Colesterol

SIMPLE LINEAR REGRESSION

```
library(car)
```

```
## Loading required package: carData
```

```
library(HH)
```

```
## Loading required package: lattice
## Loading required package: grid
## Loading required package: latticeExtra
## Loading required package: RColorBrewer
## Loading required package: multcomp
## Loading required package: mvtnorm
## Loading required package: survival
## Loading required package: TH.data
## Loading required package: MASS
##
## Attaching package: 'TH.data'
## The following object is masked from 'package:MASS':
##
##     geyser
## Loading required package: gridExtra
##
## Attaching package: 'HH'
## The following objects are masked from 'package:car':
##
##     logit, vif
```

Loading the data and printing the top part

```
setwd("G:/PiE2/2018")
colest<-read.csv2("./Dades/COL.csv")
head(colest)
```

```
##    A   H    W     C
## 1 19 174 79.9 189.5
## 2 15 151 64.5 197.5
## 3 13 133 52.0 170.5
## 4 19 173 75.5 180.5
## 5 17 163 74.0 216.5
## 6 13 135 54.9 173.5
```

We fix the number of parameters to be equal to two.

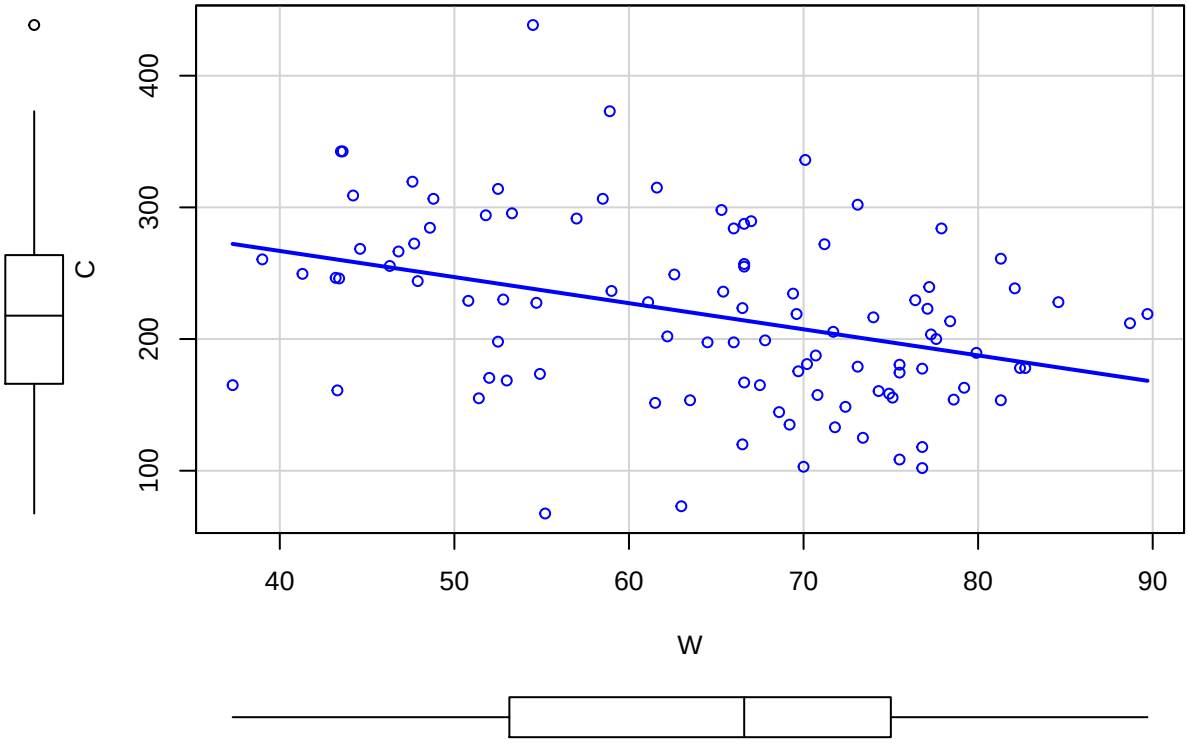
We compute the number of experimental units, in this case persons.

```
p<-2
n<-dim(colest)[1]
n
```

```
## [1] 100
```

Scatterplot of cholesterol as a function of weight

```
scatterplot(C~W, colest, smooth=F)
```



We compute the regression line for modelling the colesterol as a function of weight.

