colesterol-regmult

MULTIPLE 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

```
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
```

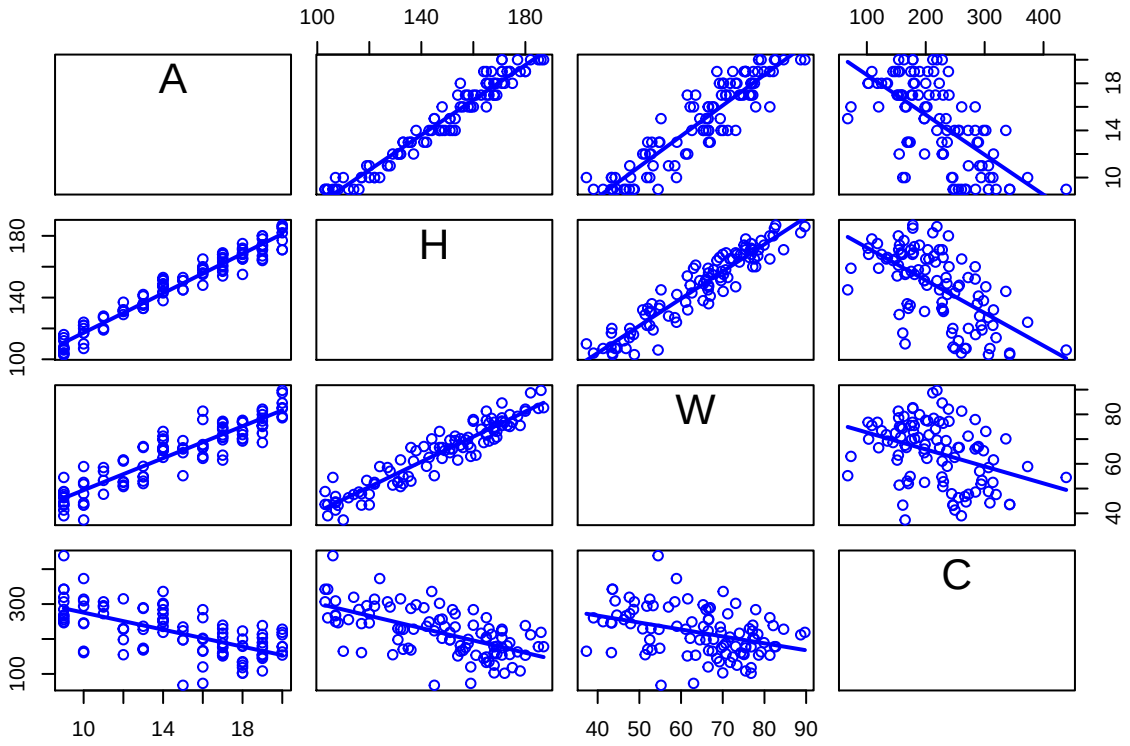
We compute the number of observations

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

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

We first perform the scatterplot of all the variables

```
scatterplotMatrix(colest,smooth=F,diagonal=F)
```



From this scatterplot we deduce that

a) there exists a strong linear relationship between height and age,

b) there exist a strong linear relationship between height and weight,

c) less clear but also important linear relationship betwee age and weight,

d) cholesterol is related with each one of the explanatory variables with a straignt line with negative slope.

We compute the multiple regression that contains the weight, the height and the age as explanatory variables.

we set the number of parameters to be equal to four.

The Model

```
mod<-lm(C~W+A+H, colest)
summary(mod)
```

```
##
## Call:
## lm(formula = C ~ W + A + H, data = colest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.608 -22.137   1.888  21.156  65.410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 490.9978    35.0517  14.008  < 2e-16 ***
## W            10.3773     0.7365  14.090  < 2e-16 ***
## A           -13.0195     3.8530  -3.379  0.00105 **
## H            -5.0989     0.7227  -7.055 2.68e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.11 on 96 degrees of freedom
## Multiple R-squared:  0.8101, Adjusted R-squared:  0.8041
## F-statistic: 136.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
p<-4
```

From the summary of the model we deduce that:

a) there exists residuals with a quite large absolute value.

b) all the parameters are significatively different from zero, meaning that the explanatory variables have a significant influence on the cholesterol level. we can seee that the weight coefficient is positive meaning that if the weight increases also do the cholesterol level. The age and the height have a negative coefficient meaning that when they increase the cholesterol level also increases.

c) The residual standard error seems to be quite big.

d) The model explains 81% of the variability in the cholesterol level, which is a good proportion.

e) The null hypothesis associated to the omnibus test is rejected meaning that the explanatory variables really capture an important part of the variability in the cholesterol level.

