R_Linear_Regression

Ulises Jose Bustamante Mora

2023-10-23

Installing libraries

```
##
## 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

## corrplot 0.92 loaded
```

Exploring the dataset

```
df = read.csv("student-mat.csv", sep = ";")
head(df)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        3       yes     no  yes         no
## 4   mother          1         3        0        no    yes  yes        yes
## 5   father          1         2        0        no    yes  yes         no
## 6   mother          1         2        0        no    yes  yes        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3
## 1        6  5  6  6
## 2        4  5  5  6
## 3       10  7  8 10
## 4        2 15 14 15
## 5        4  6 10 10
## 6       10 15 15 15
```

Getting the statistics of the dataset

```r
summary(df)
```

```
##     school              sex                 age          address
##  Length:395         Length:395         Min.   :15.0   Length:395
##  Class :character   Class :character   1st Qu.:16.0   Class :character
##  Mode  :character   Mode  :character   Median :17.0   Mode  :character
##                                        Mean   :16.7
##                                        3rd Qu.:18.0
##                                        Max.   :22.0
##    famsize            Pstatus                Medu            Fedu
##  Length:395         Length:395         Min.   :0.000   Min.   :0.000
##  Class :character   Class :character   1st Qu.:2.000   1st Qu.:2.000
##  Mode  :character   Mode  :character   Median :3.000   Median :2.000
##                                        Mean   :2.749   Mean   :2.522
##                                        3rd Qu.:4.000   3rd Qu.:3.000
##                                        Max.   :4.000   Max.   :4.000
##      Mjob               Fjob              reason            guardian
##  Length:395         Length:395         Length:395         Length:395
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    traveltime      studytime        failures        schoolsup
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:395
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character
##  Median :1.000   Median :2.000   Median :0.0000   Mode  :character
##  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
##     famsup              paid            activities          nursery
##  Length:395         Length:395         Length:395         Length:395
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##     higher            internet           romantic             famrel     
##  Length:395         Length:395         Length:395         Min.   :1.000  
##  Class :character   Class :character   Class :character   1st Qu.:4.000  
##  Mode  :character   Mode  :character   Mode  :character   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
```

More technical information about the dataset

```
str(df)
```

```
## ’data.frame’:    395 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ 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      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ 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 : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ 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 ...
```

Let´s check if the dataset has nulls values

```
any(is.na(df))
```

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

Correlation analysis

First getting all the numeric columns

```
dfNum = sapply(df, is.numeric)
head(dfNum)
```

```
##  school     sex     age address famsize Pstatus
##   FALSE   FALSE    TRUE   FALSE   FALSE   FALSE
```

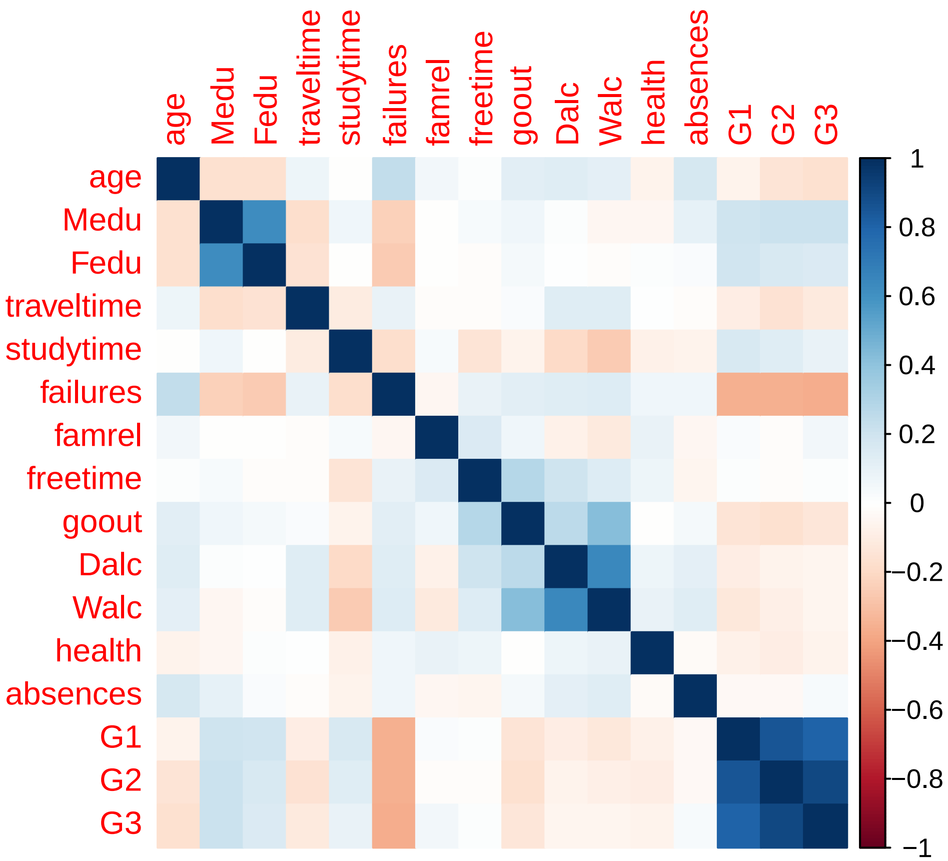
Correlation of the dataset

```
dfCor = cor(df[,dfNum])
head(dfCor)
```

```
##                     age        Medu         Fedu  traveltime    studytime
## age         1.000000000 -0.16365842 -0.163438069  0.07064072 -0.004140037
## Medu       -0.163658419  1.00000000  0.623455112 -0.17163930  0.064944137
## Fedu       -0.163438069  0.62345511  1.000000000 -0.15819405 -0.009174639
## traveltime  0.070640721 -0.17163930 -0.158194054  1.00000000 -0.100909119
## studytime  -0.004140037  0.06494414 -0.009174639 -0.10090912  1.000000000
## failures    0.243665377 -0.23667996 -0.250408444  0.09223875 -0.173563031
##                failures       famrel    freetime       goout         Dalc
## age          0.24366538  0.053940096  0.01643439  0.12696388  0.131124605
## Medu        -0.23667996 -0.003914458  0.03089087  0.06409444  0.019834099
## Fedu        -0.25040844 -0.001369727 -0.01284553  0.04310467  0.002386429
## traveltime   0.09223875 -0.016807986 -0.01702494  0.02853967  0.138325309
## studytime   -0.17356303  0.039730704 -0.14319841 -0.06390368 -0.196019263
## failures     1.00000000 -0.044336626  0.09198747  0.12456092  0.136046931
##                    Walc       health    absences          G1          G2
## age          0.11727605 -0.062187369  0.17523008 -0.06408150 -0.1434740
## Medu        -0.04712346 -0.046877829  0.10028482  0.20534100  0.2155272
## Fedu        -0.01263102  0.014741537  0.02447289  0.19026994  0.1648934
## traveltime   0.13411575  0.007500606 -0.01294378 -0.09303999 -0.1531980
## studytime   -0.25378473 -0.075615863 -0.06270018  0.16061192  0.1358800
## failures     0.14196203  0.065827282  0.06372583 -0.35471761 -0.3558956
##                      G3
## age         -0.16157944
## Medu         0.21714750
## Fedu         0.15245694
## traveltime  -0.11714205
## studytime    0.09781969
## failures    -0.36041494
```

