                0.1822742  0.0824677   2.210 0.028052 *
## G2                0.9372785  0.0720688  13.005  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.022 on 235 degrees of freedom
```

```
## Multiple R-squared:  0.8389, Adjusted R-squared:  0.8108
## F-statistic: 29.85 on 41 and 235 DF,  p-value: < 2.2e-16
```

```
summary(modelo_2)
```

```
##
## Call:
## lm(formula = G3 ~ G1 + G2, data = treino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5263 -0.3459  0.3384  1.0072  3.7443
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.82670    0.42938  -4.254 2.88e-05 ***
## G1           0.14304    0.07407   1.931   0.0545 .
## G2           0.99226    0.06570  15.104  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.078 on 274 degrees of freedom
## Multiple R-squared:  0.8015, Adjusted R-squared:  0.8001
## F-statistic: 553.3 on 2 and 274 DF,  p-value: < 2.2e-16
```

```
summary(modelo_3)
```

```
##
## Call:
## lm(formula = G3 ~ absences, data = treino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.262  -2.262   0.473   3.506   9.605
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.26240    0.33936   30.24   <2e-16 ***
## absences     0.03307    0.03376    0.98    0.328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.648 on 275 degrees of freedom
## Multiple R-squared:  0.003478,   Adjusted R-squared:  -0.0001457
## F-statistic: 0.9598 on 1 and 275 DF,  p-value: 0.3281
```

```
summary(modelo_4)
```

```
##
## Call:
## lm(formula = G3 ~ Medu, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4766  -1.8349   0.5234   3.1651   8.5234
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.1932     0.7250  11.300  < 2e-16 ***
## Medu          0.8208     0.2440   3.364 0.000877 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.563 on 275 degrees of freedom
## Multiple R-squared:  0.03953,    Adjusted R-squared:  0.03604
## F-statistic: 11.32 on 1 and 275 DF,  p-value: 0.0008769
```

```
summary(modelo_5)
```

```
##
## Call:
## lm(formula = G3 ~ Fedu, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3620  -2.0879   0.5492   3.2751   9.2751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.8138     0.7041  12.517   <2e-16 ***
## Fedu          0.6371     0.2519   2.529    0.012 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.603 on 275 degrees of freedom
## Multiple R-squared:  0.02273,    Adjusted R-squared:  0.01917
## F-statistic: 6.395 on 1 and 275 DF,  p-value: 0.012
```

```
summary(modelo_6)
```

```
##
## Call:
## lm(formula = G3 ~ failures, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1716  -1.1716   0.0703   2.8284   9.0703
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.1716     0.2841  39.327  < 2e-16 ***
## failures     -2.2419     0.3495  -6.415 6.14e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.343 on 275 degrees of freedom
## Multiple R-squared:  0.1302, Adjusted R-squared:  0.127
## F-statistic: 41.15 on 1 and 275 DF,  p-value: 6.136e-10
```

```
summary(modelo_7)
```

```
##
## Call:
## lm(formula = G3 ~ goout, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7304  -1.8914   0.4956   3.2696   9.1086
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.3434     0.8150  15.146   <2e-16 ***
## goout        -0.6130     0.2483  -2.468   0.0142 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.606 on 275 degrees of freedom
## Multiple R-squared:  0.02168,    Adjusted R-squared:  0.01812
## F-statistic: 6.093 on 1 and 275 DF,  p-value: 0.01418
```