All the sentences just mentioned will be true if the hypothesis of normality, homocedasticity and indepdendence of the residuals are verified. These hypothessis need to be checked by means of the residual analysis.

The puntual estimation of σ^2

The puntual estimation of $\sigma^2 = (30.11)^2 = 906,61$ which is equal to the the mean residual sum of squares.

Omnibus test

This test corresponds to $H_0 = \beta_1 = \beta_2 = \beta_3 = 0$ vs $H_1 : notH_0$. Thus we are testing if globaly our model is explaining a significant part of the variability in the cholesterol.

Based on the '-value associated to the F test, we reject the null hypothesis and conclude that our model is useful to explain an important part of the variability in the response variable.

Let us follow the steps required in the exercice, and previous to that

let us compute the sums of squares associated to the regression.

We do the anova/ANOVA analysis

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: C
##           Df Sum Sq Mean Sq F value    Pr(>F)
## W          1  62396   62396  68.826 6.686e-13 ***
## A          1 263670  263670 290.841 < 2.2e-16 ***
## H          1  45123   45123  49.773 2.676e-10 ***
## Residuals 96  87031     907
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting sums of squares of the anova are the ones called of type I. These sums depend on the order in which the factors have been introduced in the model. The sum of the type I sums of squares gives place to the total variability in the cholesterol level variable.

```
Anova(mod)
```

```
## Anova Table (Type II tests)
##
## Response: C
##           Sum Sq Df F value    Pr(>F)
## W         179985  1 198.533 < 2.2e-16 ***
## A          10351  1  11.418  0.001052 **
## H          45123  1  49.773 2.676e-10 ***
## Residuals  87031 96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting sums of squares of the Anova are the ones that are called type III. These sums do not depend on the order in which te factors are introduced in the model. If they are significant it means that the variable has a real influence on the response variable.

The type III sums of squares are in concordance with the t-values obtained in the summary of the model.

We can observe that for any explanatory variable, the t-value squared is equal tot he F-valued of the Anova table. For example, with respect tot he variable weight we have that $(14.09)^2 = 906.61$. The same is true for the rest of variables in the mdoel.

We can also check that $\sqrt{(87031/96)} = 30.11$ which is the the standard error estimation.

Important: Given that all the variables are significative, we do not suppress any of them inspite of the fact that two of them are highly correlated.

We compute the confidence intervals for the parameters

```
confint(mod,level=0.99)
```

```
##                   0.5 %     99.5 %
## (Intercept) 398.881272 583.114304
## W             8.441792  12.312821
## A           -23.145228  -2.893732
## H            -6.998311  -3.199551
```

The CI do not contain the zero value, thus the parameters are statically different form zero.

The same is deduced by looking at the p-values that appear in the summary

For $W = 65$, $A = 15$ and $H = 150$

We compute confidence intervals and predicted intervals in new experimental conditions. In particular in $W = 65$, $A = 15$ and $H = 150$

```
C0<-data.frame(cbind(W=c(65,75,65),A=c(15,15,12),H=c(150,150,150)), row.names=1:3)
```

```
predict(mod, C0, interval="confidence", level=.95, se.fit=T)
```

```
## $fit
##        fit      lwr      upr
## 1 205.3908 199.1668 211.6148
## 2 309.1639 294.6188 323.7089
## 3 244.4492 219.8210 269.0774
##
## $se.fit
##         1         2         3
##  3.135539  7.327533 12.407261
##
## $df
## [1] 96
##
## $residual.scale
## [1] 30.1094
```

```
C0<-data.frame(cbind(W=c(65,75,65),A=c(15,15,12),H=c(150,150,150)), row.names=1:3)
predict(mod, C0, interval="prediction", level=.95, se.fit=T)
```

```
## $fit
##        fit      lwr      upr
## 1 205.3908 145.3009 265.4807
## 2 309.1639 247.6528 370.6749
## 3 244.4492 179.8071 309.0914
##
## $se.fit
##         1         2         3
##  3.135539  7.327533 12.407261
##
## $df
## [1] 96
##
## $residual.scale
```