Visualization of the correlation

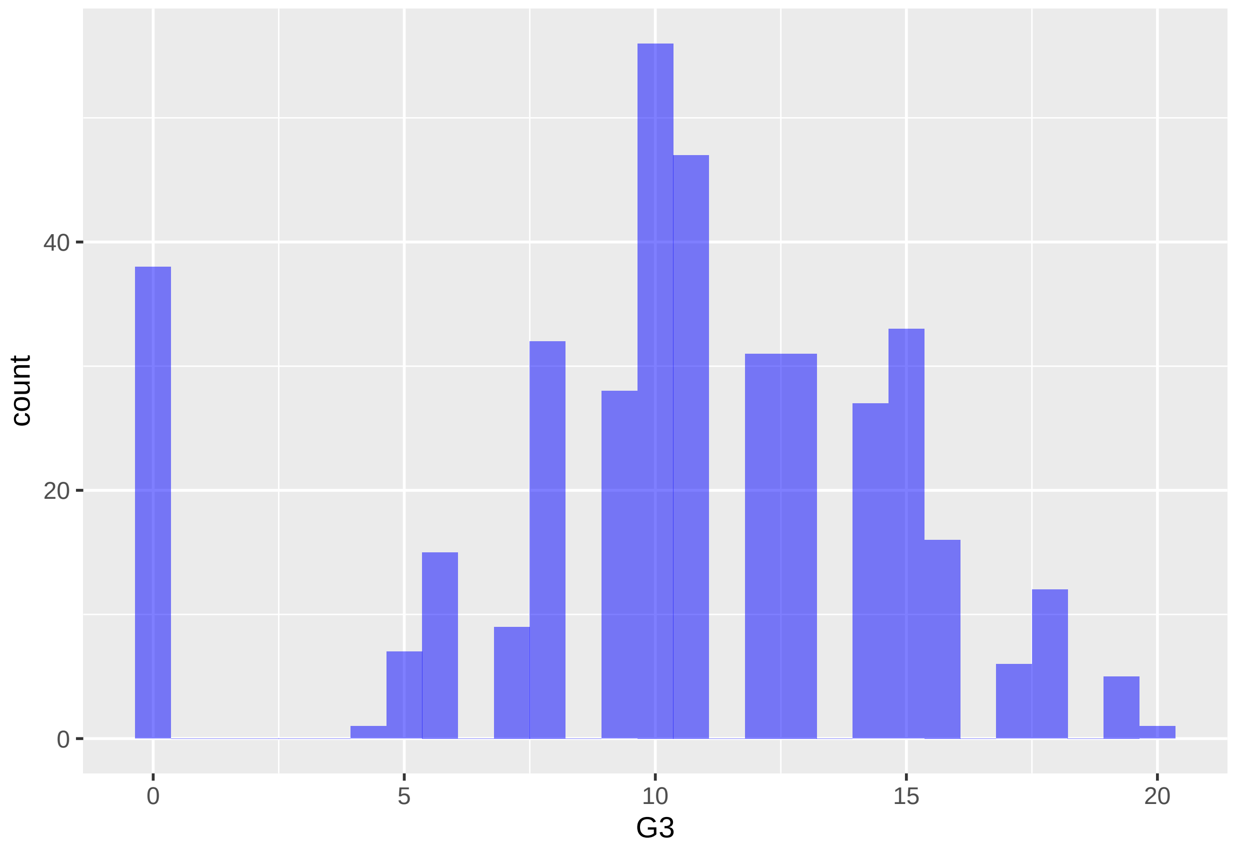
```
print(corrplot(dfCor , method = "color"))
```



```
## $corr
##                    age         Medu         Fedu   traveltime    studytime
## age        1.000000000 -0.163658419 -0.163438069  0.070640721 -0.004140037
## Medu      -0.163658419  1.000000000  0.623455112 -0.171639305  0.064944137
## Fedu      -0.163438069  0.623455112  1.000000000 -0.158194054 -0.009174639
## traveltime  0.070640721 -0.171639305 -0.158194054  1.000000000 -0.100909119
## studytime -0.004140037  0.064944137 -0.009174639 -0.100909119  1.000000000
## failures   0.243665377 -0.236679963 -0.250408444  0.092238746 -0.173563031
## famrel     0.053940096 -0.003914458 -0.001369727 -0.016807986  0.039730704
## freetime   0.016434389  0.030890867 -0.012845528 -0.017024944 -0.143198407
## goout      0.126963880  0.064094438  0.043104668  0.028539674 -0.063903675
## Dalc       0.131124605  0.019834099  0.002386429  0.138325309 -0.196019263
## Walc       0.117276052 -0.047123460 -0.012631018  0.134115752 -0.253784731
## health    -0.062187369 -0.046877829  0.014741537  0.007500606 -0.075615863
## absences   0.175230079  0.100284818  0.024472887 -0.012943775 -0.062700175
## G1        -0.064081497  0.205340997  0.190269936 -0.093039992  0.160611915
## G2        -0.143474049  0.215527168  0.164893393 -0.153197963  0.135879999
## G3        -0.161579438  0.217147496  0.152456939 -0.117142053  0.097819690
##              failures       famrel    freetime        goout         Dalc
## age        0.24366538  0.053940096  0.01643439  0.126963880  0.131124605
## Medu      -0.23667996 -0.003914458  0.03089087  0.064094438  0.019834099
## Fedu      -0.25040844 -0.001369727 -0.01284553  0.043104668  0.002386429
## traveltime  0.09223875 -0.016807986 -0.01702494  0.028539674  0.138325309
## studytime -0.17356303  0.039730704 -0.14319841 -0.063903675 -0.196019263
## failures   1.00000000 -0.044336626  0.09198747  0.124560922  0.136046931
## famrel    -0.04433663  1.000000000  0.15070144  0.064568411 -0.077594357
## freetime    0.09198747  0.150701444  1.00000000  0.285018715  0.209000848
## goout       0.12456092  0.064568411  0.28501871  1.000000000  0.266993848
## Dalc        0.13604693 -0.077594357  0.20900085  0.266993848  1.000000000
## Walc        0.14196203 -0.113397308  0.14782181  0.420385745  0.647544230
## health      0.06582728  0.094055728  0.07573336 -0.009577254  0.077179582
## absences    0.06372583 -0.044354095 -0.05807792  0.044302220  0.111908026
## G1         -0.35471761  0.022168316  0.01261293 -0.149103967 -0.094158792
## G2         -0.35589563 -0.018281347 -0.01377714 -0.162250034 -0.064120183
## G3         -0.36041494  0.051363429  0.01130724 -0.132791474 -0.054660041
##                   Walc       health    absences          G1          G2
## age         0.11727605 -0.062187369  0.17523008 -0.06408150 -0.14347405
## Medu       -0.04712346 -0.046877829  0.10028482  0.20534100  0.21552717
## Fedu       -0.01263102  0.014741537  0.02447289  0.19026994  0.16489339
## traveltime  0.13411575  0.007500606 -0.01294378 -0.09303999 -0.15319796
## studytime  -0.25378473 -0.075615863 -0.06270018  0.16061192  0.13588000
## failures    0.14196203  0.065827282  0.06372583 -0.35471761 -0.35589563
## famrel     -0.11339731  0.094055728 -0.04435409  0.02216832 -0.01828135
## freetime    0.14782181  0.075733357 -0.05807792  0.01261293 -0.01377714
## goout       0.42038575 -0.009577254  0.04430222 -0.14910397 -0.16225003
## Dalc        0.64754423  0.077179582  0.11190803 -0.09415879 -0.06412018
## Walc        1.00000000  0.092476317  0.13629110 -0.12617921 -0.08492735
## health      0.09247632  1.000000000 -0.02993671 -0.07317207 -0.09771987
## absences    0.13629110 -0.029936711  1.00000000 -0.03100290 -0.03177670
## G1         -0.12617921 -0.073172073 -0.03100290  1.00000000  0.85211807
## G2         -0.08492735 -0.097719866 -0.03177670  0.85211807  1.00000000
## G3         -0.05193932 -0.061334605  0.03424732  0.80146793  0.90486799
##                     G3
## age        -0.16157944
## Medu        0.21714750
## Fedu        0.15245694
## traveltime -0.11714205
## studytime   0.09781969
## failures   -0.36041494
## famrel      0.05136343
## freetime    0.01130724
## goout      -0.13279147
## Dalc       -0.05466004
## Walc       -0.05193932
## health     -0.06133460
## absences    0.03424732
## G1          0.80146793
## G2          0.90486799
## G3          1.00000000
## 
## $corrPos
##          xName      yName  x  y         corr
## 1          age        age  1 16  1.000000000
## 2          age       Medu  1 15 -0.163658419
## 3          age       Fedu  1 14 -0.163438069
## 4          age traveltime  1 13  0.070640721
## 5          age  studytime  1 12 -0.004140037
## 6          age   failures  1 11  0.243665377
## 7          age     famrel  1 10  0.053940096
## 8          age   freetime  1  9  0.016434389
## 9          age      goout  1  8  0.126963880
## 10         age       Dalc  1  7  0.131124605
## 11         age       Walc  1  6  0.117276052
## 12         age     health  1  5 -0.062187369
## 13         age   absences  1  4  0.175230079
## 14         age         G1  1  3 -0.064081497
## 15         age         G2  1  2 -0.143474049
## 16         age         G3  1  1 -0.161579438
## 17        Medu        age  2 16 -0.163658419
## 18        Medu       Medu  2 15  1.000000000
## 19        Medu       Fedu  2 14  0.623455112
## 20        Medu traveltime  2 13 -0.171639305
## 21        Medu  studytime  2 12  0.064944137
## 22        Medu   failures  2 11 -0.236679963
## 23        Medu     famrel  2 10 -0.003914458
## 24        Medu   freetime  2  9  0.030890867
## 25        Medu      goout  2  8  0.064094438
## 26        Medu       Dalc  2  7  0.019834099
## 27        Medu       Walc  2  6 -0.047123460
## 28        Medu     health  2  5 -0.046877829
## 29        Medu   absences  2  4  0.100284818
## 30        Medu         G1  2  3  0.205340997
## 31        Medu         G2  2  2  0.215527168
## 32        Medu         G3  2  1  0.217147496
## 33        Fedu        age  3 16 -0.163438069
## 34        Fedu       Medu  3 15  0.623455112
## 35        Fedu       Fedu  3 14  1.000000000
## 36        Fedu traveltime  3 13 -0.158194054
## 37        Fedu  studytime  3 12 -0.009174639
## 38        Fedu   failures  3 11 -0.250408444
## 39        Fedu     famrel  3 10 -0.001369727
## 40        Fedu   freetime  3  9 -0.012845528
## 41        Fedu      goout  3  8  0.043104668
## 42        Fedu       Dalc  3  7  0.002386429
## 43        Fedu       Walc  3  6 -0.012631018
## 44        Fedu     health  3  5  0.014741537
## 45        Fedu   absences  3  4  0.024472887
## 46        Fedu         G1  3  3  0.190269936
## 47        Fedu         G2  3  2  0.164893393
## 48        Fedu         G3  3  1  0.152456939
## 49  traveltime        age  4 16  0.070640721
## 50  traveltime       Medu  4 15 -0.171639305
## 51  traveltime       Fedu  4 14 -0.158194054
## 52  traveltime traveltime  4 13  1.000000000
## 53  traveltime  studytime  4 12 -0.100909119
## 54  traveltime   failures  4 11  0.092238746
## 55  traveltime     famrel  4 10 -0.016807986
## 56  traveltime   freetime  4  9 -0.017024944
## 57  traveltime      goout  4  8  0.028539674
## 58  traveltime       Dalc  4  7  0.138325309
## 59  traveltime       Walc  4  6  0.134115752
## 60  traveltime     health  4  5  0.007500606
## 61  traveltime   absences  4  4 -0.012943775
## 62  traveltime         G1  4  3 -0.093039992
## 63  traveltime         G2  4  2 -0.153197963
## 64  traveltime         G3  4  1 -0.117142053
## 65   studytime        age  5 16 -0.004140037
## 66   studytime       Medu  5 15  0.064944137
## 67   studytime       Fedu  5 14 -0.009174639
## 68   studytime traveltime  5 13 -0.100909119
## 69   studytime  studytime  5 12  1.000000000
## 70   studytime   failures  5 11 -0.173563031
## 71   studytime     famrel  5 10  0.039730704
## 72   studytime   freetime  5  9 -0.143198407
## 73   studytime      goout  5  8 -0.063903675
## 74   studytime       Dalc  5  7 -0.196019263
## 75   studytime       Walc  5  6 -0.253784731
## 76   studytime     health  5  5 -0.075615863
## 77   studytime   absences  5  4 -0.062700175
## 78   studytime         G1  5  3  0.160611915
## 79   studytime         G2  5  2  0.135879999
## 80   studytime         G3  5  1  0.097819690
## 81    failures        age  6 16  0.243665377
## 