The Model

```
mod<-lm(C~W, data=colest)
```

Summary of the model

```
summary(mod)
```

```
##
## Call:
## lm(formula = C ~ W, data = colest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -169.24  -39.81   -4.49   47.19  200.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 346.2251    33.1983   10.43  < 2e-16 ***
## W            -1.9835     0.5046   -3.93 0.000158 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.55 on 98 degrees of freedom
## Multiple R-squared:  0.1362, Adjusted R-squared:  0.1274
## F-statistic: 15.45 on 1 and 98 DF,  p-value: 0.0001581
```

From the summary and the scatterplot we see that:

a) the residuals seem to be very large,

b) the standard error is also quite large, meaning that the residuals vary a lot,

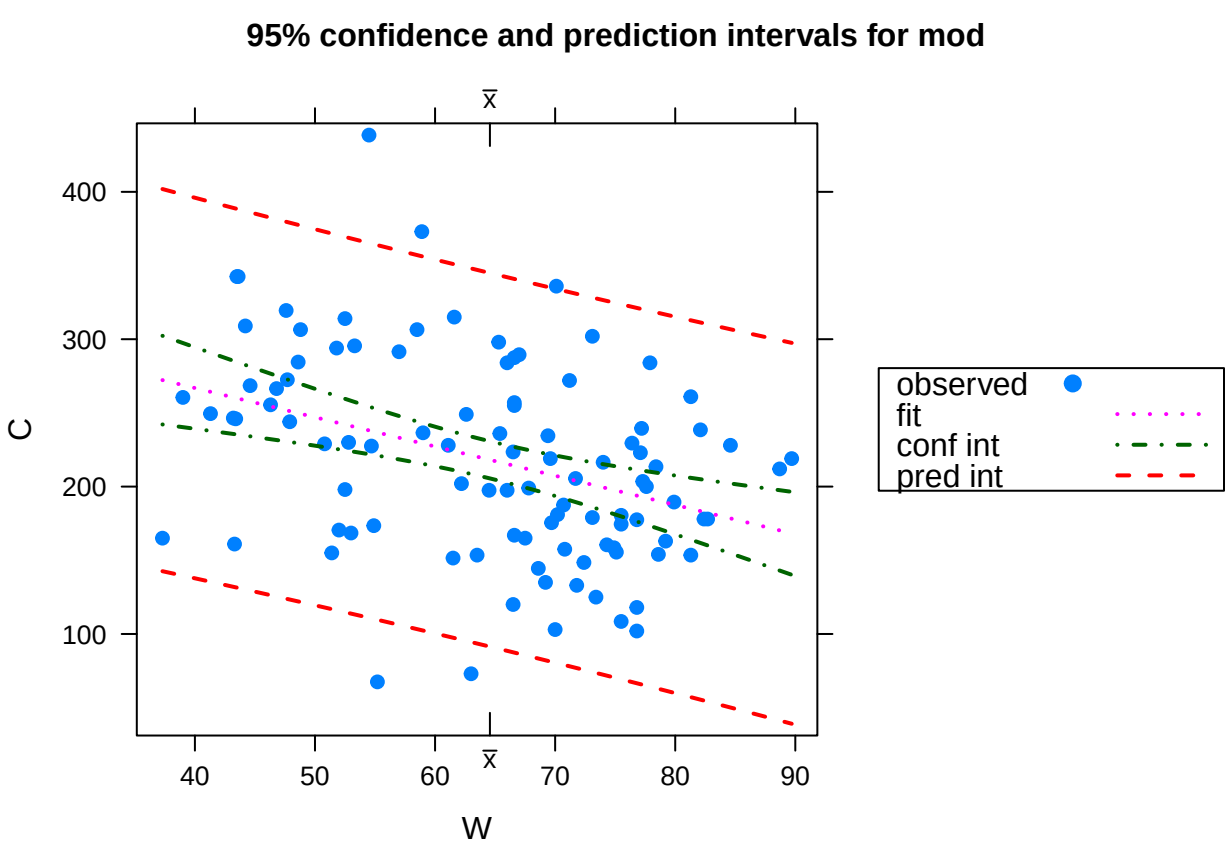
c) a very small R^2 value, the weight is just explaining a 13% of the variability in the cholesterol level,

