# LINEAR 8 - Data set: ANTROP

## INTRODUZIONE

Il dataset è costituito da alcune misure antropometriche rilevate su 248 uomini.

1. ETA': età in anni compiuti
2. PESO: peso rilevato in libbre
3. ALTEZ: altezza (cm)
4. COLLO: circonferenza del collo (cm)
5. TORACE: circonferenza toracica (cm)
6. ADDOM: circonferenza addominale (cm)
7. ANCA: circonferenza dell'anca (cm)
8. COSCIA: circonferenza della coscia (cm)
9. GINOCCH: circonferenza del ginocchio (cm)
10. CAVIGLIA: circonferenza della caviglia (cm)
11. BICIPITE: circonferenza del bicipite in estensione (cm)
12. AVANBR: circonferenza dell'avambraccio (cm)
13. POLSO: circonferenza del polso (cm)

Analisi proposte:

1. Statistiche descrittive
2. Regressione lineare
3. Diagnostiche ed analisi dei residui

```
#-- R CODE

library(pander)
library(car)
library(olsrr)
library(systemfit)
library(het.test)
panderOptions('knitr.auto.asis', FALSE)

#-- White test function
white.test <- function(lmod,data=d){
  u2 <- lmod$residuals^2
  y <- fitted(lmod)
  Ru2 <- summary(lm(u2 ~ y + I(y^2)))$r.squared
  LM <- nrow(data)*Ru2
  p.value <- 1-pchisq(LM, 2)
  data.frame("Test statistic"=LM,"P value"=p.value)
}

#-- funzione per ottenere osservazioni outlier univariate
FIND_EXTREME_OBSERVARION <- function(x,sd_factor=2){
  which(x>mean(x)+sd_factor*sd(x) | x<mean(x)-sd_factor*sd(x))
}

#-- import dei dati
ABSOLUTE_PATH <- "C:\\Users\\sbarberis\\Dropbox\\MODELLI STATISTICI"
```

```r
d <- read.csv(paste0(ABSOLUTE_PATH,"\\F. Esercizi(22) copia\\4.tutto(4)\\2.tutto\\ANTROP.TXT"),sep="\t")
d$bmi <- ((d$peso/2.2046)/(d$altez/100)^2)

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("bmi","addom","coscia","bicipite")

#-- print delle prime 6 righe del dataset
pander(head(d),big.mark=",")
```

Table 1: Table continues below

| id_sogg | eta | peso | altez | collo | torace | addom | anca | coscia |
|---|---|---|---|---|---|---|---|---|
| 1 | 23 | 154.2 | 172.1 | 36.2 | 93.1 | 85.2 | 94.5 | 59 |
| 2 | 22 | 173.2 | 183.5 | 38.5 | 93.6 | 83 | 98.7 | 58.7 |
| 3 | 22 | 154 | 168.3 | 34 | 95.8 | 87.9 | 99.2 | 59.6 |
| 4 | 26 | 184.8 | 183.5 | 37.4 | 101.8 | 86.4 | 101.2 | 60.1 |
| 5 | 24 | 184.2 | 181 | 34.4 | 97.3 | 100 | 101.9 | 63.2 |
| 6 | 24 | 210.2 | 189.9 | 39 | 104.5 | 94.4 | 107.8 | 66 |

| ginocch | caviglia | bicipite | avanbr | polso | bmi |
|---|---|---|---|---|---|
| 37.3 | 21.9 | 32 | 27.4 | 17.1 | 23.63 |
| 37.3 | 23.4 | 30.5 | 28.9 | 18.2 | 23.33 |
| 38.9 | 24 | 28.8 | 25.2 | 16.6 | 24.67 |
| 37.3 | 22.8 | 32.4 | 29.4 | 18.2 | 24.88 |
| 42.2 | 24 | 32.2 | 27.7 | 17.7 | 25.52 |
| 42 | 25.6 | 35.7 | 30.6 | 18.8 | 26.46 |

## STATISTICHE DESCRITTIVE

Si presentano innanzitutto le statistiche descrittive

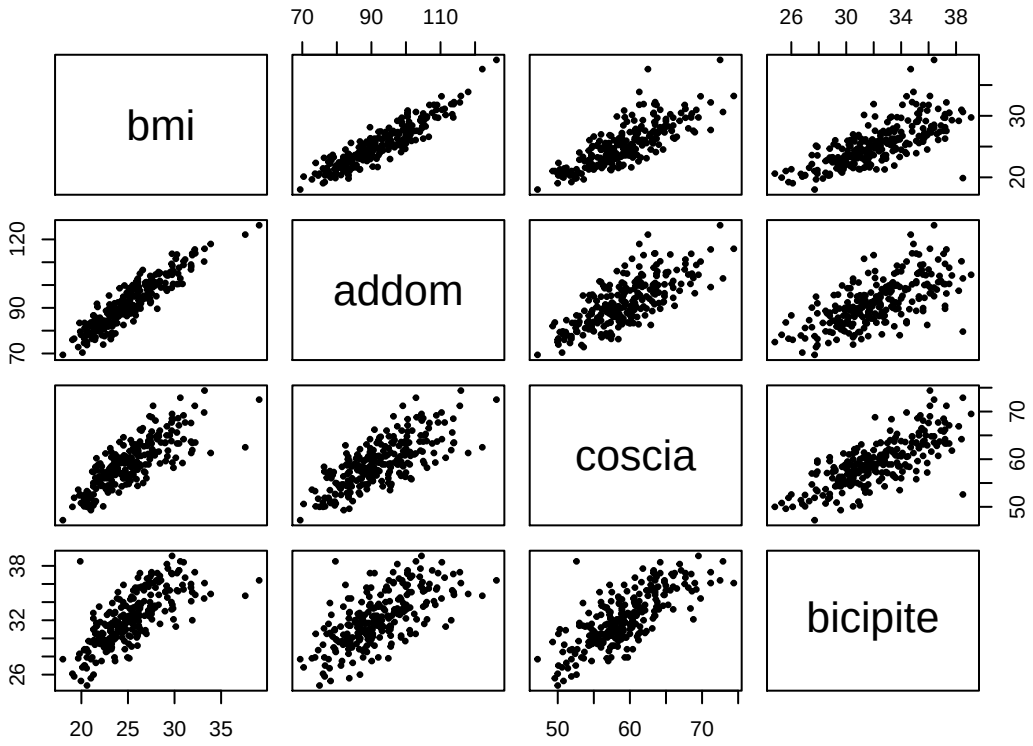
```r
#-- R CODE
pander(summary(d[,VAR_NUMERIC]),big.mark=",") #-- statistiche descrittive
```

| bmi | addom | coscia | bicipite |
|---|---|---|---|
| Min. :18.02 | Min. : 69.40 | Min. :47.20 | Min. :24.80 |
| 1st Qu.:23.03 | 1st Qu.: 84.47 | 1st Qu.:56.00 | 1st Qu.:30.20 |
| Median :25.02 | Median : 90.95 | Median :59.00 | Median :32.00 |
| Mean :25.30 | Mean : 92.31 | Mean :59.27 | Mean :32.22 |
| 3rd Qu.:27.31 | 3rd Qu.: 99.20 | 3rd Qu.:62.30 | 3rd Qu.:34.33 |
| Max. :39.08 | Max. :126.20 | Max. :74.40 | Max. :39.10 |

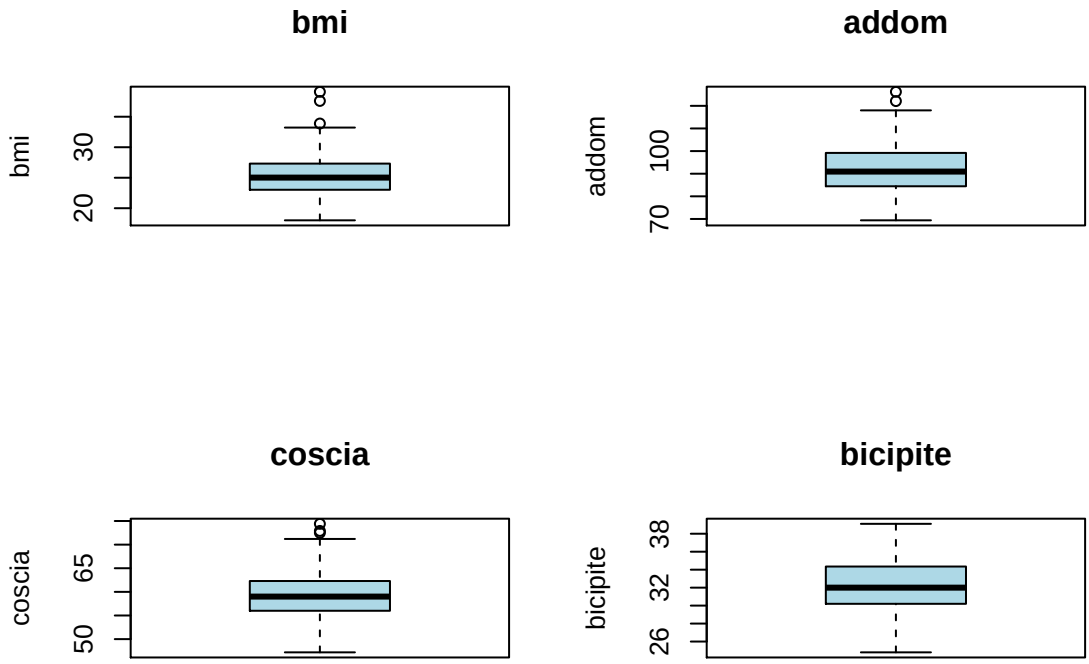
```r
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

| | bmi | addom | coscia | bicipite |
|---|---|---|---|---|
| **bmi** | 1 | 0.9142 | 0.7886 | 0.726 |
| **addom** | 0.9142 | 1 | 0.7373 | 0.6568 |
| **coscia** | 0.7886 | 0.7373 | 1 | 0.7459 |
| **bicipite** | 0.726 | 0.6568 | 0.7459 | 1 |

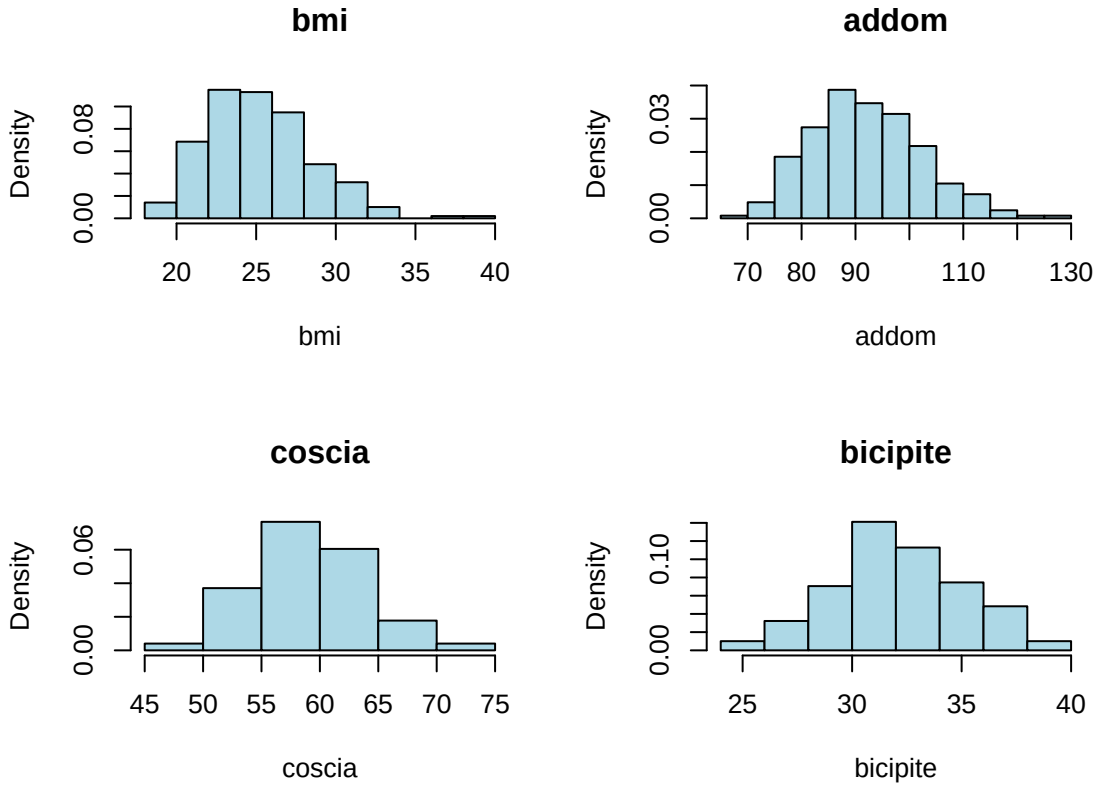
```r
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```r
par(mfrow=c(2,2))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
```




