Linear Model: Levels of cholesterol

It is well known that the excess of weight is one of the factors that has a negative influence in the cholesterol level of human beings. In an experiment with children from 9 to 20 years old, the following variables were obtained:

- Cholesterol level (C)
- Weight (W)
- Height (H)
- Age (A)

(1) Compute the regression line for modeling the cholesterol as a function of weight.

```
library(car)
dd <- read.csv2("COL.csv")
head(dd)
```

```
##    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 want to find out how are cholesterol and weight related. We are going to perform a **simple linear regression** with the following model:

$$C_i = \beta_0 + \beta_1 W_i + e_i$$

And we are going to test whether the weight has influence upon the cholesterol (by intuition, we expect to see with more weight - more cholesterol). The null model would be $C_i = \mu + e_i$, where the level of cholesterol would be constant (= same response for every weight).

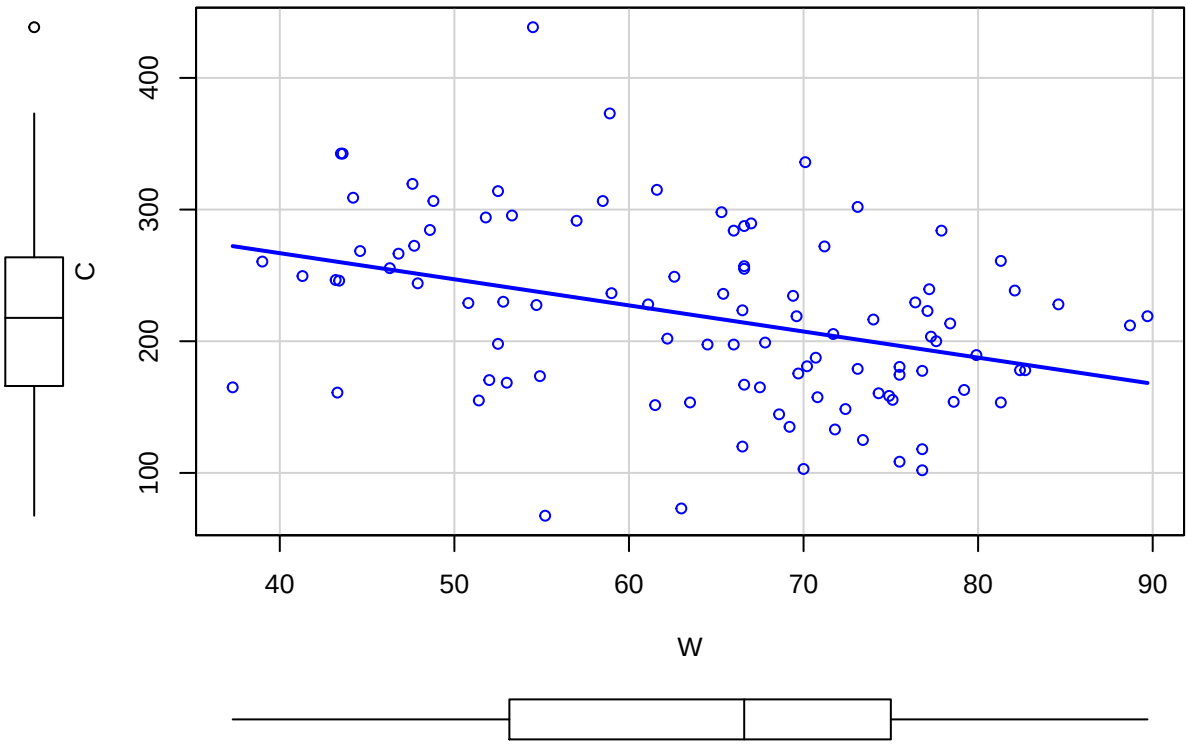
We set the number of parameters to 2 and we compute the number of **experimental units** (= people):

```
p<-2                # parameters
n<-dim(dd)[1]     # experimental units (= people)
```

Now, let's have a look on the scatterplot of the cholesterol as a function of weight:

$$C = g(W) \rightarrow C \sim W$$

```
sp(C~W, smooth=F, dd)
```



The scatter plot is correct but... we see a contradiction: with more weight we see lower levels of cholesterol! To find the origin of that paradox, we shall have a look on the summary of the linear model.

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

```
##
## Call:
## lm(formula = C ~ W, data = dd)
##
## 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 (*Residual standard error* $= 63.55$), 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.

To test whether this model explains the data or nor (if it does not, it is the null model!), we use the **Omnibus**

test:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$$

$$H_1 : \exists i | \beta_i \neq 0$$

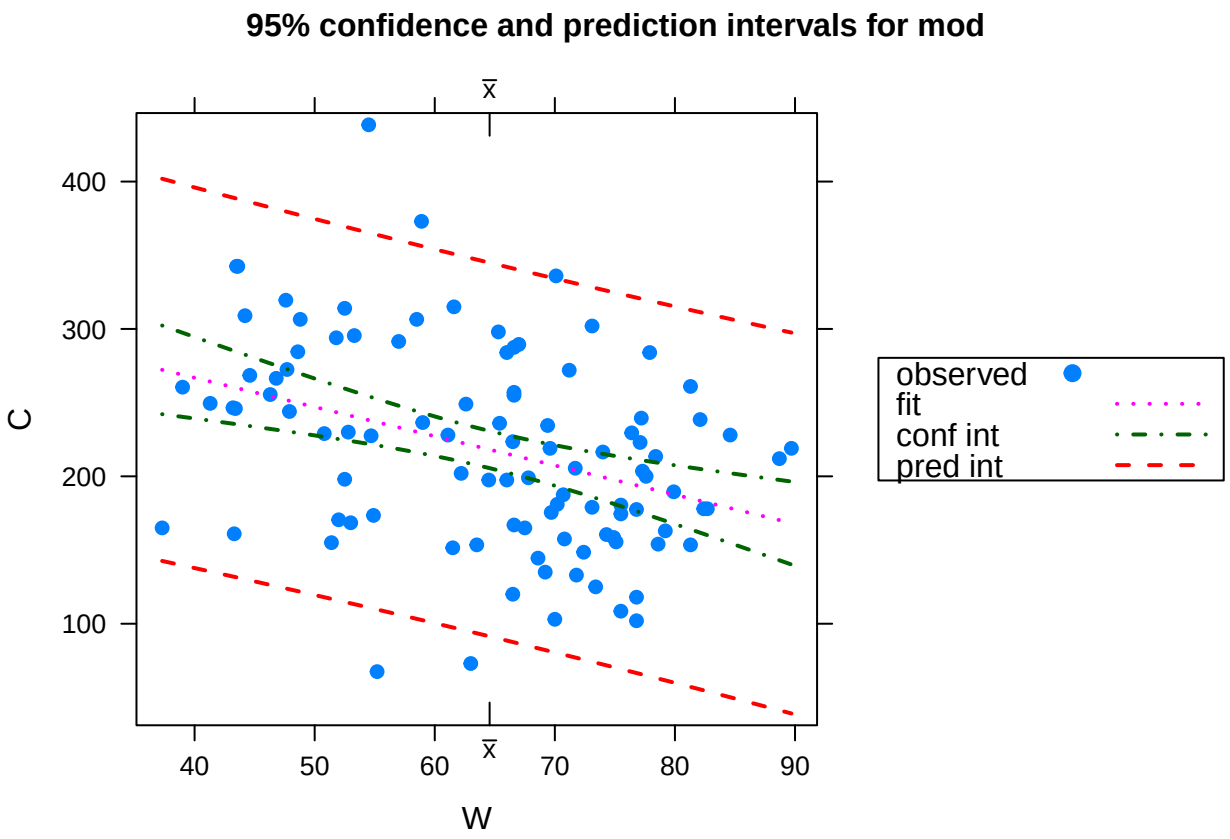
Com es contrasta aquest test? En H_0 tenim uns valors predits $\bar{Y}$ i en H_1 tenim uns valors predits $\hat{Y}_i$. Si $\bar{Y}$ i $\hat{Y}_i$ són molt iguals, aleshores el model no aporta gairebé res i ens podem quedar amb el model nul:

$$F_{p-1,n-p} = \frac{RegSS/p-1}{\hat{\sigma}^2} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2/p-1}{\hat{\sigma}^2} = 15.45$$

$(-3.93)^2$ (*T-Student*) $= 15.45$ (*Fisher*). El p-valor surt significatiu: $0.0001581 < 0.05$, per tant acceptem H_1, el model explica! Però, segurament per falta d'informació, el model no ajusta bé.