```
## [1] 30.1094
```

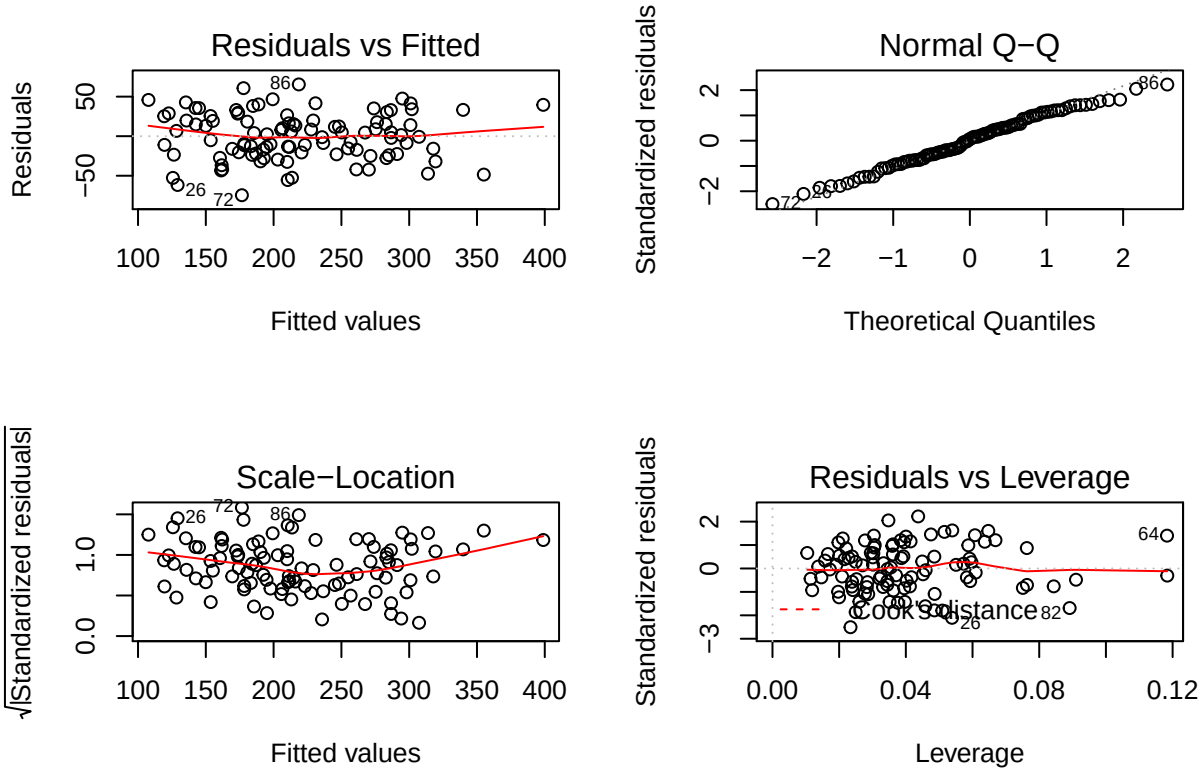
Givent hat the CI for the prediction of the mean value associated to the conditions $W = 65$, $A = 15$ and $H = 150$, does not contain the value 190, we reject the null hypothesis and conclude that the cholesterol level under these conditions is statistically different from 190.

In what follows we perform the model diagnostic.

R diagnostic

In what follows it appear the plots that R perform by defect, in order to check the hypothesis (assumptions) associated to the linear model. We want them in a two by two matrix.

```
oldpar <- par( mfrow=c(2,2))
plot(mod,ask=F)
```

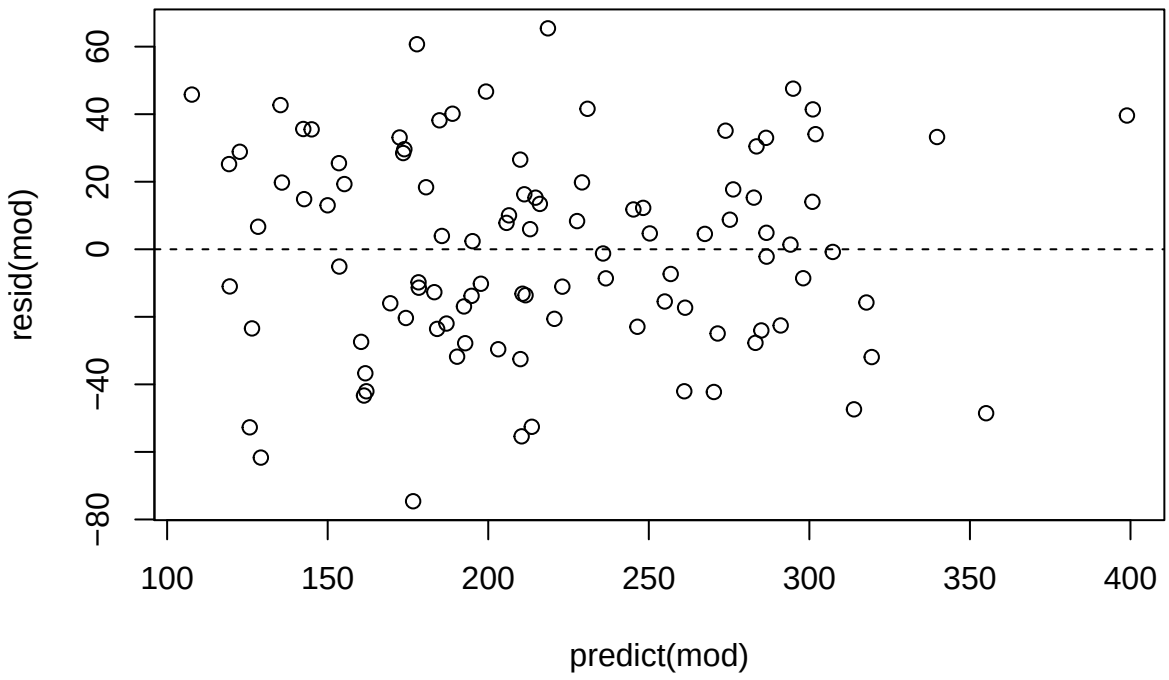