```
par(mfrow=c(2,2))
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```




## REGRESSIONE - Esempio 1

Si effettua ora la regressione multipla di "bmi" su "addom", "coscia" e "bicipite".

```
#-- R CODE
mod1 <- lm(bmi ~ addom + coscia + bicipite,d)

pander(summary(mod1),big.mark=",")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | -7.866 | 0.9545 | -8.241 | 1.051e-14 |
| **addom** | 0.2281 | 0.01139 | 20.02 | 2.105e-53 |
| **coscia** | 0.1152 | 0.02684 | 4.292 | 2.558e-05 |
| **bicipite** | 0.1639 | 0.04035 | 4.062 | 6.555e-05 |

Table 6: Fitting linear model: bmi ~ addom + coscia + bicipite

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 248 | 1.204 | 0.8731 | 0.8715 |

```
pander(anova(mod1),big.mark=",")
```

Table 7: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **addom** | 1 | 2,329 | 2,329 | 1,607 | 2.363e-109 |
| **coscia** | 1 | 80.07 | 80.07 | 55.26 | 1.783e-12 |
| **bicipite** | 1 | 23.91 | 23.91 | 16.5 | 6.555e-05 |
| **Residuals** | 244 | 353.6 | 1.449 | NA | NA |

```
pander(white.test(mod1),big.mark=",")
```

| Test.statistic | P.value |
|---|---|
| 34.82 | 2.749e-08 |

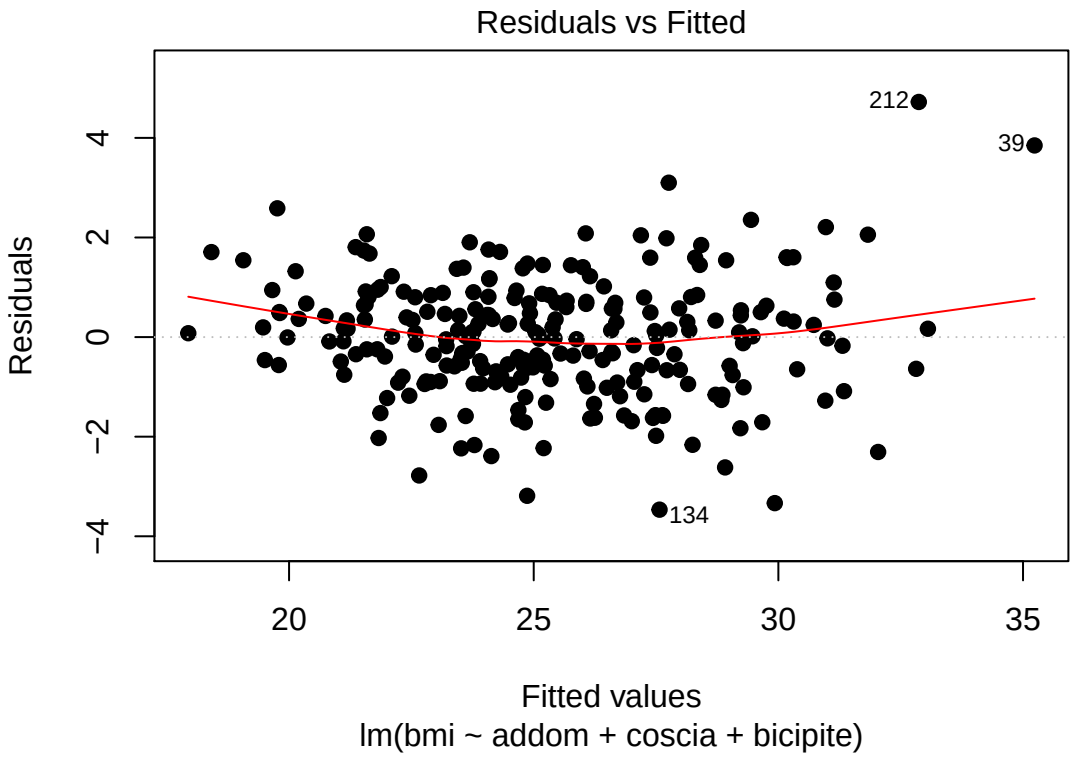
```
pander(dwtest(mod1),big.mark=",")
```

Table 9: Durbin-Watson test: `mod1`

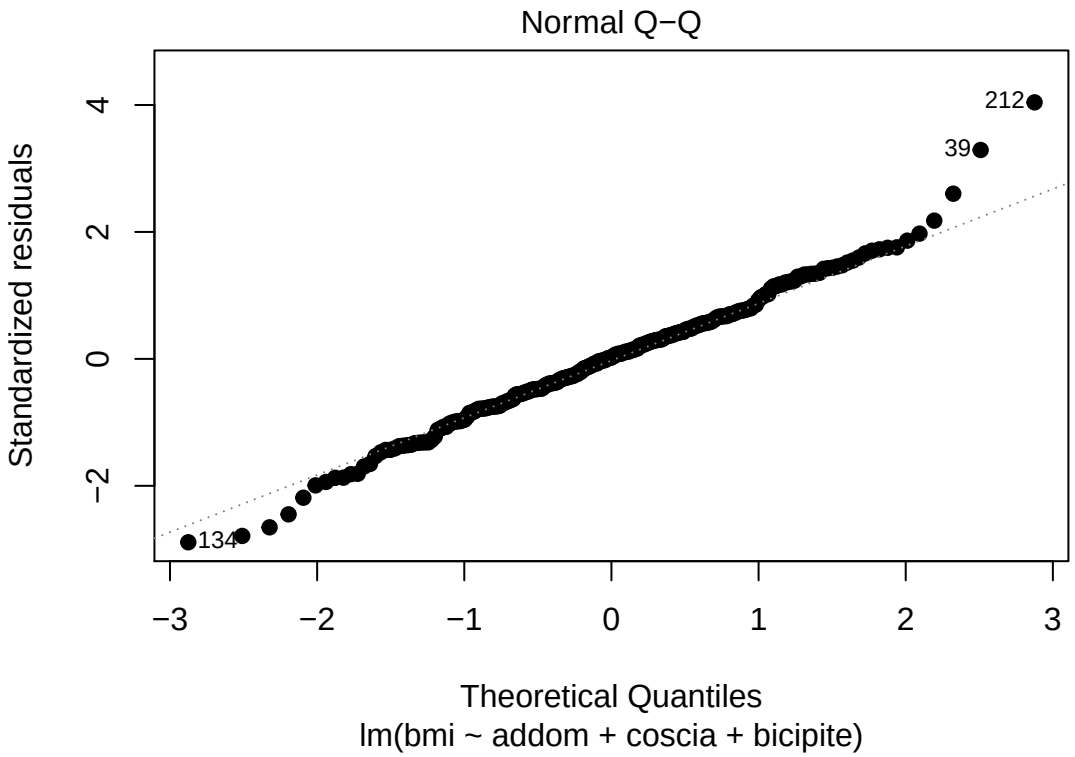
| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 1.84 | 0.09524 | true autocorrelation is greater than 0 |

Il modello risulta significativo e tutti i parametri risultano significativi. Si effettua ora una diagnostica sugli errori cominciando da rappresentazioni grafiche.

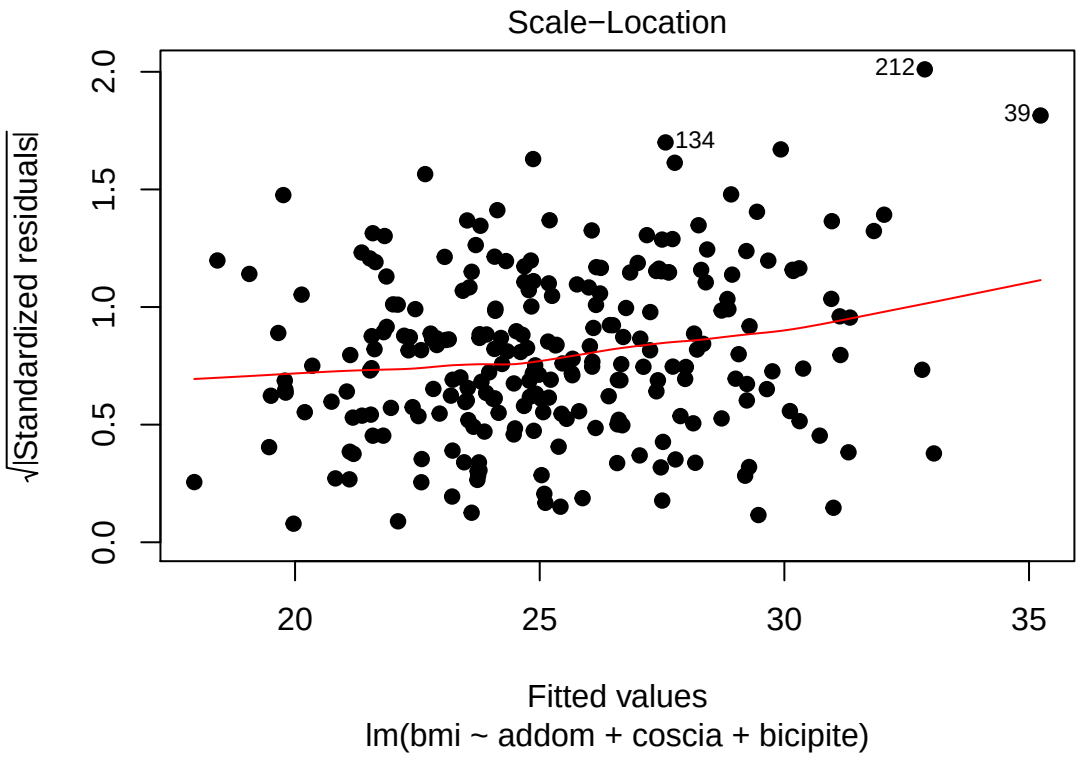
```
#-- R CODE
plot(mod1,which=1,pch=19)
```



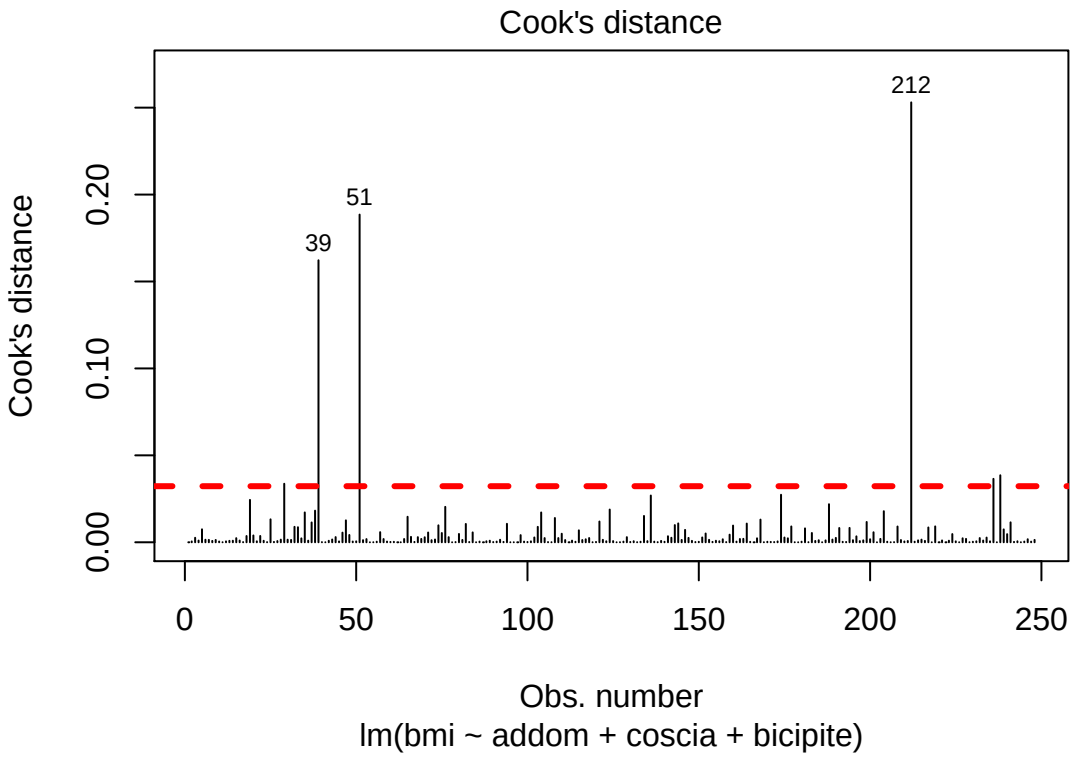