(2) Plot the regression line jointly with the confidence intervals (CI) and the prediction intervals (PI).

```
library(HH)
ci.plot(mod)
```



From this plot we see that:

(a) As the weight increases the cholesterol level decreases, which is quite contradictory.
(b) 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.
(c) 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.

(3) Perform the appropriate plots to check:

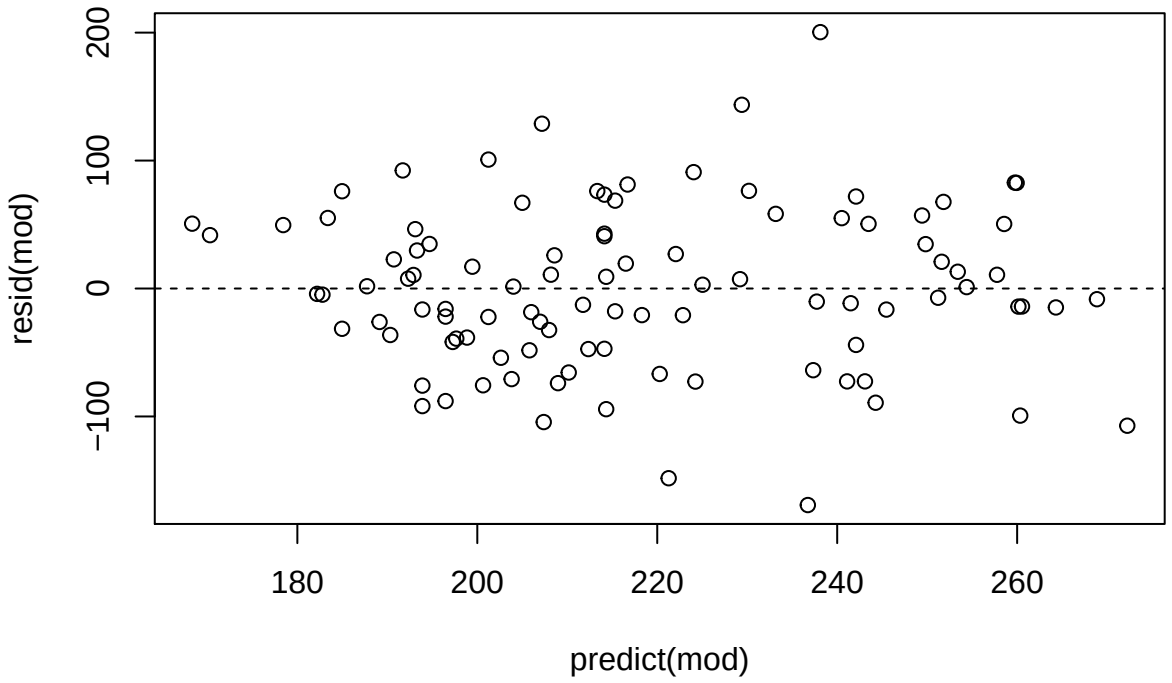
- **Tendency and homogeneity of variances**. Plot the residuals as a function of the predicted values.
- **Outliers**. Plot the studentized residuals as a function of any of the predictors, observation number ... jointly with the horizontal lines at ± 2.

- **Influence values**. Plot the dffits as a function of any of the following variables: predicted values, observation number ...; jointly with the horizontal lines at $\pm 2\sqrt{\frac{p}{n}}$.

Diagnostic: tendencies

We first plot the predicted values vs residuals.

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

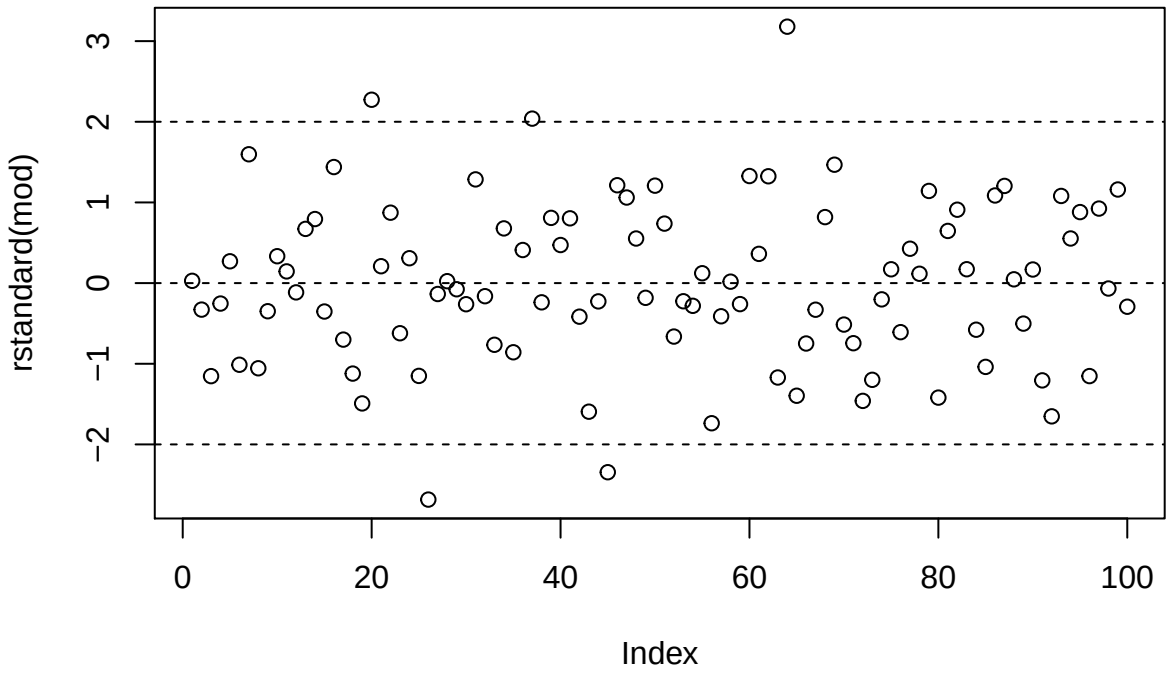


We can see that the homogenity of variances is reasonable to be assumed, there are not patters in the plot - which is good. Nevertheless, the residuals are very large.

Diagnostic: outliers

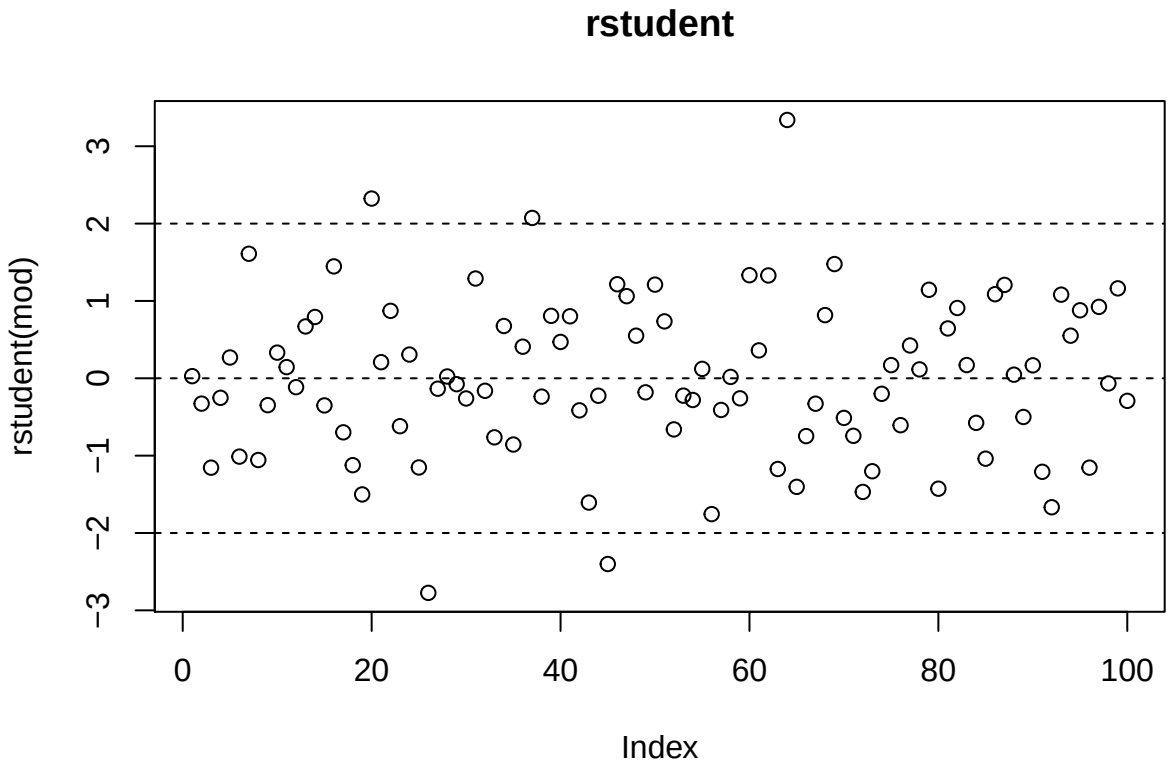
Let us plot the standarized/studentized residuals:

