# Final2

*Wen*

*12/4/2017*

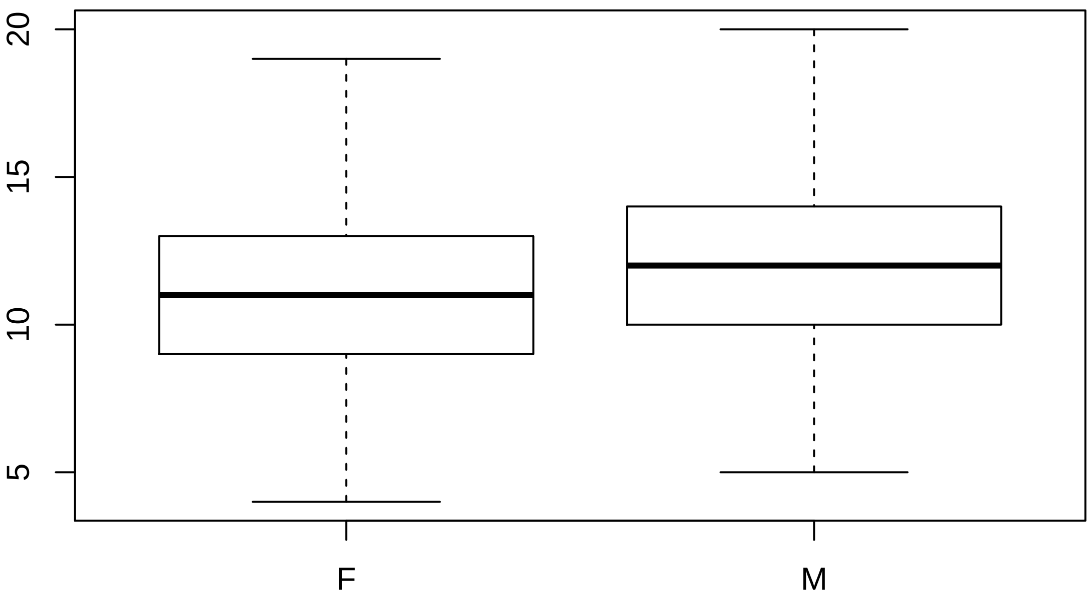
```
Alcohol <- read.csv("student-mat.csv")
```

In order to analyze the data, we first load the libraries DataComputing and mosaic: these libraries help making changes in our dataset. We also combine the variables Daily alcohol consumption (Dalc) with Weekend alcohol consumption (Walc) and create a new variable in the dataset called DWalc. Then we attach Alcohol to make easier the data analysis process.

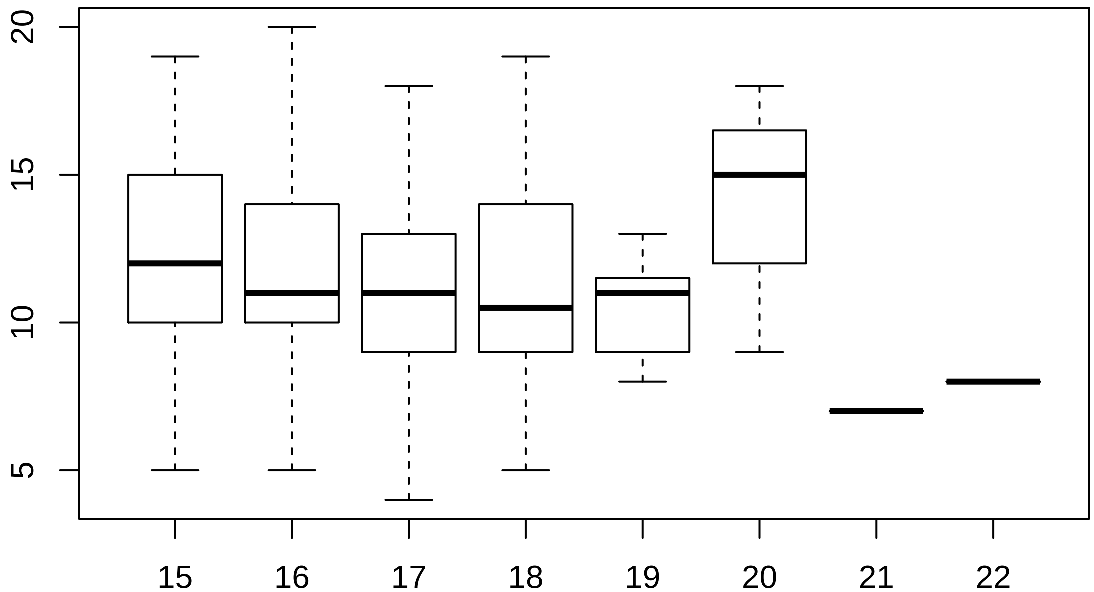
```
library(DataComputing)
library(mosaic)
Alcohol <- Alcohol %>%
  mutate(DWalc = Dalc + Walc)
```

```
attach(Alcohol)
Alcohol3 <- Alcohol[-c(which(G3=='0')),]
detach(Alcohol)
attach(Alcohol3)
```

```
boxplot(G3~sex) #males perform slightly better than females
```



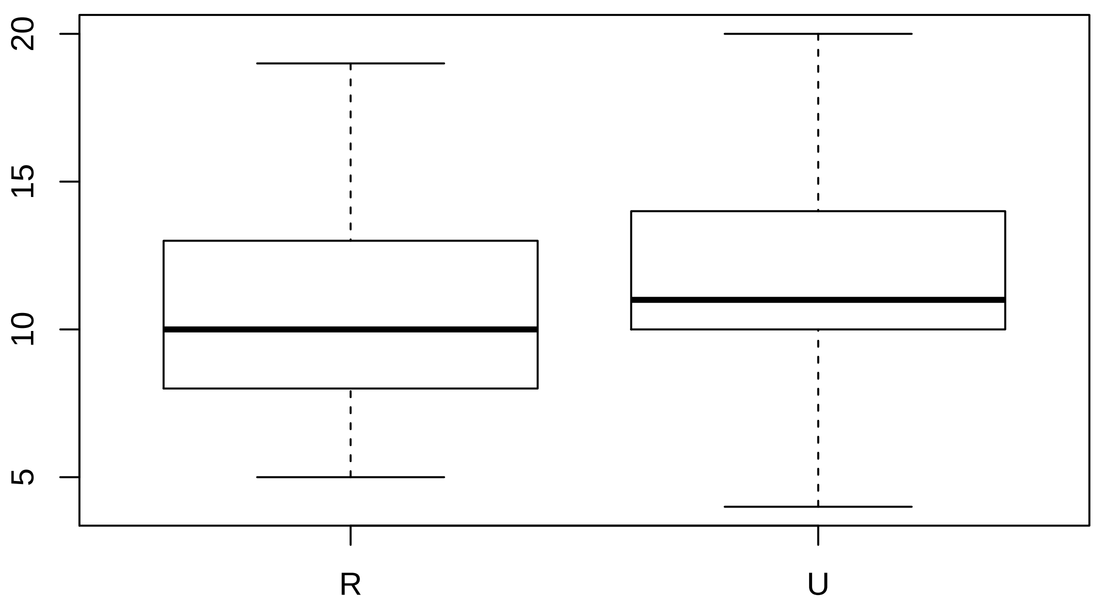

```
boxplot(G3~age) #even though 20years old students seem to perform better than the other age groups, the
```



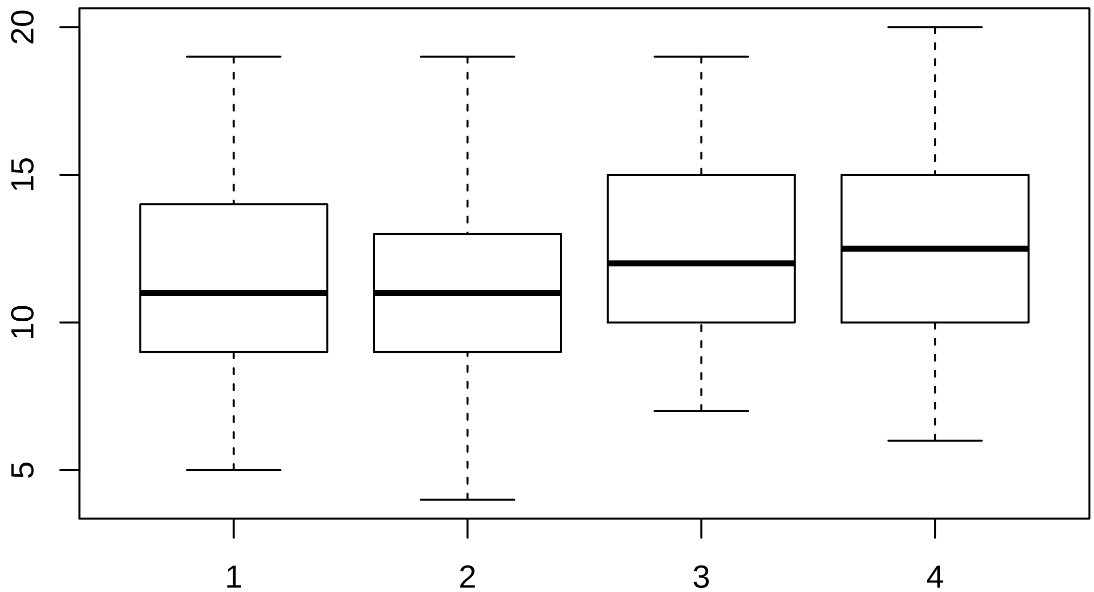

```r
table(age)
```

```
## age
## 15 16 17 18 19 20 21 22
## 76 97 90 70 19  3  1  1
```

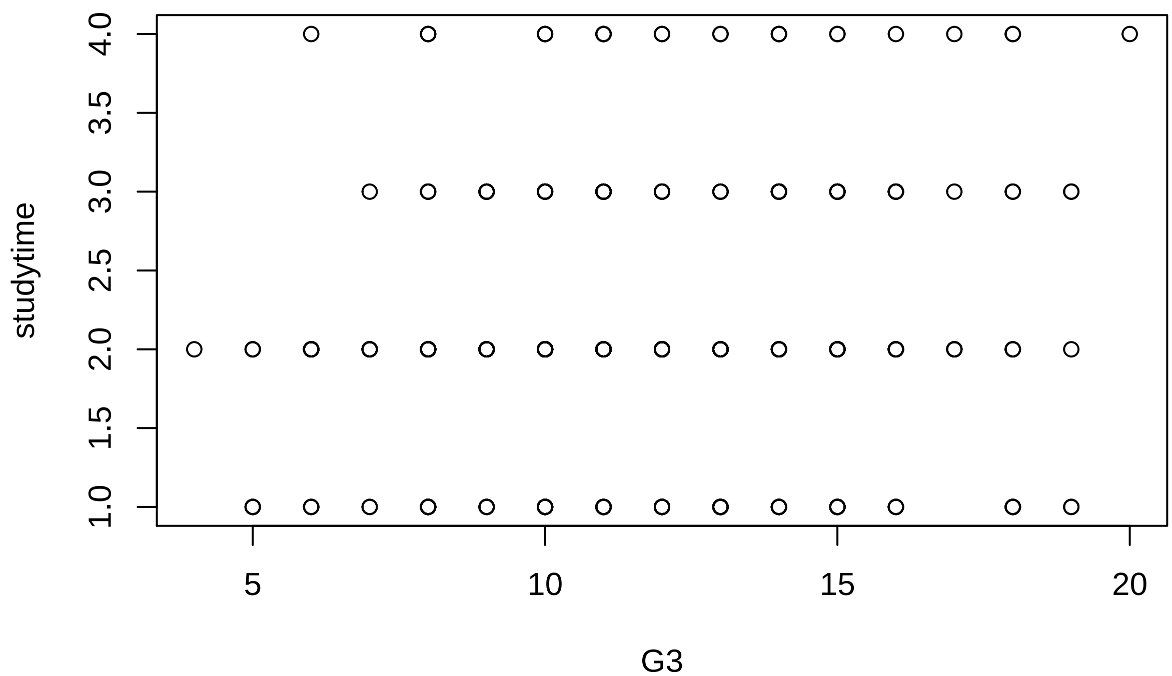
```r
boxplot(G3~address) #students coming from urban homes perform slightly better than the ones coming from
```



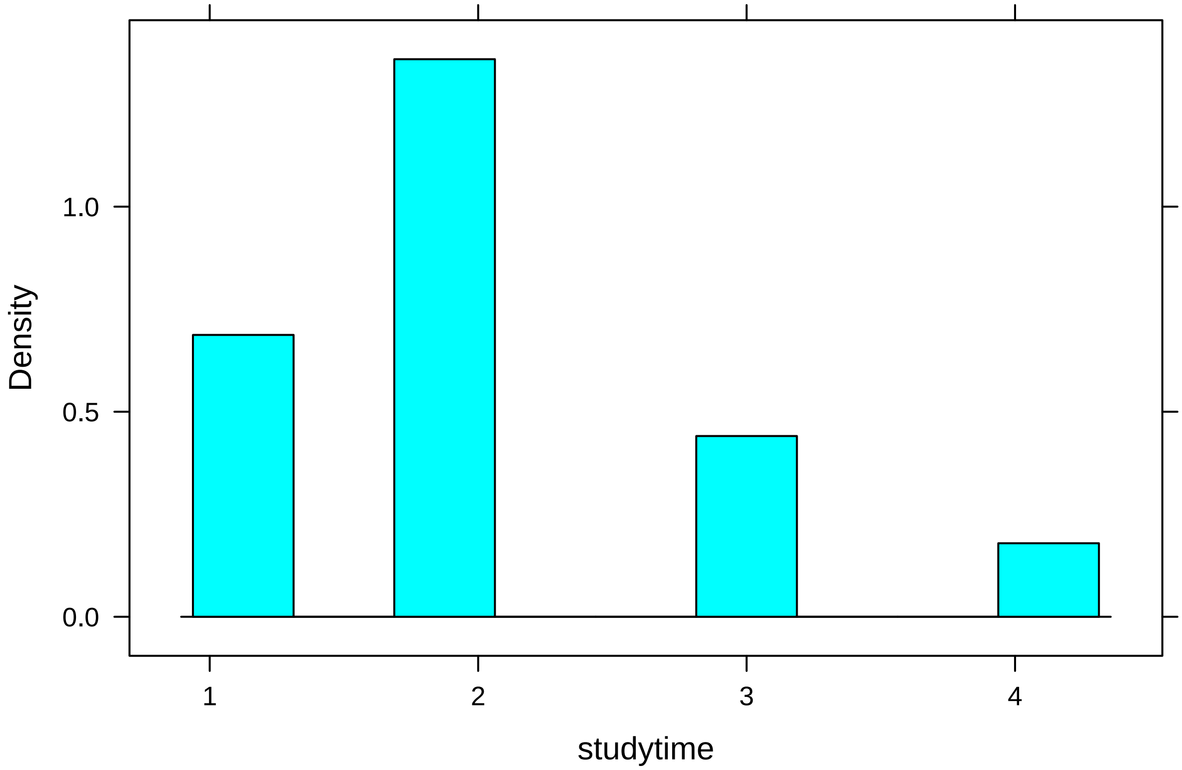

```r
boxplot(G3~studytime) #unfortunately there might be a correlation between studytime and final grades bu
```




```r
plot(G3, studytime)
```



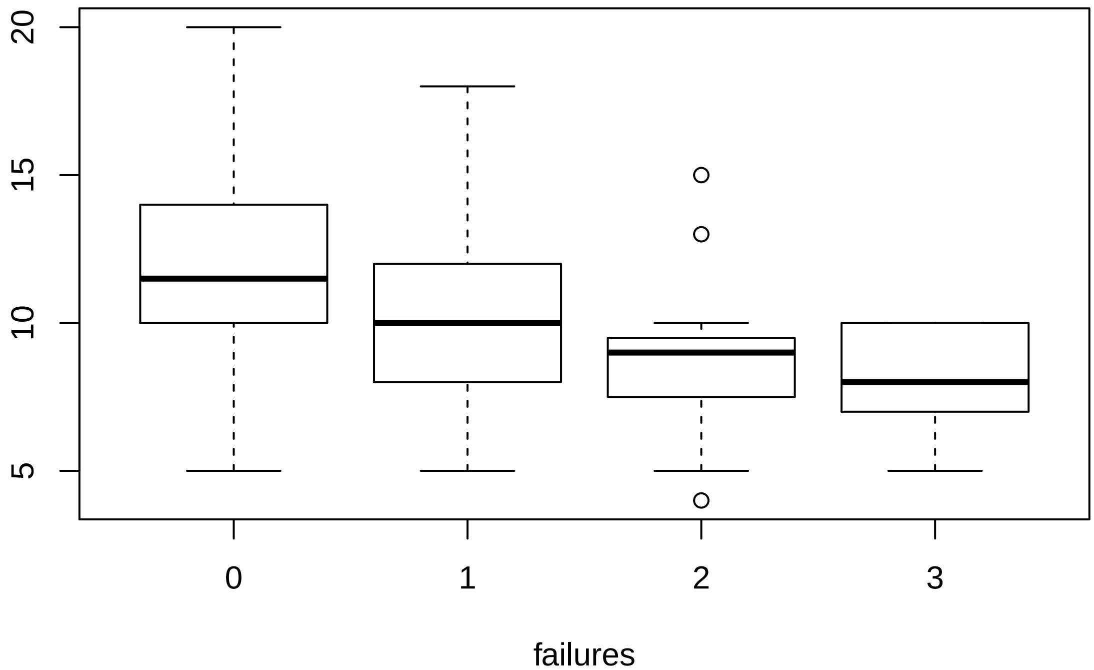

```r
histogram(studytime)
```




```r
modelstudytime = lm(G3~studytime)
summary(modelstudytime)
```

```
##
## Call:
## lm(formula = G3 ~ studytime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5031 -2.4866 -0.5031  2.0051  7.9886
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.5197     0.4503  23.360   <2e-16 ***
## studytime     0.4917     0.2043   2.407   0.0166 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.206 on 355 degrees of freedom
## Multiple R-squared:  0.01606,    Adjusted R-squared:  0.01329
## F-statistic: 5.794 on 1 and 355 DF,  p-value: 0.01659
```

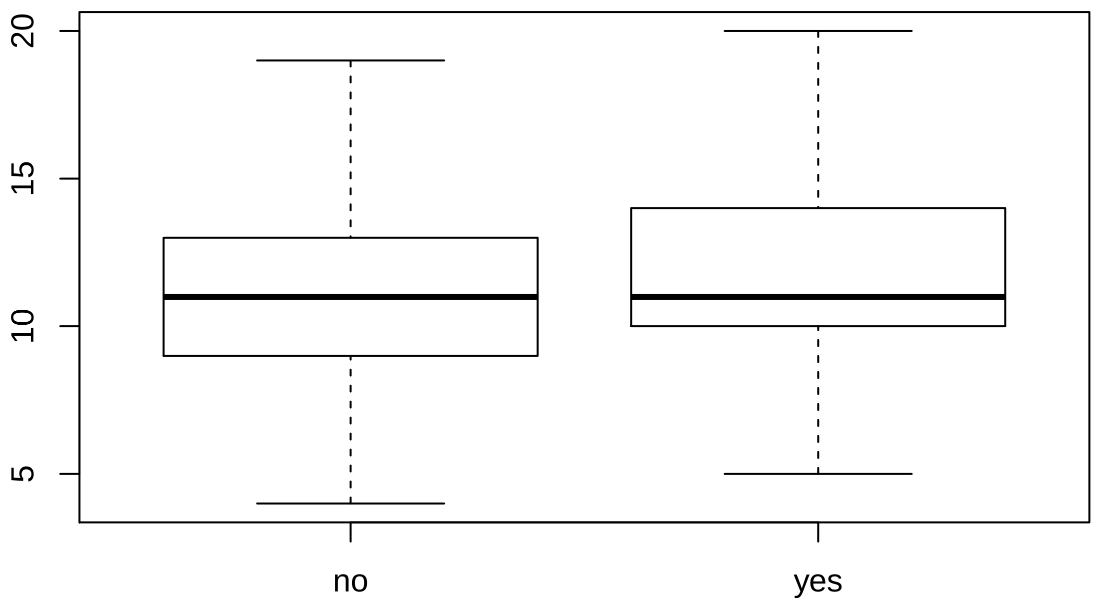
```r
boxplot(G3~failures, xlab = "failures") #it can be seen that as the number of failures in past classes
```