```
par(oldpar)
```

We do not observe patthers in the plot of the residuals versus fitted neither in the square root of the standarized residuals versus fitted. Moreover the residuals may be assumed to be normal distributed, since they fit with the straignht line in the qqplot.

Diagnostic: tendences

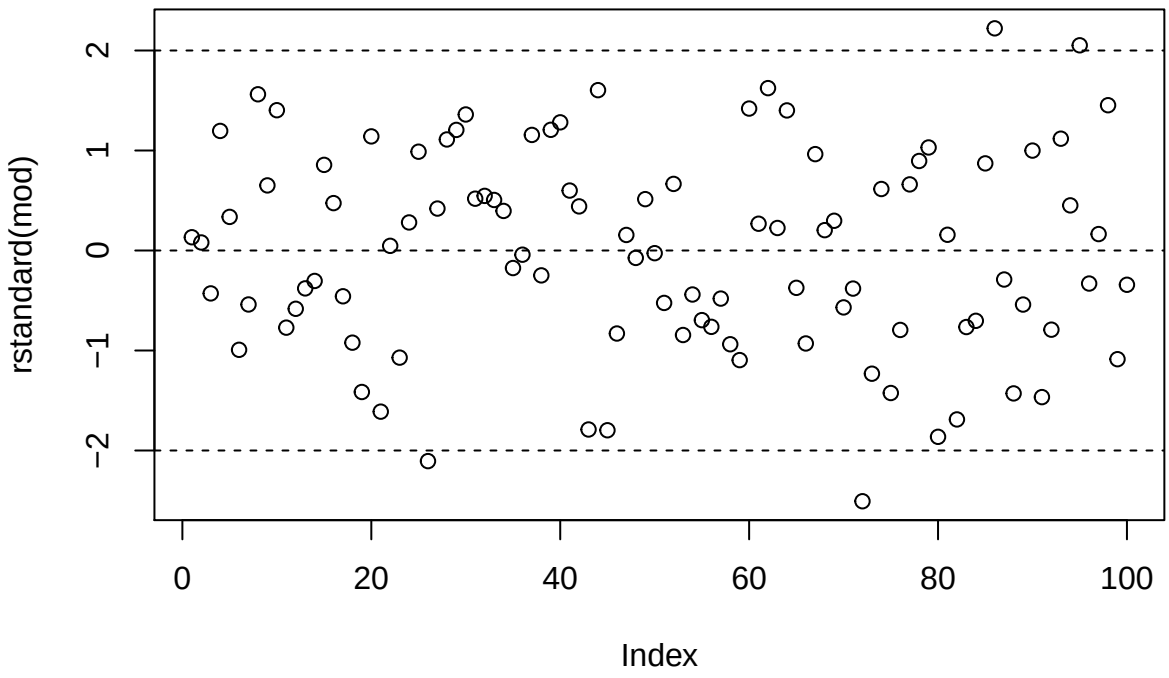
```
plot(predict(mod),resid(mod))
abline(h=0,lty=2)
```



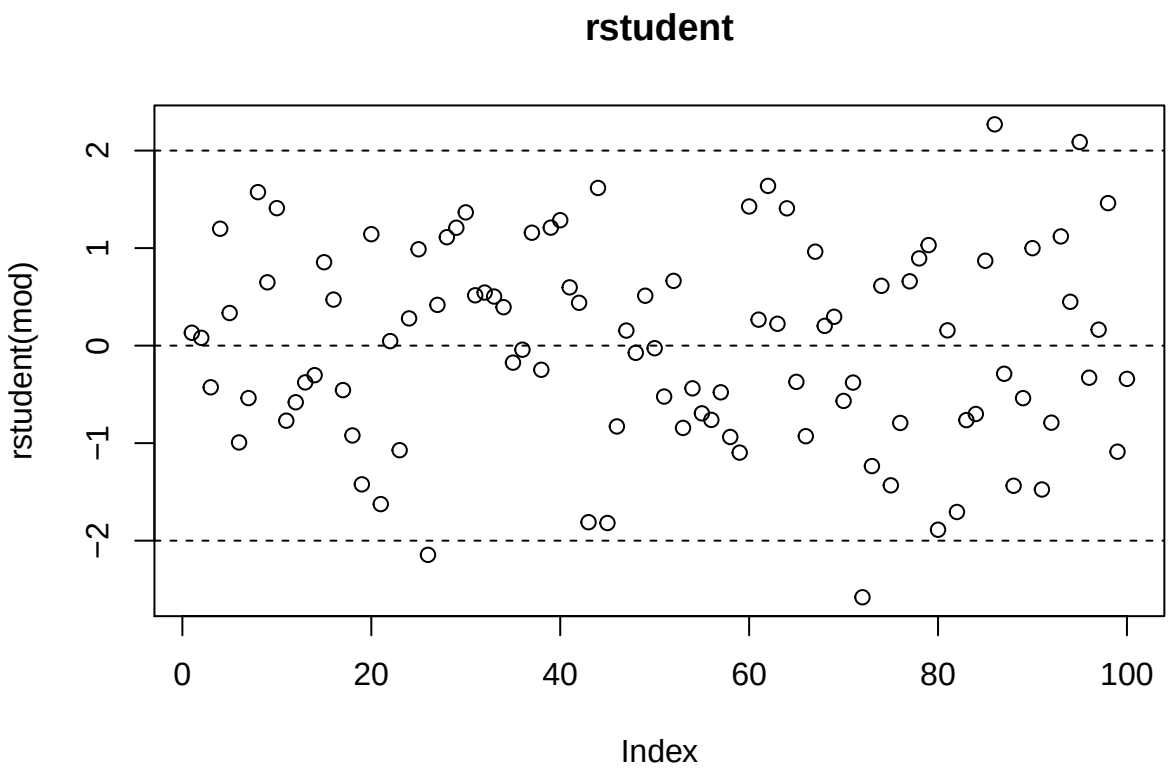