```
plot(mod1,which=2,pch=19)
```



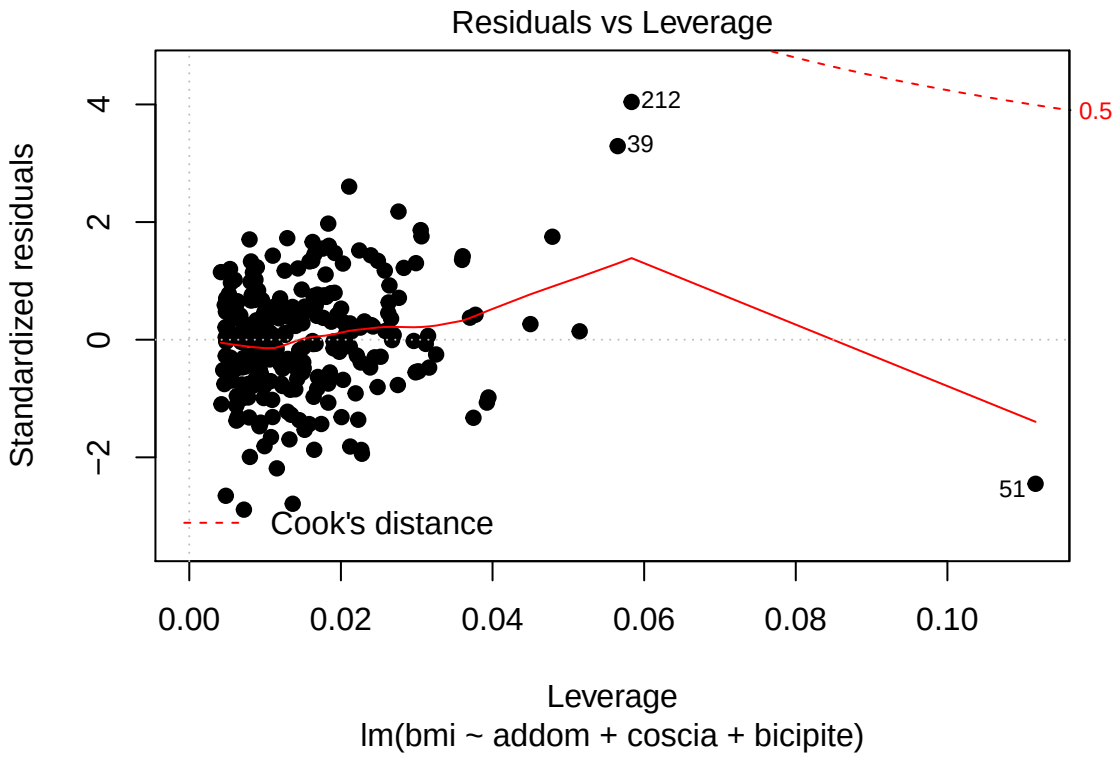

```
plot(mod1,which=3,pch=19)
```



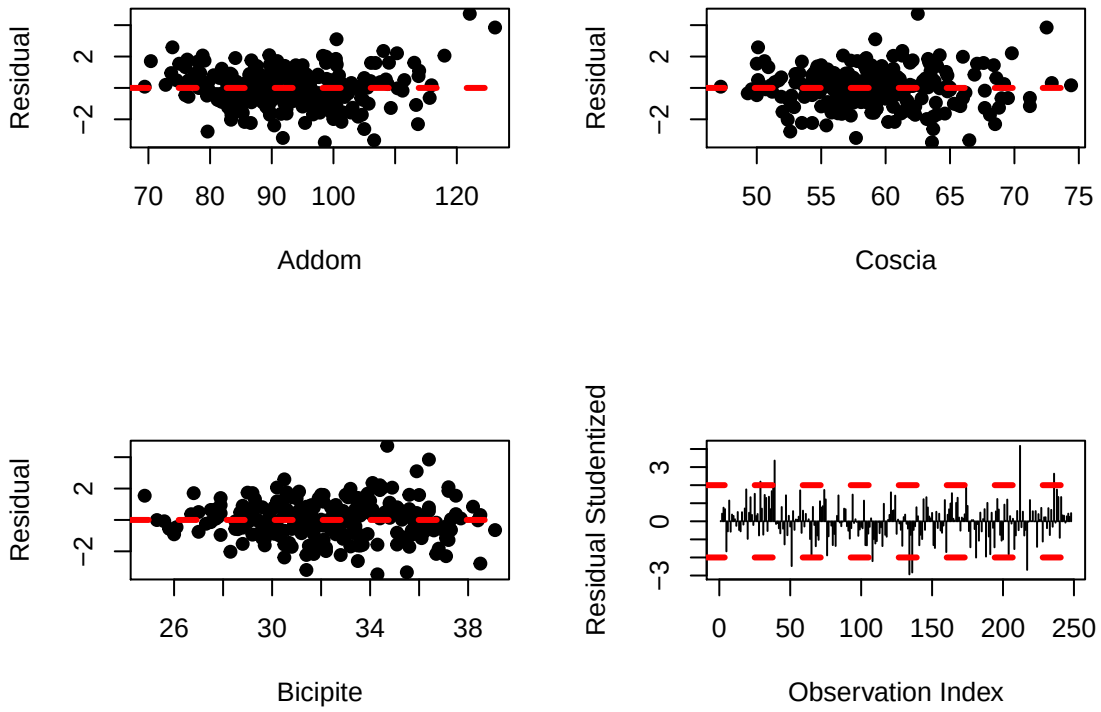

```
plot(mod1,which=4,pch=19)
abline(h=2*4/nrow(d),col=2,lwd=3,lty=2)
```




```
plot(mod1,which=5,pch=19)
```




```r
#-- R CODE
par(mfrow=c(2,2))
plot(d$addom,resid(mod1),pch=19,xlab="Addom",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d$coscia,resid(mod1),pch=19,xlab="Coscia",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d$bicipite,resid(mod1),pch=19,xlab="Bicipite",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(1:nrow(d),rstudent(mod1),pch=19,xlab="Observation Index",ylab="Residual Studentized",type="h")
abline(h=2,lwd=3,lty=2,col=2)
abline(h=-2,lwd=3,lty=2,col=2)
```

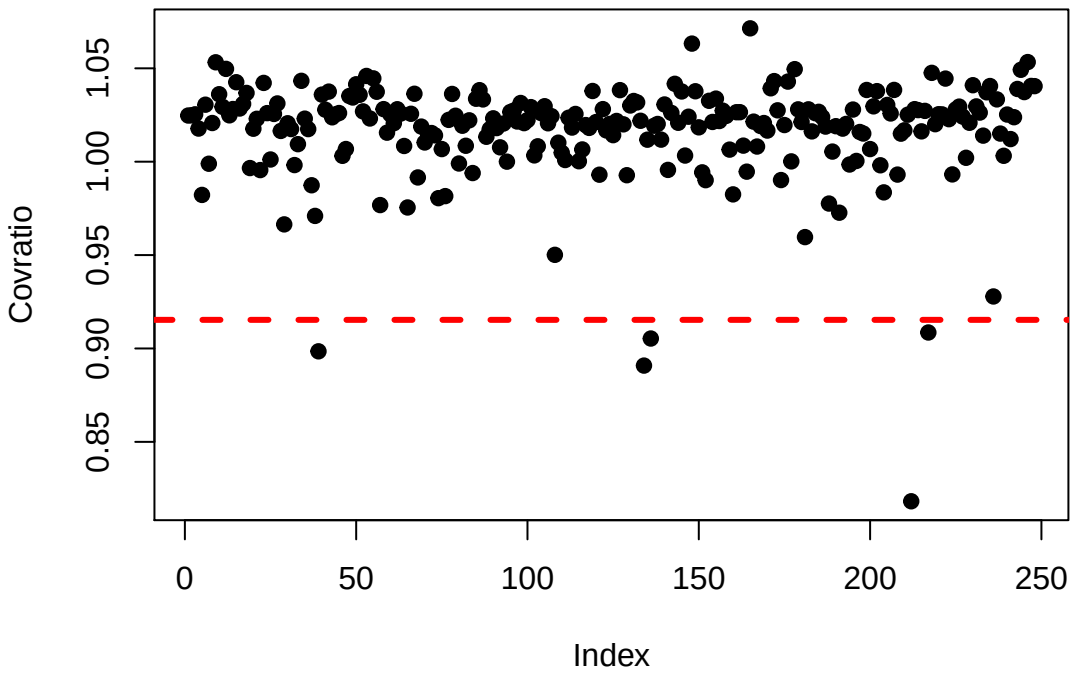



Nel Q-Q plot e dalla distribuzione teorica dei residui si vede una distribuzione normale poiché i quantili della distribuzione teorica normale si sovrappongono a quelli della distribuzione empirica. Tuttavia nella parte superiore del Q-Q plot si osserva la presenza di valori anomali in quanto i quantili della distribuzione teorica normale in questa parte non si sovrappongono a quelli della distribuzione empirica.

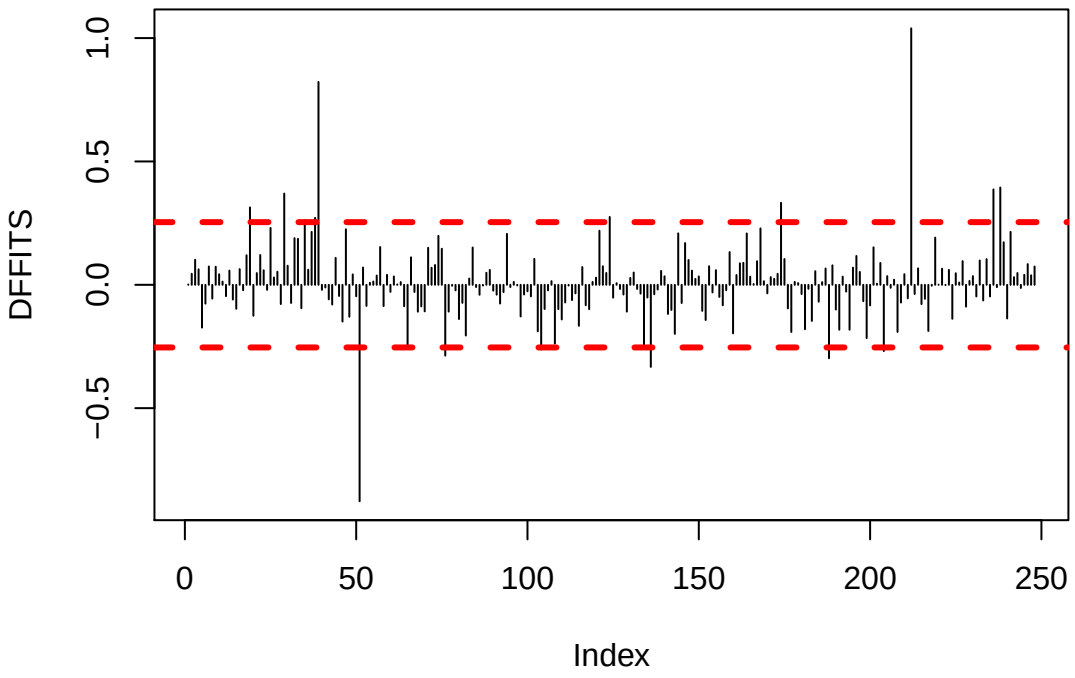
La presenza di valori anomali è confermata dal grafico residui-valori previsti.

Per verificare se effettivamente esistono tali outlier si considera nel grafico il grafico dei leverage e delle distanze di Cook che mostrano con chiarezza la presenza di valori al di fuori dei limiti di tolleranza. Per ulteriore verifica si considera quindi il grafico dei DFFITS.