```
summary(modelo_8)
```

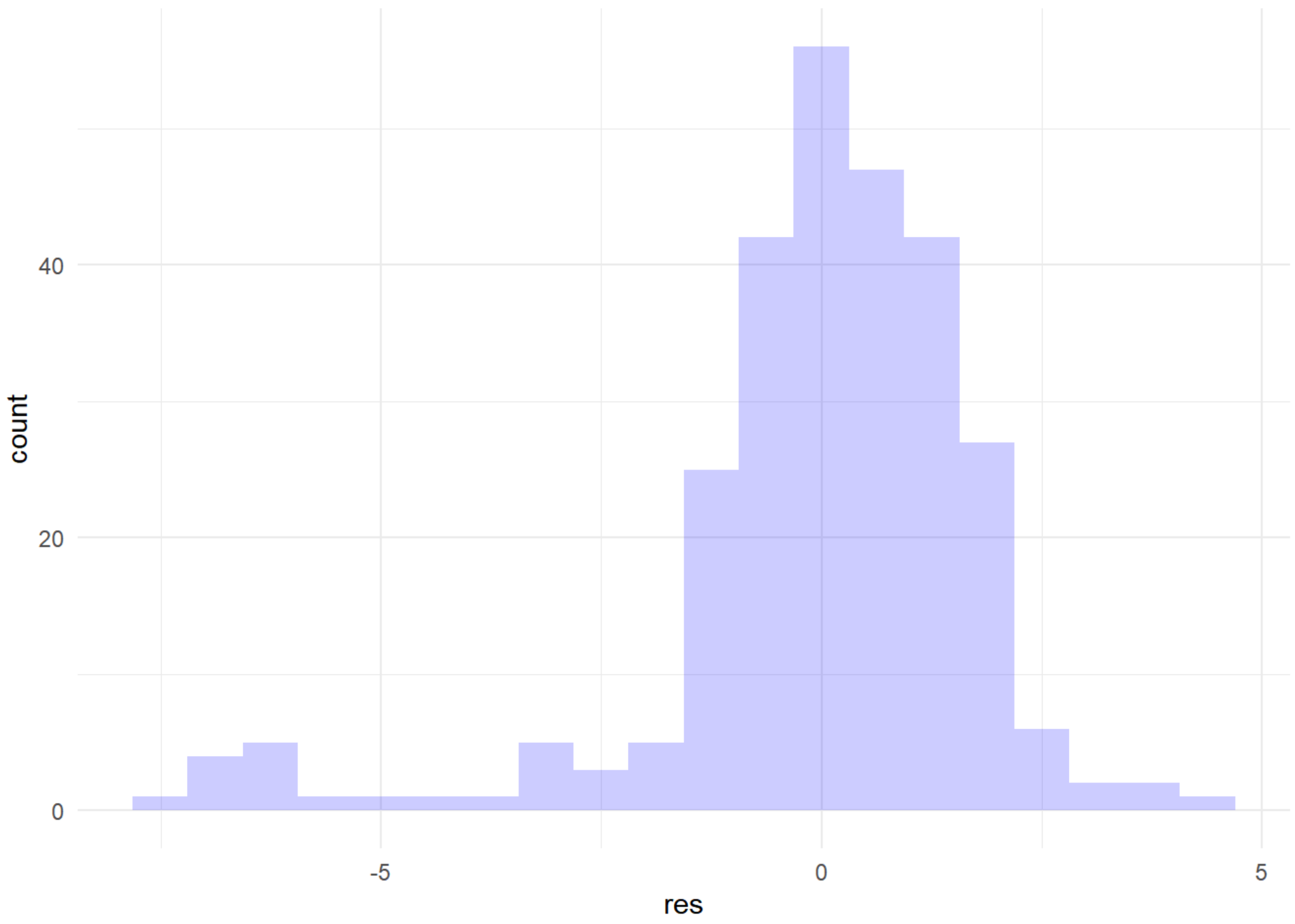
```
##
## Call:
## lm(formula = G3 ~ Walc, data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7092  -1.8927   0.4949   3.2908   9.2908
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.9133     0.5764  18.935   <2e-16 ***
## Walc         -0.2041     0.2227  -0.916     0.36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.649 on 275 degrees of freedom
## Multiple R-squared:  0.003045,   Adjusted R-squared:  -0.0005807
## F-statistic: 0.8398 on 1 and 275 DF,  p-value: 0.3602
```

12. Visualizando as taxas de erro (resíduos) do modelo escolhido