d) the two parameters are significatively different from zero, meaning that the weight has an influence on the response variable,

e) the fact that the weight coefficient is negative implies that as the weight increases the cholesterol leves decreases, which is the contrary of what we should expect.

We plot the regression line with PI and CI

```
ci.plot(mod)
```



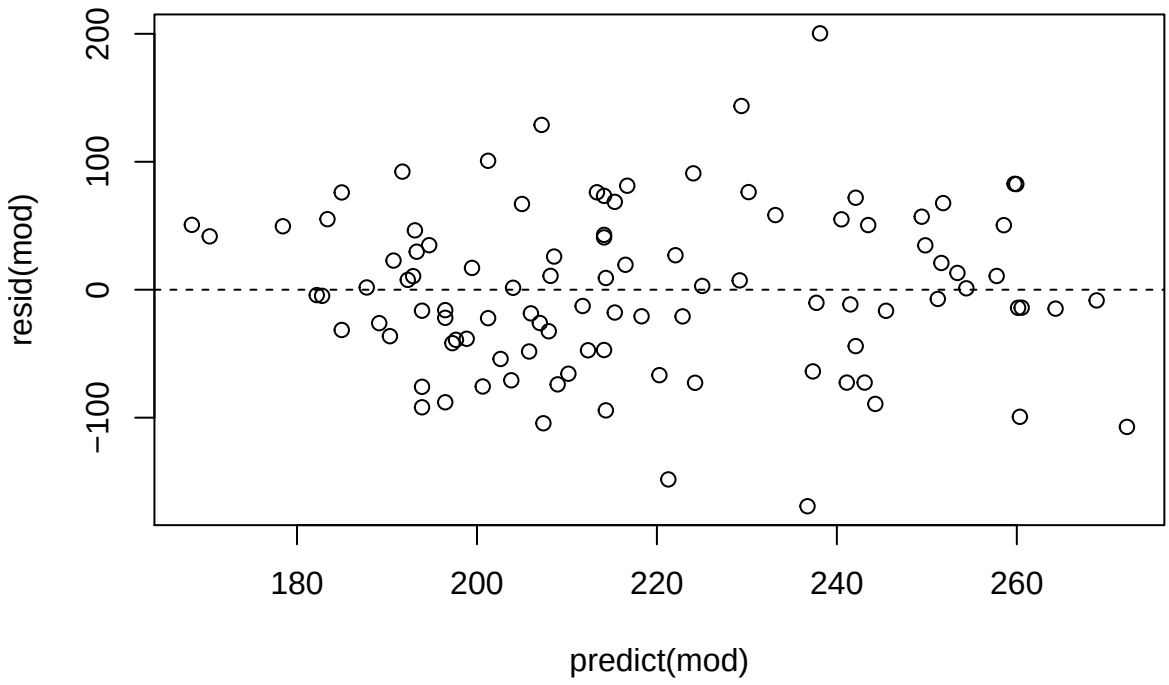
From this plot we see that:

1) As the weight increases the cholesterol level decreases, which is quite contradictory.
2) The PI (intervals for the predicted values, in red) are wider than the CI (intervals for the mean values, in green), as it has to be.
3) Both intervals get wider as the distance from the gravity center of the dots cloud increases.