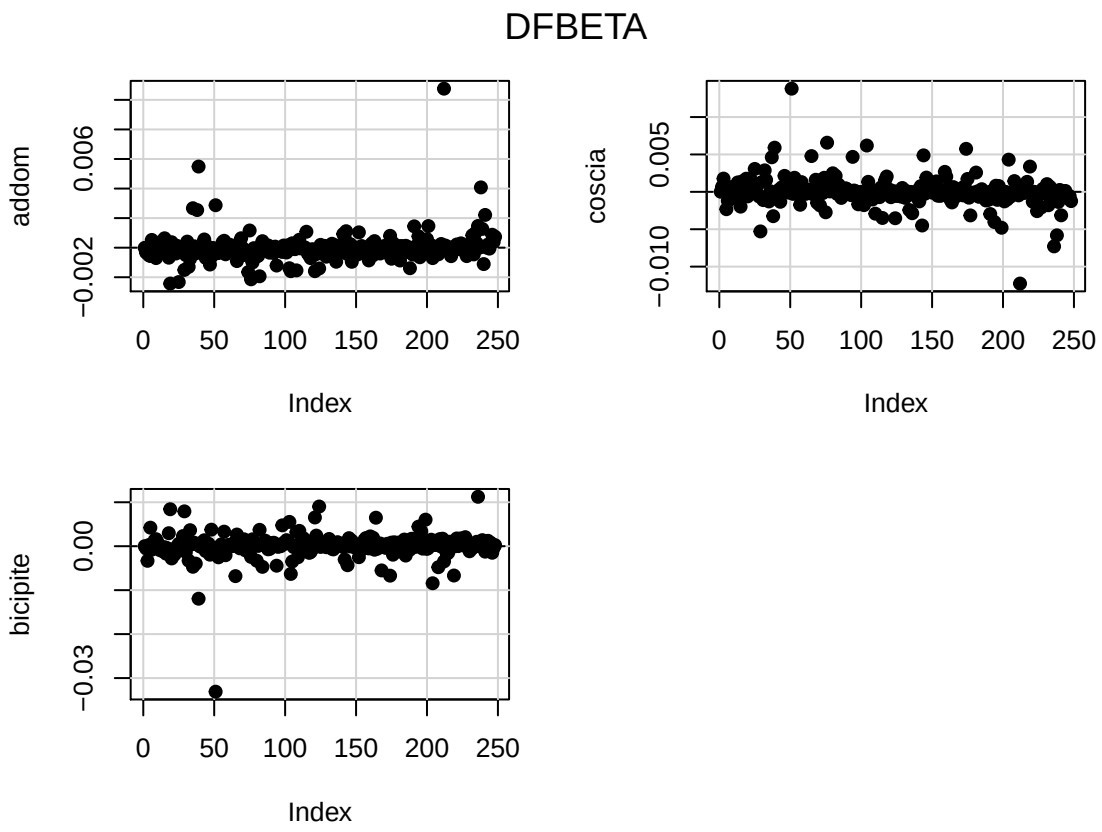
```r
#-- R CODE

plot(covratio(mod1),pch=19,ylab="Covratio")
abline(h=1-3*7/nrow(d),lwd=3,col=2,lty=2)
abline(h=1+3*7/nrow(d),lwd=3,col=2,lty=2)
```




```
plot(dffits(mod1),pch=19,ylab="DFFITS",type="h")
abline(h=2*sqrt(4/nrow(d)),lwd=3,col=2,lty=2)
abline(h=-2*sqrt(4/nrow(d)),lwd=3,col=2,lty=2)
```




```
dfbetaPlots(mod1,pch=19,main="DFBETA")
```




Anche in questo caso si vedono punti al di fuori delle soglie di tolleranza. Facendo riferimento a questa misura dei DFFITS si elencano i punti influenti o outlier: 19, 29, 35, 38, 39, 51, 76, 104, 124, 136, 174, 188, 204, 212, 236, 238.

```
#-- R CODE
d1 <- d[-c(19, 29, 35, 38, 39, 51, 76, 104, 124, 136, 174, 188, 204, 212, 236, 238),]
mod1 <- lm(bmi ~ addom + coscia + bicipite,d1)

pander(summary(mod1),big.mark=",")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | -7.215 | 0.8374 | -8.616 | 1.168e-15 |
| **addom** | 0.2042 | 0.01087 | 18.79 | 3.343e-48 |
| **coscia** | 0.1139 | 0.02486 | 4.583 | 7.544e-06 |
| **bicipite** | 0.2127 | 0.03824 | 5.562 | 7.443e-08 |

Table 11: Fitting linear model: bmi ~ addom + coscia + bicipite

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 232 | 1.01 | 0.8892 | 0.8878 |

```
pander(anova(mod1),big.mark=",")
```

Table 12: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **addom** | 1 | 1,748 | 1,748 | 1,714 | 5.06e-108 |
| **coscia** | 1 | 87.71 | 87.71 | 85.99 | 1.428e-17 |
| **bicipite** | 1 | 31.55 | 31.55 | 30.94 | 7.443e-08 |
| **Residuals** | 228 | 232.6 | 1.02 | NA | NA |

```
pander(white.test(mod1),big.mark=",")
```