1) We do not observe any pattern in the residuals as a function of the predicted values.
2) We also see that the variability is quite constant over all the range, so the homocedasticity property is not rejected.

Diagnostic: OUTLIERS (rstudent)

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




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

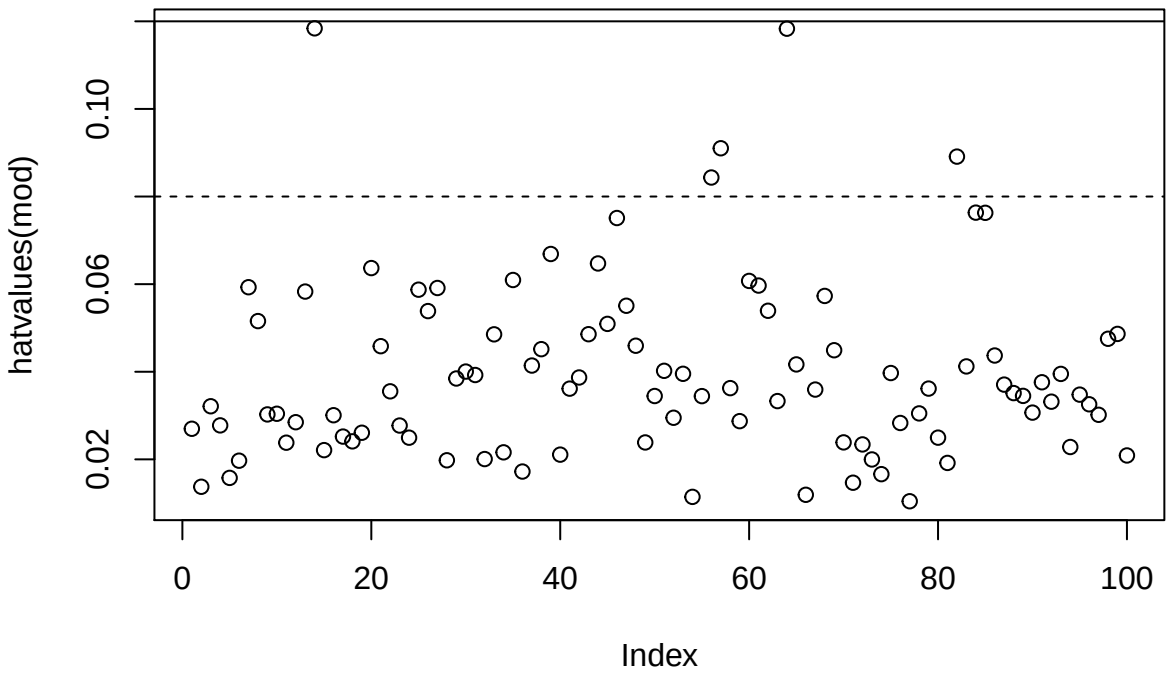