```
res <- residuals(modelo_1)
res <- as.data.frame(res)
head(res)
```

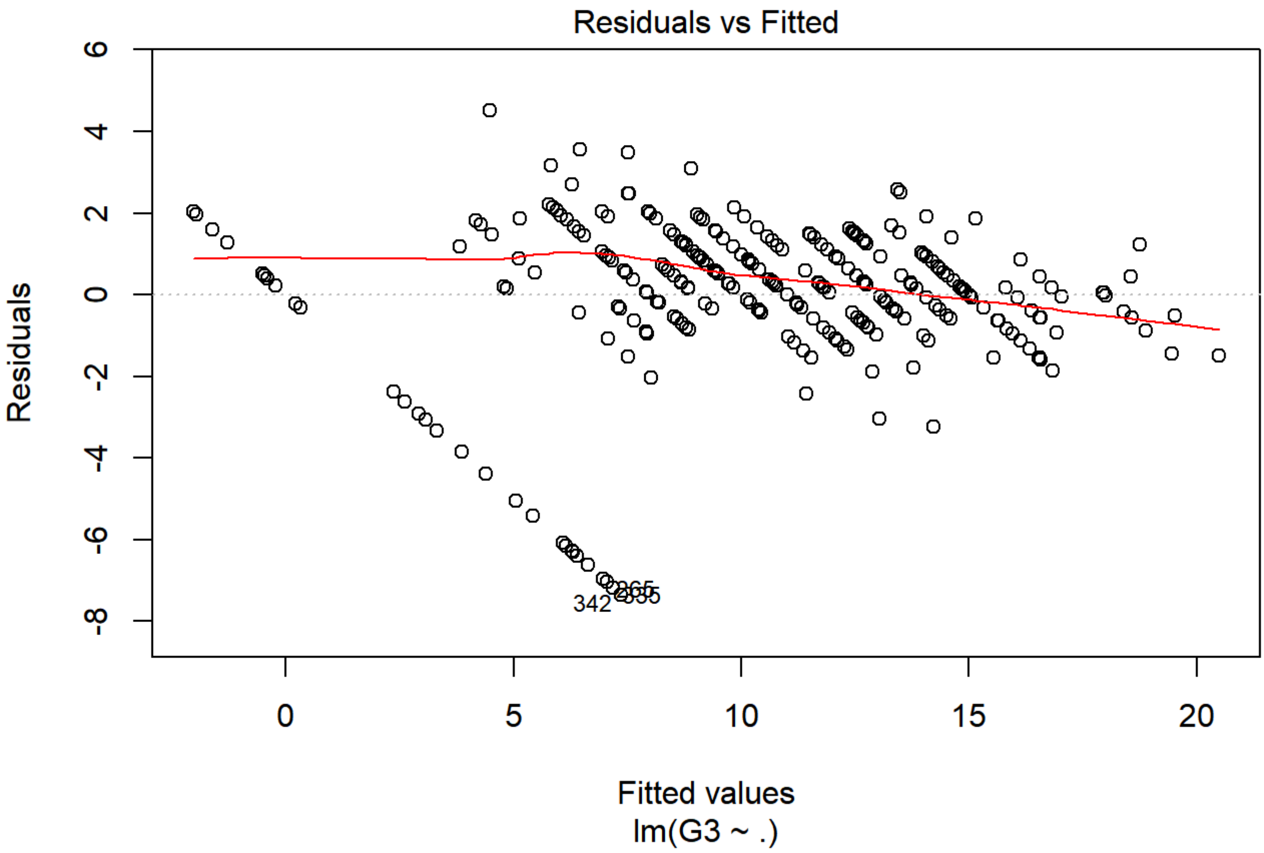
```
##          res
## 1  0.8732349
## 2  1.4725042
## 3  1.5695031
## 6 -1.8529704
## 8  1.7205089
## 9  0.4403052
```