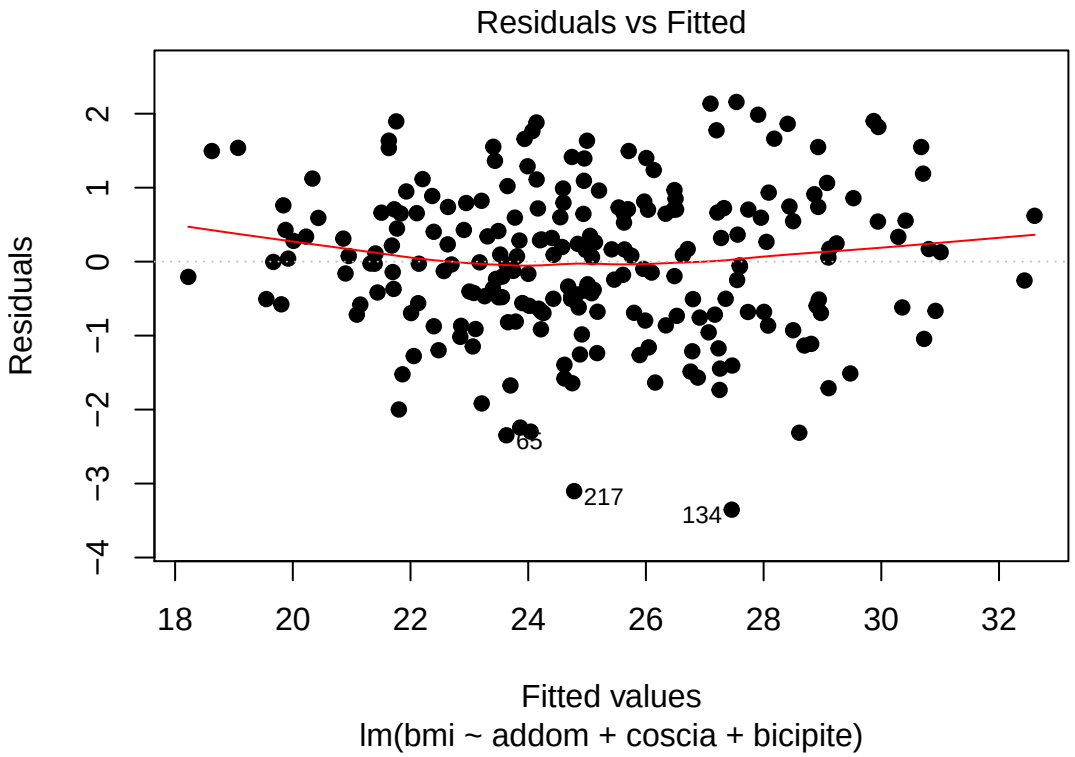
| Test.statistic | P.value |
|---|---|
| 3.984 | 0.1364 |

```
pander(dwtest(mod1),big.mark=",")
```

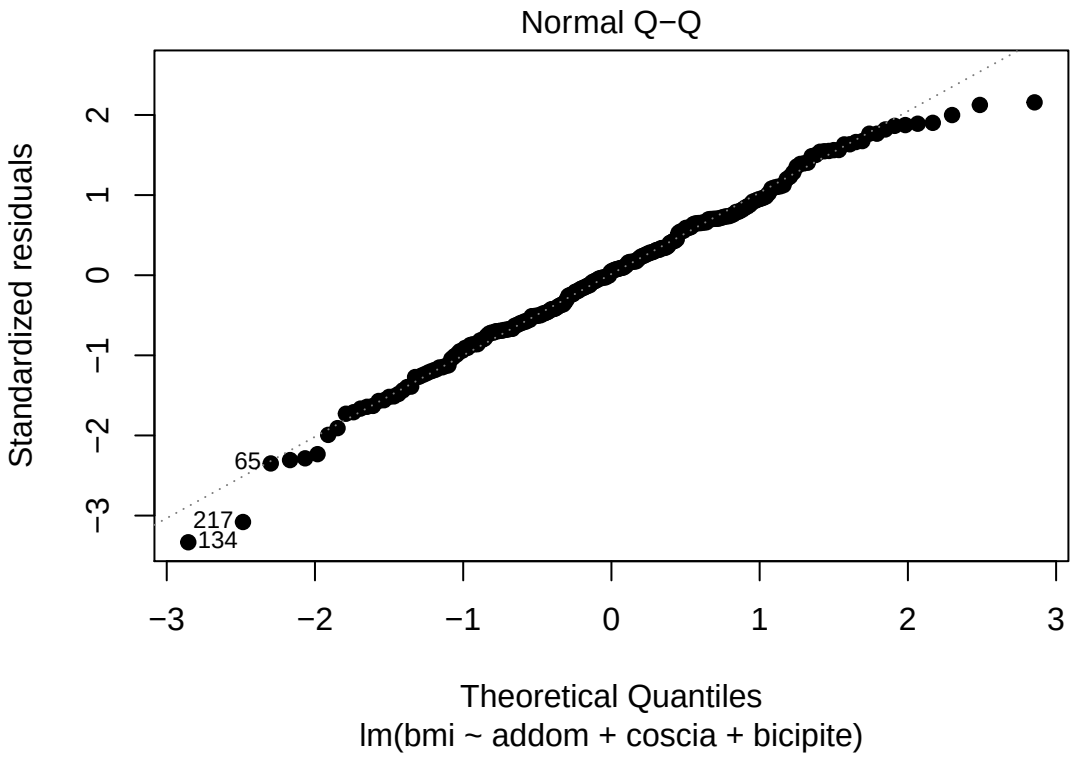
Table 14: Durbin-Watson test: `mod1`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 1.824 | 0.08216 | true autocorrelation is greater than 0 |

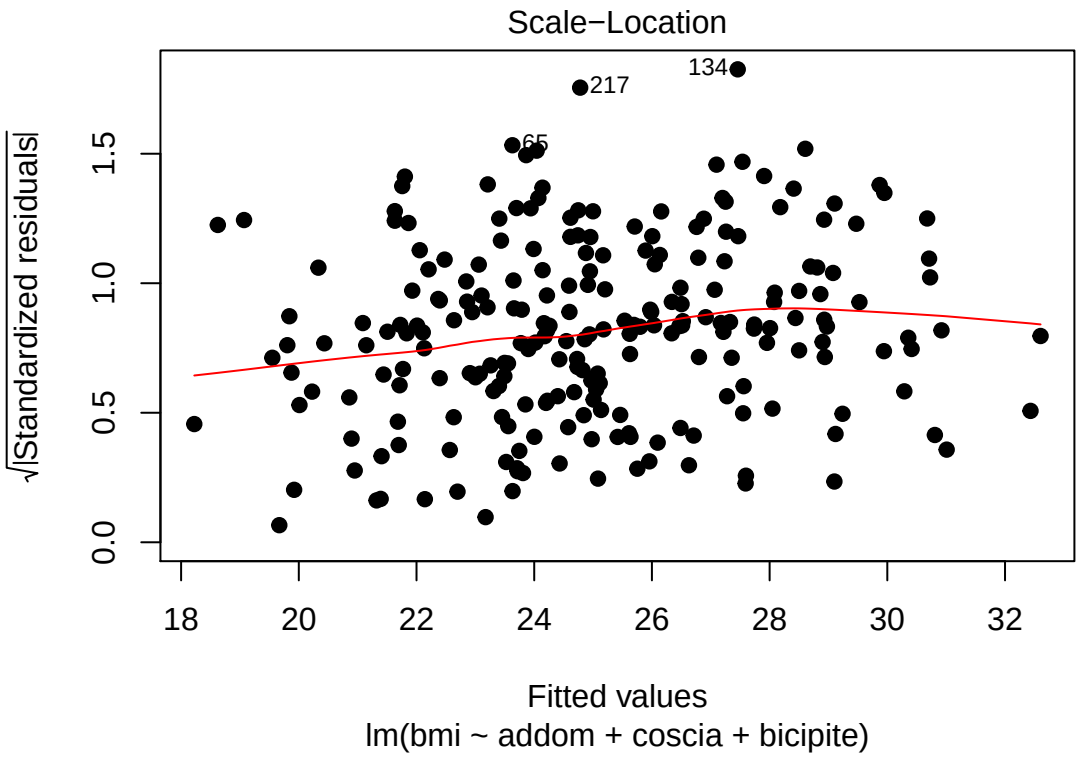
```
#-- R CODE
plot(mod1,which=1,pch=19)
```