To check if the hypothesis of the Linear Model are verified, we perform several residual graphics.

We first plot the predicted values vs residuals

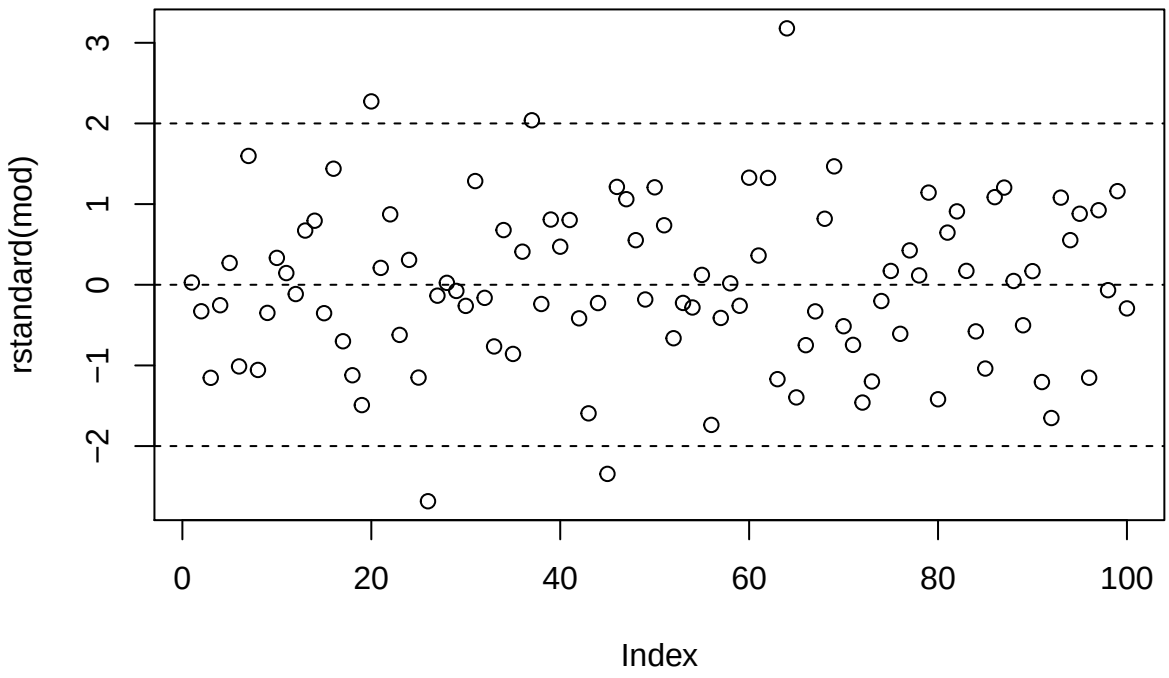
```
plot(predict(mod),resid(mod))
abline(h=0,lty=2)
```