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



Amb els residus de Student, els valors extranys surten encara una mica més apartats (= més lluny de 0), més marcats i podem detectar els possibles outliers. *Nota*: no s'han de fer les dues gràfiques, amb una qualsevol ja n'hi ha prou!

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



We observe that:

(a) Both residuals plots are quite reasonable. As a consequence since the standar error estimation is large,

the standarized and studentized residuals become relatively small.

(b) Four residuals (less than a 5%) lie out of the interval (-2,2) which is not a problem.

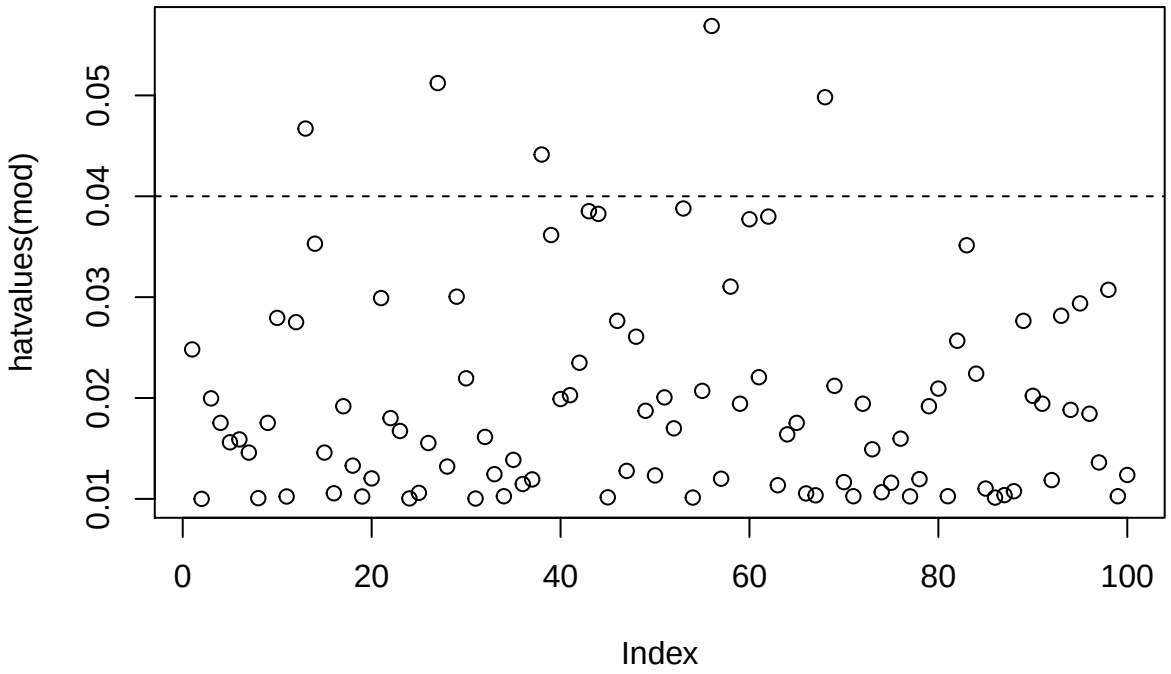
(c) There are no patterns.

(d) 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

More diagnostics: in what follows we compute the **leverage** of the observations. Remind that the leverage just depends on the values of the X matrix (it checks whether the data are far from the center of gravity of the explanatory variables).

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

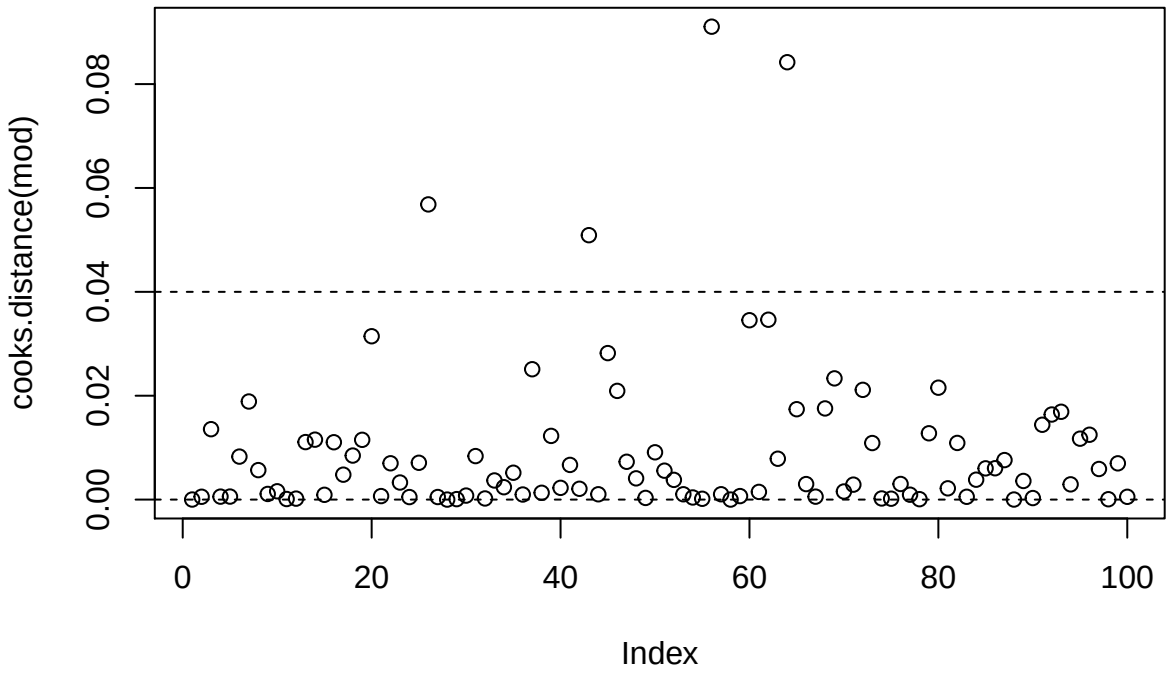


We observe that there are no values, with a leverage larger than $3p/n = 0.06$. Només hi ha 5 dades que són més grans que 2 cops la mitjana, que podriem reconsiderar, però si els deixem no passa res, perquè no hi ha res extraordinari. Si hi haguéssin valors extranys, això seria un problema del dissenys (de com hem escollit les dades per fer la regressió (el pes)), no de la regressió (el colesterol).