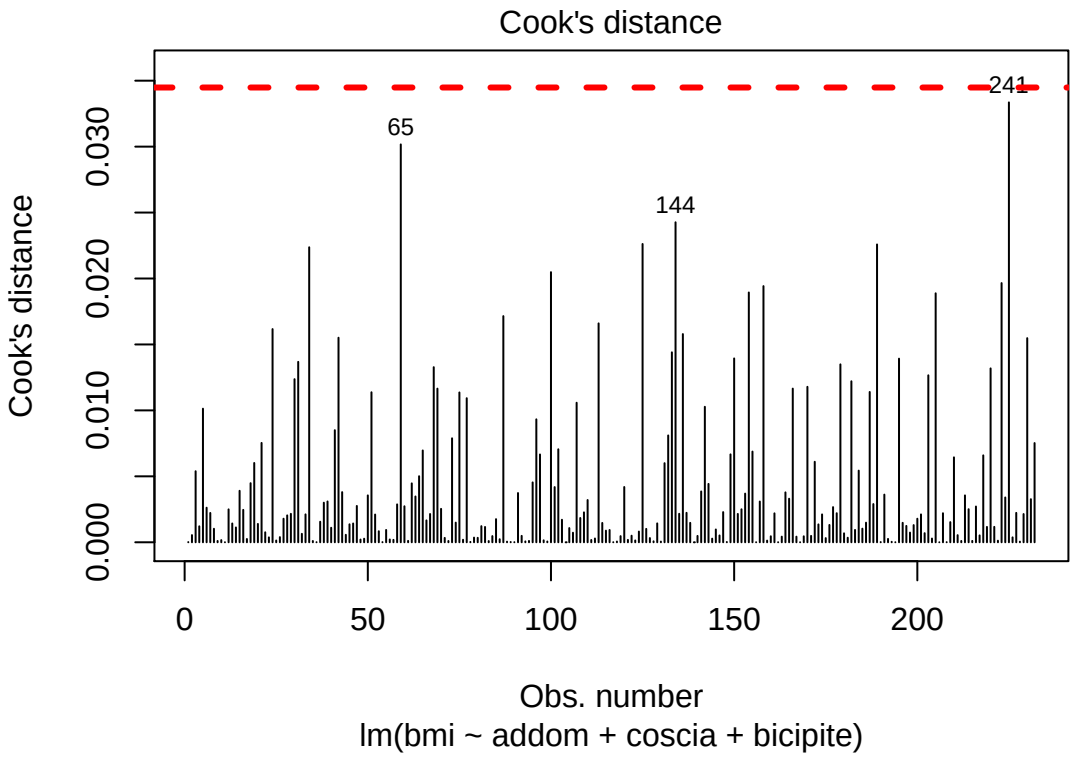

```
plot(mod1,which=2,pch=19)
```



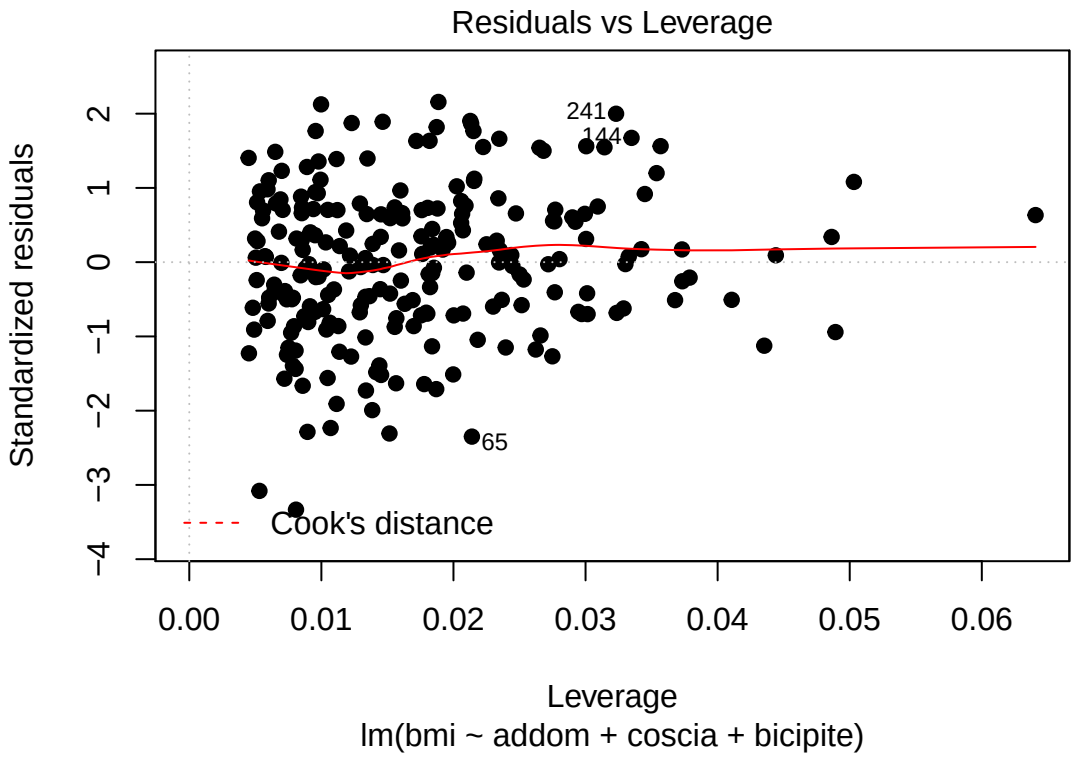

```
plot(mod1,which=3,pch=19)
```



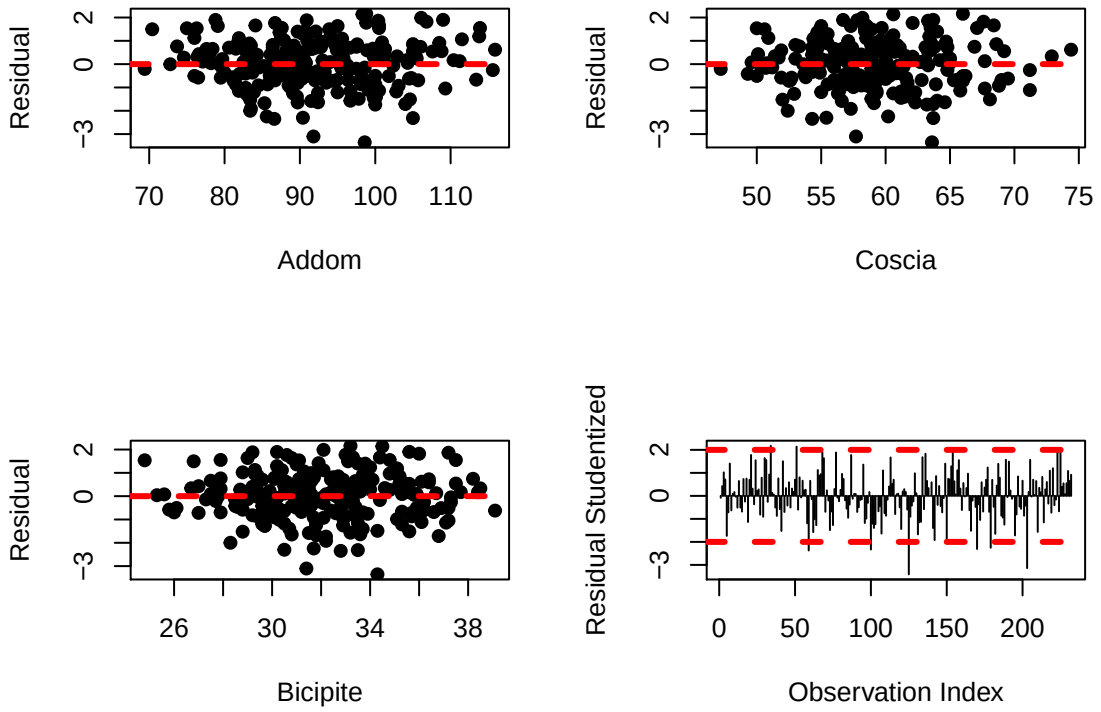

```
plot(mod1,which=4,pch=19)
abline(h=2*4/nrow(d1),col=2,lwd=3,lty=2)
```




```
plot(mod1,which=5,pch=19)
```

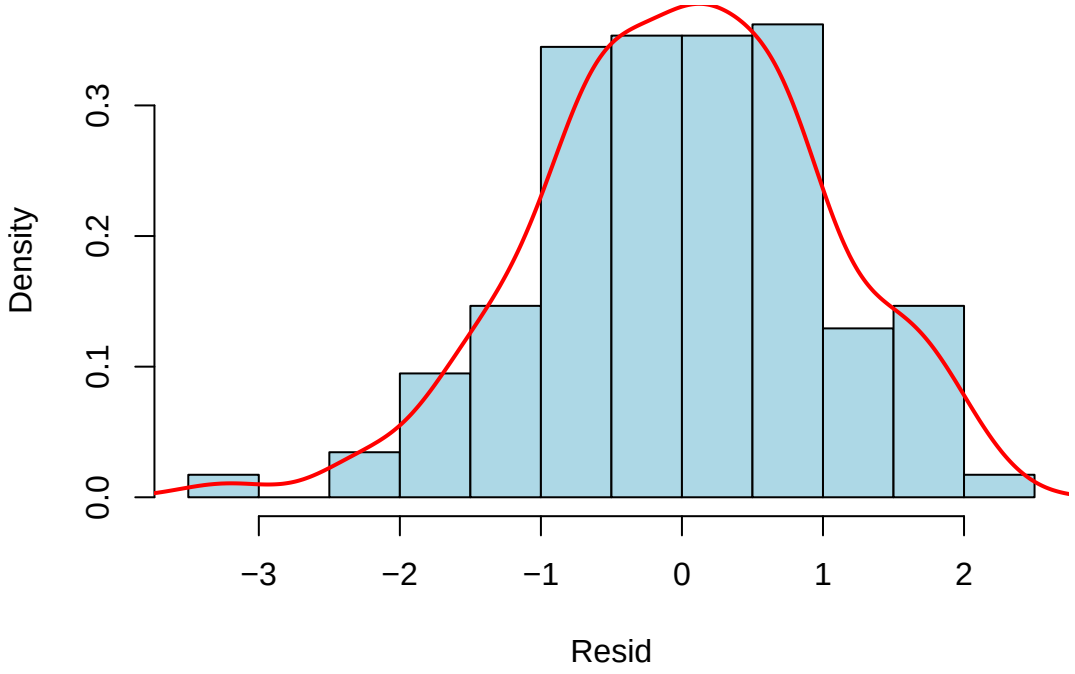



```r
#-- R CODE
par(mfrow=c(2,2))
plot(d1$addom,resid(mod1),pch=19,xlab="Addom",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d1$coscia,resid(mod1),pch=19,xlab="Coscia",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(d1$bicipite,resid(mod1),pch=19,xlab="Bicipite",ylab="Residual")
abline(h=0,lwd=3,lty=2,col=2)

plot(1:nrow(d1),rstudent(mod1),pch=19,xlab="Observation Index",ylab="Residual Studentized",type="h")
abline(h=2,lwd=3,lty=2,col=2)
abline(h=-2,lwd=3,lty=2,col=2)
```




Per verificare più precisamente la normalità dei residui si propongono i test sulla normalità dei medesimi. Tutti i test accettano l'ipotesi nulla di normalità dei residui.

```
#-- R CODE

hist(resid(mod1),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1)),col=2,lwd=2)
```




```
pander(shapiro.test(resid(mod1)))
```

Table 15: Shapiro-Wilk normality test: `resid(mod1)`

| Test statistic | P value |
|---|---|
| 0.9907 | 0.1472 |

```
pander(ks.test(resid(mod1),"pnorm"))
```

Table 16: One-sample Kolmogorov-Smirnov test: `resid(mod1)`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 0.03307 | 0.9615 | two-sided |

Si verifica l'omoschedasticità e l'incorrelazione dei residui. Già i grafici fanno intuire che i residui sono omoschedastici e l'incorrelati tuttavia si effettua comunque una verifica con gli appositi test.

```
#-- R CODE
pander(white.test(mod1),big.mark=",")
```

| Test.statistic | P.value |
|---|---|
| 3.984 | 0.1364 |

```
pander(dwtest(mod1),big.mark=",")
```

Table 18: Durbin-Watson test: `mod1`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 1.824 | 0.08216 | true autocorrelation is greater than 0 |

## REGRESSIONE - Esempio 2

In un secondo esempio si considerano 5 variabili "peso" (var.dip) e "eta", "altezza", "circonferenza toracica", "circonferenza addominale".