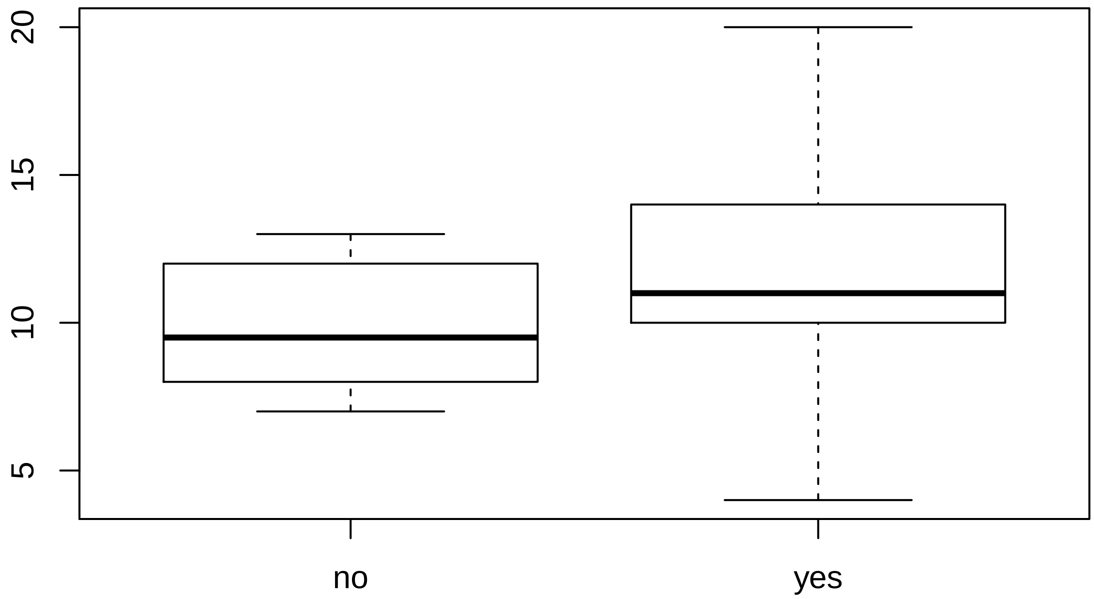

```r
table(failures)
```

```
## failures
##   0   1   2   3
## 294  40  12  11
```

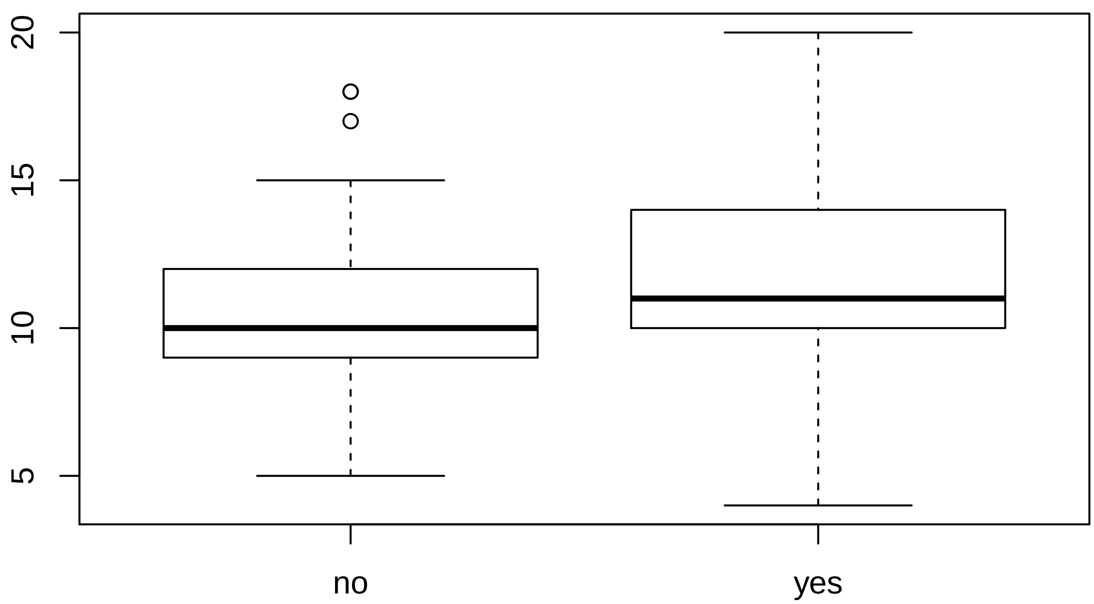
```r
boxplot(G3~activities) #there might be a small positive correlation between participating to activities
```




```r
boxplot(G3~higher) #students who plan to go to higher education tend to have higher grades
```



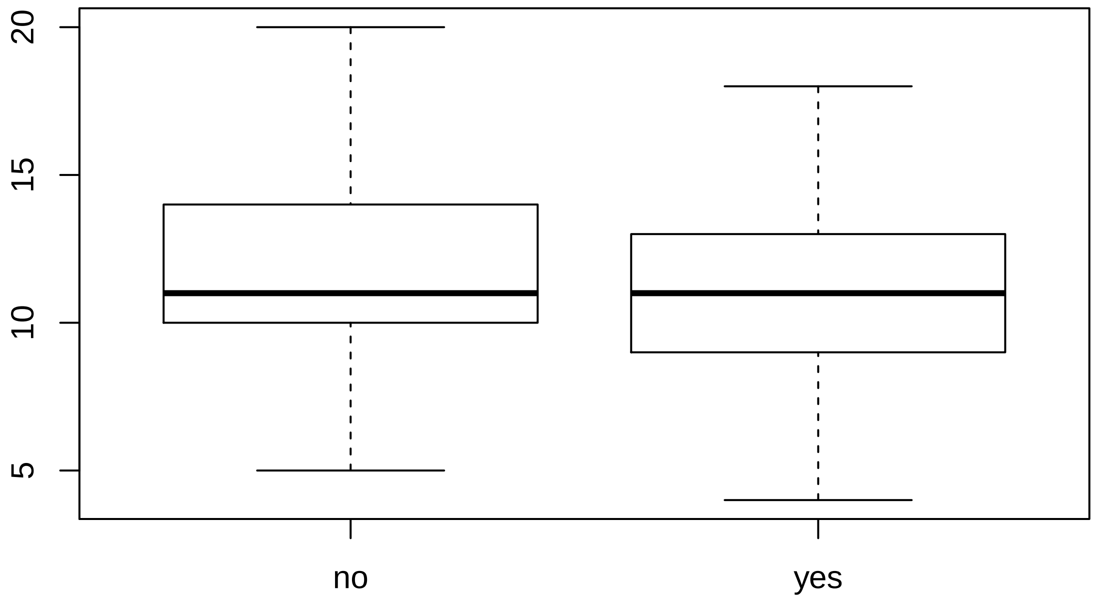

```
boxplot(G3~internet) #seems that students who have access to internet have slightly higher grades
```



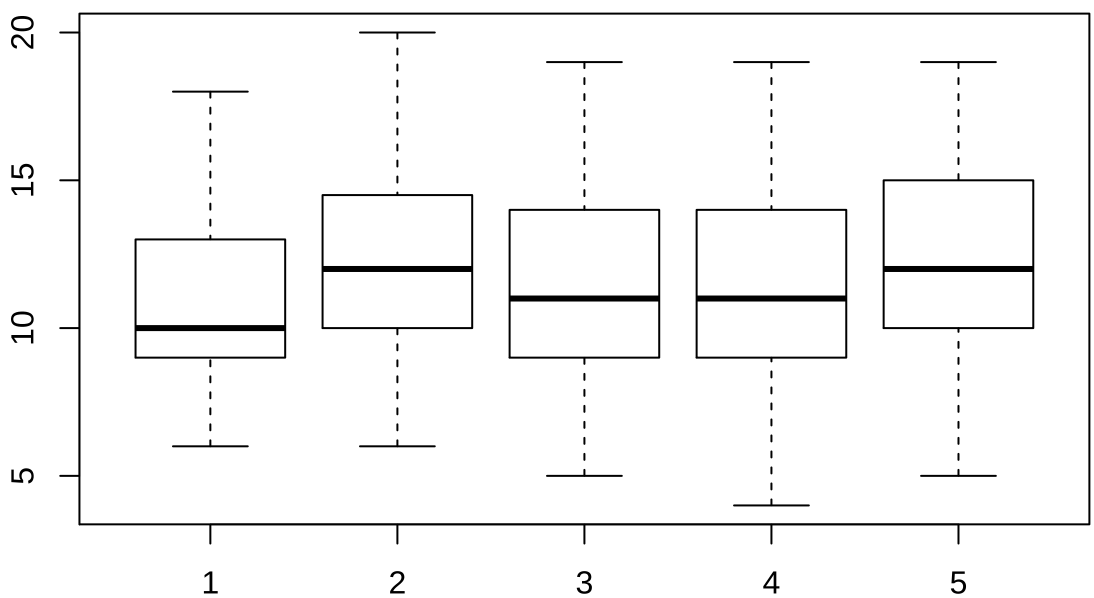

```
table(internet)
```

```
## internet
##  no yes
##  58 299
```

```
boxplot(G3~romantic) #seems that students who are not involved in a relationship have slighly higher gr
```



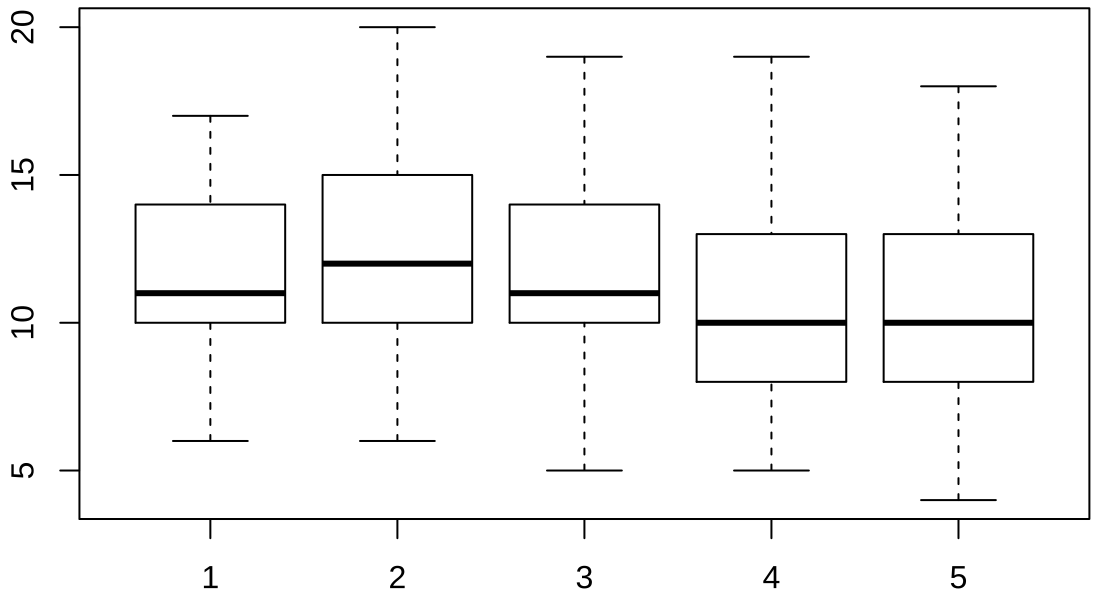

```r
boxplot(G3~freetime) #not easy to judge
```



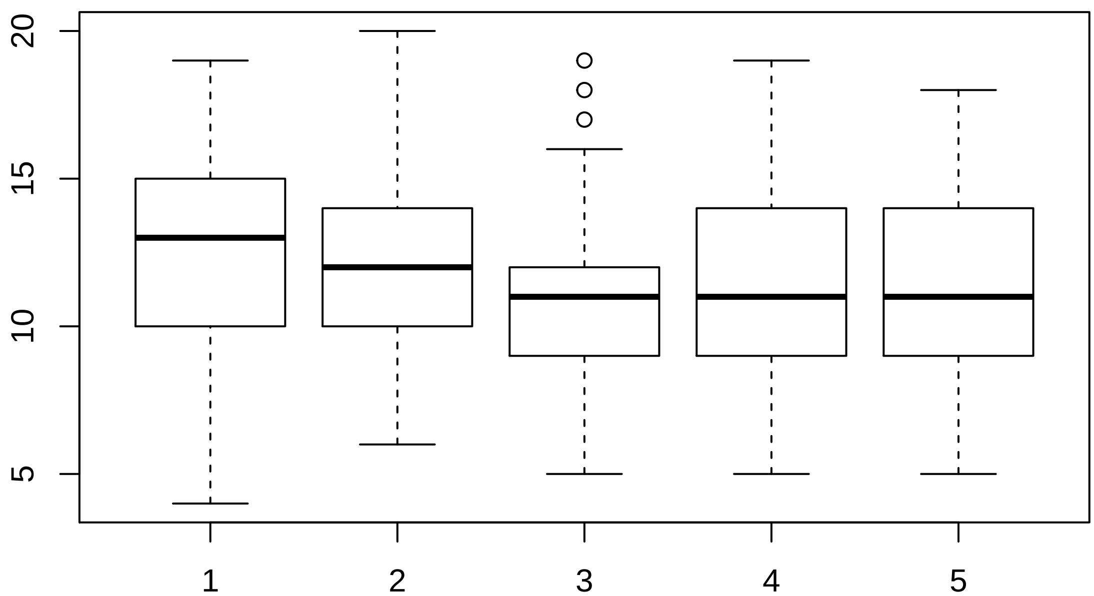

```r
table(freetime)
```

```
## freetime
##   1   2   3   4   5
##  17  60 136 106  38
```

```r
boxplot(G3~goout) #seems the frequency of going out might suggest a lower grade
```



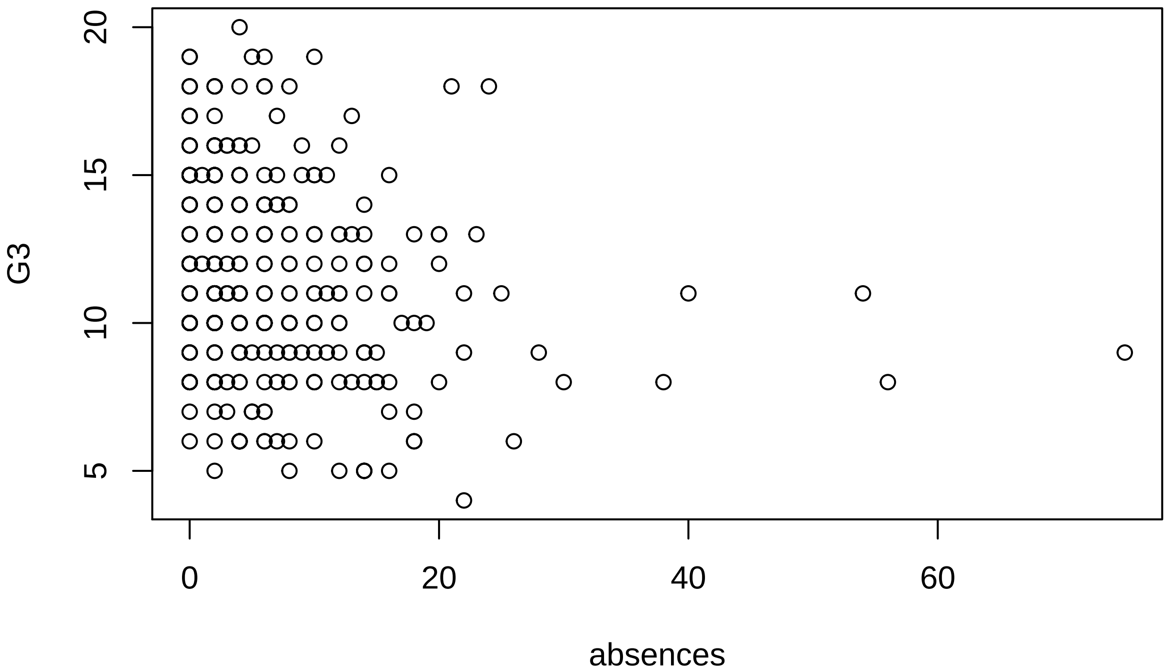

```r
boxplot(G3~health) #it is very scary but seems like students who have very bad health might perform bet
```



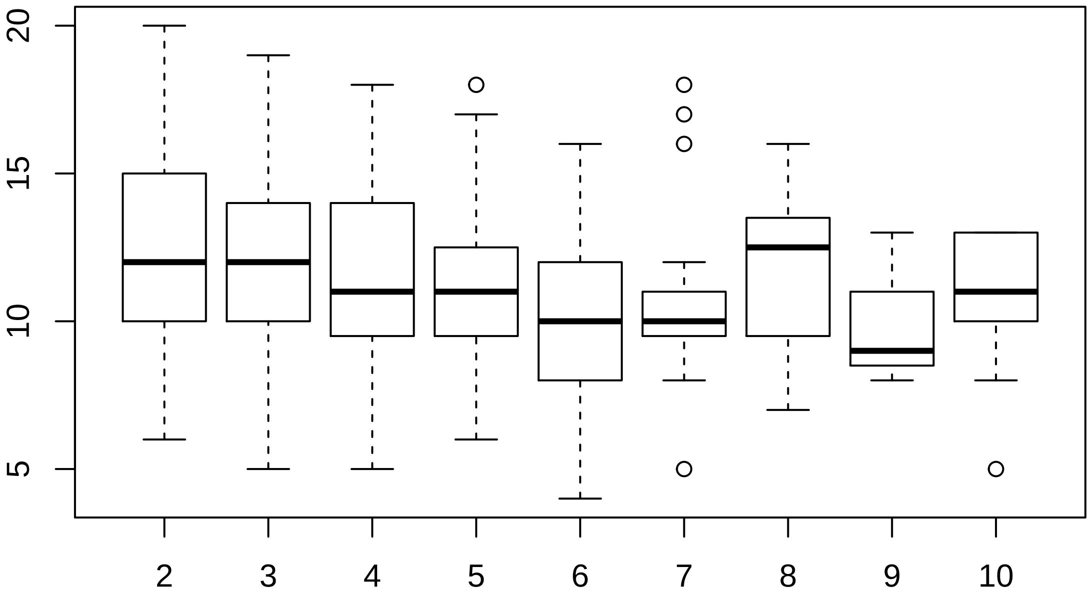

```r
table(health)
```

```
## health
##   1   2   3   4   5
##  45  38  83  58 133
```

```r
plot(absences, G3) #students with a lot of absences can't get high grades
```