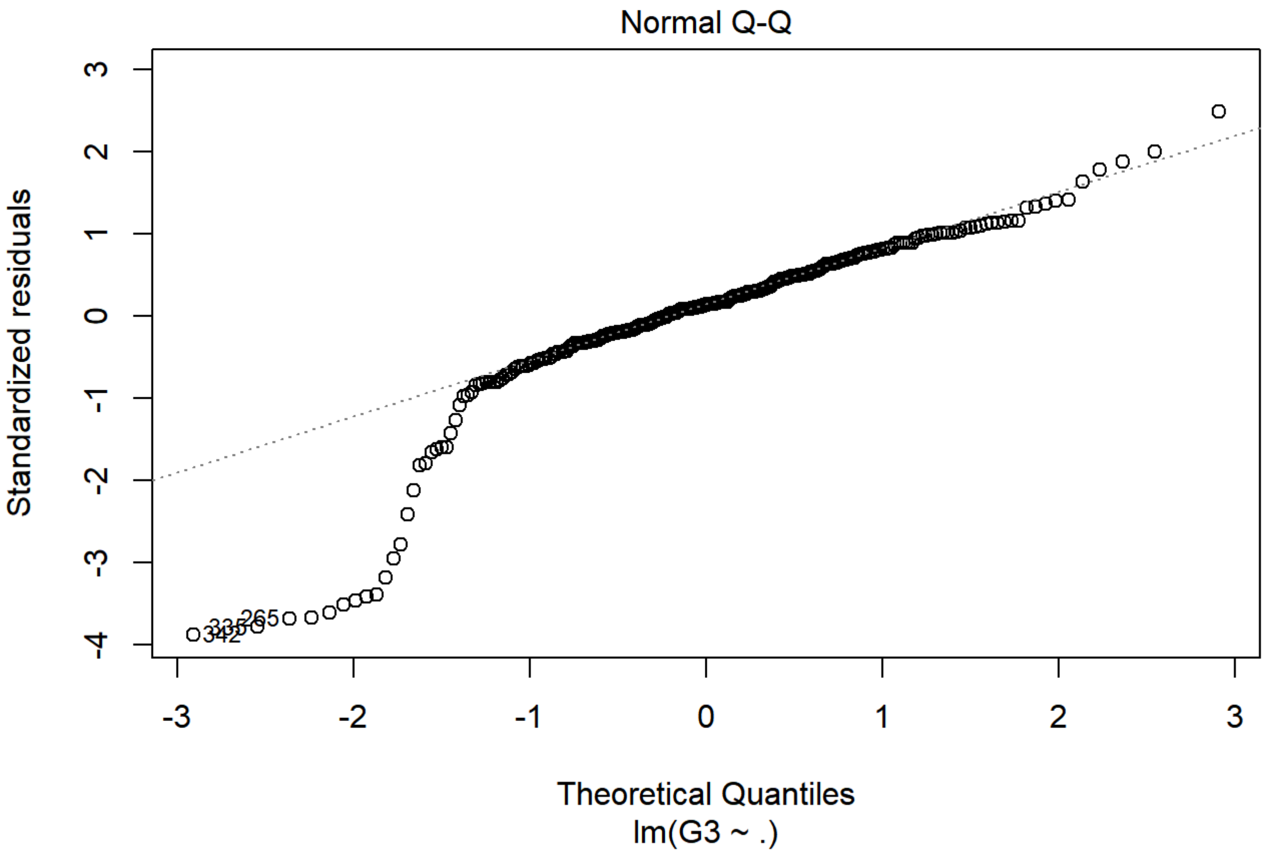
```
ggplot(res, aes(res)) +
  geom_histogram(bins = 20,
                 alpha = 0.20,
                 fill = 'blue') +
  theme_minimal()
```