82    failures       Medu  6 15 -0.236679963
## 83    failures       Fedu  6 14 -0.250408444
## 84    failures traveltime  6 13  0.092238746
## 85    failures  studytime  6 12 -0.173563031
## 86    failures   failures  6 11  1.000000000
## 87    failures     famrel  6 10 -0.044336626
## 88    failures   freetime  6  9  0.091987471
## 89    failures      goout  6  8  0.124560922
## 90    failures       Dalc  6  7  0.136046931
## 91    failures       Walc  6  6  0.141962030
## 92    failures     health  6  5  0.065827282
## 93    failures   absences  6  4  0.063725833
## 94    failures         G1  6  3 -0.354717613
## 95    failures         G2  6  2 -0.355895635
## 96    failures         G3  6  1 -0.360414940
## 97      famrel        age  7 16  0.053940096
## 98      famrel       Medu  7 15 -0.003914458
## 99      famrel       Fedu  7 14 -0.001369727
## 100     famrel traveltime  7 13 -0.016807986
## 101     famrel  studytime  7 12  0.039730704
## 102     famrel   failures  7 11 -0.044336626
## 103     famrel     famrel  7 10  1.000000000
## 104     famrel   freetime  7  9  0.150701444
## 105     famrel      goout  7  8  0.064568411
## 106     famrel       Dalc  7  7 -0.077594357
## 107     famrel       Walc  7  6 -0.113397308
## 108     famrel     health  7  5  0.094055728
## 109     famrel   absences  7  4 -0.044354095
## 110     famrel         G1  7  3  0.022168316
## 111     famrel         G2  7  2 -0.018281347
## 112     famrel         G3  7  1  0.051363429
## 113   freetime        age  8 16  0.016434389
## 114   freetime       Medu  8 15  0.030890867
## 115   freetime       Fedu  8 14 -0.012845528
## 116   freetime traveltime  8 13 -0.017024944
## 117   freetime  studytime  8 12 -0.143198407
## 118   freetime   failures  8 11  0.091987471
## 119   freetime     famrel  8 10  0.150701444
## 120   freetime   freetime  8  9  1.000000000
## 121   freetime      goout  8  8  0.285018715
## 122   freetime       Dalc  8  7  0.209000848
## 123   freetime       Walc  8  6  0.147821813
## 124   freetime     health  8  5  0.075733357
## 125   freetime   absences  8  4 -0.058077922
## 126   freetime         G1  8  3  0.012612930
## 127   freetime         G2  8  2 -0.013777139
## 128   freetime         G3  8  1  0.011307240
## 129      goout        age  9 16  0.126963880
## 130      goout       Medu  9 15  0.064094438
## 131      goout       Fedu  9 14  0.043104668
## 132      goout traveltime  9 13  0.028539674
## 133      goout  studytime  9 12 -0.063903675
## 134      goout   failures  9 11  0.124560922
## 135      goout     famrel  9 10  0.064568411
## 136      goout   freetime  9  9  0.285018715
## 137      goout      goout  9  8  1.000000000
## 138      goout       Dalc  9  7  0.266993848
## 139      goout       Walc  9  6  0.420385745
## 140      goout     health  9  5 -0.009577254
## 141      goout   absences  9  4  0.044302220
## 142      goout         G1  9  3 -0.149103967
## 143      goout         G2  9  2 -0.162250034
## 144      goout         G3  9  1 -0.132791474
## 145       Dalc        age 10 16  0.131124605
## 146       Dalc       Medu 10 15  0.019834099
## 147       Dalc       Fedu 10 14  0.002386429
## 148       Dalc traveltime 10 13  0.138325309
## 149       Dalc  studytime 10 12 -0.196019263
## 150       Dalc   failures 10 11  0.136046931
## 151       Dalc     famrel 10 10 -0.077594357
## 152       Dalc   freetime 10  9  0.209000848
## 153       Dalc      goout 10  8  0.266993848
## 154       Dalc       Dalc 10  7  1.000000000
## 155       Dalc       Walc 10  6  0.647544230
## 156       Dalc     health 10  5  0.077179582
## 157       Dalc   absences 10  4  0.111908026
## 158       Dalc         G1 10  3 -0.094158792
## 159       Dalc         G2 10  2 -0.064120183
## 160       Dalc         G3 10  1 -0.054660041
## 161       Walc        age 11 16  0.117276052
## 162       Walc       Medu 11 15 -0.047123460
## 163       Walc       Fedu 11 14 -0.012631018
## 164       Walc traveltime 11 13  0.134115752
## 165       Walc  studytime 11 12 -0.253784731
## 166       Walc   failures 11 11  0.141962030
## 167       Walc     famrel 11 10 -0.113397308
## 168       Walc   freetime 11  9  0.147821813
## 169       Walc      goout 11  8  0.420385745