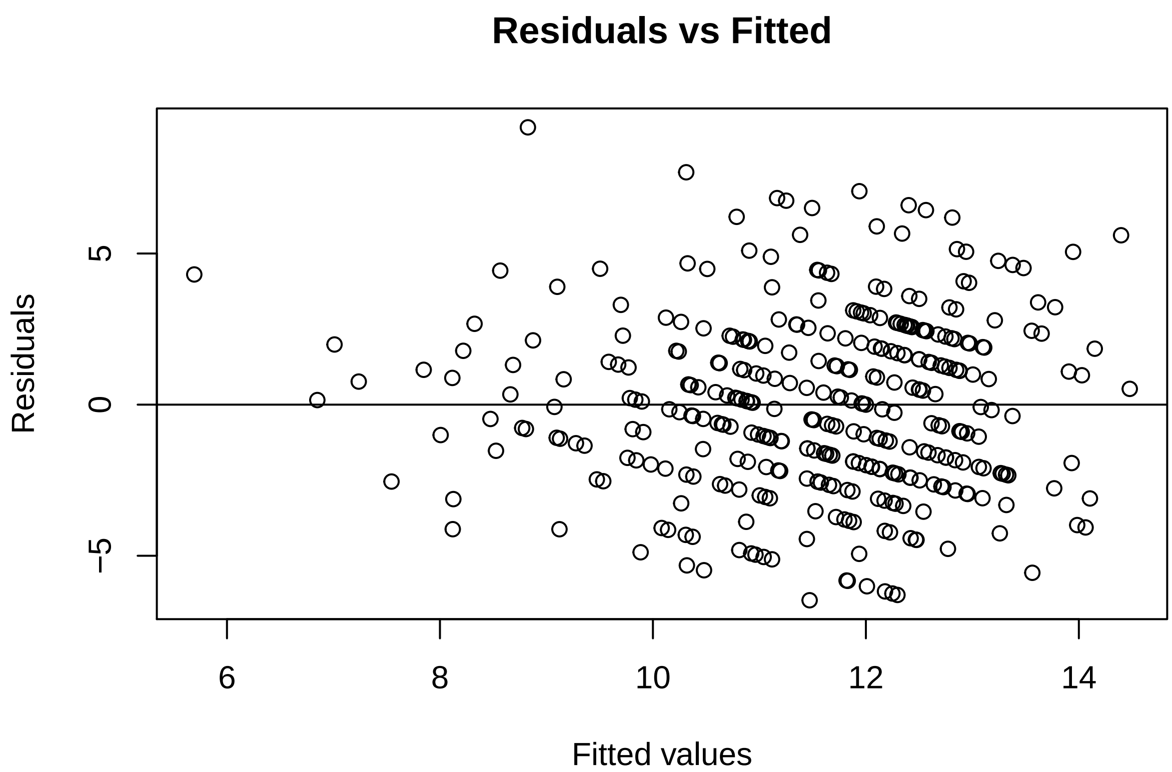

```
DWalc <- Dalc + Walc
boxplot(G3~DWalc) #worth exploring more
```



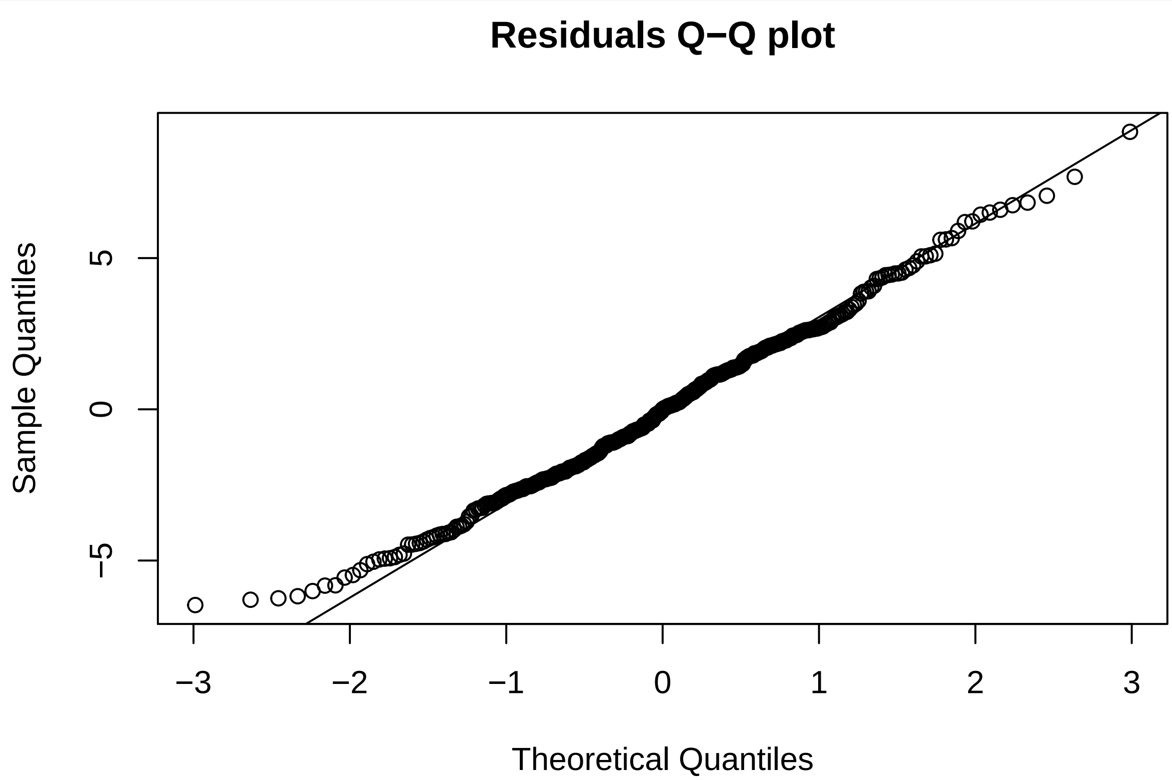

**Regression analysis**

```
lmfit3 = lm(G3~sex+age+address+studytime+failures+activities+higher+internet+romantic+freetime+goout+hea
```

```
residuals=lmfit3$residuals
y_hat=lmfit3$fitted.values
#Not linear
plot(y_hat,residuals,xlab='Fitted values',ylab='Residuals',main='Residuals vs Fitted')
abline(h=0)
```




```r
#Not normal
qqnorm(residuals,main='Residuals Q-Q plot')
qqline(residuals)
```



```r
shapiro.test(residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.99348, p-value = 0.1262
```

Four assumptions L I N E

```r
cor(Alcohol[,c(3, 14, 25, 26, 29, 30, 34)])
```

```
##                   age    studytime    freetime        goout       health
## age        1.000000000 -0.004140037  0.01643439  0.126963880 -0.062187369
## studytime -0.004140037  1.000000000 -0.14319841 -0.063903675 -0.075615863
## freetime   0.016434389 -0.143198407  1.00000000  0.285018715  0.075733357
## goout      0.126963880 -0.063903675  0.28501871  1.000000000 -0.009577254
## health    -0.062187369 -0.075615863  0.07573336 -0.009577254  1.000000000
## absences   0.175230079 -0.062700175 -0.05807792  0.044302220 -0.029936711
## DWalc      0.134972274 -0.252697870  0.18975355  0.392682938  0.094662362
##              absences       DWalc
## age        0.17523008  0.13497227
## studytime -0.06270018 -0.25269787
## freetime  -0.05807792  0.18975355
## goout      0.04430222  0.39268294
## health    -0.02993671  0.09466236
## absences   1.00000000  0.13868748
## DWalc      0.13868748  1.00000000
```

```r
require(leaps)
X1=model.matrix(G3~sex+age+address+studytime+failures+activities+higher+internet+romantic+freetime+goou
R2=vector("numeric",14)
  for(j in 1:14){
    y_tmp=X1[,1+j]
    x_tmp=as.matrix(X1[,-c(1,1+j)])
    lm_fit=lm(y_tmp~x_tmp)
    R2[j]=summary(lm_fit)$r.squared
}
VIF=1/(1-R2)
names(VIF)=colnames(X1)[-1]
VIF
```

```
##          sexM          age      addressU     studytime      failures
##      1.293529     1.243493      1.116295      1.212939      1.208512
## activitiesyes     higheryes   internetyes   romanticyes      freetime
##      1.098679     1.146835      1.123540      1.091266      1.178560
##         goout        health      absences         DWalc
##      1.352941      1.044548      1.156869      1.465652
```

### Outliers

```r
residuals=lmfit3$residuals
sigma_hat=summary(lmfit3)$sigma
X1=model.matrix(G3~sex+age+address+studytime+failures+activities+higher+internet+romantic+freetime+goou
H=X1%*%solve(t(X1)%*%X1)%*%t(X1)
```

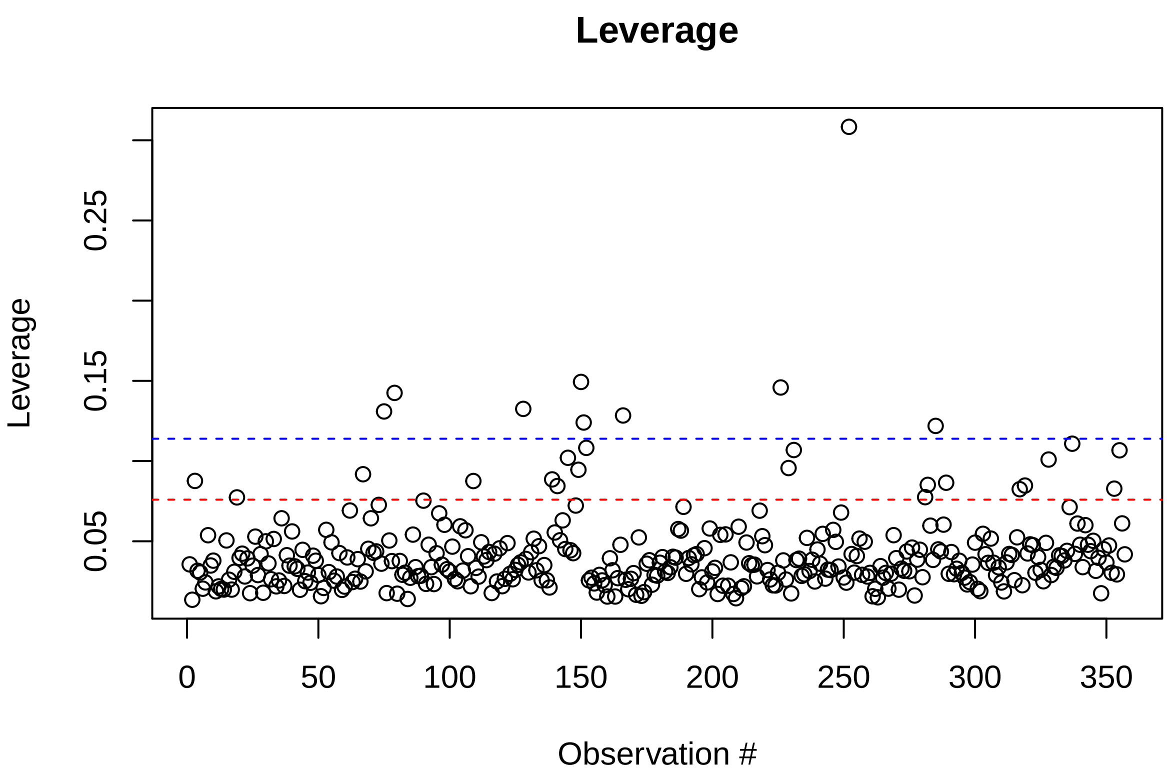
```r
h=diag(H)
r=residuals/(sigma_hat*sqrt(1-h))
p=15
n=395
thresh2=2*p/n
thresh3=3*p/n
which(h>thresh2) #showing the points of the leverage
```

```
##   3  19  67  75  79 109 128 139 141 145 149 150 151 152 166 226 229 231
##   3  19  67  75  79 109 128 139 141 145 149 150 151 152 166 226 229 231
## 252 281 282 285 289 317 319 328 337 353 355
## 252 281 282 285 289 317 319 328 337 353 355
```

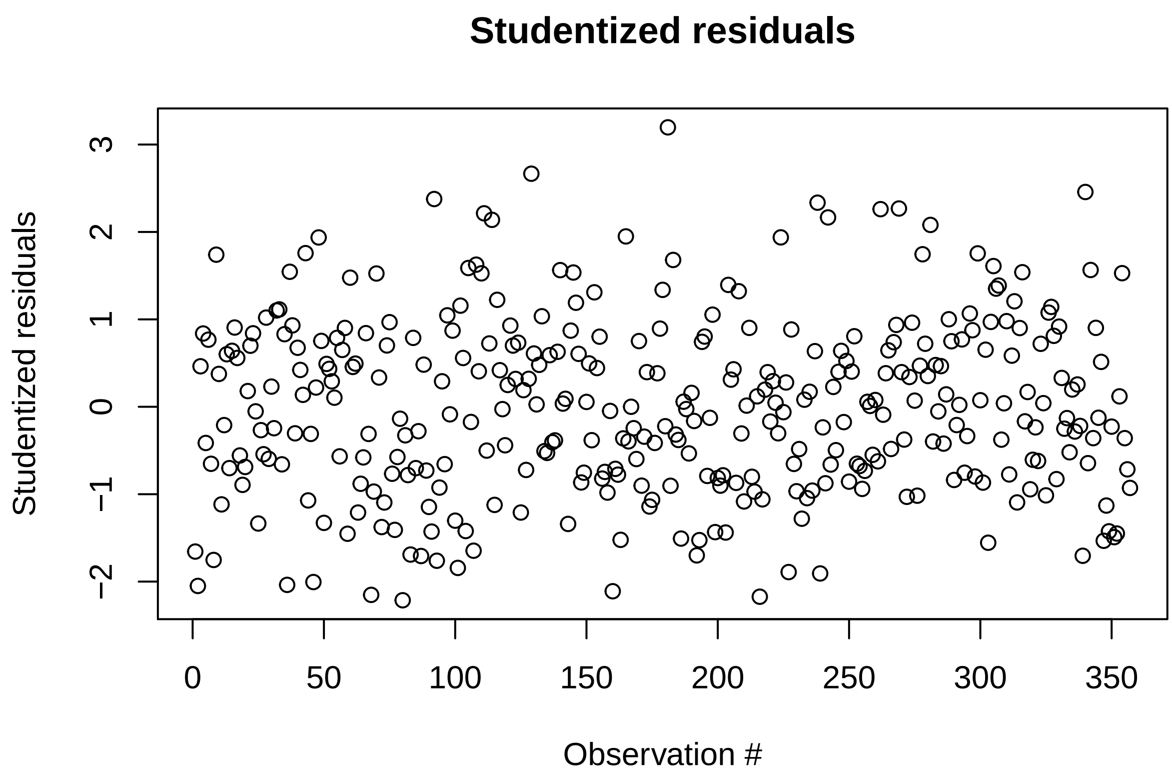
```r
which(h>thresh3)
```

```
##  75  79 128 150 151 166 226 252 285
##  75  79 128 150 151 166 226 252 285
```

```r
plot(h,xlab='Observation #',ylab='Leverage',main='Leverage')
abline(h=thresh2,lty=2,col="red")
abline(h=thresh3,lty=2,col="blue")
```




```r
t=r*sqrt((n-p-1)/(n-p-r^2))
plot(t,xlab='Observation #',ylab='Studentized residuals',main='Studentized residuals')
```



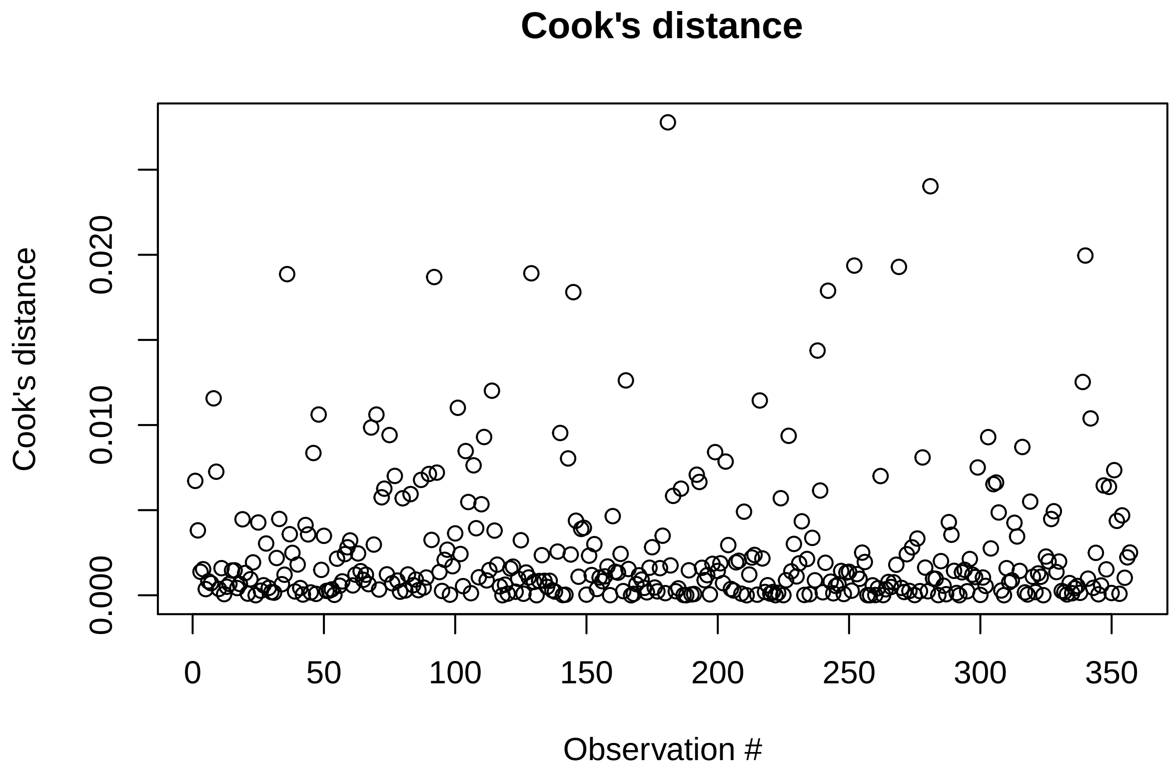

```
which(t< (-2))
```

```
##   2  36  46  68  80 178 235
##   2  36  46  68  80 160 216
```

```
which(t > 2)
```

```
##  92 111 114 130 199 261 266 287 294 307 375
##  92 111 114 129 181 238 242 262 269 281 340
```

```
D=(1/p)*r^2*h/(1-h)
plot(D,xlab='Observation #',ylab='Cook\'s distance',main='Cook\'s distance')
```




```
which(D>0.015)
```

```
##  36  92 130 158 199 266 277 294 307 375
##  36  92 129 145 181 242 252 269 281 340
```

### Model selection

We use Best model selection method

– describe –

```
subset=regsubsets(G3~sex+age+address+studytime+failures+activities+higher+internet+romantic+freetime+go

sum_subset=summary(subset)
sum_subset$which
```