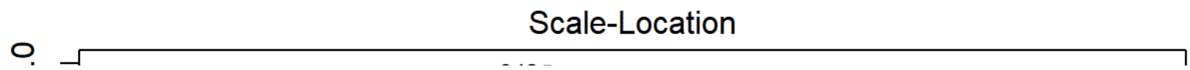


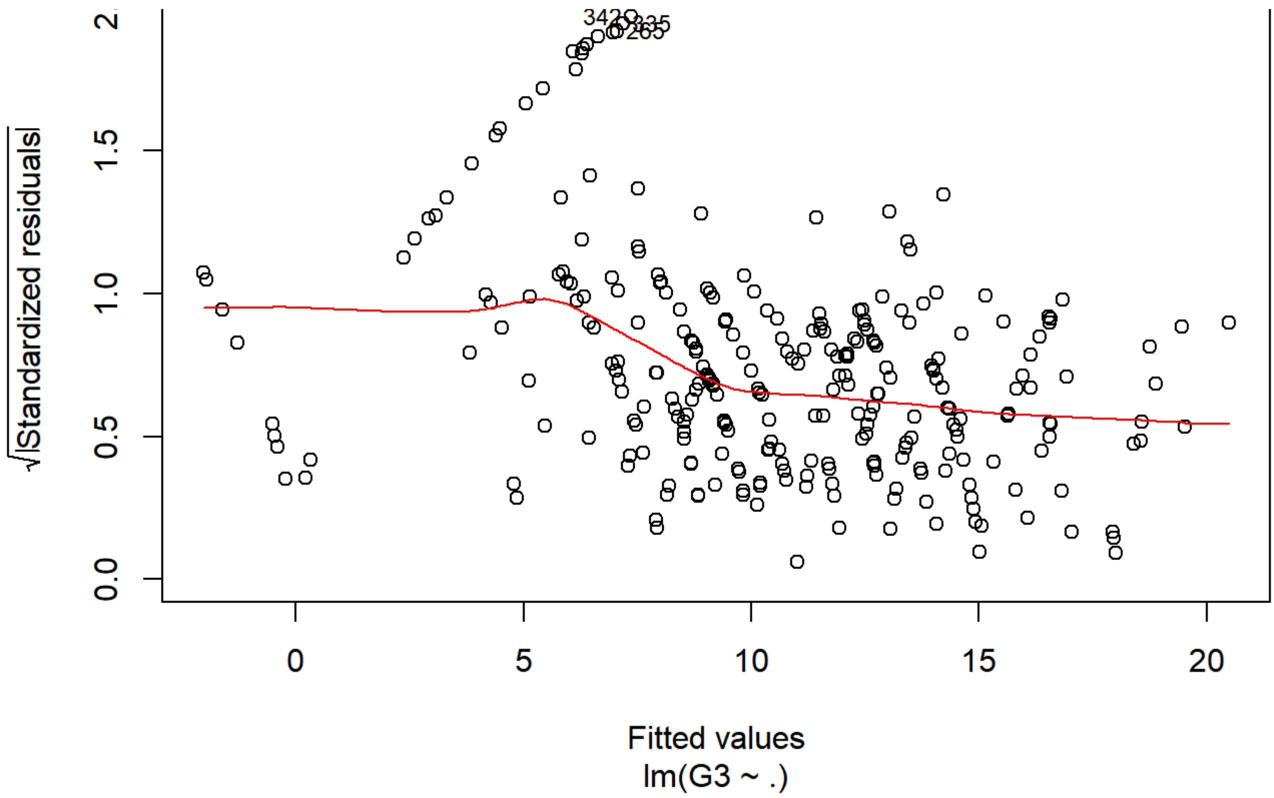
13. Plot do modelo

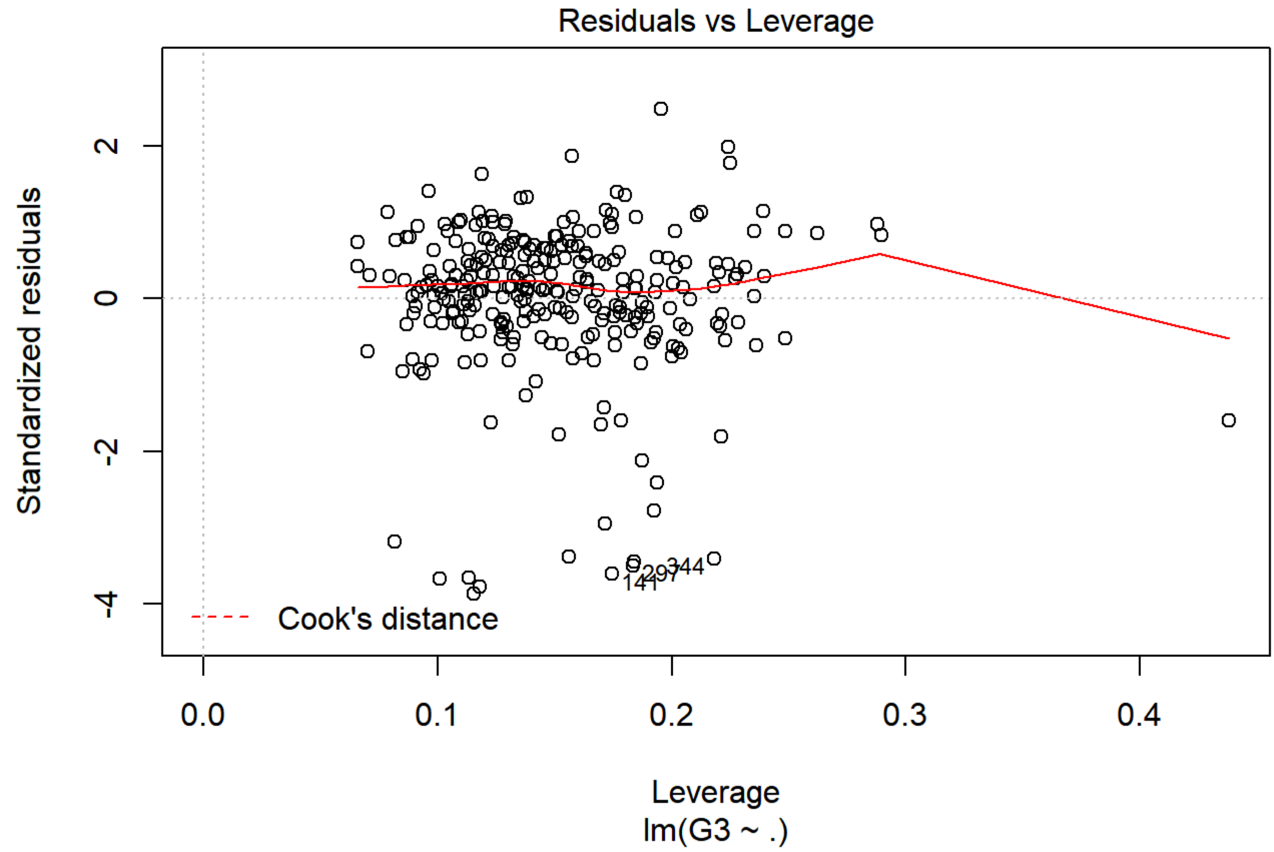
```
plot(modelo_1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(G3 ~ .)
342
265
335