## 170       Walc       Dalc 11  7  0.647544230
## 171       Walc       Walc 11  6  1.000000000
## 172       Walc     health 11  5  0.092476317
## 173       Walc   absences 11  4  0.136291101
## 174       Walc         G1 11  3 -0.126179208
## 175       Walc         G2 11  2 -0.084927353
## 176       Walc         G3 11  1 -0.051939324
## 177     health        age 12 16 -0.062187369
## 178     health       Medu 12 15 -0.046877829
## 179     health       Fedu 12 14  0.014741537
## 180     health traveltime 12 13  0.007500606
## 181     health  studytime 12 12 -0.075615863
## 182     health   failures 12 11  0.065827282
## 183     health     famrel 12 10  0.094055728
## 184     health   freetime 12  9  0.075733357
## 185     health      goout 12  8 -0.009577254
## 186     health       Dalc 12  7  0.077179582
## 187     health       Walc 12  6  0.092476317
## 188     health     health 12  5  1.000000000
## 189     health   absences 12  4 -0.029936711
## 190     health         G1 12  3 -0.073172073
## 191     health         G2 12  2 -0.097719866
## 192     health         G3 12  1 -0.061334605
## 193   absences        age 13 16  0.175230079
## 194   absences       Medu 13 15  0.100284818
## 195   absences       Fedu 13 14  0.024472887
## 196   absences traveltime 13 13 -0.012943775
## 197   absences  studytime 13 12 -0.062700175
## 198   absences   failures 13 11  0.063725833
## 199   absences     famrel 13 10 -0.044354095
## 200   absences   freetime 13  9 -0.058077922
## 201   absences      goout 13  8  0.044302220
## 202   absences       Dalc 13  7  0.111908026
## 203   absences       Walc 13  6  0.136291101
## 204   absences     health 13  5 -0.029936711
## 205   absences   absences 13  4  1.000000000
## 206   absences         G1 13  3 -0.031002901
## 207   absences         G2 13  2 -0.031776704
## 208   absences         G3 13  1  0.034247316
## 209         G1        age 14 16 -0.064081497
## 210         G1       Medu 14 15  0.205340997
## 211         G1       Fedu 14 14  0.190269936
## 212         G1 traveltime 14 13 -0.093039992
## 213         G1  studytime 14 12  0.160611915
## 214         G1   failures 14 11 -0.354717613
## 215         G1     famrel 14 10  0.022168316
## 216         G1   freetime 14  9  0.012612930
## 217         G1      goout 14  8 -0.149103967
## 218         G1       Dalc 14  7 -0.094158792
## 219         G1       Walc 14  6 -0.126179208
## 220         G1     health 14  5 -0.073172073
## 221         G1   absences 14  4 -0.031002901
## 222         G1         G1 14  3  1.000000000
## 223         G1         G2 14  2  0.852118066
## 224         G1         G3 14  1  0.801467932
## 225         G2        age 15 16 -0.143474049
## 226         G2       Medu 15 15  0.215527168
## 227         G2       Fedu 15 14  0.164893393
## 228         G2 traveltime 15 13 -0.153197963
## 229         G2  studytime 15 12  0.135879999
## 230         G2   failures 15 11 -0.355895635
## 231         G2     famrel 15 10 -0.018281347
## 232         G2   freetime 15  9 -0.013777139
## 233         G2      goout 15  8 -0.162250034
## 234         G2       Dalc 15  7 -0.064120183
## 235         G2       Walc 15  6 -0.084927353
## 236         G2     health 15  5 -0.097719866
## 237         G2   absences 15  4 -0.031776704
## 238         G2         G1 15  3  0.852118066
## 239         G2         G2 15  2  1.000000000
## 240         G2         G3 15  1  0.904867989
## 241         G3        age 16 16 -0.161579438
## 242         G3       Medu 16 15  0.217147496
## 243         G3       Fedu 16 14  0.152456939
## 244         G3 traveltime 16 13 -0.117142053
## 245         G3  studytime 16 12  0.097819690
## 246         G3   failures 16 11 -0.360414940
## 247         G3     famrel 16 10  0.051363429
## 248         G3   freetime 16  9  0.011307240
## 249         G3      goout 16  8 -0.132791474
## 250         G3       Dalc 16  7 -0.054660041
## 251         G3       Walc 16  6 -0.051939324
## 252         G3     health 16  5 -0.061334605
## 253         G3   absences 16  4  0.034247316
## 254         G3         G1 16  3  0.801467932
## 255         G3         G2 16  2  0.904867989
## 256         G3         G3 16  1  1.000000000
##
## $arg
## $arg$type
## [1] "full"
```