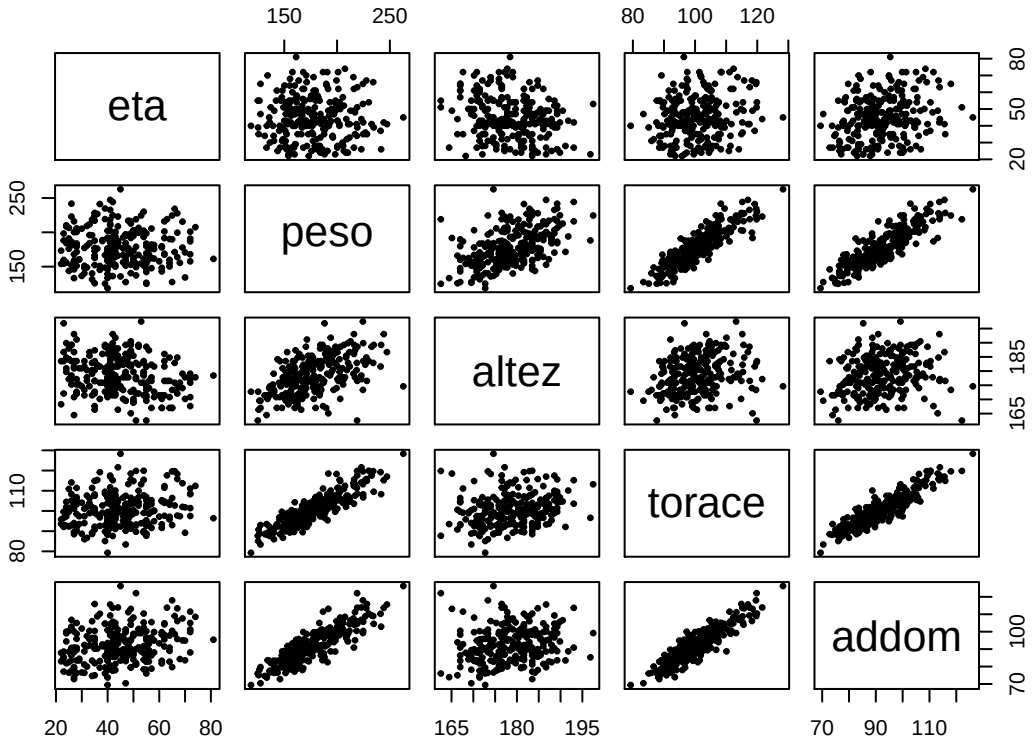
```
#-- R CODE
VAR_NUMERIC <- c("eta","peso","altez","torace","addom")
pander(summary(d[,VAR_NUMERIC]),big.mark=",") #-- statistiche descrittive
```

| eta | peso | altez | torace | addom |
|---|---|---|---|---|
| Min. :22.00 | Min. :118.5 | Min. :162.6 | Min. : 79.30 | Min. : 69.40 |
| 1st Qu.:35.75 | 1st Qu.:158.2 | 1st Qu.:173.4 | 1st Qu.: 94.15 | 1st Qu.: 84.47 |
| Median :43.00 | Median :176.1 | Median :177.8 | Median : 99.60 | Median : 90.95 |
| Mean :44.85 | Mean :178.1 | Mean :178.6 | Mean :100.67 | Mean : 92.31 |
| 3rd Qu.:54.00 | 3rd Qu.:196.8 | 3rd Qu.:183.5 | 3rd Qu.:105.30 | 3rd Qu.: 99.20 |
| Max. :81.00 | Max. :262.8 | Max. :197.5 | Max. :128.30 | Max. :126.20 |

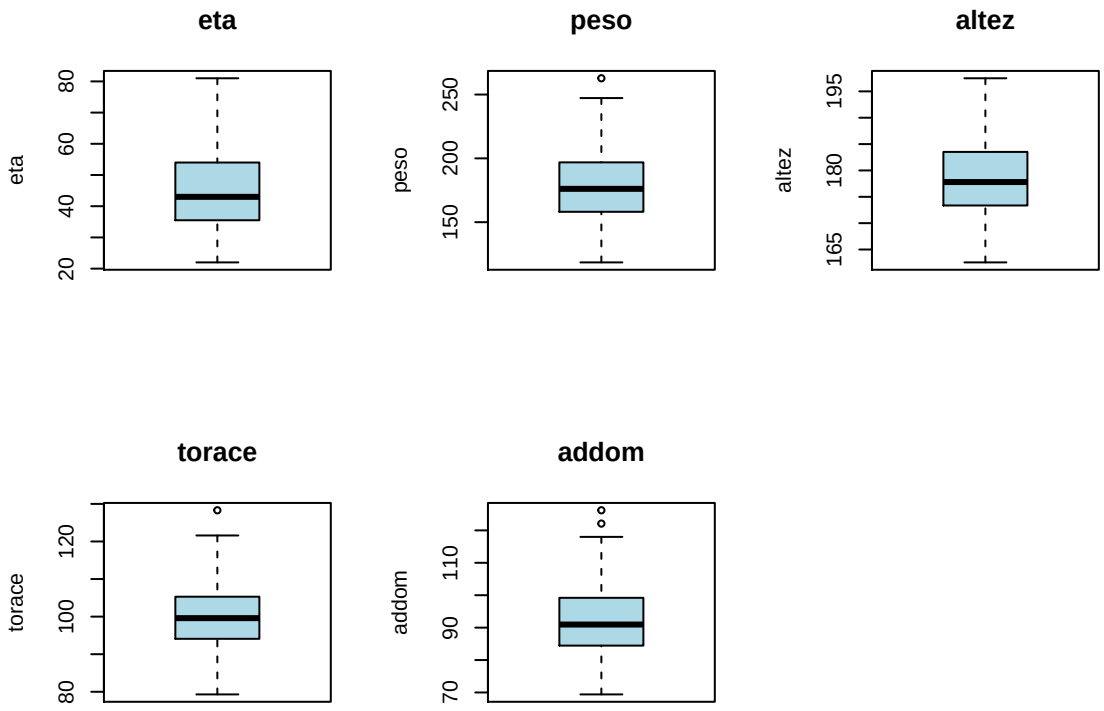
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

| | eta | peso | altez | torace | addom |
|---|---|---|---|---|---|
| **eta** | 1 | -0.01269 | -0.2363 | 0.1848 | 0.2452 |
| **peso** | -0.01269 | 1 | 0.5136 | 0.8914 | 0.8742 |
| **altez** | -0.2363 | 0.5136 | 1 | 0.2241 | 0.1886 |
| **torace** | 0.1848 | 0.8914 | 0.2241 | 1 | 0.9103 |
| **addom** | 0.2452 | 0.8742 | 0.1886 | 0.9103 | 1 |

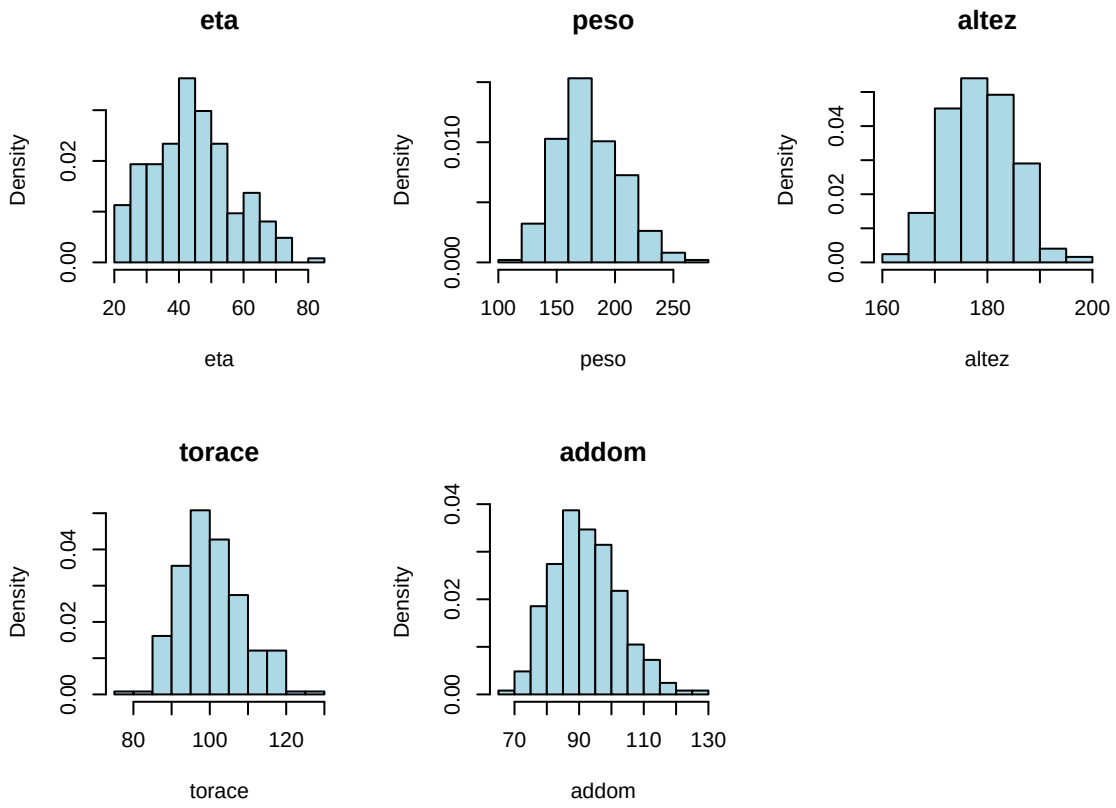
```
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```
par(mfrow=c(2,3))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
par(mfrow=c(2,3))
```




```
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```




Si osserva la forte correlazione esistente tra circonferenza torace e addominale. Si regrediscono ora le 4 variabili esplicative sulla variabile dipendente.