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(G3 ~ .)
342
335
265


Scale-Location





14. Prevendo as notas finais

```
previsao_G3 <- predict(modelo_1, teste)
as.data.frame(head(previsao_G3))
```

```
##    head(previsao_G3)
## 4          12.542098
## 5           9.109655
## 7          11.941119
## 15         15.006421
## 17         13.095782
## 29         11.636579
```

15. Comparando os dados previstos com os reais

```
comparacao <- cbind(as.integer(previsao_G3), teste$G3)
class(comparacao)
```

```
## [1] "matrix"
```

```
comparacao <- as.data.frame(comparacao)
colnames(comparacao) <- c("Previsto", "Real")
head(comparacao)
```

```
##   Previsto Real
## 1       12   15
## 2        9   10
## 3       11   11
## 4       15   16
## 5       13   14
## 6       11   11
```

16. Tratando valores negativos

```
tratamento <- function(x){
  if (x < 0) {
    return(0)
  } else{
    return(x)
  }
}

comparacao$Previsto <- sapply(comparacao$Previsto, tratamento)
View(comparacao)
```

17. Calculando o erro médio

17.1. MSE:

```
mse <- mean((comparacao$Real - comparacao$Previsto)^2)
print(mse) # Distancia dos valores previstos para os valores observados
```

```
## [1] 3.618644
```

18. Calculando R-Squared

```
SSE = sum((comparacao$Previsto - comparacao$Real)^2)
SST = sum((mean(df$G3) - comparacao$Real)^2)
```

18.1. R-Squared

```
R2 = 1 - (SSE/SST)
R2*100 # Percentual da precisão do modelo criado
```

```
## [1] 81.49076
```