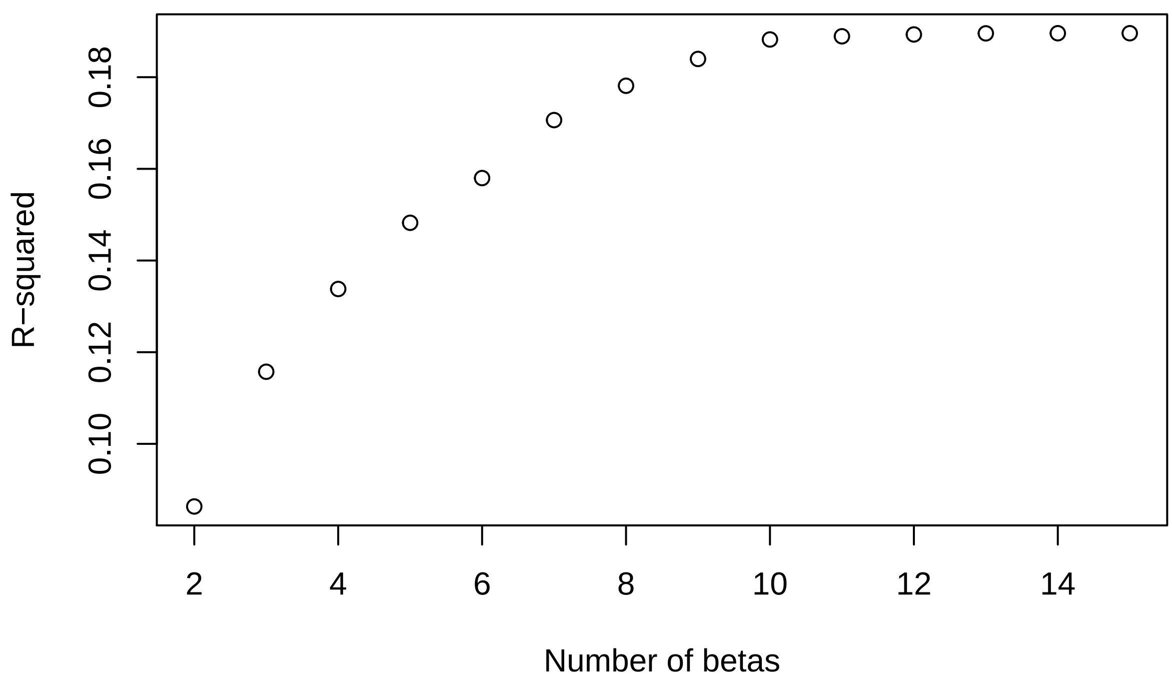
```
##    (Intercept)  sexM   age addressU studytime failures activitiesyes
## 1         TRUE FALSE FALSE    FALSE     FALSE     TRUE         FALSE
## 2         TRUE FALSE FALSE    FALSE     FALSE     TRUE         FALSE
## 3         TRUE FALSE FALSE    FALSE     FALSE     TRUE         FALSE
## 4         TRUE FALSE FALSE    FALSE     FALSE     TRUE         FALSE
## 5         TRUE  TRUE FALSE    FALSE     FALSE     TRUE         FALSE
## 6         TRUE  TRUE FALSE     TRUE      TRUE     TRUE         FALSE
## 7         TRUE  TRUE FALSE     TRUE      TRUE     TRUE         FALSE
## 8         TRUE  TRUE FALSE     TRUE      TRUE     TRUE         FALSE
## 9         TRUE  TRUE FALSE     TRUE      TRUE     TRUE         FALSE
## 10        TRUE  TRUE FALSE     TRUE      TRUE     TRUE         FALSE
## 11        TRUE  TRUE FALSE     TRUE      TRUE     TRUE         FALSE
## 12        TRUE  TRUE FALSE     TRUE      TRUE     TRUE          TRUE
## 13        TRUE  TRUE  TRUE     TRUE      TRUE     TRUE          TRUE
## 14        TRUE  TRUE  TRUE     TRUE      TRUE     TRUE          TRUE
```

```
##    higheryes internetyes romanticyes freetime goout health absences DWalc
## 1      FALSE       FALSE       FALSE    FALSE FALSE  FALSE    FALSE FALSE
## 2      FALSE       FALSE       FALSE    FALSE FALSE  FALSE     TRUE FALSE
## 3      FALSE       FALSE       FALSE    FALSE  TRUE  FALSE     TRUE FALSE
## 4      FALSE        TRUE       FALSE    FALSE  TRUE  FALSE     TRUE FALSE
## 5      FALSE        TRUE       FALSE    FALSE FALSE  FALSE     TRUE  TRUE
## 6      FALSE       FALSE       FALSE    FALSE  TRUE  FALSE     TRUE FALSE
## 7      FALSE       FALSE       FALSE    FALSE  TRUE   TRUE     TRUE FALSE
## 8      FALSE        TRUE       FALSE    FALSE  TRUE   TRUE     TRUE FALSE
## 9      FALSE        TRUE       FALSE    FALSE  TRUE   TRUE     TRUE  TRUE
## 10      TRUE        TRUE       FALSE    FALSE  TRUE   TRUE     TRUE  TRUE
## 11      TRUE        TRUE       FALSE     TRUE  TRUE   TRUE     TRUE  TRUE
## 12      TRUE        TRUE       FALSE     TRUE  TRUE   TRUE     TRUE  TRUE
## 13      TRUE        TRUE       FALSE     TRUE  TRUE   TRUE     TRUE  TRUE
## 14      TRUE        TRUE        TRUE     TRUE  TRUE   TRUE     TRUE  TRUE
```

```
p_full=15
p=2:p_full
RSS_p=sum_subset$rss
totalSS=sum((G3-mean(G3))^2)
R2_p=1-RSS_p/totalSS
R2_p
```

```
##  [1] 0.0863366 0.1157351 0.1337869 0.1482315 0.1579957 0.1706373 0.1781268
##  [8] 0.1839771 0.1882308 0.1889238 0.1893234 0.1895638 0.1895889 0.1895898
```

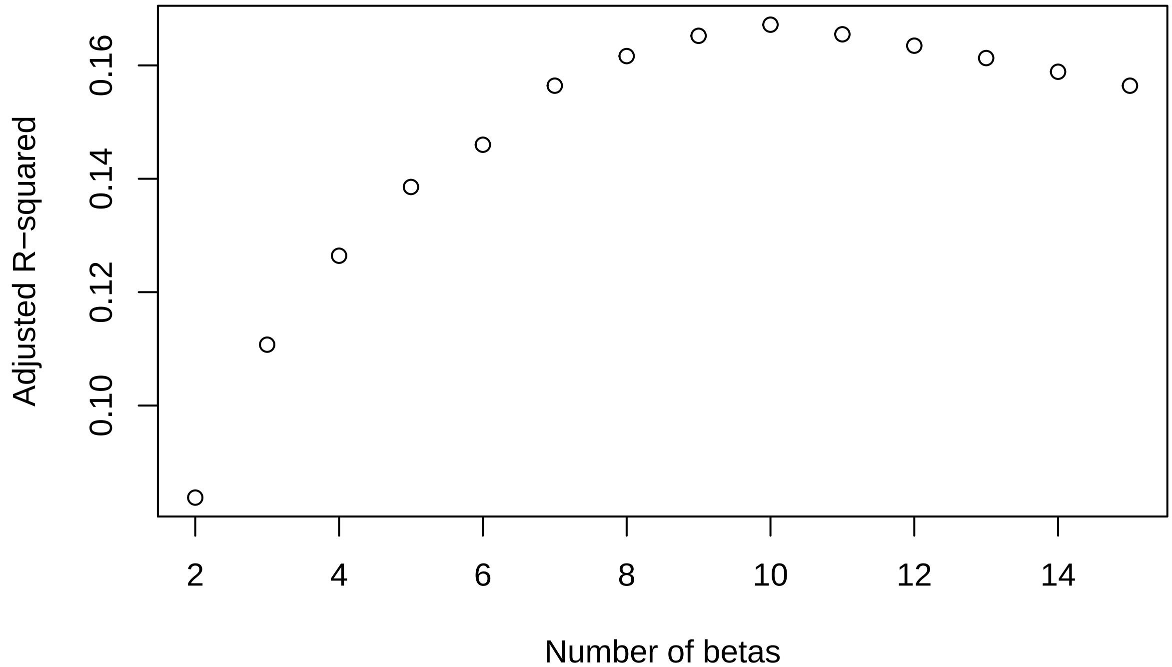
```
plot(p,R2_p,xlab="Number of betas",ylab="R-squared")
```




```
n=nrow(Alcohol3)
R2_adj=1-(RSS_p/(n-p))/(totalSS/(n-1))
R2_adj
```

```
##  [1] 0.0837629 0.1107392 0.1264253 0.1385523 0.1460013 0.1564197 0.1616423
##  [8] 0.1652179 0.1671762 0.1654822 0.1634757 0.1612928 0.1588736 0.1564151
```

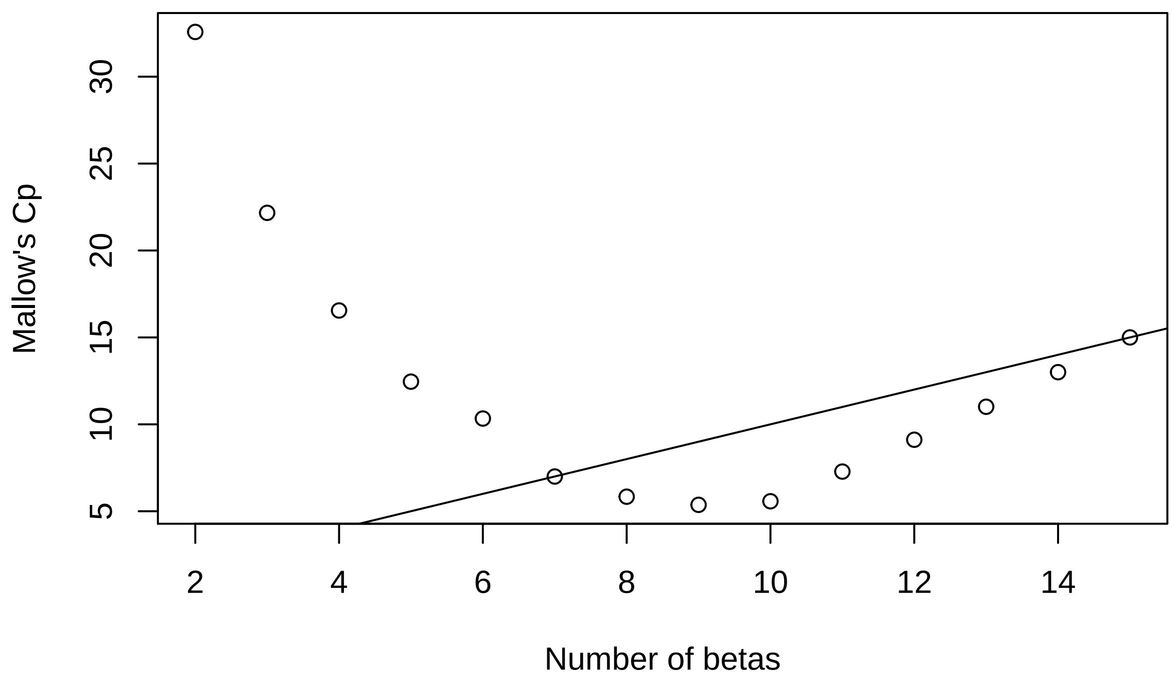
```r
plot(p,R2_adj,xlab="Number of betas",ylab="Adjusted R-squared") #10 best model (9 pred)
```




```r
sigma_hat_full=summary(lmfit3)$sigma
C_p=RSS_p/(sigma_hat_full^2)+2*p-n
C_p
```

```
##  [1] 32.573724 22.167324 16.549302 12.453550 10.332990  6.998100  5.837475
##  [8]  5.368606  5.573517  7.281074  9.112428 11.010956 13.000370 15.000000
```

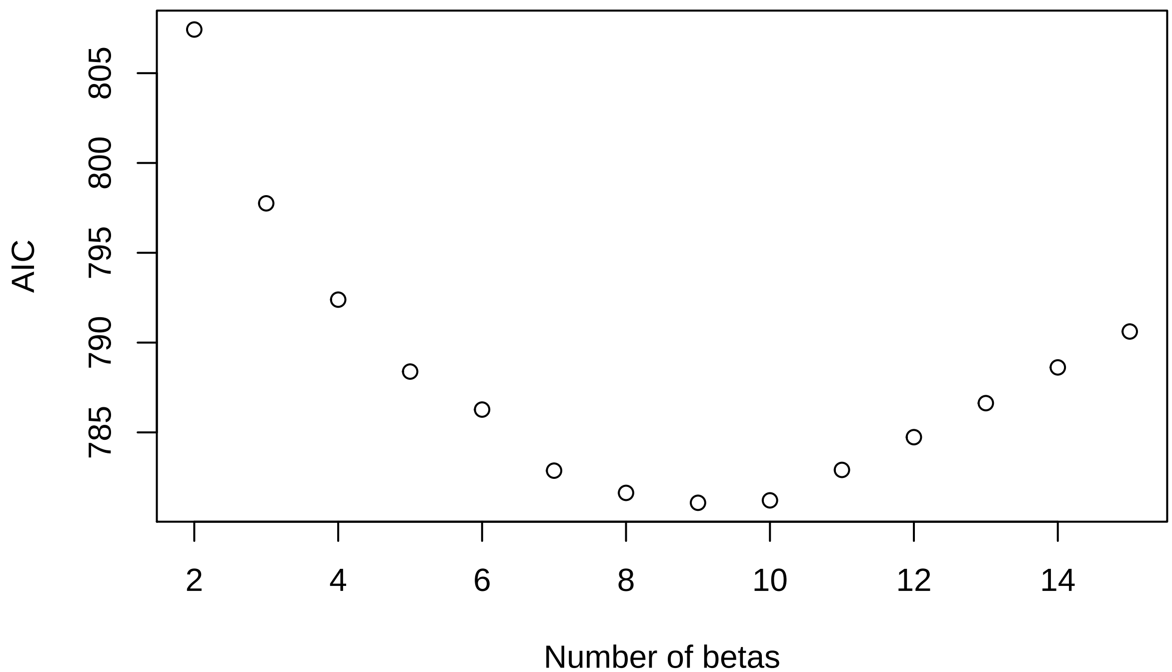
```r
plot(p,C_p,xlab="Number of betas",ylab="Mallow's Cp") #7 (6 pred)
abline(0,1) # what should be set for this one?
```



```r
aic_p=n*log(RSS_p/n)+2*p
aic_p
```

```
##  [1] 807.4291 797.7532 792.3898 788.3864 786.2704 782.8698 781.6313
##  [8] 781.0810 781.2152 782.9103 784.7344 786.6285 788.6174 790.6170
```

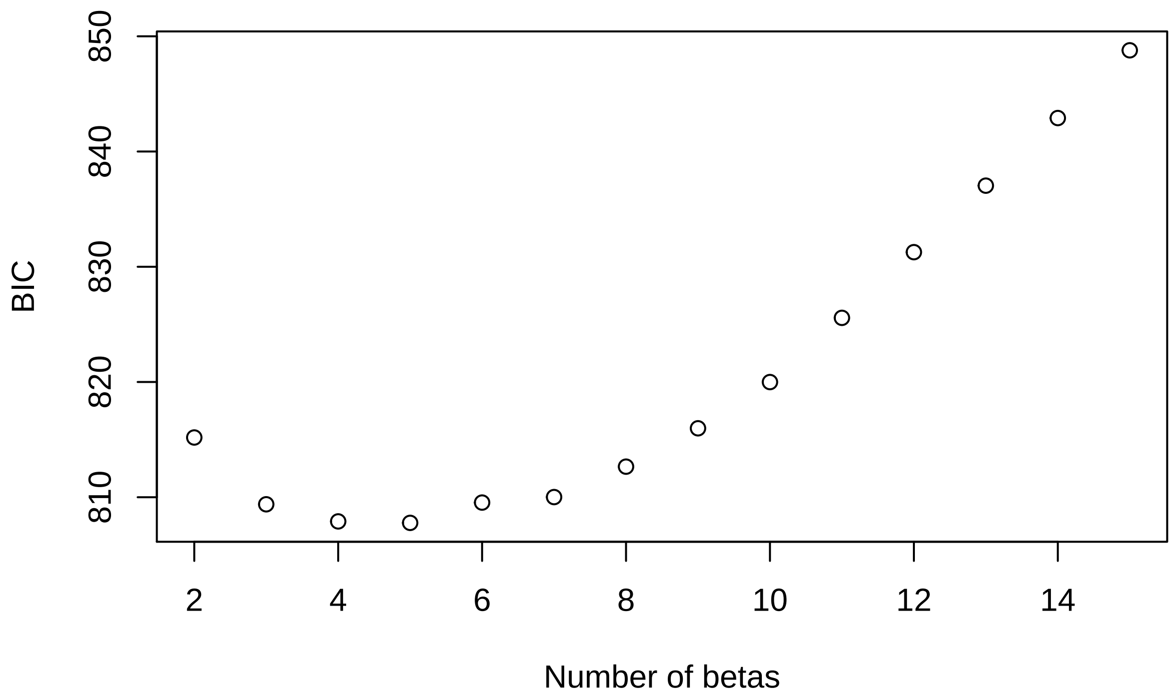
```r
plot(p,aic_p,xlab="Number of betas",ylab="AIC") #9 (8 predictors)
```




```r
bic_p=n*log(RSS_p/n)+p*log(n)
bic_p
```

```
##  [1] 815.1845 809.3864 807.9008 807.7751 809.5368 810.0139 812.6532
##  [8] 815.9806 819.9926 825.5654 831.2672 837.0390 842.9057 848.7831
```

```r
plot(p,bic_p,xlab="Number of betas",ylab="BIC") #5 (4 predictors)
```




```r
cbind(sum_subset$which,R2_adj,C_p,aic_p,bic_p)
```

```
##    (Intercept) sexM age addressU studytime failures activitiesyes
## 1            1    0   0        0         0        1             0
## 2            1    0   0        0         0        1             0
## 3            1    0   0        0         0        1             0
## 4            1    0   0        0         0        1             0
## 5            1    1   0        0         0        1             0
## 6            1    1   0        1         1        1             0
## 7            1    1   0        1         1        1             0
## 8            1    1   0        1         1        1             0
## 9            1    1   0        1         1        1             0
## 10           1    1   0        1         1        1             0
## 11           1    1   0        1         1        1             0
## 12           1    1   0        1         1        1             1
## 13           1    1   1        1         1        1             1
## 14           1    1   1        1         1        1             1
##    higheryes internetyes romanticyes freetime goout health absences DWalc
## 1          0           0           0        0     0      0        0     0
## 2          0           0           0        0     0      0        1     0
## 3          0           0           0        0     1      0        1     0
## 4          0           1           0        0     1      0        1     0
## 5          0           1           0        0     0      0        1     1
## 6          0           0           0        0     1      0        1     0
## 7          0           0           0        0     1      1        1     0
## 8          0           1           0        0     1      1        1     0
## 9          0           1           0        0     1      1        1     1
## 10         1           1           0        0     1      1        1     1
## 11         1           1           0        1     1      1        1     1
## 12         1           1           0        1     1      1        1     1
## 13         1           1           0        1     1      1        1     1
## 14         1           1           1        1     1      1        1     1
```