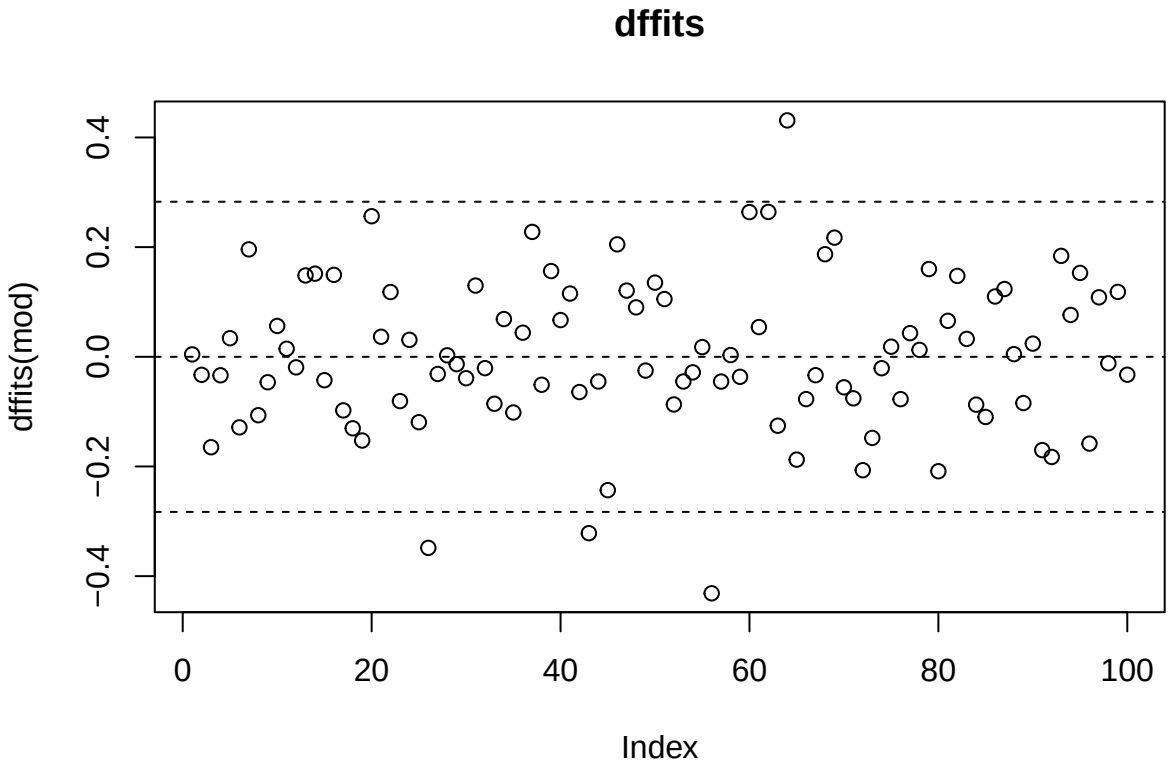
Diagnostic: influential values (cook.distance, dffits)

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

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



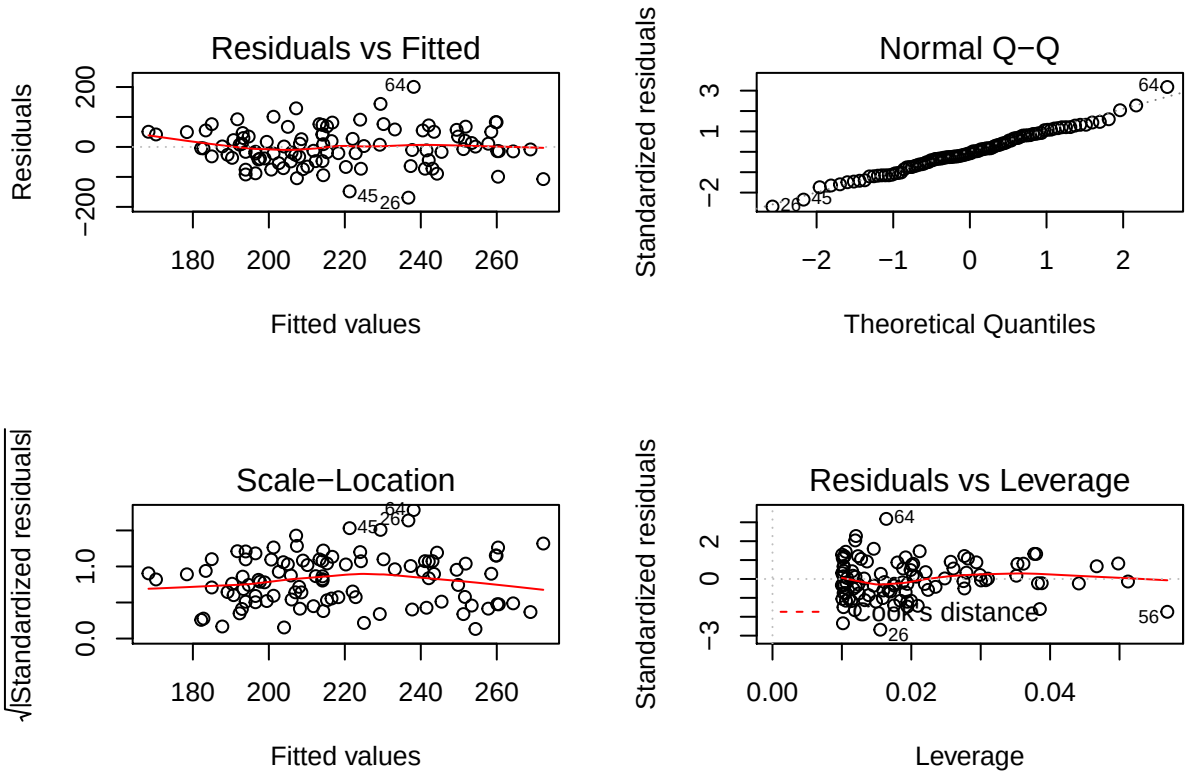

```
plot(dffits(mod),main="dffits")
abline(h=c(-2*sqrt(p/n),0,2*sqrt(p/n)),lty=2)
```



Si per algún el dffit és molt gran, vol dir que aquell punt és molt influent. Els límits que normalment s'utilitzen pels dffits són $\pm 2\sqrt{\frac{p}{n}}$, perquè per una mostra en aquest interval hi ha el 95% dels dffits.

R diagnostic

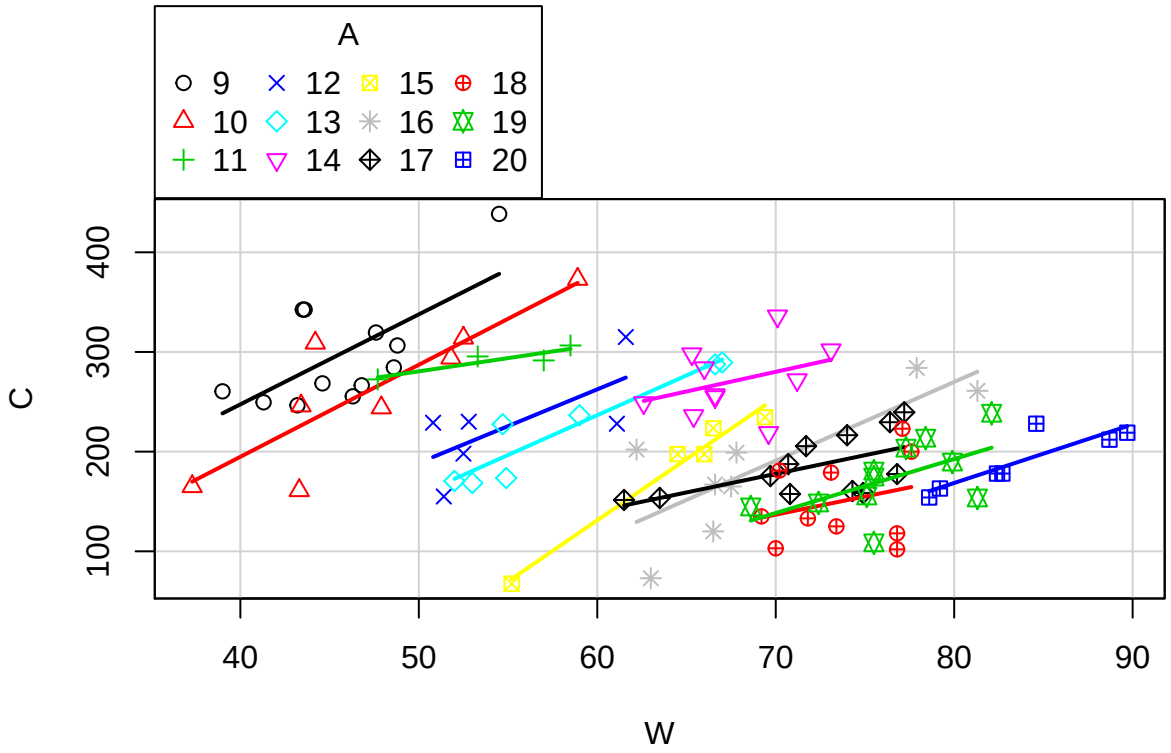
```
oldpar <- par( mfrow=c(2,2))
plot(mod,ask=F)
```




```
par(oldpar)
```

(4) Perform the dispersion plot of C as a function of W, jointly with the regression lines for ages. Is there any contradiction?