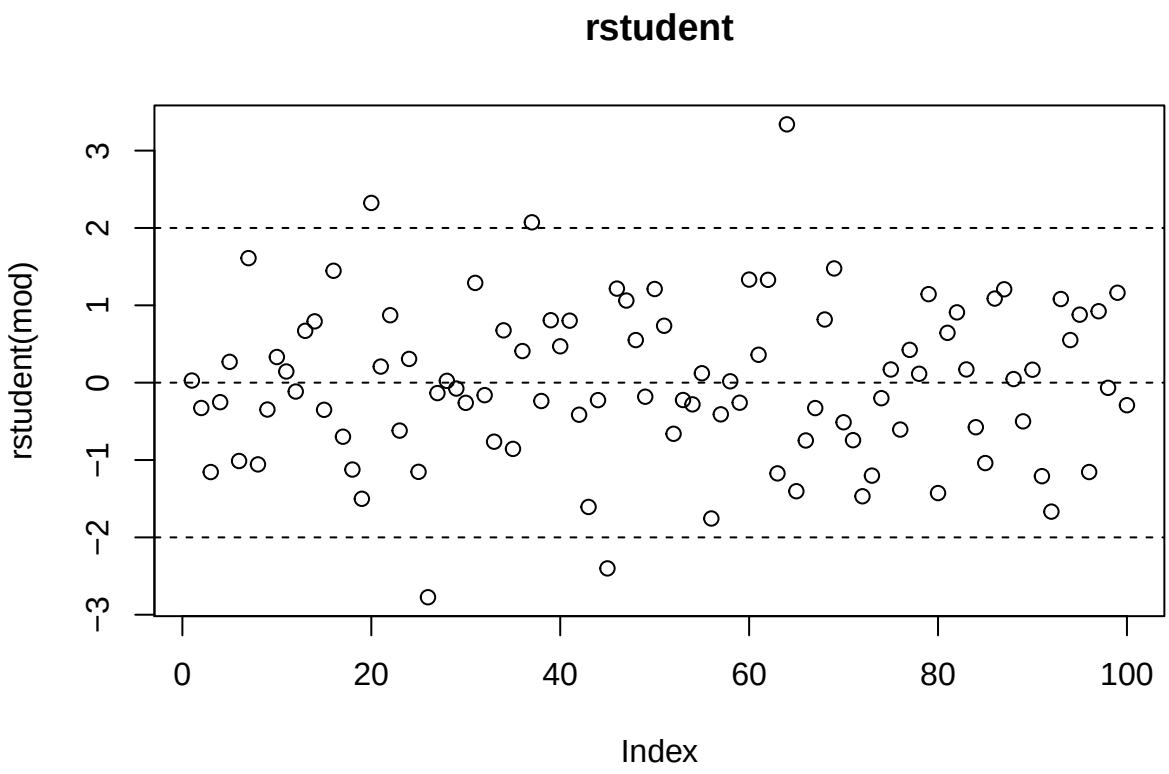
We can see that the homogenity of variances is reasonable to be assumed, there are not patters in the plot which is good. Nevertheless, the residuals are very large.

Let us plot the standarized/studentized residuals

```
plot(rstandard(mod))
abline(h=c(-2,0,2),lty=2)
```




```
plot(rstudent(mod),main="rstudent")
abline(h=c(-2,0,2),lty=2)
```