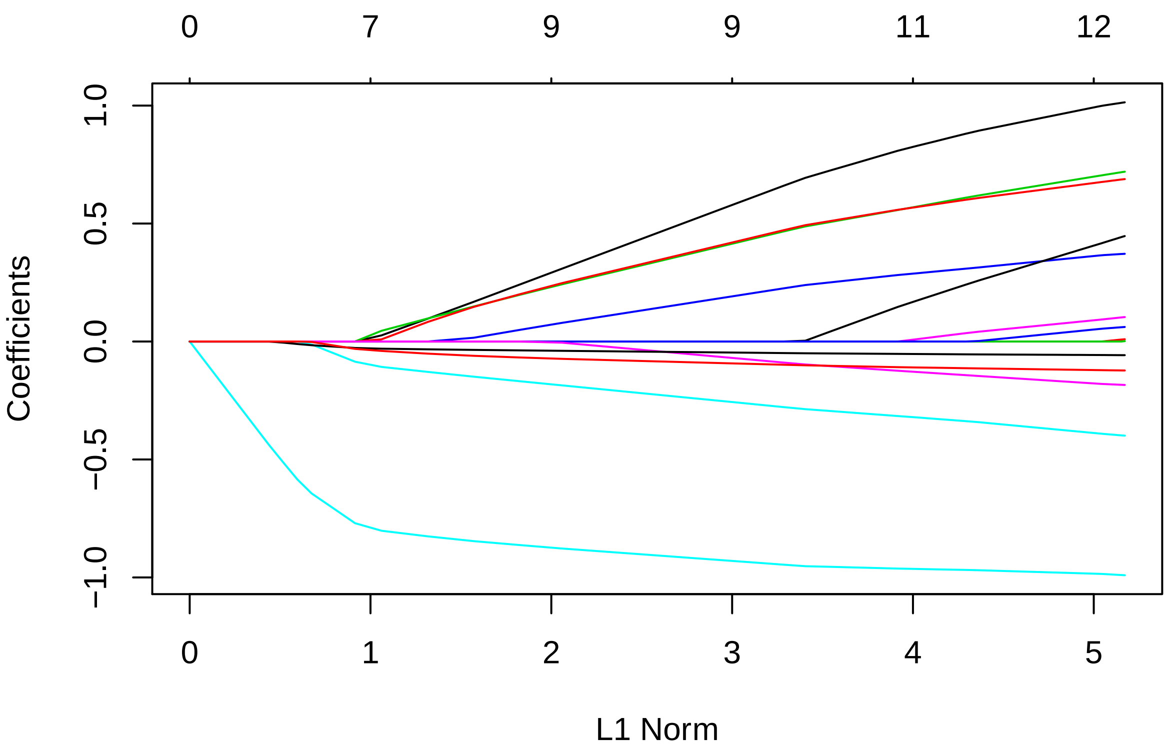
```
##         R2_adj       C_p    aic_p    bic_p
## 1  0.0837629 32.573724 807.4291 815.1845
## 2  0.1107392 22.167324 797.7532 809.3864
## 3  0.1264253 16.549302 792.3898 807.9008
## 4  0.1385523 12.453550 788.3864 807.7751
## 5  0.1460013 10.332990 786.2704 809.5368
## 6  0.1564197  6.998100 782.8698 810.0139
## 7  0.1616423  5.837475 781.6313 812.6532
## 8  0.1652179  5.368606 781.0810 815.9806
## 9  0.1671762  5.573517 781.2152 819.9926
## 10 0.1654822  7.281074 782.9103 825.5654
## 11 0.1634757  9.112428 784.7344 831.2672
## 12 0.1612928 11.010956 786.6285 837.0390
## 13 0.1588736 13.000370 788.6174 842.9057
## 14 0.1564151 15.000000 790.6170 848.7831
```

```
#install.packages("glmnet")

#How to use Lasso?
library(glmnet)
```

```
## Loading required package: foreach
```

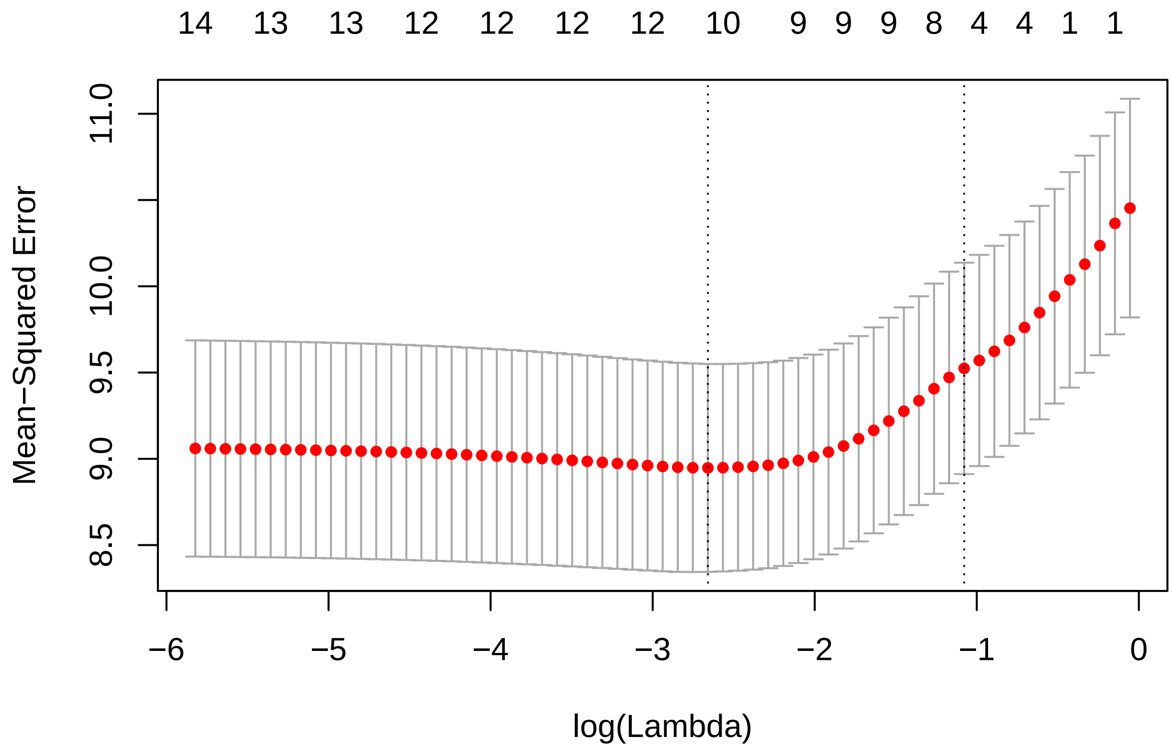
```
## Loaded glmnet 2.0-13
```

```
y=Alcohol3$G3
X2 <- X1[,-1]
lasso_fit=glmnet(X2,y,alpha=1)
plot(lasso_fit)
```




```
k=10
cv_lasso=cv.glmnet(X2,y,nfolds=k)
```

```r
plot(cv_lasso)
```




```r
#look for cv and
```

What model do we want?

```r
model4 <- lm(G3~failures+internet+goout+absences)
model6 <- lm(G3~sex+address+studytime+failures+goout+absences)
model8 <- lm(G3~sex+address+studytime+failures+internet+goout+health+absences)
model9 <- lm(G3~sex+address+studytime+failures+internet+goout+health+absences+DWalc)

summary(model4)
```

```
##
## Call:
## lm(formula = G3 ~ failures + internet + goout + absences)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6410 -2.0633 -0.2019  2.0584  8.7954
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.75177    0.57905  22.022  < 2e-16 ***
## failures    -1.13581    0.24262  -4.681 4.07e-06 ***
## internetyes  1.06802    0.43714   2.443  0.01505 *
## goout       -0.43912    0.14771  -2.973  0.00315 **
## absences    -0.07179    0.01975  -3.634  0.00032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.996 on 352 degrees of freedom
## Multiple R-squared:  0.1482, Adjusted R-squared:  0.1386
## F-statistic: 15.31 on 4 and 352 DF,  p-value: 1.477e-11
```

```r
summary(model6)
```

```
##
## Call:
## lm(formula = G3 ~ sex + address + studytime + failures + goout +
##     absences)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6914 -2.1781 -0.1012  2.0101  9.4287
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.40984    0.75534  15.106  < 2e-16 ***
## sexM         0.92311    0.33103   2.789  0.00558 **
## addressU     0.92404    0.38284   2.414  0.01631 *
## studytime    0.47715    0.20007   2.385  0.01762 *
## failures    -1.10759    0.24106  -4.595 6.06e-06 ***
## goout       -0.44101    0.14602  -3.020  0.00271 **
## absences    -0.05700    0.01959  -2.910  0.00384 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.965 on 350 degrees of freedom
## Multiple R-squared:  0.1706, Adjusted R-squared:  0.1564
## F-statistic:    12 on 6 and 350 DF,  p-value: 2.89e-12
```

```r
summary(model8)
```

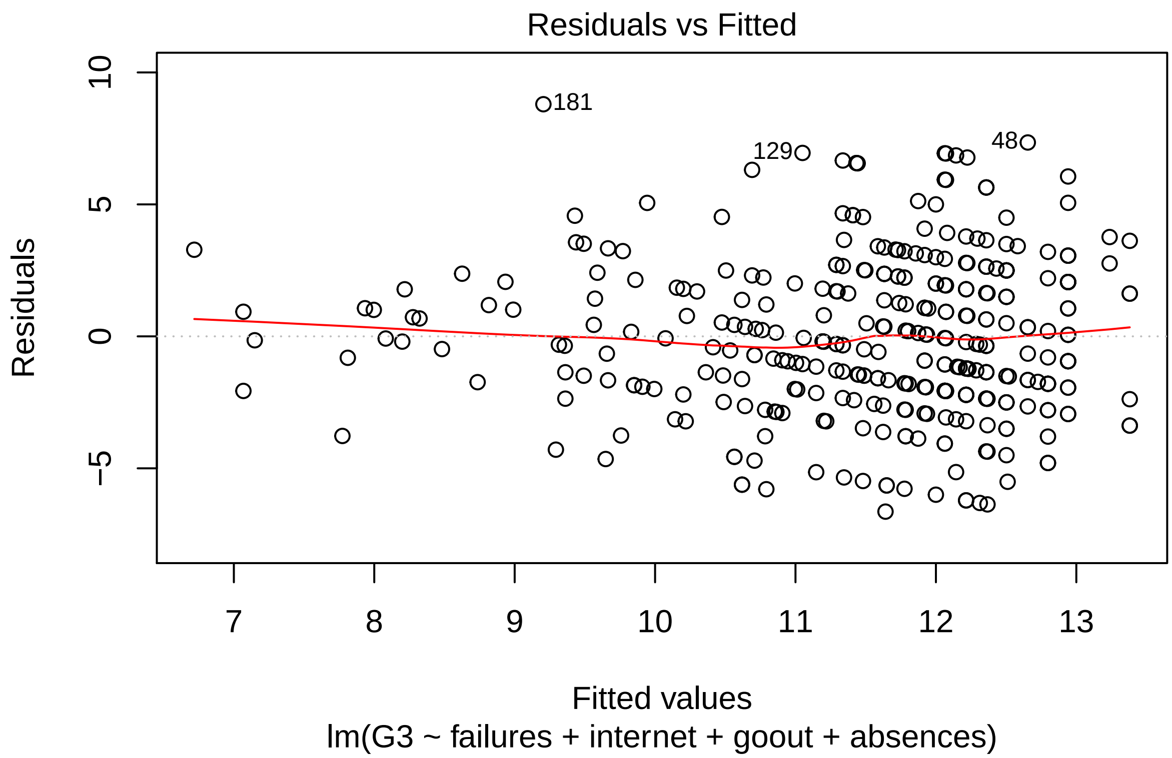
```
##
## Call:
## lm(formula = G3 ~ sex + address + studytime + failures + internet +
##     goout + health + absences)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4259 -2.1399 -0.1436  2.0680  9.0768
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.80473    0.88005  13.414  < 2e-16 ***
## sexM         0.93674    0.33371   2.807  0.00528 **
## addressU     0.79003    0.38933   2.029  0.04320 *
## studytime    0.43055    0.20037   2.149  0.03234 *
## failures    -1.05296    0.24109  -4.367 1.66e-05 ***
## internetyes  0.70070    0.44361   1.580  0.11512
## goout       -0.46716    0.14586  -3.203  0.00149 **
## health      -0.19427    0.11284  -1.722  0.08602 .
## absences    -0.06219    0.01968  -3.161  0.00171 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

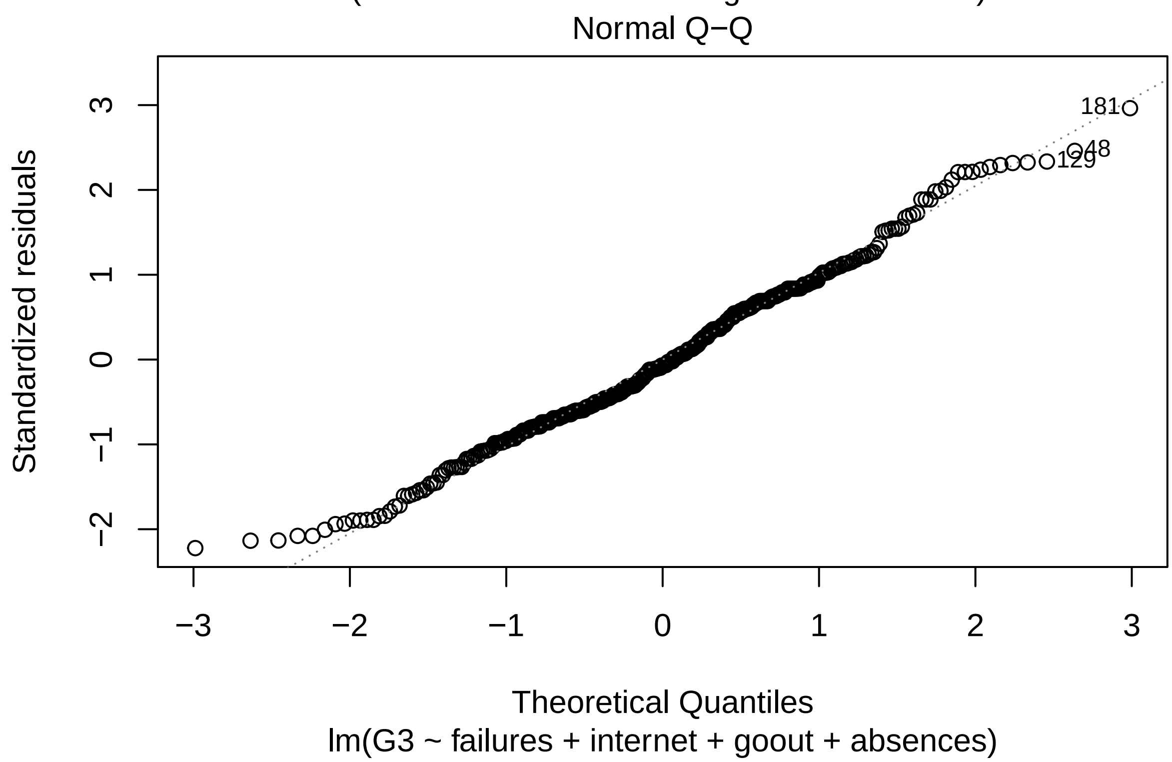
```
##
## Residual standard error: 2.949 on 348 degrees of freedom
## Multiple R-squared:  0.184,  Adjusted R-squared:  0.1652
## F-statistic: 9.807 on 8 and 348 DF,  p-value: 2.644e-12
```

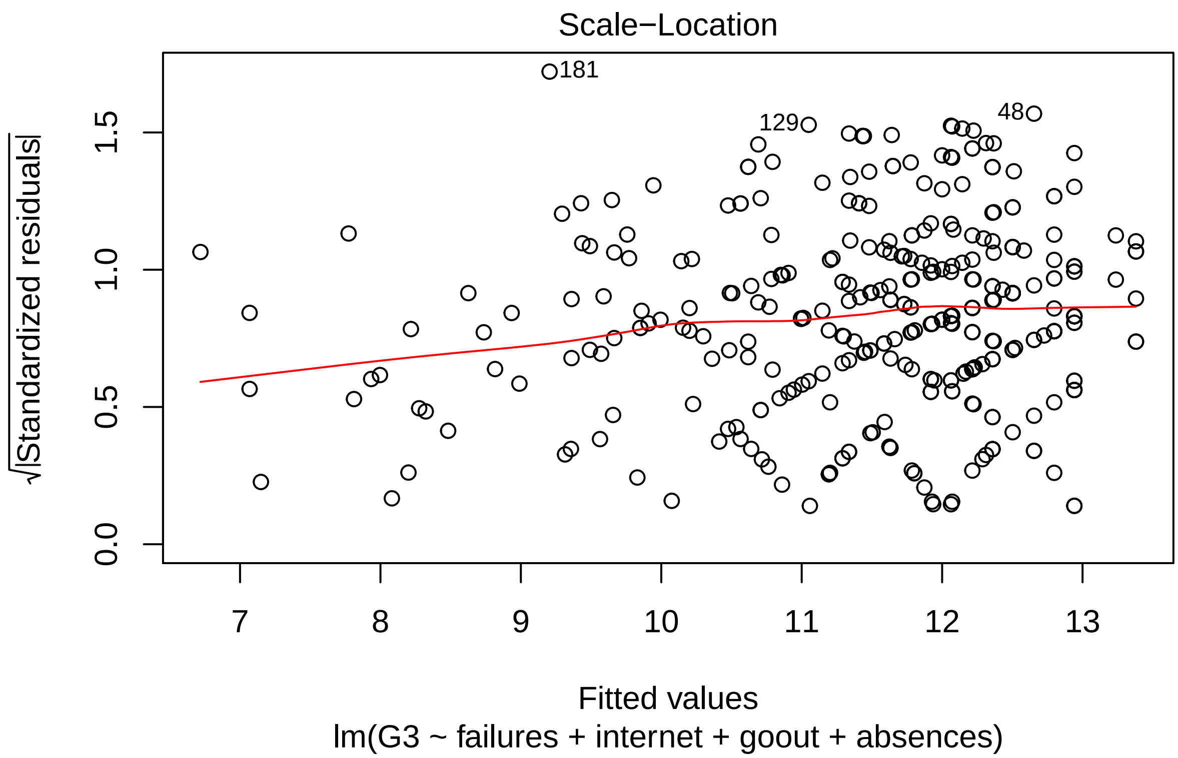
```r
summary(model9)
```

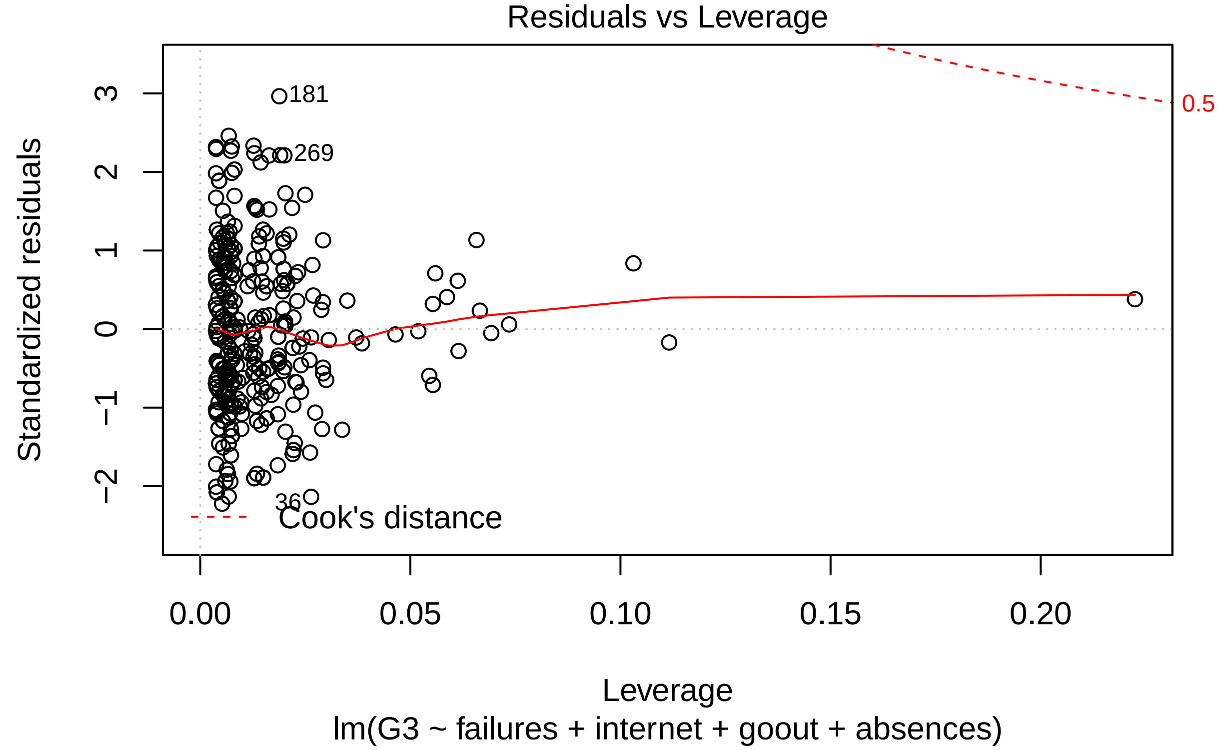
```
##
## Call:
## lm(formula = G3 ~ sex + address + studytime + failures + internet +
##     goout + health + absences + DWalc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4572 -2.1208  0.0222  2.0531  9.1072
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.02801    0.89447  13.447  < 2e-16 ***
## sexM         1.04261    0.34244   3.045  0.00251 **
## addressU     0.70992    0.39338   1.805  0.07199 .
## studytime    0.38434    0.20305   1.893  0.05921 .
## failures    -1.02035    0.24202  -4.216 3.18e-05 ***
## internetyes  0.71536    0.44322   1.614  0.10744
## goout       -0.37775    0.16007  -2.360  0.01883 *
## health      -0.18097    0.11314  -1.600  0.11061
## absences    -0.05929    0.01977  -2.999  0.00291 **
## DWalc       -0.12571    0.09323  -1.348  0.17840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.946 on 347 degrees of freedom
## Multiple R-squared:  0.1882, Adjusted R-squared:  0.1672
## F-statistic:  8.94 on 9 and 347 DF,  p-value: 3.781e-12
```