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



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 has 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.

La informació que falta és que el nivell de colesterol dels nens és més elevat que en els adults, i en l'época d'adolescencia, el colesterol baixa, però el pes puja i aquí és on trobem la contradicció si tenim en compte l'edat.

(5) Perform a **multiple linear regression** to model the cholesterol level as a function of weight, height and age.

That is that one assumes that:

$$C_i = \beta_) + \beta_1 W_i + \beta_2 A_i + \beta_3 H_i + e_i$$

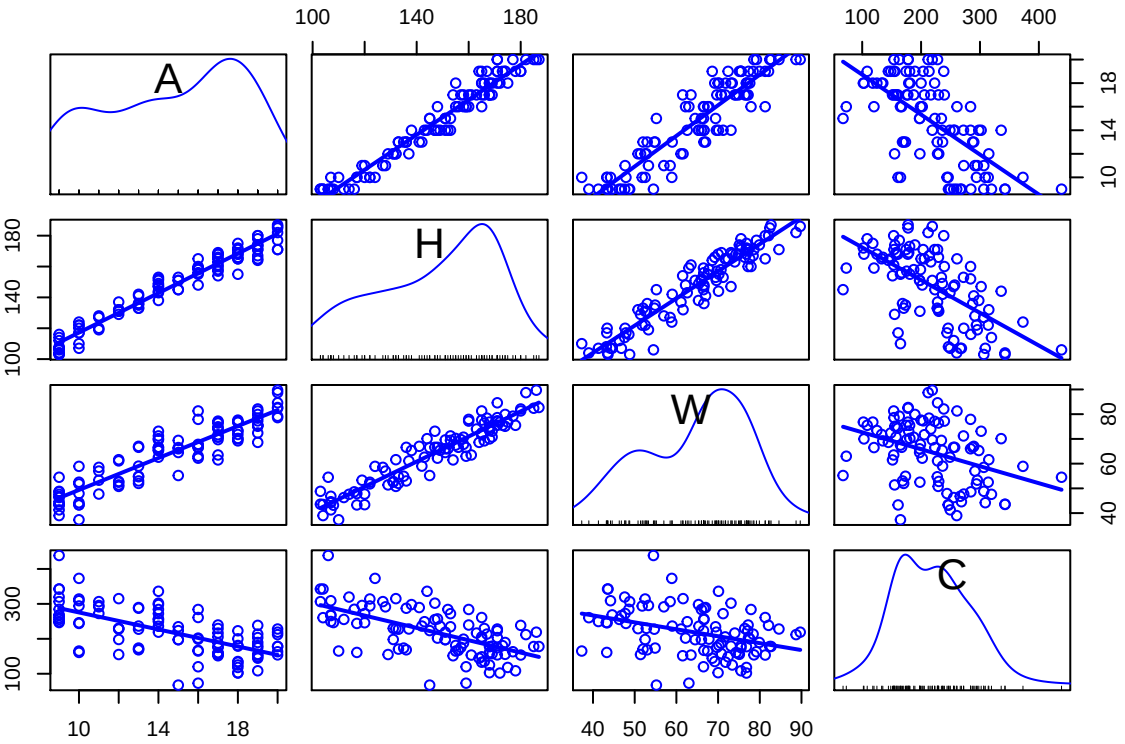
with $e_i \sim N(0, \sigma^2)$. In matrix notation this is equal to: $Y = X\beta + e$.

First of all, we compute the number of observations and we set the number of parameters to be equal to four:

```
p<-4                # parameters
n<-dim(dd)[1]       # experimental units (= people)
```

Secondly, we plot the scatterplot of the variables (descriptive):

```
scatterplotMatrix(dd, smooth=F, diagonal=T)
```



From this scatterplot we deduce that:

(a) There exists a strong linear relationship between height and age,
(b) There exists a strong linear relationship between height and weight,
(c) Less clear but also important linear relationship between age and weight,
(d) Cholesterol is related with each one of the explanatory variables with a straignt line with negative slope.

Let's define the linear model:

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

```
##
## Call:
## lm(formula = C ~ W + A + H, data = dd)
##
## 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
```

From the summary of the model we deduce that:

(a) There exist 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 see that the weight coefficient is positive meaning that when the weight increases, the cholesterol level also does. The age and the height have a negative coefficient meaning that when they increase, the cholesterol level decreases.
(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**.

(f) The puntual estimation of $\sigma^2 = (30.11)^2 = 906.61$ which is equal to the the mean residual sum of squares.
(g) **Omnibus test**: this test corresponds to $H_0 = \beta_1 = \beta_2 = \beta_3 = 0$ vs $H1 : notH_0$. Thus we are testing if globaly our model is explaining a significant part of the variability in the cholesterol.

Based on the p-value associated to the F test (2.2e-16 < 0.05), we reject the null hypothesis and conclude that our model is useful to explain an important part of the variability in the response variable.

$$F_{p-1,n-p} = \frac{RegSS/p-1}{\hat{\sigma}^2} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2/p-1}{\hat{\sigma}^2} = 136.5$$

(6) For each regression parameter compute its punctual and confidence interval estimations (95%) and perform the test $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$.

```
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 from zero. The same is deduced by looking at the p-values that appear in the summary (the tests have already been performed in the *summary*).

This intervals show us, that, for instance, $\beta_1 = 10.3773$, with probability 95% is the range of $[8.441792, 12.312821]$.

(7) Define the hypothesis of the ANOVA test, using the sums of squares of type I and type III.

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 (SS1, sum of squares 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.

To compare $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ it checks this:

$$H_0 : C = \beta_0 + e$$

$$H_1 : C = \beta_0 + \beta_1 W + e$$

Sense tenir en compte l'edat i l'alçada, el pes afecta el nivell de colesterol?

To compare $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$ it checks this:

$$H_0 : C = \beta_0 + \beta_1 W + e$$

$$H_1 : C = \beta_0 + \beta_1 W + \beta_2 A + e$$

Sense tenir en compte l'alçada, l'edat afecta el nivell de colesterol, tenint en compte que el pes hi pot intervenir? Dient-ho d'una altra forma, l'edat ens aporta alguna cosa que no ens aport el pes?

To compare $H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$ it checks this:

$$H_0 : C = \beta_0 + \beta_1 W + \beta_2 A + e$$

$$H_1 : C = \beta_0 + \beta_1 W + \beta_2 A + \beta_3 H + e$$

Sabent que les altres dues variables poden intervenir, l'alçada ens aporta alguna cosa? L'alçada ens aporta alguna cosa que no ens aporten el pes i l'edat?

I aquests tests depenen de l'ordre en què hem posat les variables! En general, aquest sistema no es fa servir, a no ser que tinguem un ordre clar de preferències entre les variables (principals i accessòries). Si totes les variables són igual d'importants, l'ordre no ha d'intervenir.

Observació: els nivells de significació del test *anova* no tenen res a veure, en un prinicipi, amb els del *summary*. I aquests valors canvien amb l'ordre.

```
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 the 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 to the F-value of the Anova table. For example, with respect to the variable weight we have that $(14.09)^2 = 906.61$. The same is true for the rest of variables in the model. We can also check that $\left(\frac{87031}{96}\right) = 30.11$ which is the the *standard error estimation*.