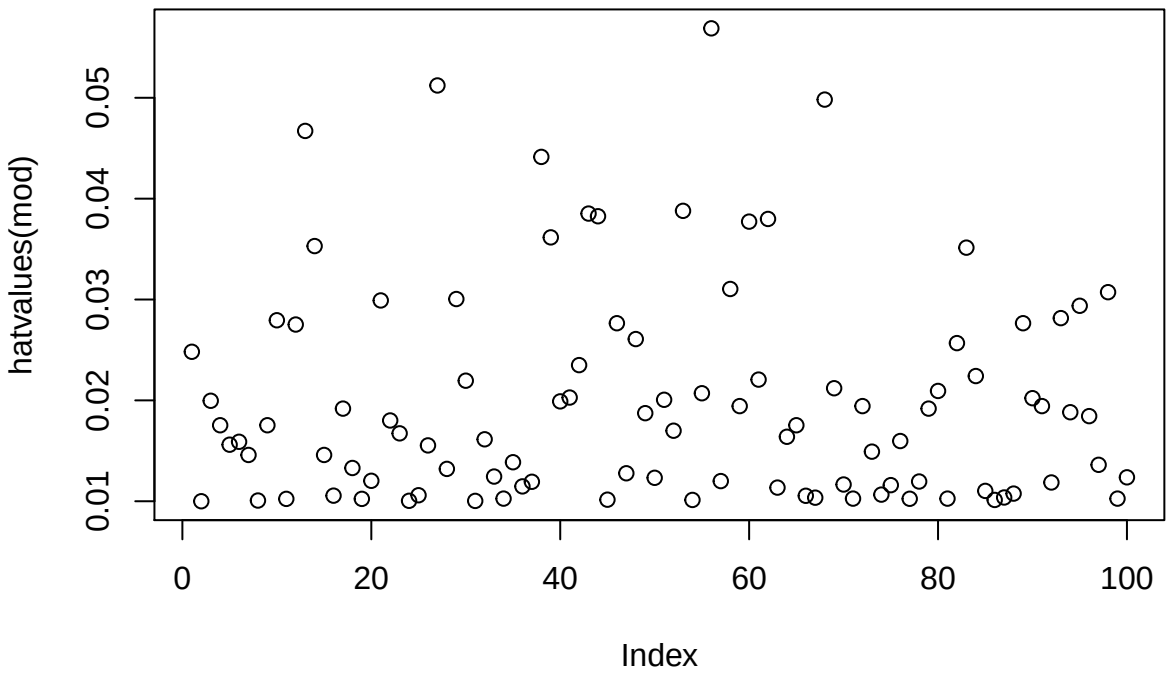
We observe that:

1) Both residuals plots are quite reasonable. As a consequence that the standar error estimation is large, the standarized and studentized residuals become relatively small.
2) Four residuals (les than a 5%) lie out of the interval (-2,2) which is not a problem.
3) There are nou patterns.
4) From the studentized residual plot one can identify the observations that are possible outliers. The studentized residuals just show a possible observation that could be an outlier, which is the one that has a residual larger than 3.

Diagnostic: LEVERAGE

In what follows we compute the leverage of the observations Remind that the leverage just depend on the values of the X matrix.

```
plot(hatvalues(mod))
abline(h=c(0,3*(p/n)),lty=2)
```