```r
plot(model4)
```


Residuals vs Fitted
Residuals
Fitted values
lm(G3 ~ failures + internet + goout + absences)
181
129
48
10
5
0
−5
7
8
9
10
11
12
13


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(G3 ~ failures + internet + goout + absences)
181
48
129
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3



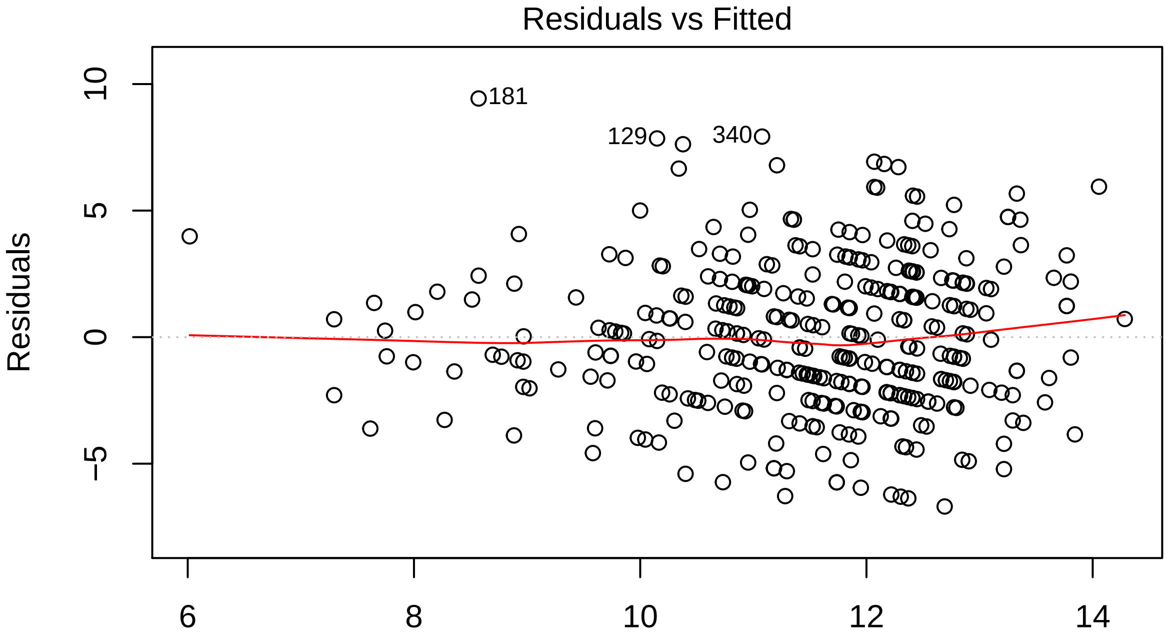



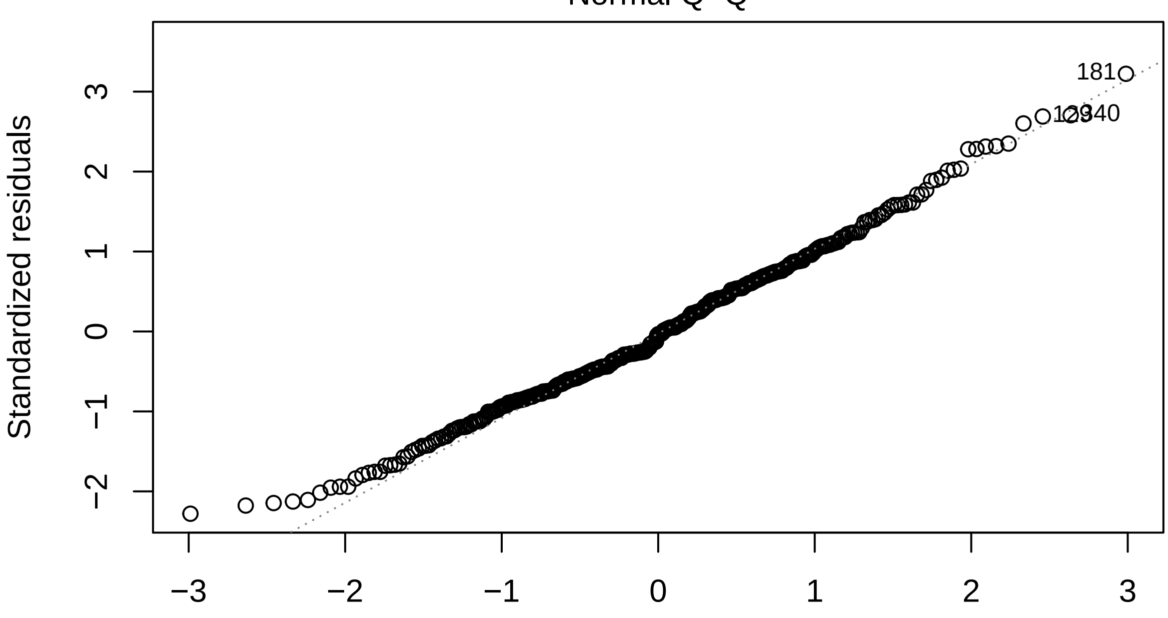

```
residuals4=model4$residuals
shapiro.test(residuals4)
```

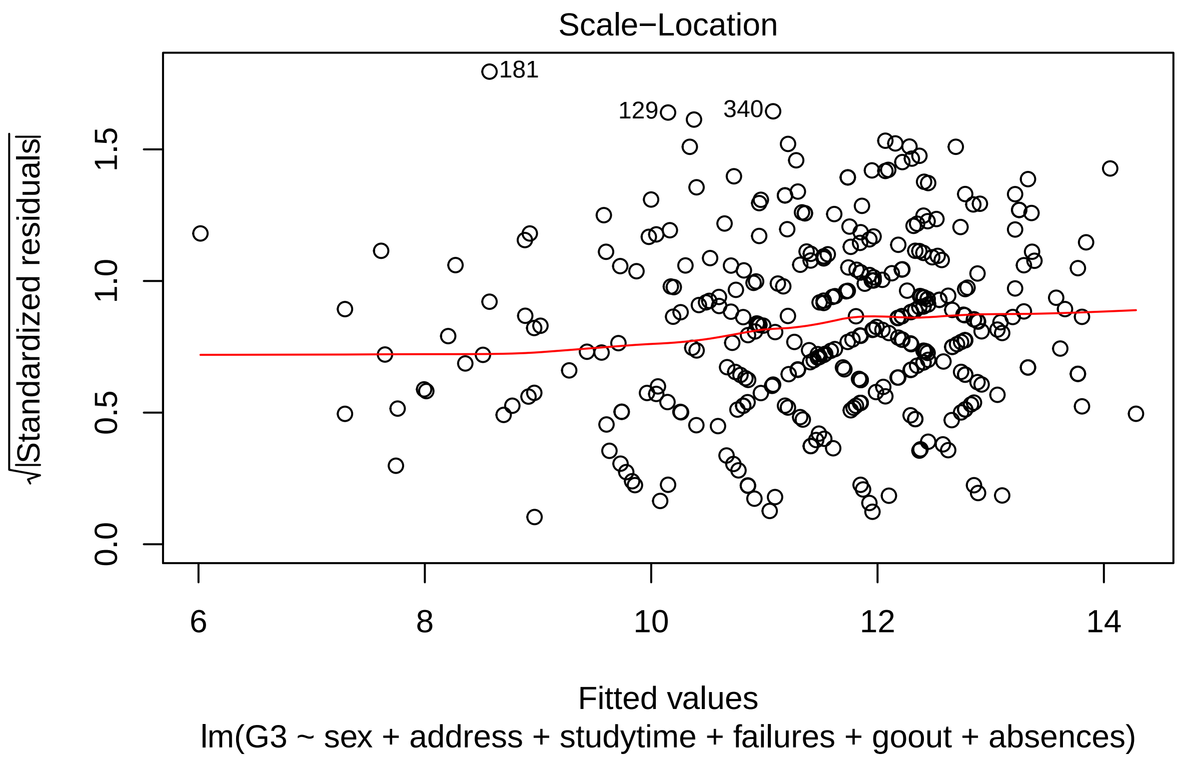
```
##
##  Shapiro-Wilk normality test
##
```

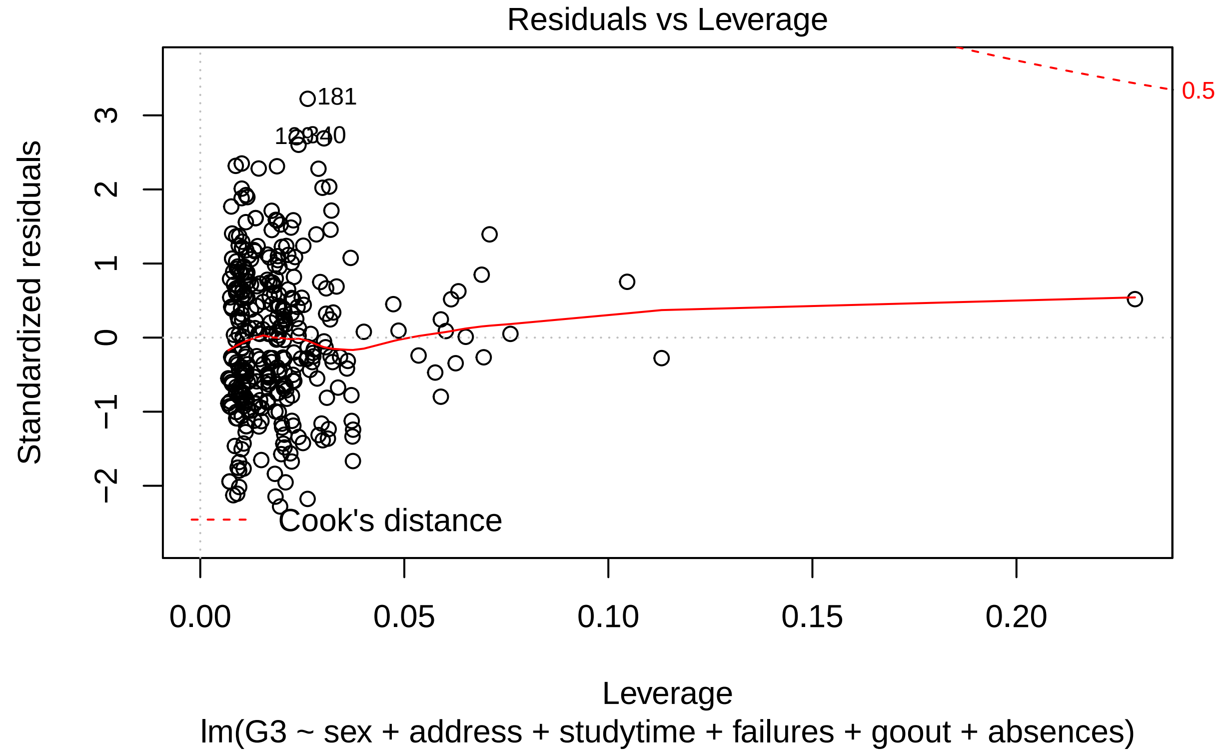
```
## data:  residuals4
## W = 0.99061, p-value = 0.0224
```

```r
plot(model6)
```





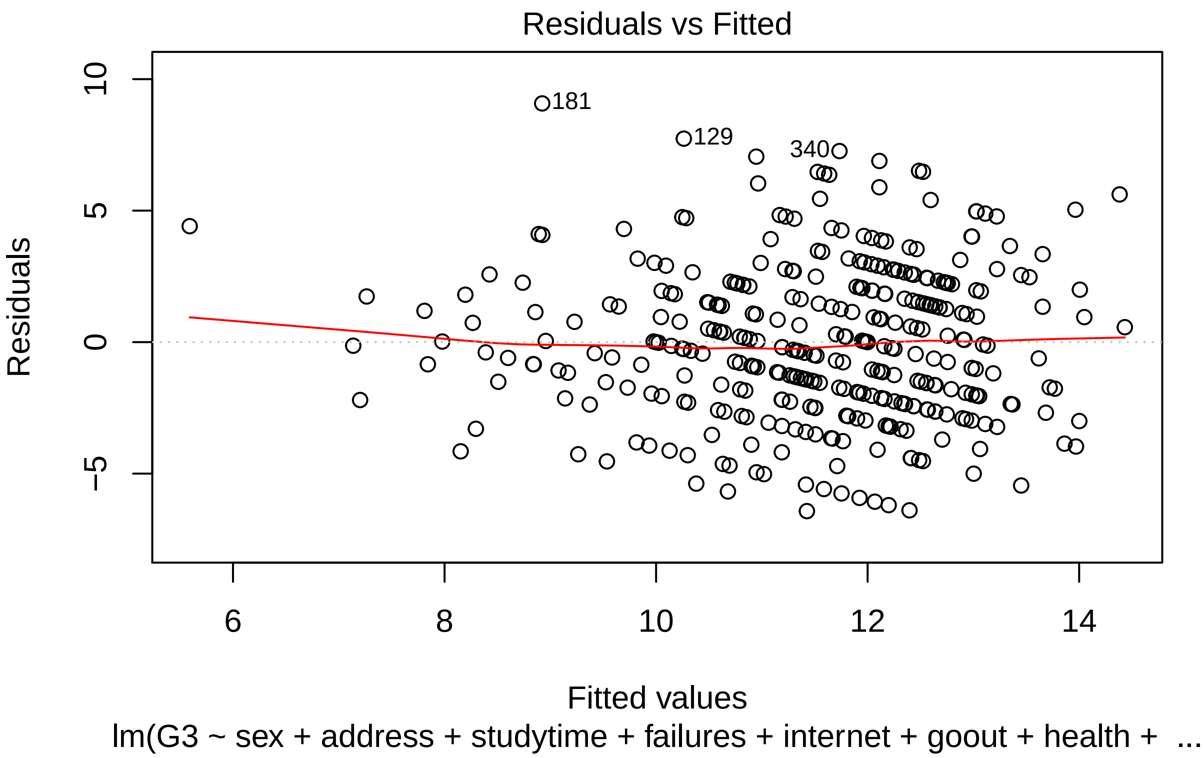



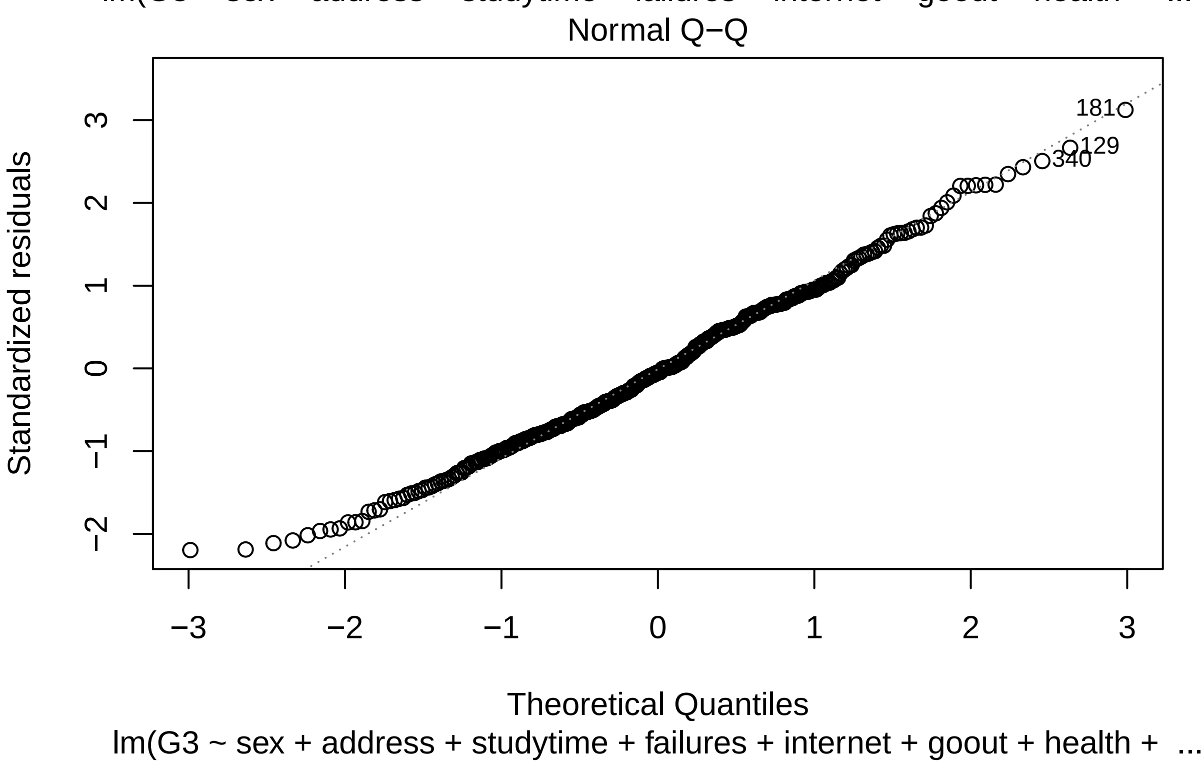