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

Anova en majúscules fa la suma de quadrats de tipus 2 (o de tipus 3), SS2 i SS3. La idea, en aquest cas, és contrastar:

$$H_0 : C = \beta_0 + \beta_2 A + \beta_3 H + e$$

$$H_1 : C = \beta_0 + \beta_1 W + \beta_2 A + \beta_3 H + e$$

El que fem és treure del model el pes, i comparar el model sense la variable pes i el model complet. Si tenim totes les variables en el model, la variable pes ens aporta alguna cosa més o no? I per tota la resta de variables, és de fa manera equivalent, per exemple, per l'edat:

$$H_0 : C = \beta_0 + \beta_1 W + \beta_3 H + e$$

$$H_1 : C = \beta_0 + \beta_1 W + \beta_2 A + \beta_3 H + e$$

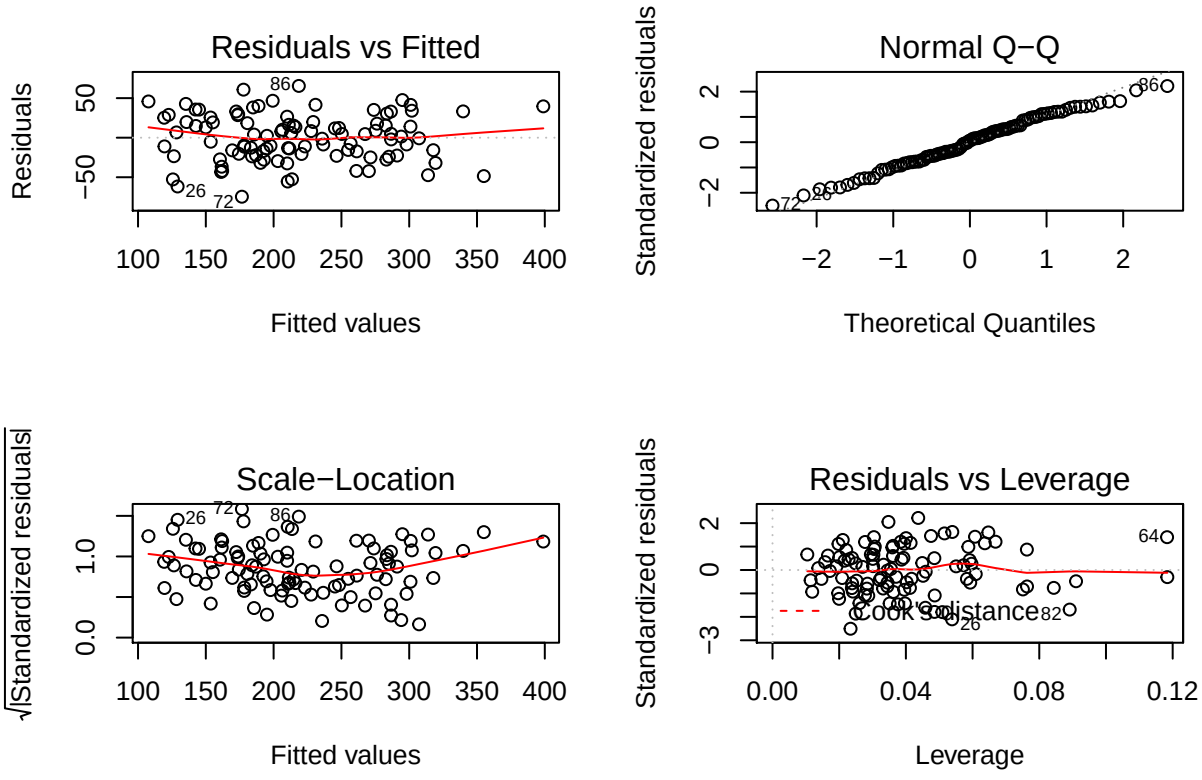
I aquí no importa l'ordre en què hem posat les variables.

(8) Study the model diagnostics.

R diagnostic

The plots that R performs by default, we can 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)
```

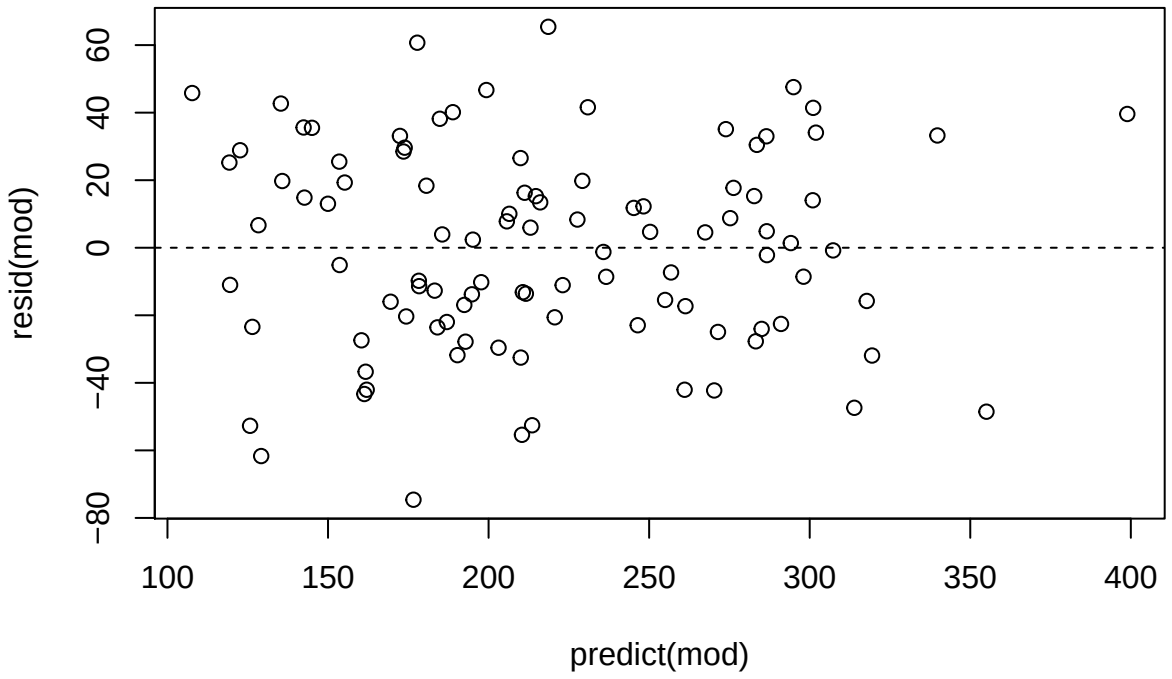