From the standarized residuals we observe that less than a 5% do not belong to the interval $(-2, 2)$.

We do not observe observations susceptible to be outliers

Diagnostic: LEVERAGE

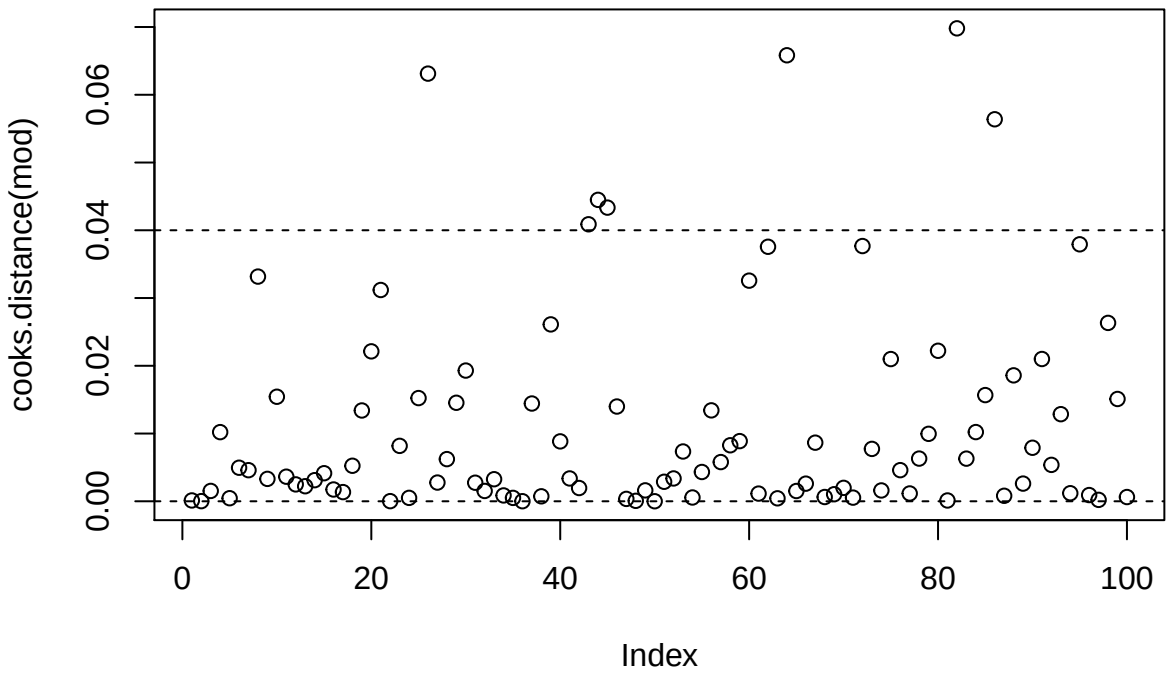
```
plot(hatvalues(mod))
abline(h=c(0,2*mean(hatvalues(mod))),lty=2)
abline(h=c(0,3*p/n))
```



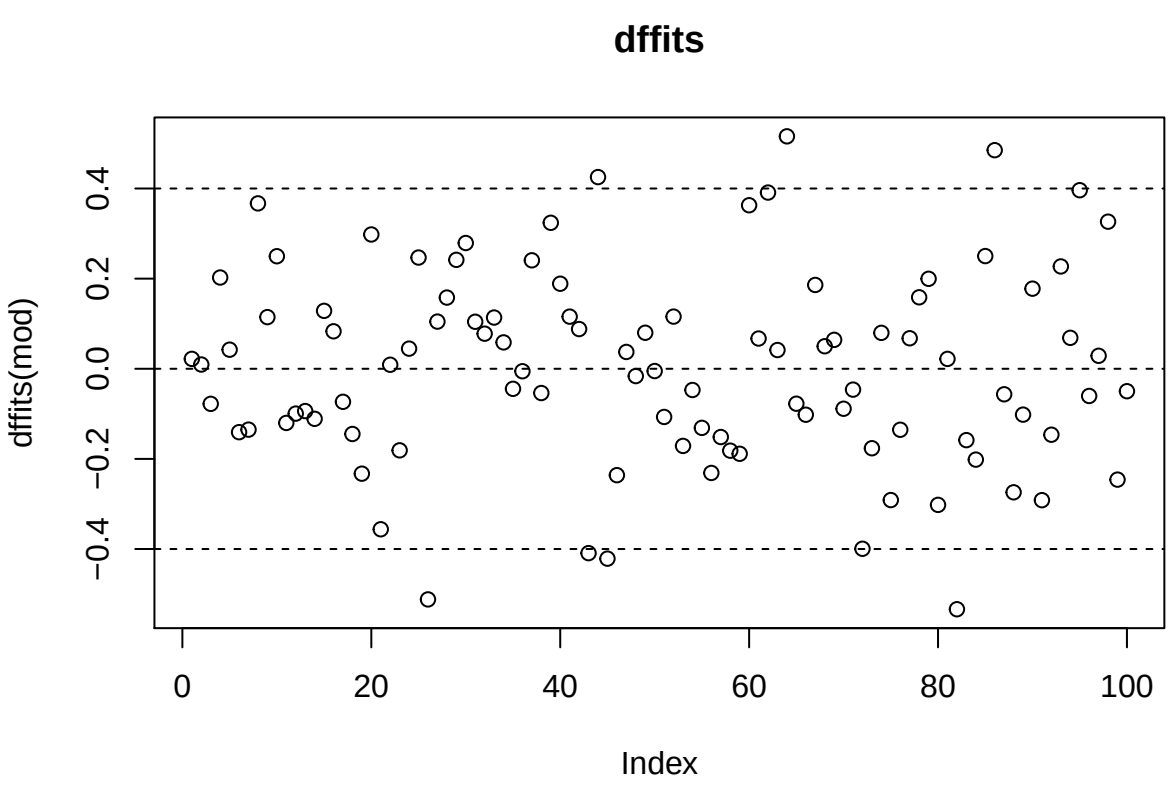
There are no values with a leverage larger than $3 * p/n$.

Diagnostic: Influential values (dffits)