```
residuals6=model6$residuals
shapiro.test(residuals6)
```

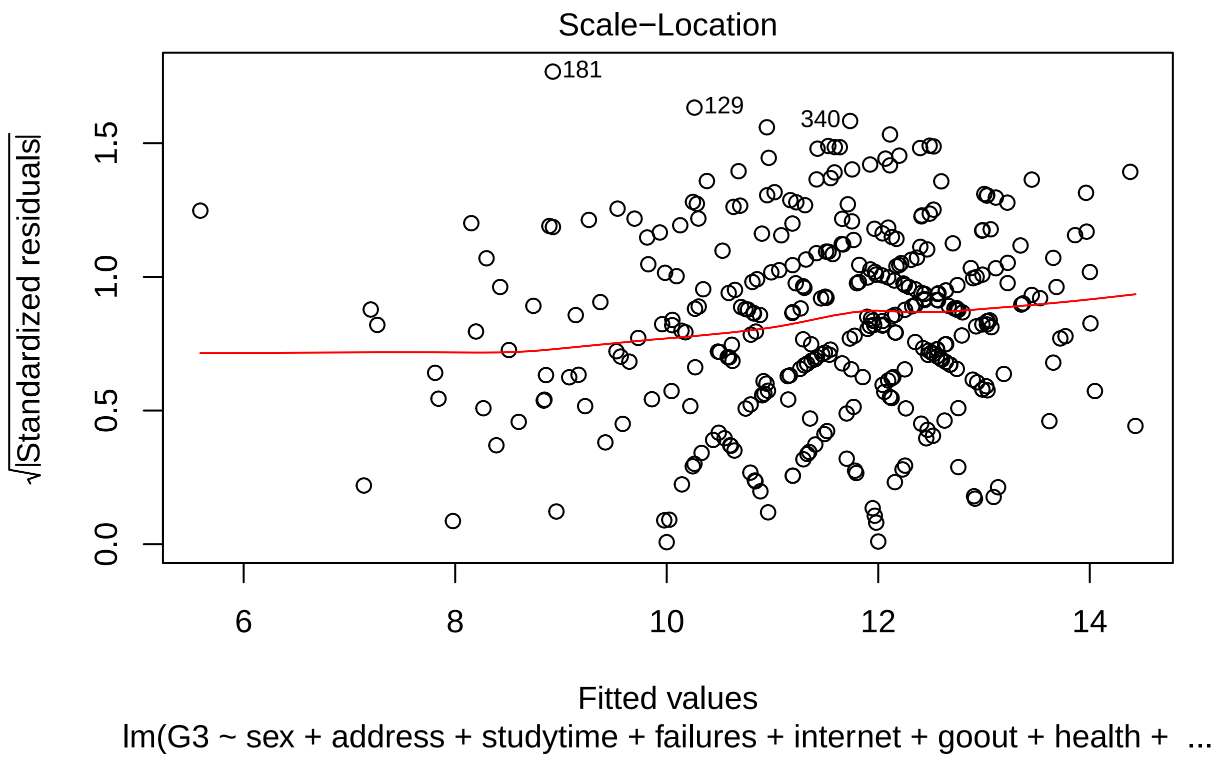
```
##
##  Shapiro-Wilk normality test
##
```

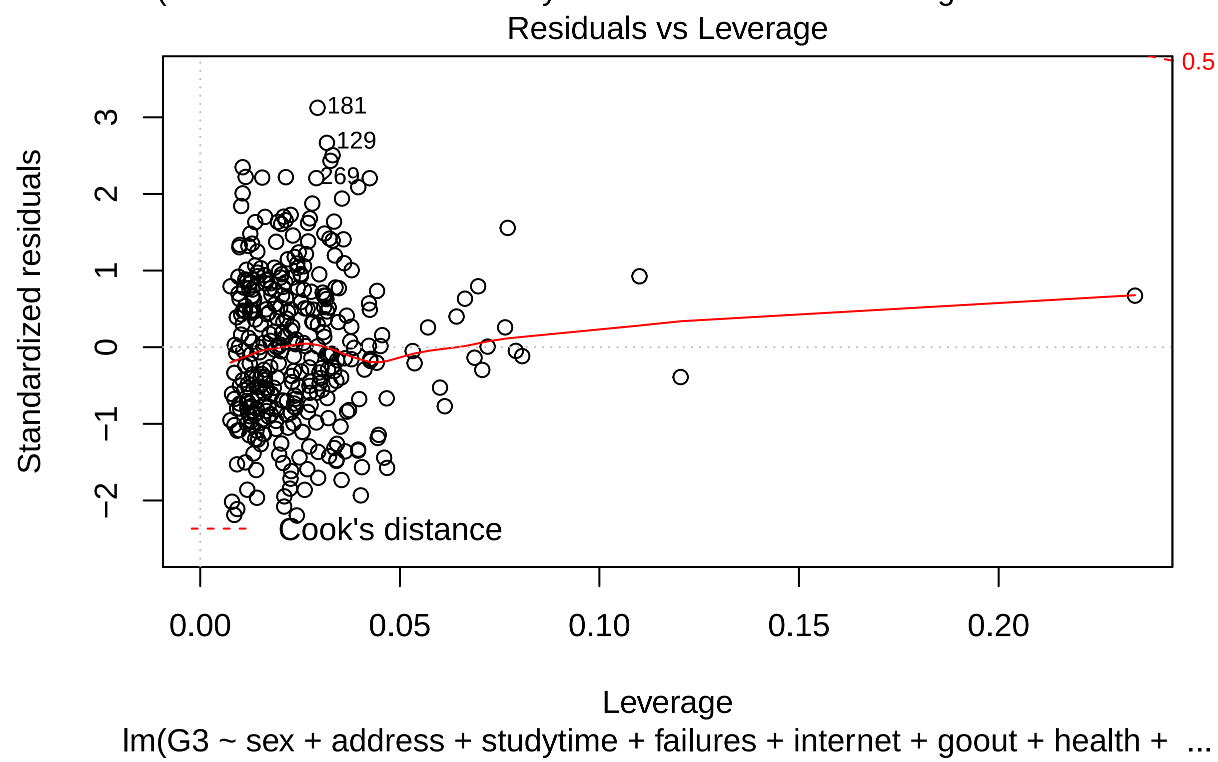
```
## data:  residuals6
## W = 0.99243, p-value = 0.06694
```

```r
plot(model8)
```





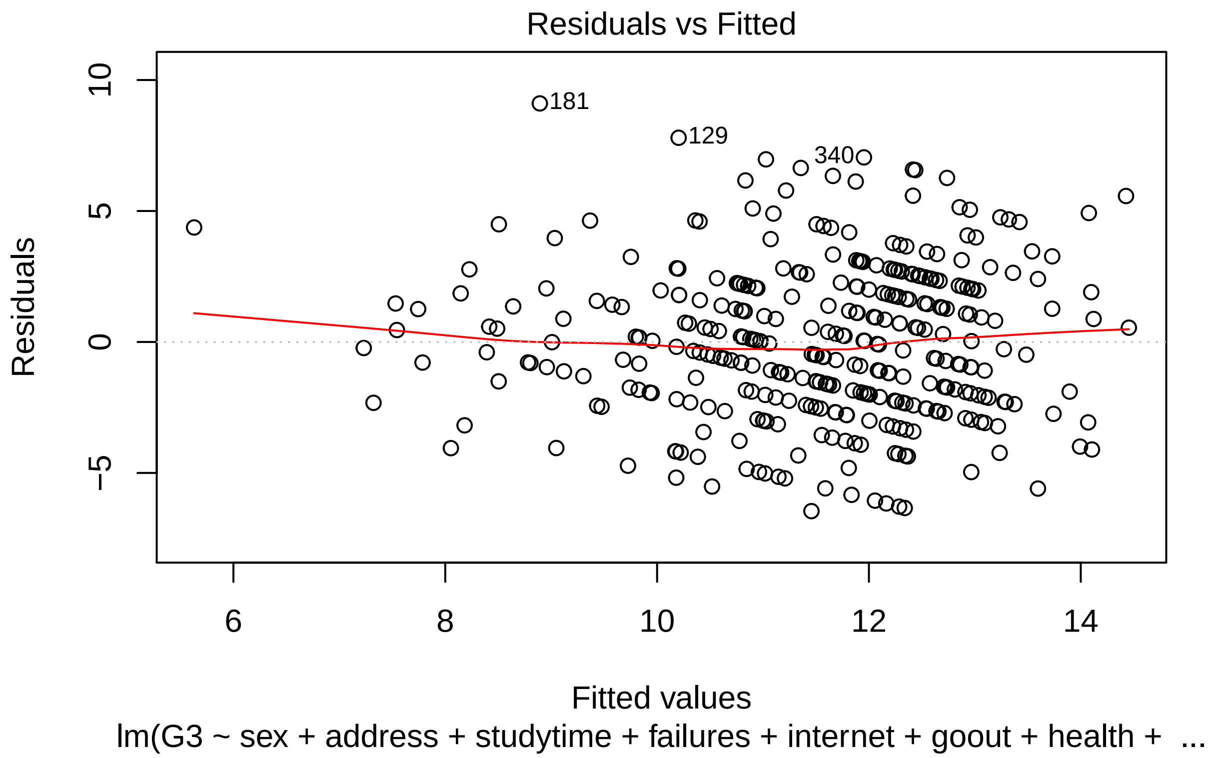



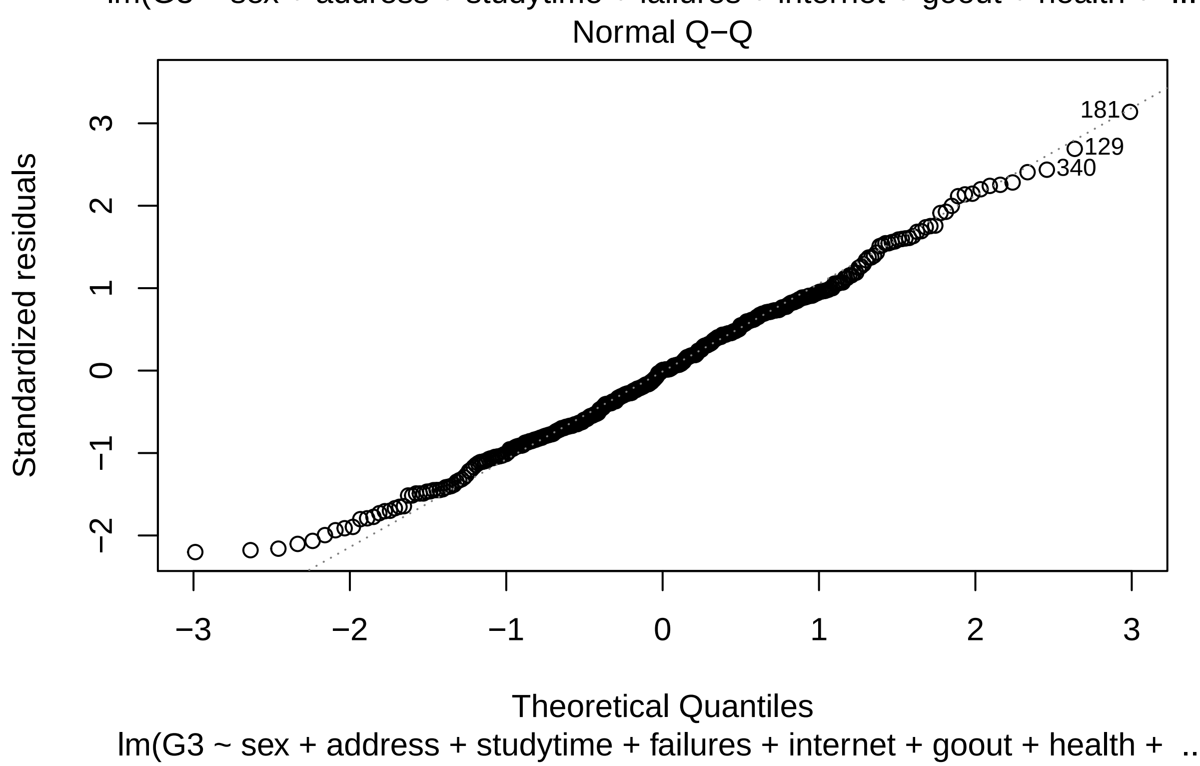

```
residuals8=model8$residuals
shapiro.test(residuals8)
```

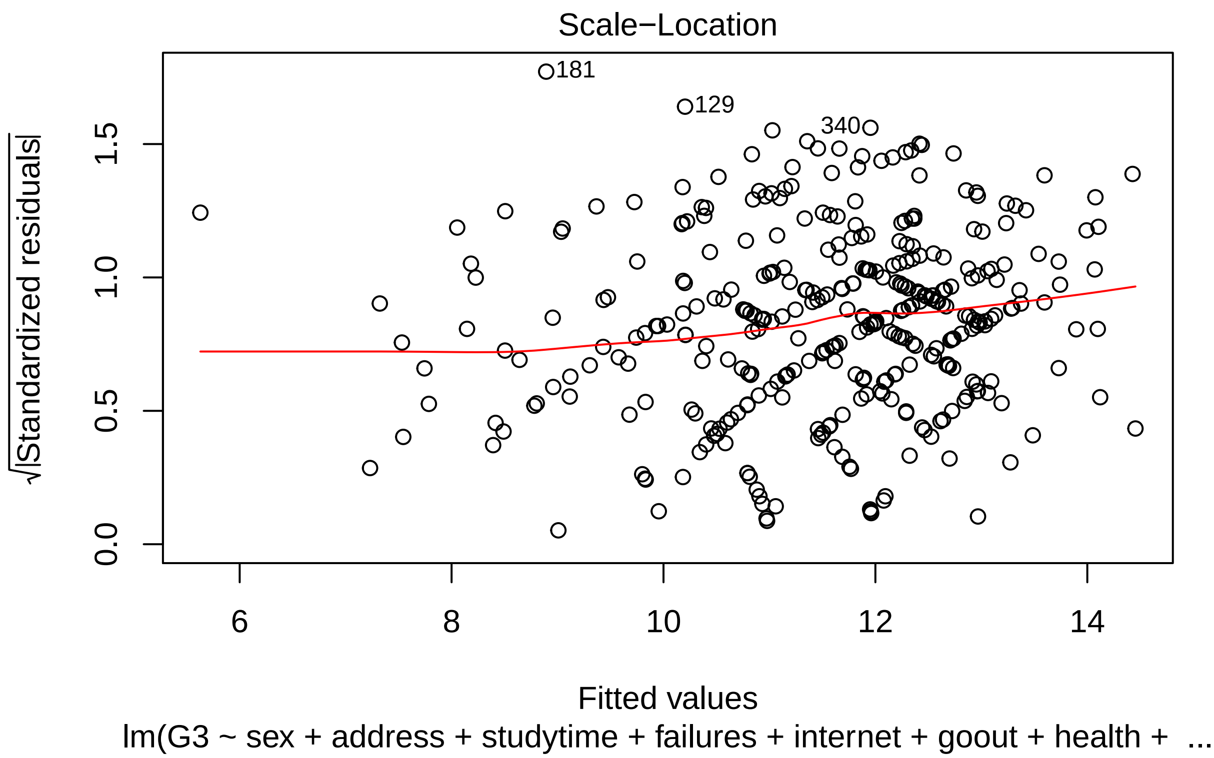
```
##
##  Shapiro-Wilk normality test
##
```

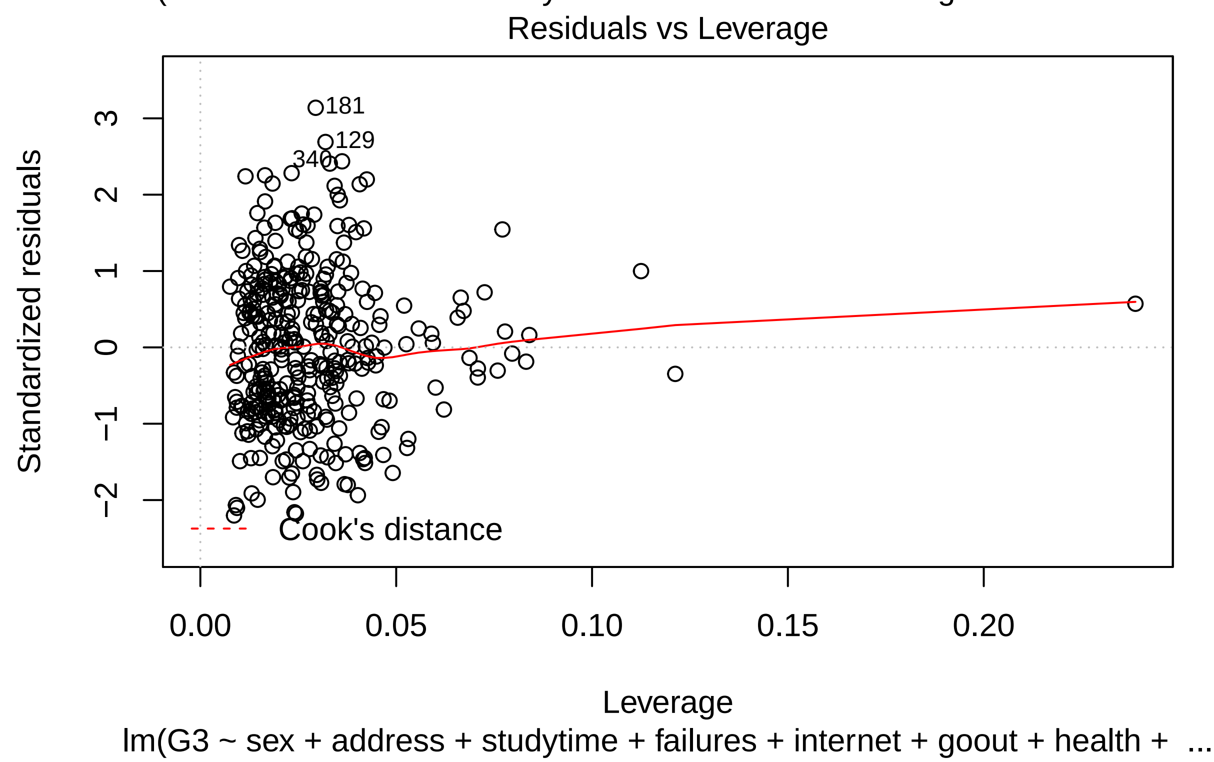
```
## data:  residuals8
## W = 0.99279, p-value = 0.08368
```

```r
plot(model9)
```






```
residuals9=model9$residuals
shapiro.test(residuals9)
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  residuals9
## W = 0.99332, p-value = 0.1149
```

```
require(leaps)
X1=model.matrix(G3~failures+internet+goout+absences)
R2=vector("numeric",4)
  for(j in 1:4){
    y_tmp=X1[,1+j]
    x_tmp=as.matrix(X1[,-c(1,1+j)])
    lm_fit=lm(y_tmp~x_tmp)
    R2[j]=summary(lm_fit)$r.squared
}
VIF=1/(1-R2)
names(VIF)=colnames(X1)[-1]
VIF
```