```
par(oldpar)
```

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 within the straight line in the *qqplot*.

Diagnostic: tendencies

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

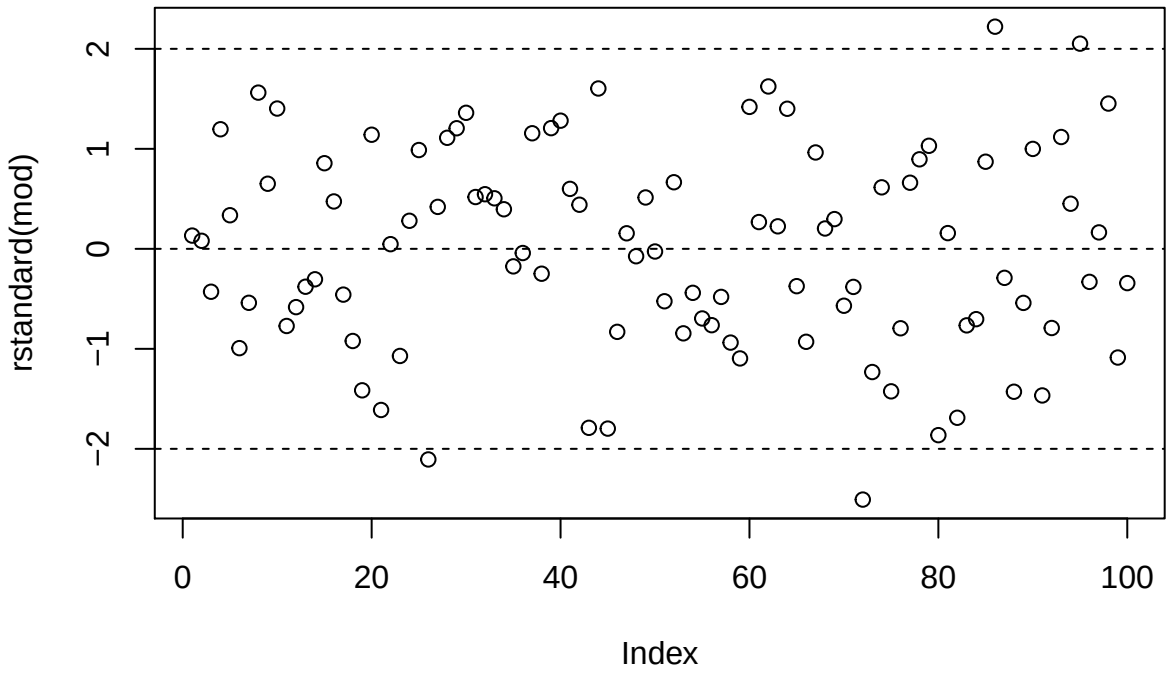


(a) We do not observe any pattern in the residuals as a function of the predicted values.
(b) We also see that the variability is quite constant over all the range, so the homocedasticity property is not rejected.

Diagnostic: outliers

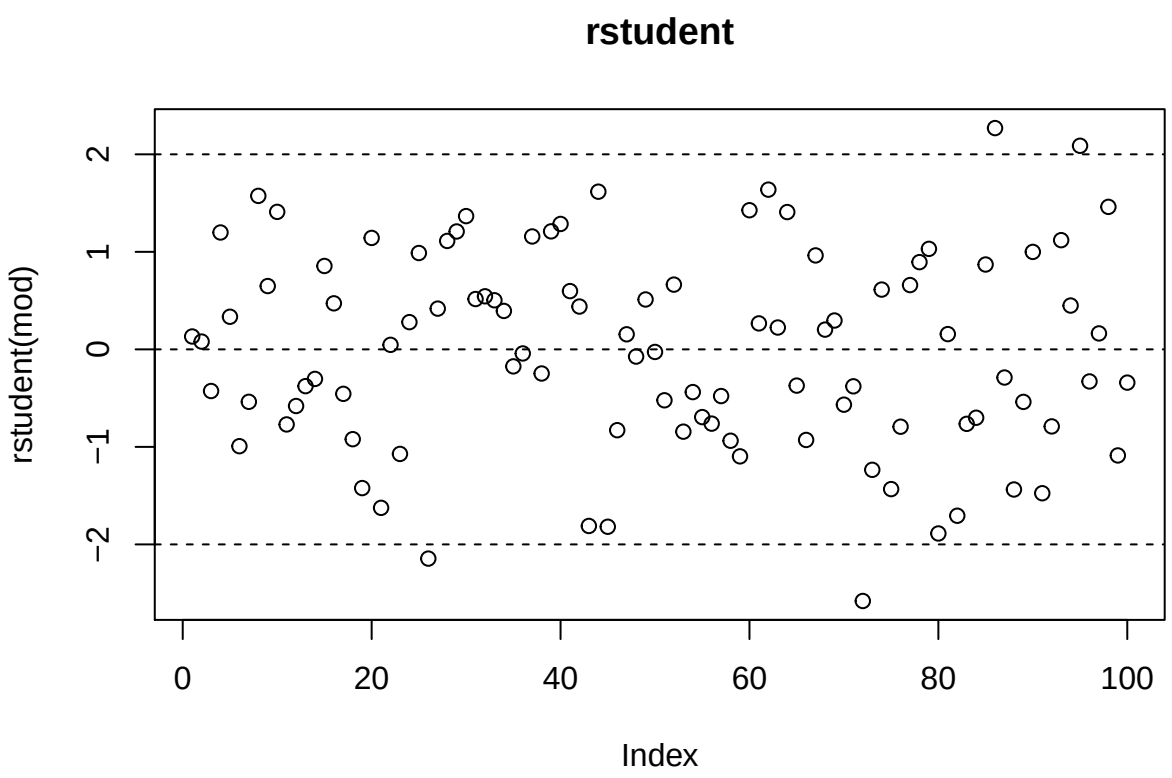
We plot the standarized or studentised residuals (choose one!):

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



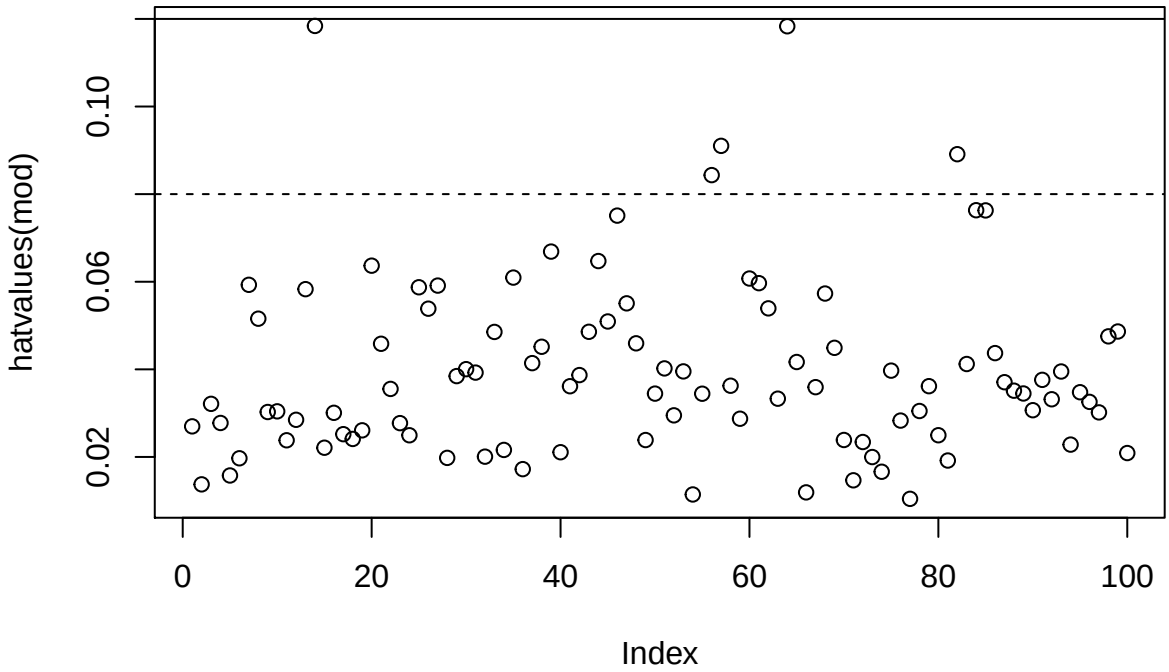
From the standarized residuals we observe that less than a 5% do not belong to the interval (-2, 2). We do not observe observations susceptible to be outliers.

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



Diagnostic: leverage

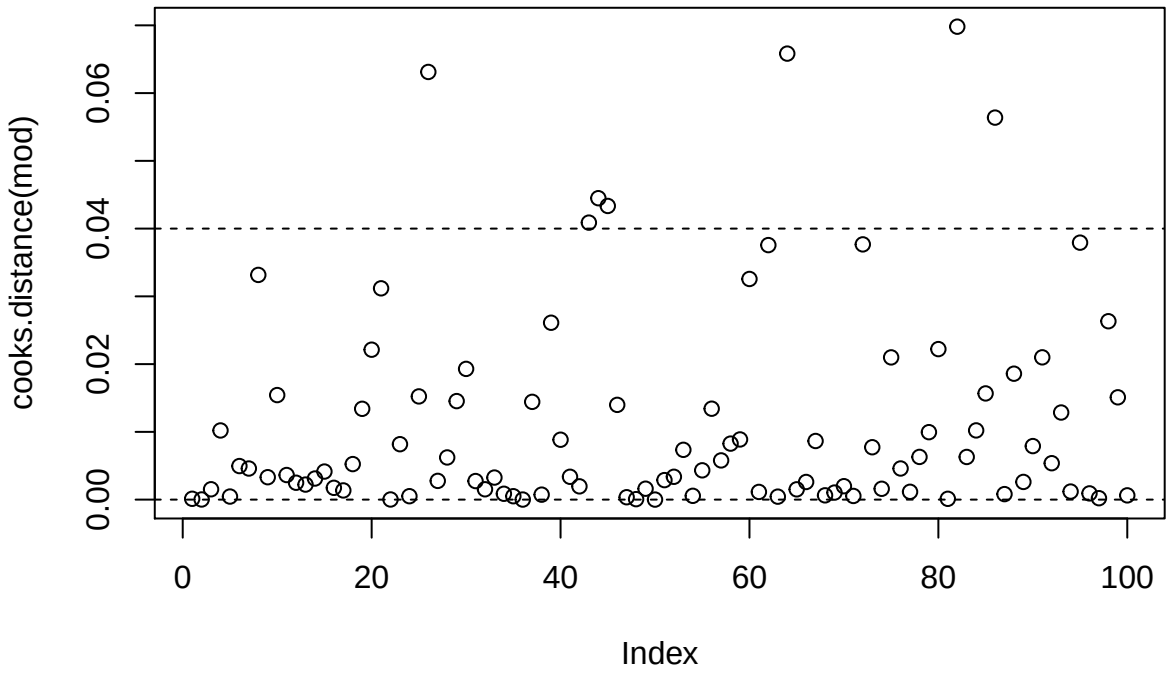
```
plot(hatvalues(mod))
abline(h=c(0,2*mean(hatvalues(mod))),lty=2)
abline(h=c(0,3*p/n))
```



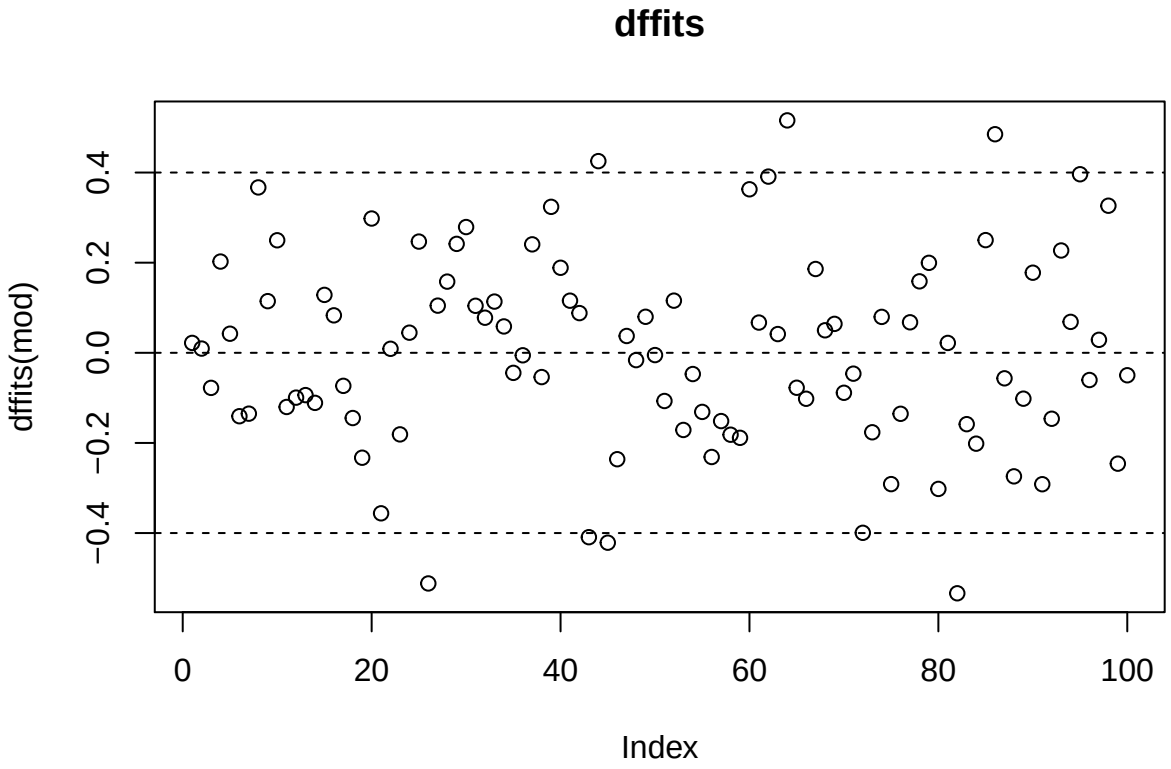
There are no values with a leverage larger than $3 \cdot p/n$.

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 enough to consider any of the observations as an influential observation.

Diagnostic: collinearity

It is important to compute the VIF values for each explanatory variable. VIF values larger than 5 indicate that this 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 the largest VIF value. Instead of supressing it, we will try to redefine some variables in order to see if it is possible to break the linear dependence between them.

(9) Assume that $W_0 = -10 + 0.5 \cdot H$ is a pattern of weight as a function of height, and that the excess of weight is computed as: $EW = W - W_0$. Compute the regression:

$$C_i = \beta_0^{'} + \beta_1^{'} EW_i + \beta_2^{'} A_i + \beta_3^{'} H_i + e_i$$

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: it is the **excess** of weight instead and not just the 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 relation 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: $W_0 = -10 + 0.5 \cdot H$ and We define the excess of weight as $EW = W - W_0$. We will model the cholesterol level considering the excess of weight instead of the weight itself.

I is for an internal computation (of the formula):

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