```
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)
```



There are no values with a cook distance larger than one, which are clearly the influential values.

The difference of the fitted values with and without each one of the observations is not large enought to consider any of the observations as an influential observation.

Diagnostic: Colinearity

It is important to compute the VIF values for each explanatory variable. VIF values larger than 5 indicate that his variable is largely correlated with the others.

```
vif(mod)
```

```
##         W         A         H
##  9.489406 20.904776 31.695499
```

We clearly see that the variable H is the one that has a large VIF value. Instead of supress it, we will try to redefine some variables in order to see if it is possible to break the linear dependence between them.

It is important to include any additional information that we have with respect to the data set. In this particular case, it is reasonable to think that what will really influence the cholesterol level is the excess of weight instead of the proper weight of the person.

Thus, it is important to look for a pattern of the behaviour of the weight as a function of the height, from which we can compute the excess of zero. Moreover, the excess of weight will no longer be related with the height neither with the age or at least, the realtion will be less strong.

In what follows we assume that the weight is described as a function of the height by means of the equation: $W_0 = -10 + 0.5 * H$ **and We define the exces of weight as** $EW = W - W0$**.**

we will model the cholesterol level considering the excess of weight instead of the weight itself.

```
newmod<-lm(C~I(W-(-10+0.5*H))+A+H, colest)
summary(newmod)
```

```
##
## Call:
## lm(formula = C ~ I(W - (-10 + 0.5 * H)) + A + H, data = colest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.608 -22.137   1.888  21.156  65.410
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            387.22473   33.69605  11.492  < 2e-16 ***
## I(W - (-10 + 0.5 * H))  10.37731    0.73649  14.090  < 2e-16 ***
## A                      -13.01948    3.85300  -3.379  0.00105 **
## H                        0.08972    0.58736   0.153  0.87891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.11 on 96 degrees of freedom
## Multiple R-squared:  0.8101, Adjusted R-squared:  0.8041
## F-statistic: 136.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

What changes in this model are the parameter estimations as a consequence that we have changed the design matrix. We also see that the H is not significant.

It doesn't change the *sd*, the R^2 and the F-value.