```
##    failures internetyes       goout    absences
##    1.053616    1.034247    1.029620    1.037410
```

```
X1=model.matrix(G3~sex+address+studytime+failures+goout+absences)
R2=vector("numeric",6)
  for(j in 1:6){
    y_tmp=X1[,1+j]
    x_tmp=as.matrix(X1[,-c(1,1+j)])
    lm_fit=lm(y_tmp~x_tmp)
    R2[j]=summary(lm_fit)$r.squared
}
VIF=1/(1-R2)
names(VIF)=colnames(X1)[-1]
VIF
```

```
##      sexM  addressU studytime  failures     goout  absences
##  1.111313  1.016530  1.122055  1.062127  1.027516  1.041748
```

```
X1=model.matrix(G3~sex+address+studytime+failures+internet+goout+health+absences)
R2=vector("numeric",8)
  for(j in 1:8){
    y_tmp=X1[,1+j]
    x_tmp=as.matrix(X1[,-c(1,1+j)])
    lm_fit=lm(y_tmp~x_tmp)
    R2[j]=summary(lm_fit)$r.squared
}
VIF=1/(1-R2)
names(VIF)=colnames(X1)[-1]
VIF
```

```
##        sexM    addressU   studytime    failures internetyes       goout
##    1.141282    1.062368    1.137321    1.073622    1.099133    1.036159
##      health    absences
##    1.025368    1.062528
```

```
X1=model.matrix(G3~sex+address+studytime+failures+internet+goout+health+absences+DWalc)
R2=vector("numeric",9)
  for(j in 1:9){
    y_tmp=X1[,1+j]
    x_tmp=as.matrix(X1[,-c(1,1+j)])
    lm_fit=lm(y_tmp~x_tmp)
```

```
    R2[j]=summary(lm_fit)$r.squared
}
VIF=1/(1-R2)
names(VIF)=colnames(X1)[-1]
VIF
```

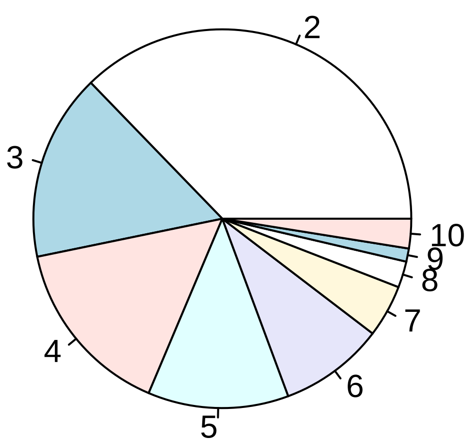
```
##        sexM    addressU   studytime    failures internetyes       goout
##    1.204600    1.087163    1.170673    1.084449    1.099795    1.250805
##      health    absences       DWalc
##    1.033224    1.075272    1.447646
```

### PART B: If Alcohol was the main predictor...

#### Alcohol vs G3

First, we will show the distribution of how much people drink per week

```
Alcohol4<- Alcohol3 %>%
  group_by(DWalc) %>%
  summarise(count = n())
pie(Alcohol4$count, labels= c('2','3','4','5','6','7','8','9','10'))
```




2 = very little to no drink 10 = drink a lot a lot

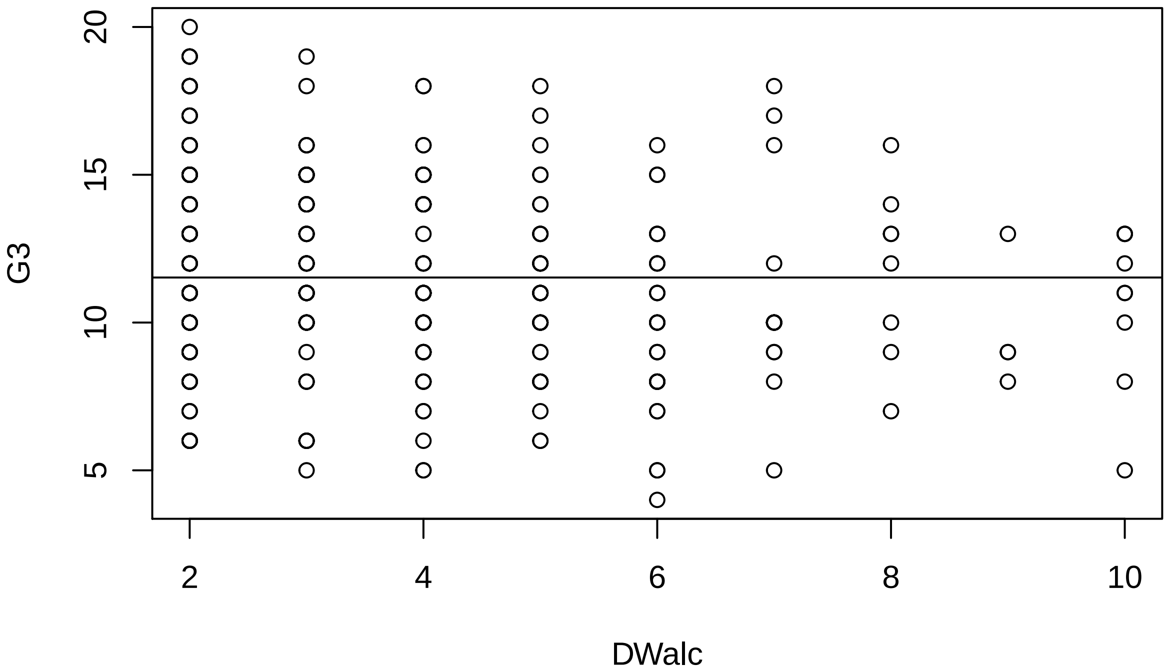
We see that most of people drink at least once a week.

Then we calculate the mean and compare it with the others

```
meanG3 <- mean(G3)
plot(DWalc, G3, data=Alcohol)
```

```
## Warning in plot.window(...): "data" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not
## a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not
## a graphical parameter

## Warning in box(...): "data" is not a graphical parameter

## Warning in title(...): "data" is not a graphical parameter
```

```r
abline(meanG3, 0)
```



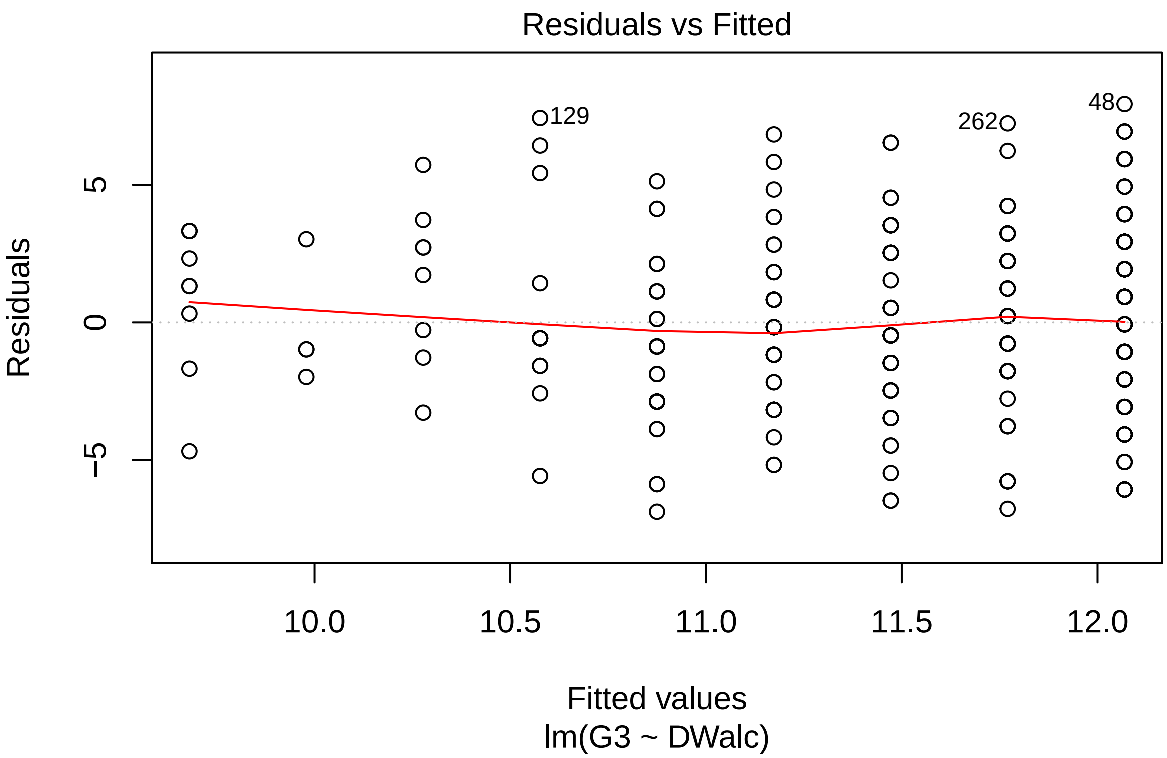

From the scatterplot, seems people who drink less alcohol (especially minimal 2) score above average and also obtain highest grades.

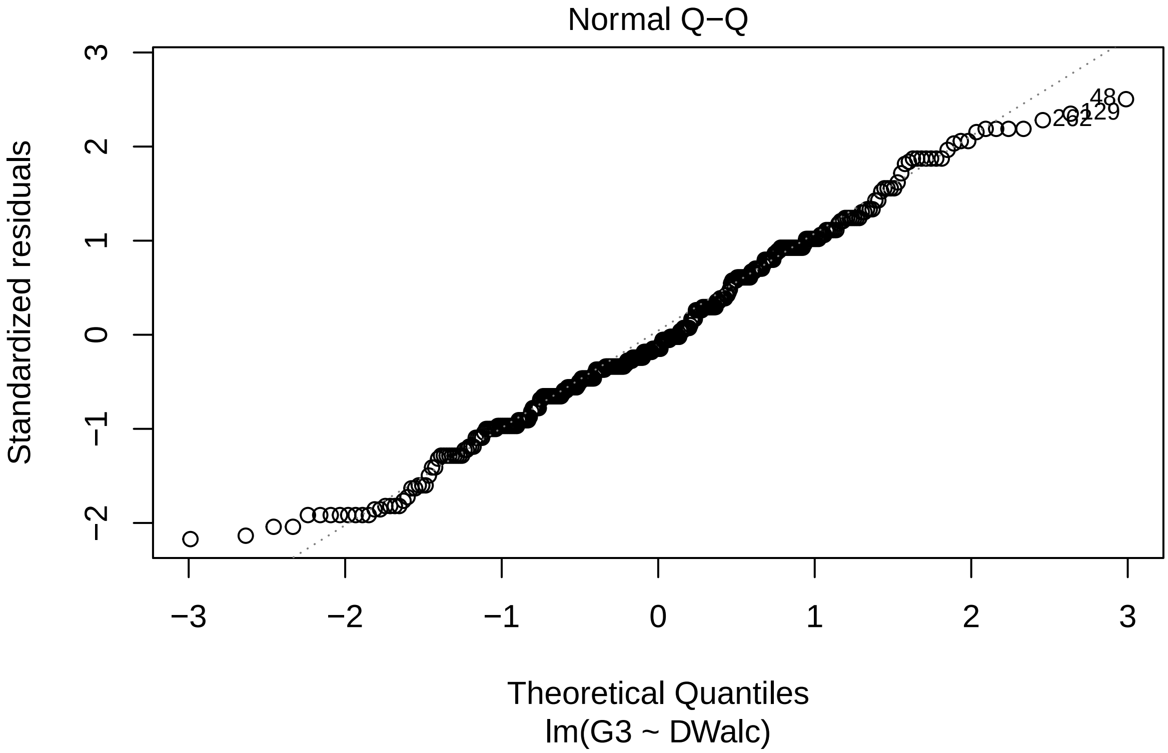
We still compute our model with simple linear regression and use T-test

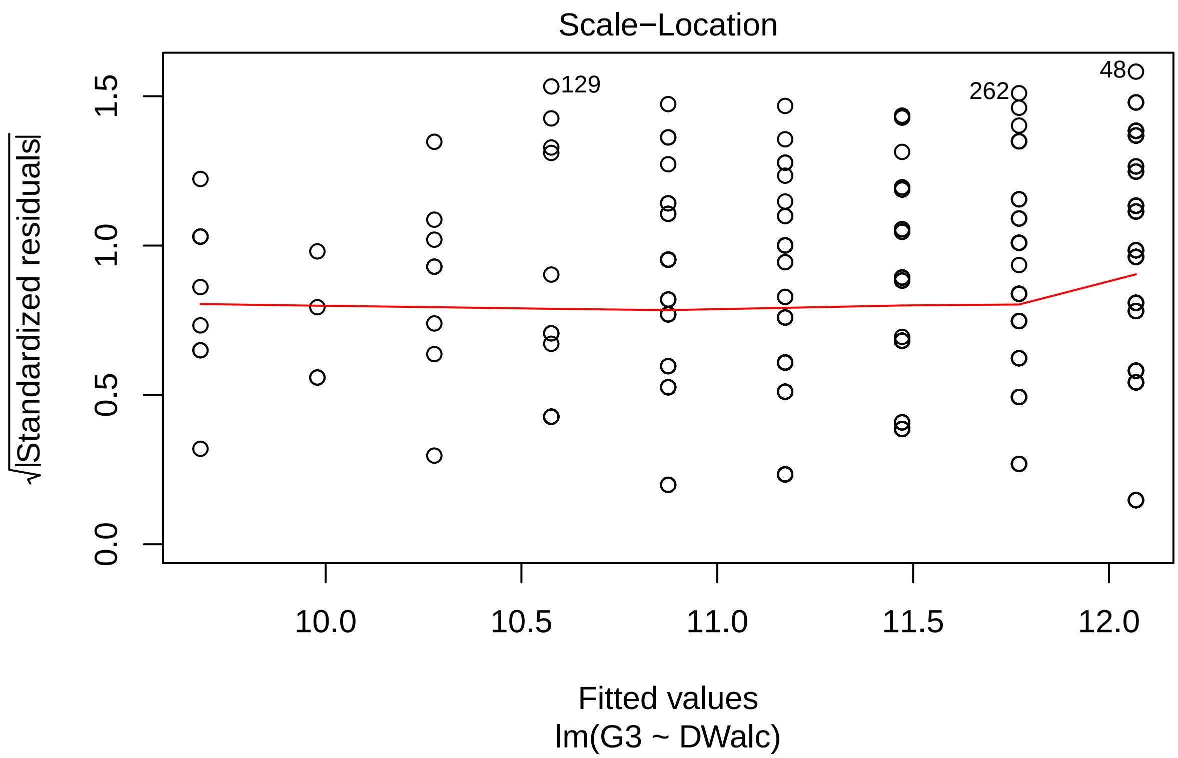
```r
alcoholmodel <- lm(G3~DWalc)
summary(alcoholmodel)
```

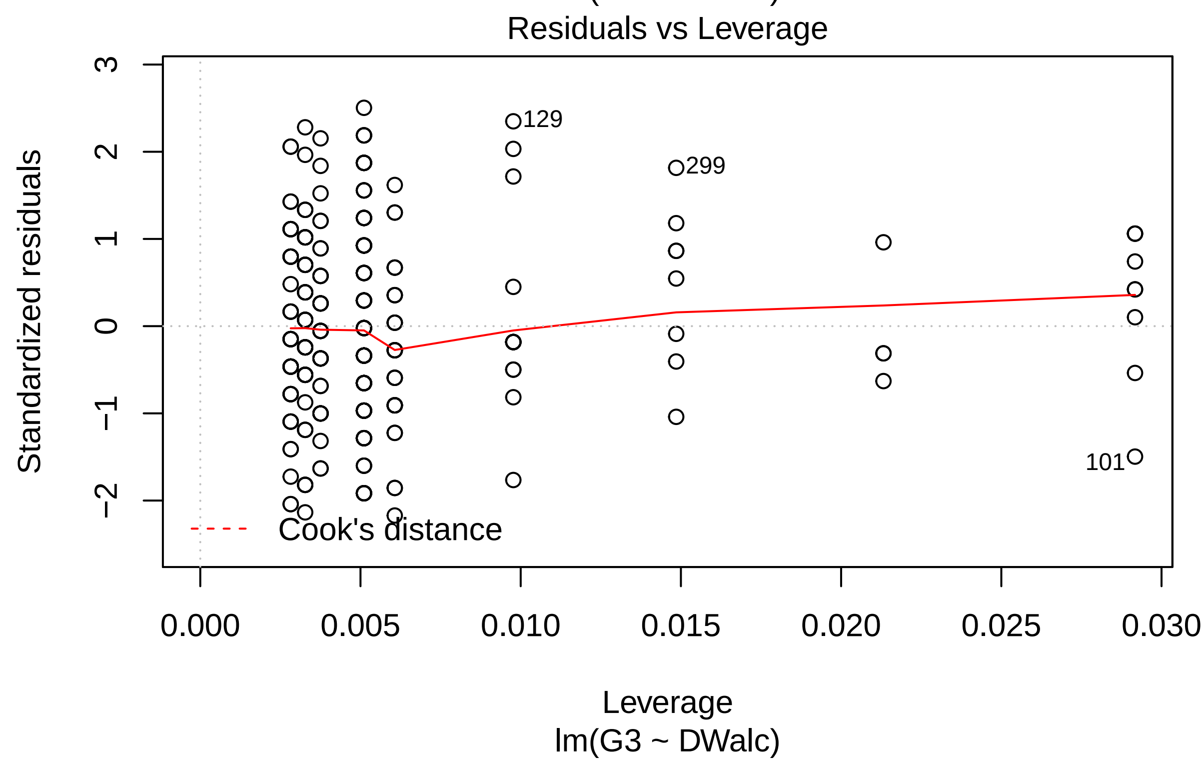
```
##
## Call:
## lm(formula = G3 ~ DWalc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.875 -2.069 -0.472  2.320  7.931
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.66629    0.36113  35.074   <2e-16 ***
## DWalc       -0.29858    0.08354  -3.574    4e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.176 on 355 degrees of freedom
## Multiple R-squared:  0.03474,    Adjusted R-squared:  0.03202
## F-statistic: 12.78 on 1 and 355 DF,  p-value: 0.0003997
```

```r
plot(alcoholmodel)
```


Residuals vs Fitted
Residuals
5
0
−5
129
262
48
10.0
10.5
11.0
11.5
12.0
Fitted values
lm(G3 ~ DWalc)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
48
262
129
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(G3 ~ DWalc)




We reject! Alcohol is significant

**Alcohol vs G3 considering the other variables**

Even though Alcohol vs G3 was bad, we still want to compare the model without alcohol with the full model

```r
modelnoalcohol <- lm(G3~sex+age+address+studytime+failures+activities+higher+internet+romantic+freetime+
anova(lmfit3, modelnoalcohol)
```

```
## Analysis of Variance Table
##
## Model 1: G3 ~ sex + age + address + studytime + failures + activities +
##     higher + internet + romantic + freetime + goout + health +
##     absences + DWalc
## Model 2: G3 ~ sex + age + address + studytime + failures + activities +
##     higher + internet + romantic + freetime + goout + health +
##     absences
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    342 3005.8
## 2    343 3020.9 -1   -15.023 1.7092  0.192
```

Alcohol is not significant considering all the others