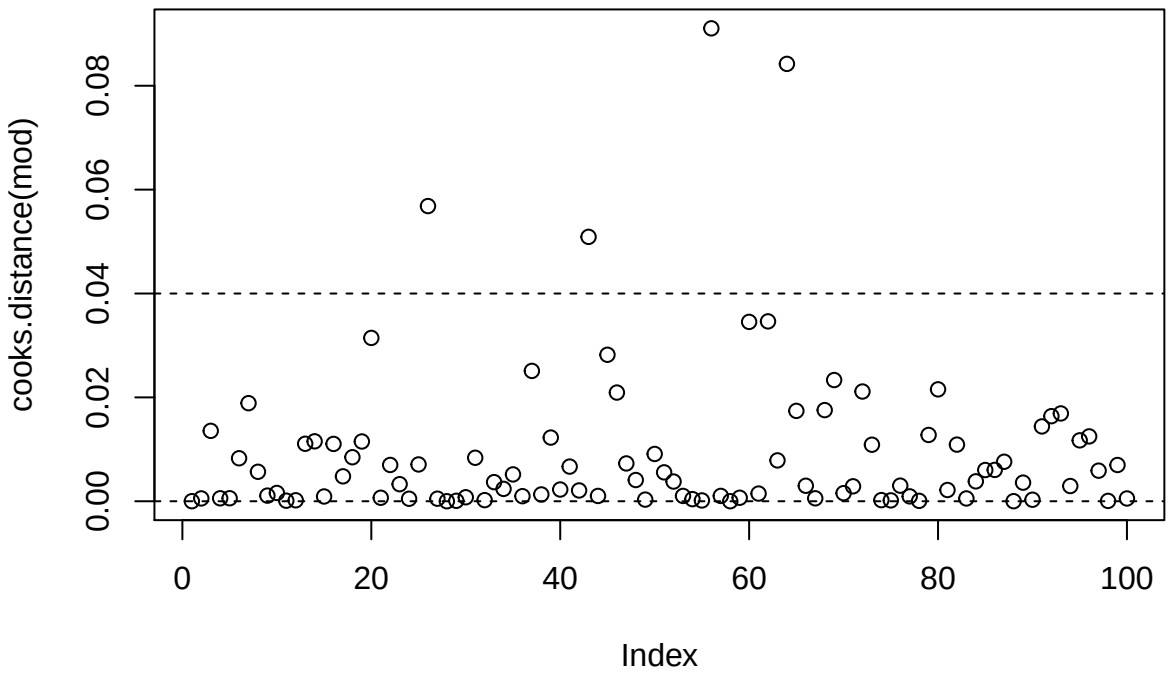


We observe that there are no values, with a leverage larger than $3p/n$.

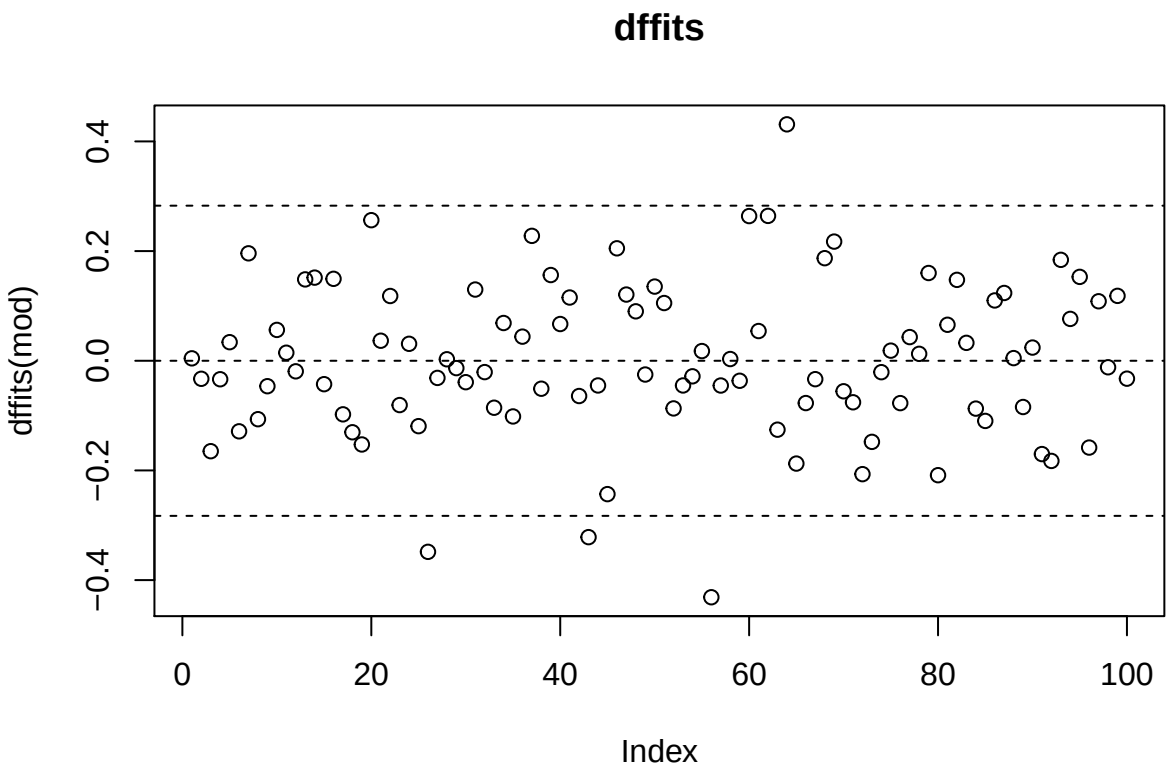
Diagnostic: Influential observations (dffits, cooks.distance)