```
##
## Call:
## lm(formula = C ~ I(W - (-10 + 0.5 * H)) + A + H, data = dd)
##
## 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. The standard deviation does not change, nor the R^2 and the F-value.

Let us compute the VIF for the new model:

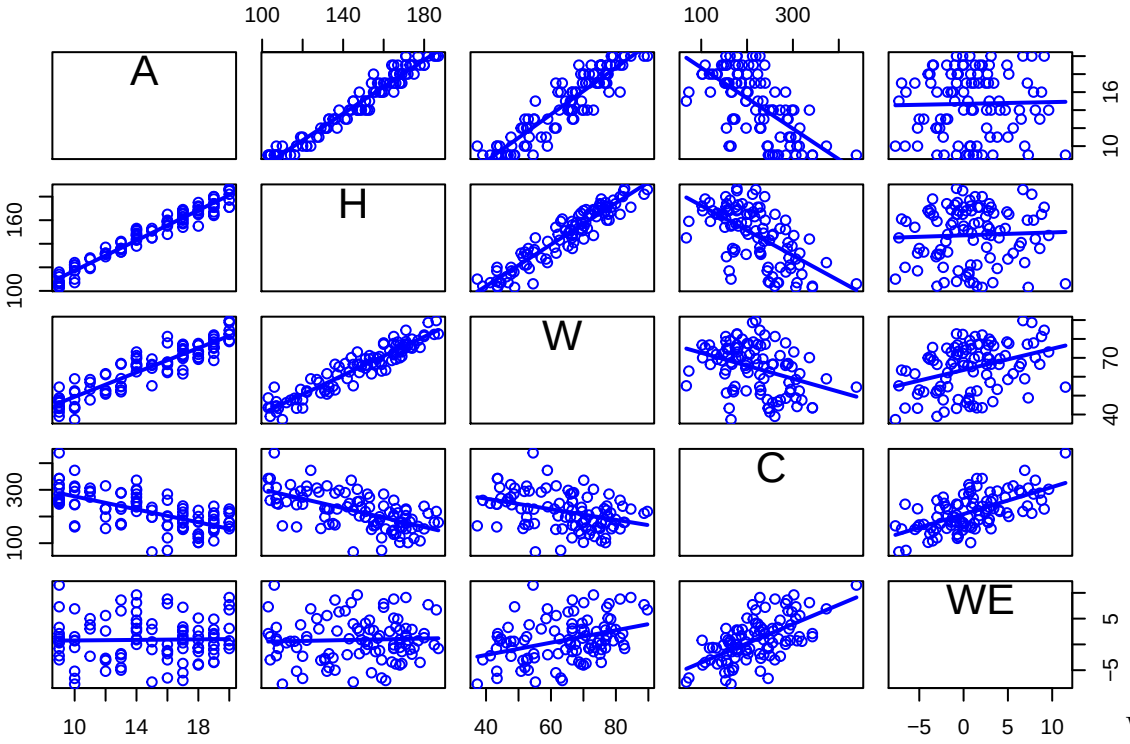
```
vif(newmod)
```

```
## 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:

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



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

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, dd)
summary(renewmod)
```

```
##
## Call:
## lm(formula = C ~ I(W - (-10 + 0.5 * H)) + A, data = dd)
##
## 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 the collinearity trouble.