Let us compute the VIF for the newmodel

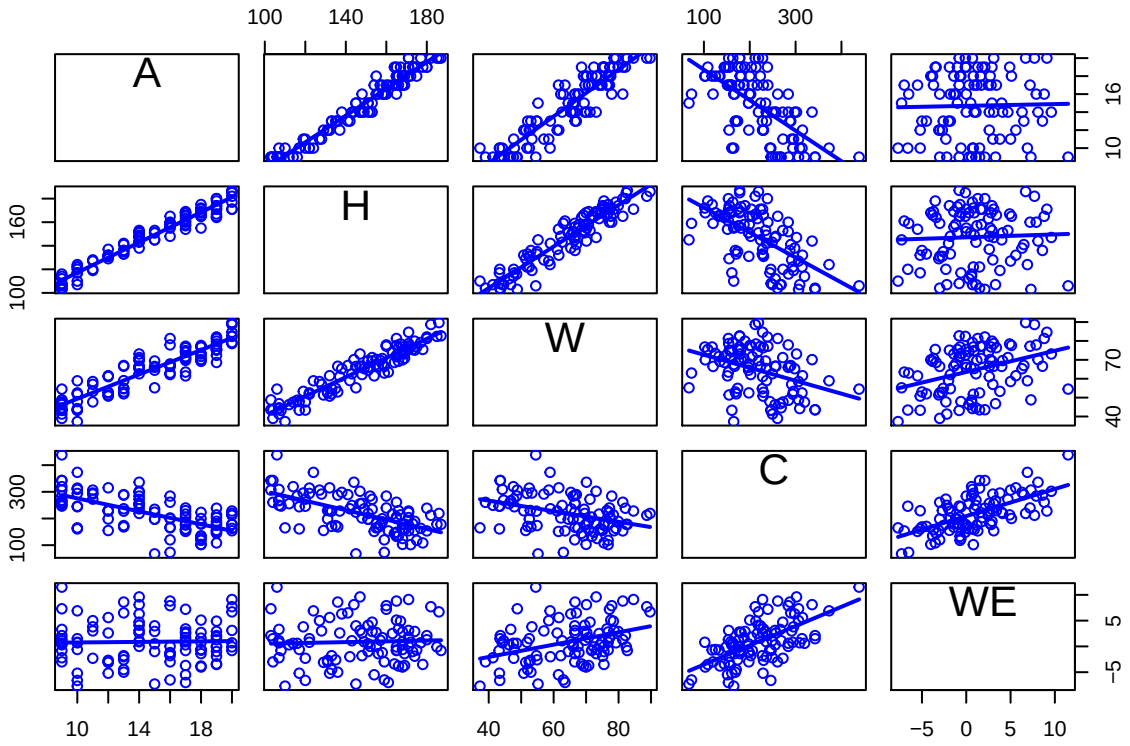
```
vif(newmod)
```

```
## I(W - (-10 + 0.5 * H))                      A                      H
##               1.009937              20.904776              20.933520
```

The VIF values continue being large, but the values have changed. Observe that the Excess of weight is no longer lineary related with the other two variables.

Let us do the scattered plot with the new dataframe that contains the excess of weight

```
colest$WE<-colest$W-0.5*colest$H+10
scatterplotMatrix(colest,smooth=F,diagonal=F)
```



We now see that the excess of weight is not any more correlated with the age.

Since Height is not significant and highly correlated with age, we suppres it and we model only with the excess of weight and the age

```
renewmod<-lm(C~I(W-(-10+0.5*H))+A, colest)
summary(renewmod)
```

```
##
## Call:
## lm(formula = C ~ I(W - (-10 + 0.5 * H)) + A, data = colest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.286 -22.638   1.755  20.935  66.244
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             391.9885    12.6975   30.87   <2e-16 ***
## I(W - (-10 + 0.5 * H))  10.3882     0.7294   14.24   <2e-16 ***
```

```
## A                     -12.4452     0.8387  -14.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.96 on 97 degrees of freedom
## Multiple R-squared:   0.81,  Adjusted R-squared:  0.8061
## F-statistic: 206.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
vif(renewmod)
```

```
## I(W - (-10 + 0.5 * H))                      A
##               1.000527               1.000527
```

Modelling with the excess of weight and the age, we obtain approximatly the same values of the *sd* and the R^2. The VIF values are both small and thus, we do not have a colinearity problem.

In what follows we center in some values that are near the sample

means, which are: $W = 65$**,** $A = 15$ **and** $H = 150$**.**

```
centermod<-lm(C~I(W-65)+I(A-15)+I(H-150), colest)
summary(centermod)
```

```
##
## Call:
## lm(formula = C ~ I(W - 65) + I(A - 15) + I(H - 150), data = colest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.608 -22.137   1.888  21.156  65.410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 205.3908     3.1355  65.504  < 2e-16 ***
## I(W - 65)    10.3773     0.7365  14.090  < 2e-16 ***
## I(A - 15)   -13.0195     3.8530  -3.379  0.00105 **
## I(H - 150)   -5.0989     0.7227  -7.055 2.68e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.11 on 96 degrees of freedom
## Multiple R-squared:  0.8101, Adjusted R-squared:  0.8041
## F-statistic: 136.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

Comparing this summary with the one obtained in model *mod*, it can be seen that the only parameter estimation that changes is the one of the intercept. parameter interpretation will change since we have tranlated the explanatory variables. The values for the *sd*, the R^2 and the F do not change.