EDA

```
ggplot(df, aes(x=G3)) + geom_histogram(bins = 29, alpha = 0.5, fill = "blue")
```



# Starting the Liner Regression model

Sampling the dataset

```
sample = sample.split(df$G3, SplitRatio = 0.7)
trainDF = subset(df, sample == T)
testDF = subset(df, sample == F)
```

Model

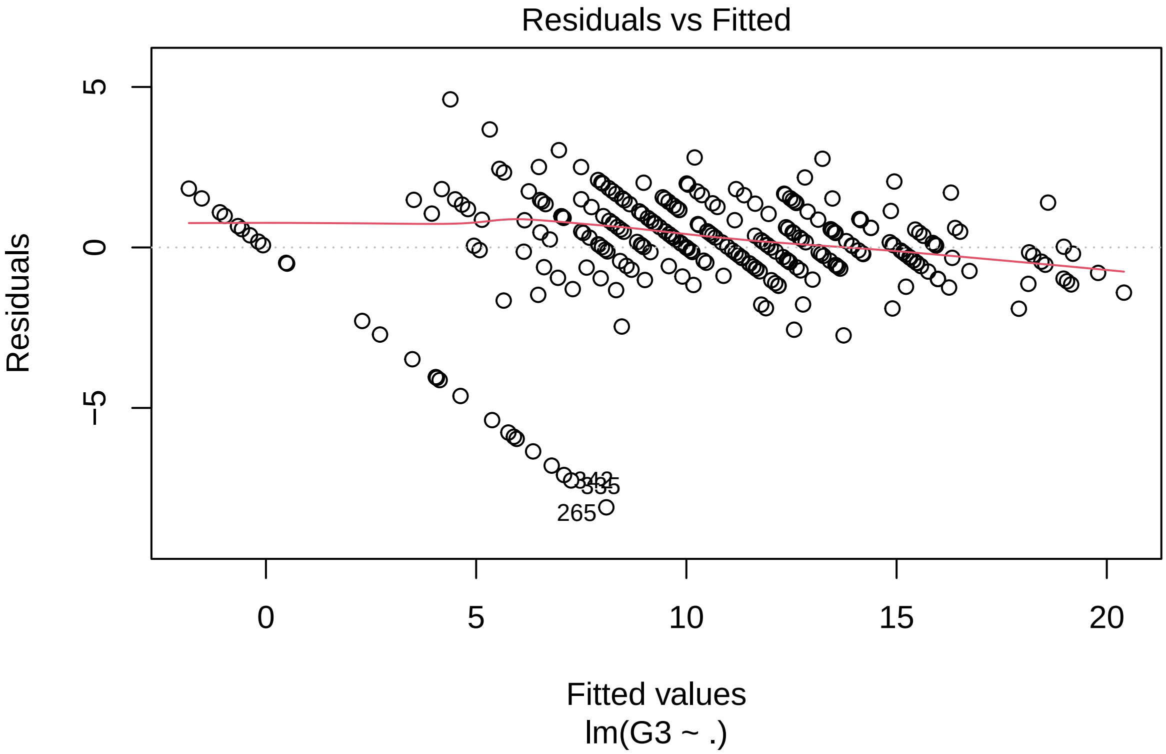
```
model = lm(G3~., data = trainDF)
print(summary(model))
```

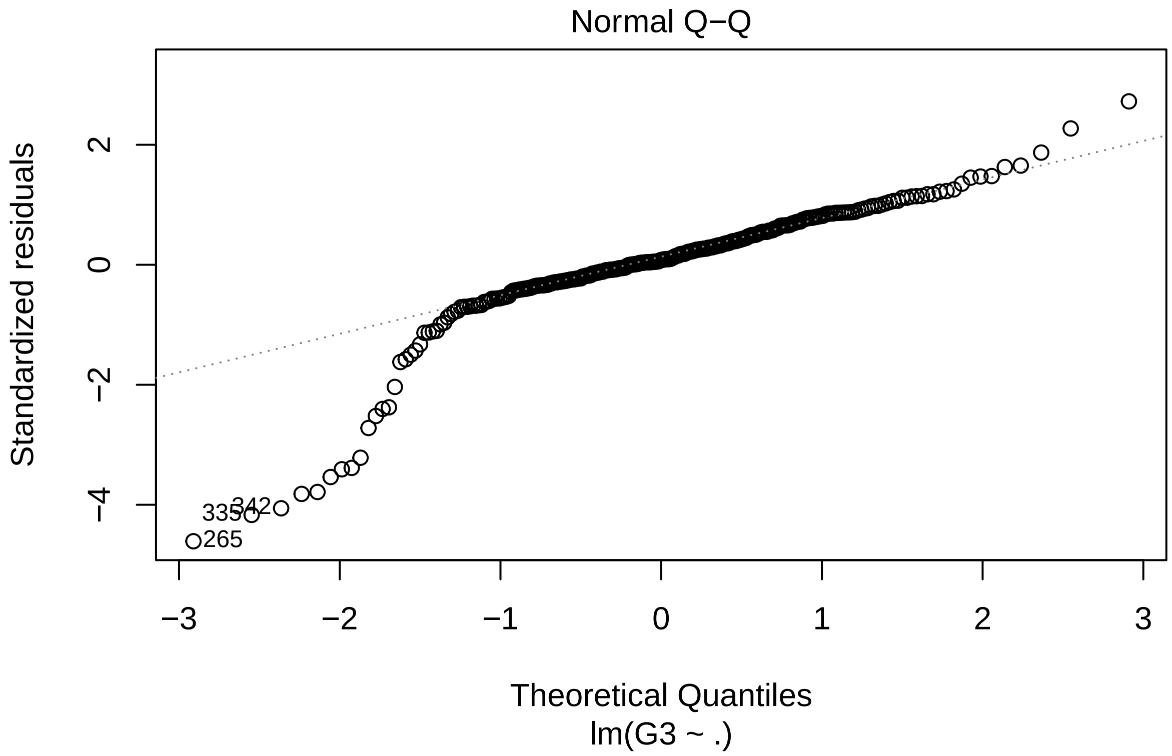
```
##
## Call:
## lm(formula = G3 ~ ., data = trainDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0963 -0.5069  0.1381  0.9719  4.6130
##
## Coefficients:
```