We first compute the Cook's distance and later the dffits values

```
plot(cooks.distance(mod))
abline(h=c(0,4/n),lty=2)
```




```
plot(dffits(mod),main="dffits")
abline(h=c(-2*sqrt(p/n),0,2*sqrt(p/n)),lty=2)
```

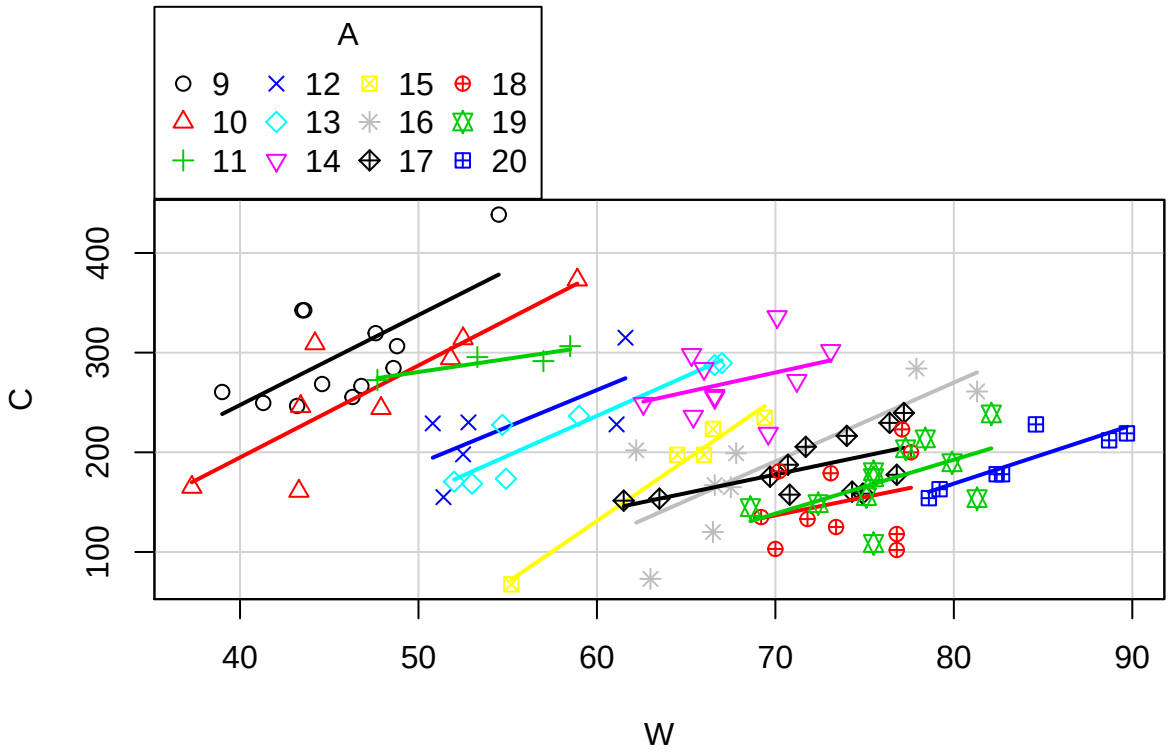


In what follows and in order to see if the age has any influence in the cholesterol level, split the experimental units in groups, grouping those that have the same age, and we compute a regression line for each group.

We perform a simple regression for each group of age

In the following sentence we use sp which is an abreviation of scatterplot

```
sp(C~W|A,smooth=F,col=1:20, data=colest)
```



In this plot of the regression lines for ages we deduce that the way in which the weight affects the cholesterol level depends on the age of the person. Now we see that the cholesterol level increases with the weight, since the regression line have a positive slope. We also see that for some ages the regression line predicts quite accurately and for some others it is not very good.

In conclusion, the simple regresion model that only contains the weight is too simple, and we need to consider a multiple regression model. At least the age should be included.