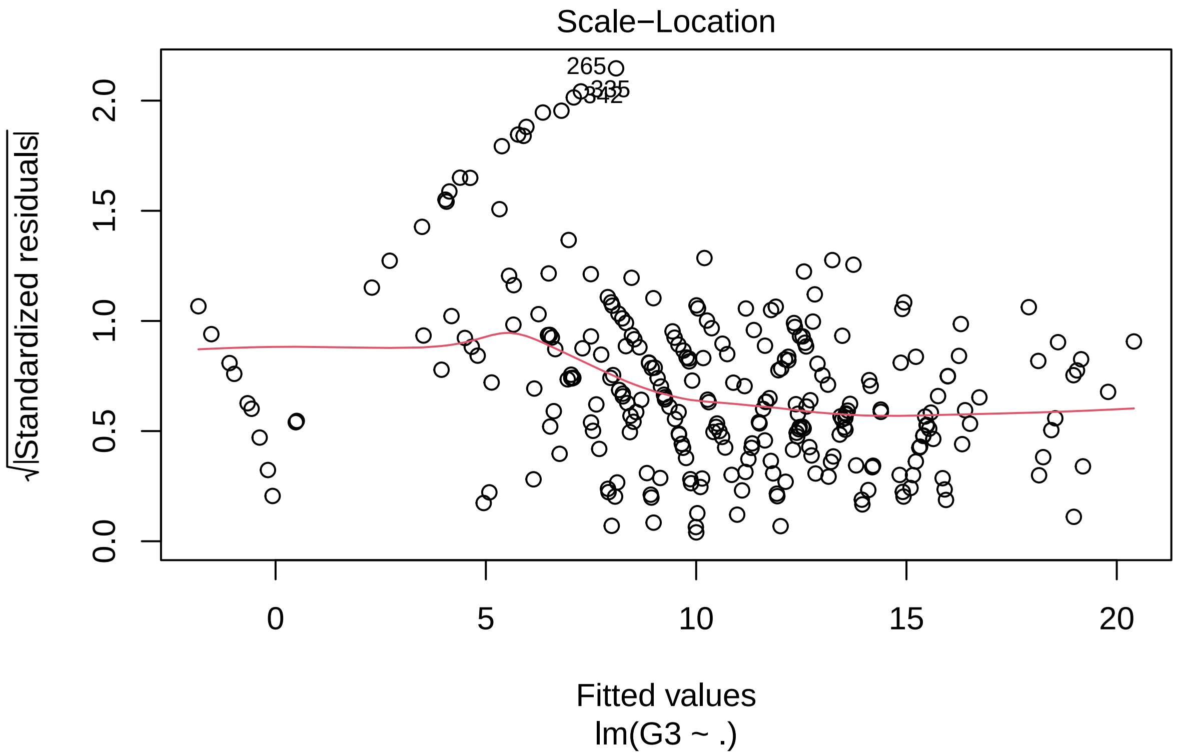
```
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       1.37882    2.55034   0.541 0.589267    
## schoolMS          1.02815    0.42853   2.399 0.017209 *  
## sexM              0.10407    0.28143   0.370 0.711881    
## age              -0.30119    0.12202  -2.468 0.014286 *  
## addressU          0.21346    0.32950   0.648 0.517727    
## famsizeLE3        0.27039    0.27202   0.994 0.321229    
## PstatusT         -0.41647    0.39849  -1.045 0.297045    
## Medu              0.01778    0.18031   0.099 0.921511    
## Fedu             -0.09603    0.15461  -0.621 0.535156    
## Mjobhealth       -0.47031    0.63105  -0.745 0.456851    
## Mjobother         0.16359    0.40413   0.405 0.686003    
## Mjobservices      0.05339    0.46038   0.116 0.907780    
## Mjobteacher      -0.29721    0.57497  -0.517 0.605708    
## Fjobhealth        0.22177    0.76335   0.291 0.771669    
## Fjobother        -0.26376    0.54931  -0.480 0.631555    
## Fjobservices     -0.59705    0.56869  -1.050 0.294853    
## Fjobteacher      -0.33395    0.68904  -0.485 0.628366    
## reasonhome       -0.10819    0.30140  -0.359 0.719944    
## reasonother       0.07961    0.47340   0.168 0.866599    
## reasonreputation  0.27940    0.33521   0.834 0.405403    
## guardianmother    0.39300    0.29801   1.319 0.188536    
## guardianother    -0.19306    0.52111  -0.370 0.711356    
## traveltime        0.11369    0.18772   0.606 0.545334    
## studytime        -0.01264    0.15527  -0.081 0.935186    
## failures         -0.16167    0.21025  -0.769 0.442697    
## schoolsupyes     -0.14228    0.39369  -0.361 0.718114    
## famsupyes         0.27790    0.27616   1.006 0.315303    
## paidyes           0.26819    0.26754   1.002 0.317153    
## activitiesyes    -0.09473    0.24373  -0.389 0.697869    
## nurseryyes       -0.29019    0.30724  -0.945 0.345871    
## higheryes        -0.10054    0.69190  -0.145 0.884594    
## internetyes       0.16807    0.35056   0.479 0.632075    
## romanticyes      -0.20254    0.26192  -0.773 0.440127    
## famrel            0.39879    0.13936   2.862 0.004596 ** 
## freetime          0.06749    0.13681   0.493 0.622272    
## goout            -0.04584    0.13049  -0.351 0.725703    
## Dalc             -0.28589    0.18176  -1.573 0.117084    
## Walc              0.16147    0.13751   1.174 0.241480    
## health            0.09311    0.10032   0.928 0.354264    
## absences          0.07247    0.01977   3.666 0.000305 ***
## G1                0.16857    0.07216   2.336 0.020333 *  
## G2                0.93765    0.05992  15.649  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.861 on 235 degrees of freedom
## Multiple R-squared:  0.8611, Adjusted R-squared:  0.8368 
## F-statistic: 35.53 on 41 and 235 DF,  p-value: < 2.2e-16
```

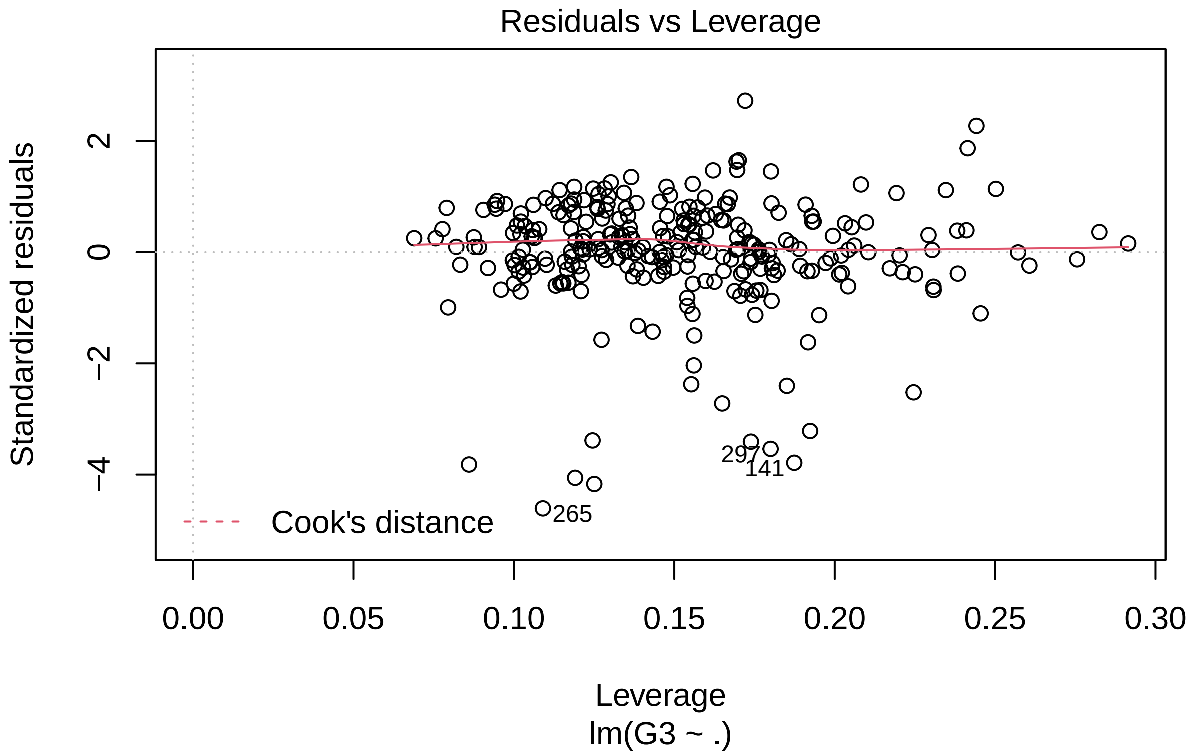
Plotting the model

```r
plot(model)
```




Normal Q−Q
Standardized residuals
2
0
−2
−4
335
342
265
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(G3 ~ .)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
5
10
15
20
265
335
342
Fitted values
lm(G3 ~ .)



Making predictions

```
preG3 = predict(model, testDF)
results = cbind(preG3, testDF$G3)
colnames(results) = c("predicted", "real")
results = as.data.frame(results)
head(results)
```

```
##   predicted real
## 1  4.258866    6
## 3  7.869486   10
## 4 13.237897   15
## 5  9.043884   10
## 6 17.286363   15
## 7 11.736511   11
```

Changing negative values

```
to_zero = function(x){
  if (x<0){
    return(0)
  } else{
    return(x)
  }
}

results$predicted = sapply(results$predicted, to_zero)
```

Metrics of our model (Mean squared error and root mean squared error)

```
mse = mean((results$real - results$predicted)^2)
print(paste0("Mean squared error ", round(mse,2)))
```

```
## [1] "Mean squared error 4.7"
```

```
rmse = mse^0.5
print(paste0("Root mean squared error ", round(rmse,2)))
```

```
## [1] "Root mean squared error 2.17"
```

Explaining the variance of our model

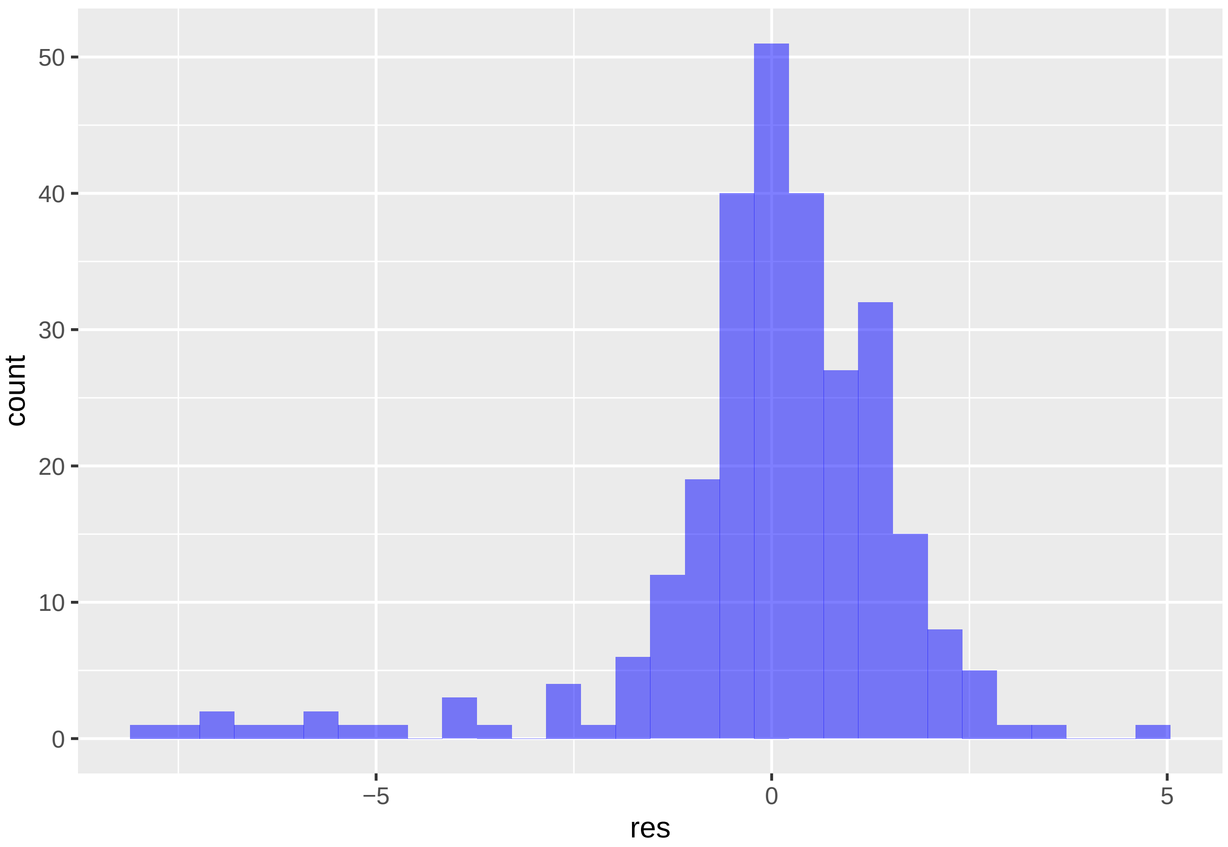
```
SSE = sum((results$predicted - results$real)^2)
SST = sum((mean(df$G3) - results$real)^2)

R2 = 1 - SSE/SST
print(paste0("We can explain the variance of the data in a ", round(R2,2)))
```

```
## [1] "We can explain the variance of the data in a 0.77"
```

Getting the residuals of the model

```
res = residuals(model)
res = as.data.frame(res)
ggplot(res, aes(res)) + geom_histogram(fill = "blue", alpha = 0.5)
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```


50
40
30
20
10
0
count
−5
0
5
res