```
#-- R CODE
mod1 <- lm(peso ~ eta + altez + torace + addom, d)

pander(summary(mod1),big.mark=",")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | -283.5 | 12.62 | -22.47 | 2.998e-61 |
| **eta** | -0.2886 | 0.03648 | -7.912 | 8.958e-14 |
| **altez** | 1.221 | 0.06864 | 17.79 | 6.761e-46 |
| **torace** | 1.447 | 0.1266 | 11.43 | 1.678e-24 |
| **addom** | 1.201 | 0.102 | 11.77 | 1.273e-25 |

Table 22: Fitting linear model: peso ~ eta + altez + torace + addom

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 248 | 6.674 | 0.9405 | 0.9395 |

```
pander(anova(mod1),big.mark=",")
```

Table 23: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **eta** | 1 | 29.29 | 29.29 | 0.6575 | 0.4182 |
| **altez** | 1 | 50,209 | 50,209 | 1,127 | 3.075e-93 |
| **torace** | 1 | 114,620 | 114,620 | 2,573 | 2.823e-131 |
| **addom** | 1 | 6,176 | 6,176 | 138.6 | 1.273e-25 |
| **Residuals** | 243 | 10,825 | 44.55 | NA | NA |

```
pander(white.test(mod1),big.mark=",")
```

| Test.statistic | P.value |
|---|---|
| 2.282 | 0.3195 |

```
pander(dwtest(mod1),big.mark=",")
```

Table 25: Durbin-Watson test: `mod1`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 2.044 | 0.6072 | true autocorrelation is greater than 0 |

E' respinta dal test F l'ipotesi nulla che il modello nel suo complesso non spieghi la variabile dipendente. Inoltre il modello ha un ottimo fitting (0.9405) e tutte le variabili risultano significative. Per quanto riguarda la collinearità vediamo che l'indice di tolleranza riporta valori bassi per le variabili circonferenza addominale e toracica; tuttavia la variance inflation si mantiene ampiamente sotto la soglia di 20.
Per prendere una decisione analizziamo le altre diagnostiche di collinearità.

```
#-- R CODE
pander(ols_eigen_cindex(mod1),big.mark=",")
```

Table 26: Table continues below

| Eigenvalue | Condition Index | intercept | eta | altez | torace |
|---|---|---|---|---|---|
| 4.932 | 1 | 4.614e-05 | 0.002439 | 4.877e-05 | 4.499e-05 |
| 0.05751 | 9.261 | 0.0009619 | 0.8689 | 0.001506 | 0.0008409 |
| 0.008704 | 23.81 | 0.03046 | 0.0178 | 0.02893 | 0.01328 |
| 0.0007542 | 80.87 | 0.04577 | 0.001988 | 0.2561 | 0.7989 |
| 0.0005915 | 91.32 | 0.9228 | 0.1089 | 0.7134 | 0.1869 |

| addom |
|---|
| 8.182e-05 |
| 0.001091 |
| 0.1027 |

| addom |
|---|
| 0.6886 |
| 0.2075 |

```
pander(ols_vif_tol(mod1),big.mark=",")
```

| Variables | Tolerance | VIF |
|---|---|---|
| eta | 0.8539 | 1.171 |
| altez | 0.8697 | 1.15 |
| torace | 0.168 | 5.953 |
| addom | 0.1652 | 6.052 |

Gli ultimi due autovalori hanno valori molto piccoli e condition index ampiamente oltre la soglia di 30. Il 4 autovalore spiega le varibili "circonf", "addom" e "torace" rispettivamente per il 68.8% e 79.8%. Se ne deduce che i due regressori sono correlati (è confermato il suggerimento dato dall'indice di tolleranza).

Si ristima quindi modello lineare escludendo la variabile esplicativa circonferenza toracica.

```
#-- R CODE
mod1 <- lm(peso ~ eta + altez + addom, d)

pander(summary(mod1),big.mark=",")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | -248.4 | 15.15 | -16.4 | 2.968e-41 |
| **eta** | -0.3139 | 0.04506 | -6.967 | 2.995e-11 |
| **altez** | 1.304 | 0.08447 | 15.44 | 5.719e-38 |
| **addom** | 2.251 | 0.05483 | 41.05 | 1.525e-111 |

Table 30: Fitting linear model: peso ~ eta + altez + addom

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 248 | 8.26 | 0.9085 | 0.9073 |

```
pander(anova(mod1),big.mark=",")
```

Table 31: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **eta** | 1 | 29.29 | 29.29 | 0.4293 | 0.5129 |
| **altez** | 1 | 50,209 | 50,209 | 735.9 | 1.276e-75 |
| **addom** | 1 | 114,974 | 114,974 | 1,685 | 1.525e-111 |
| **Residuals** | 244 | 16,647 | 68.23 | NA | NA |

```
pander(white.test(mod1),big.mark=",")
```

| Test.statistic | P.value |
|---|---|
| 1.896 | 0.3875 |

```
pander(dwtest(mod1),big.mark=",")
```

Table 33: Durbin-Watson test: `mod1`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 1.84 | 0.0901 | true autocorrelation is greater than 0 |

```
#-- R CODE
pander(ols_eigen_cindex(mod1),big.mark=",")
```

| Eigenvalue | Condition Index | intercept | eta | altez | addom |
|---|---|---|---|---|---|
| 3.937 | 1 | 7.688e-05 | 0.003864 | 7.73e-05 | 0.0006788 |
| 0.05444 | 8.504 | 0.001834 | 0.8623 | 0.00253 | 0.009936 |
| 0.007582 | 22.79 | 0.02601 | 0.03077 | 0.02159 | 0.9741 |
| 0.0006166 | 79.91 | 0.9721 | 0.1031 | 0.9758 | 0.01524 |

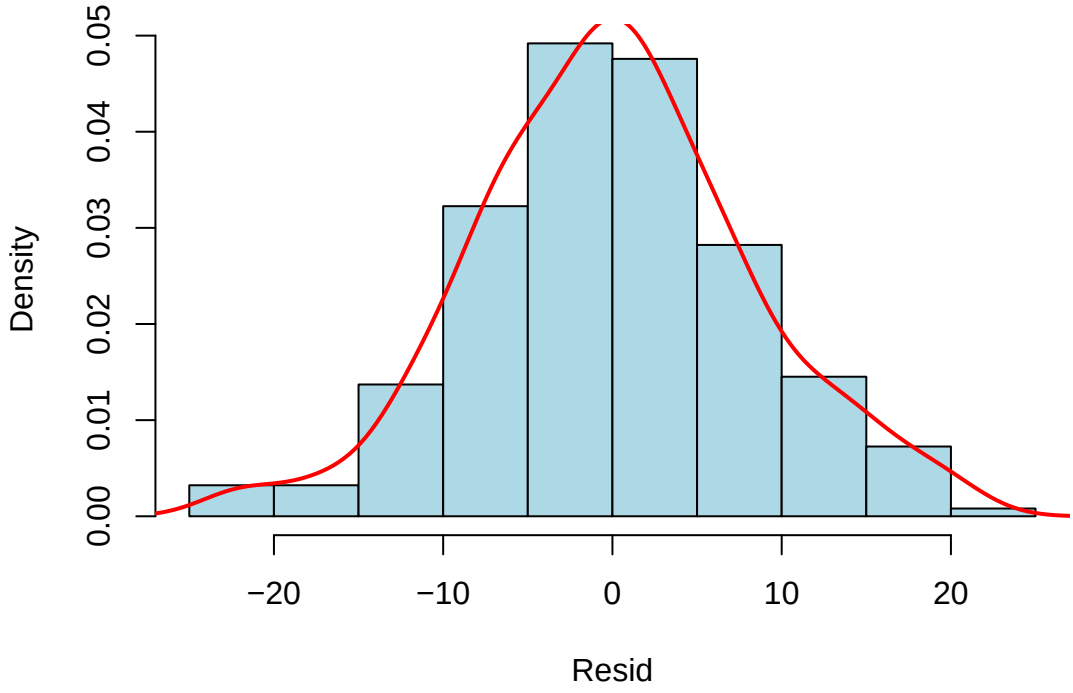
```
pander(ols_vif_tol(mod1),big.mark=",")
```

| Variables | Tolerance | VIF |
|---|---|---|
| eta | 0.8571 | 1.167 |
| altez | 0.8795 | 1.137 |
| addom | 0.8755 | 1.142 |

Il modello è ancora significativo con un elevatissimo $R^2$ (0.9085) e tutte le variabili risultano ancora significative. In questo caso però l'indice di tolleranza ha sempre valori prossimi a 1 e la variance inflation è molto bassa. Il condition index è elevato per gli ultimi 2 valori ma vediamo come ogni autovalore spieghi una proporzione di varianza elevata per variabili esplicative diverse (il 2° età, il 3° circonf. addom, il 4° altezza). Il problema di col inearità è quindi risolto.

Affrontiamo ora il problema della normalità degli errori. Si consideri innanzitutto la distribuzione dei residui

```
#-- R CODE
hist(resid(mod1),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1)),col=2,lwd=2)
```




```
pander(shapiro.test(resid(mod1)))
```

Table 36: Shapiro-Wilk normality test: `resid(mod1)`

| Test statistic | P value |
|---|---|
| 0.9942 | 0.4596 |

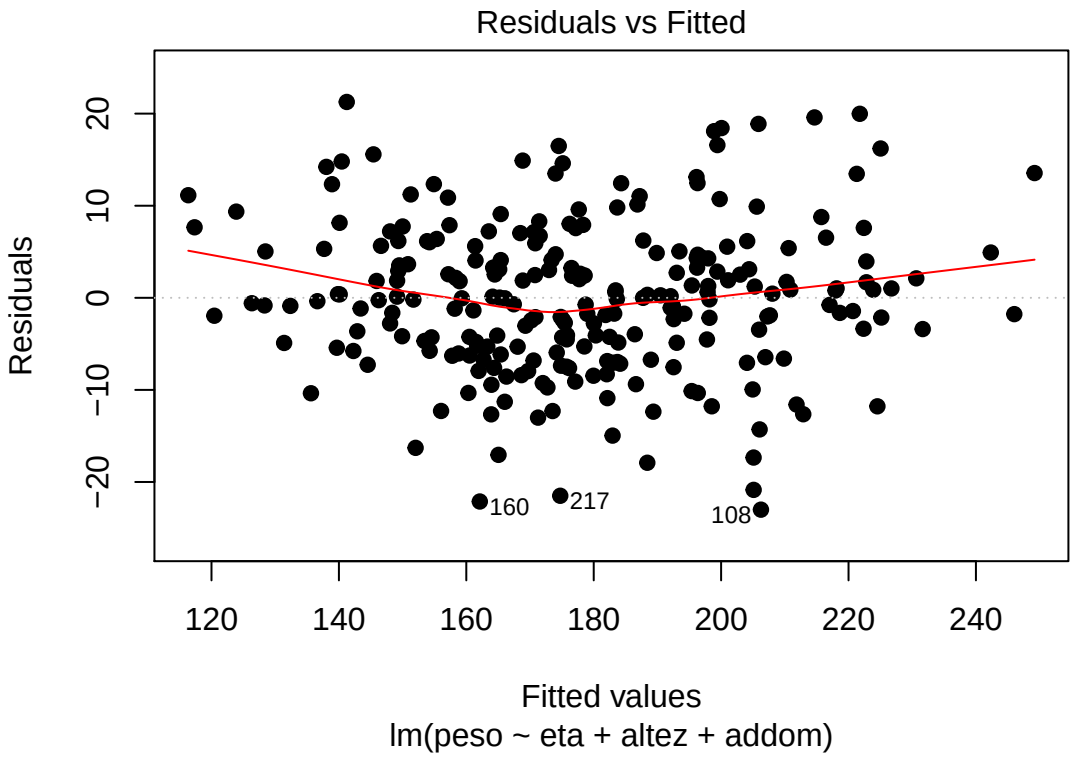
```
pander(ks.test(resid(mod1),"pnorm"))
```

Table 37: One-sample Kolmogorov-Smirnov test: `resid(mod1)`

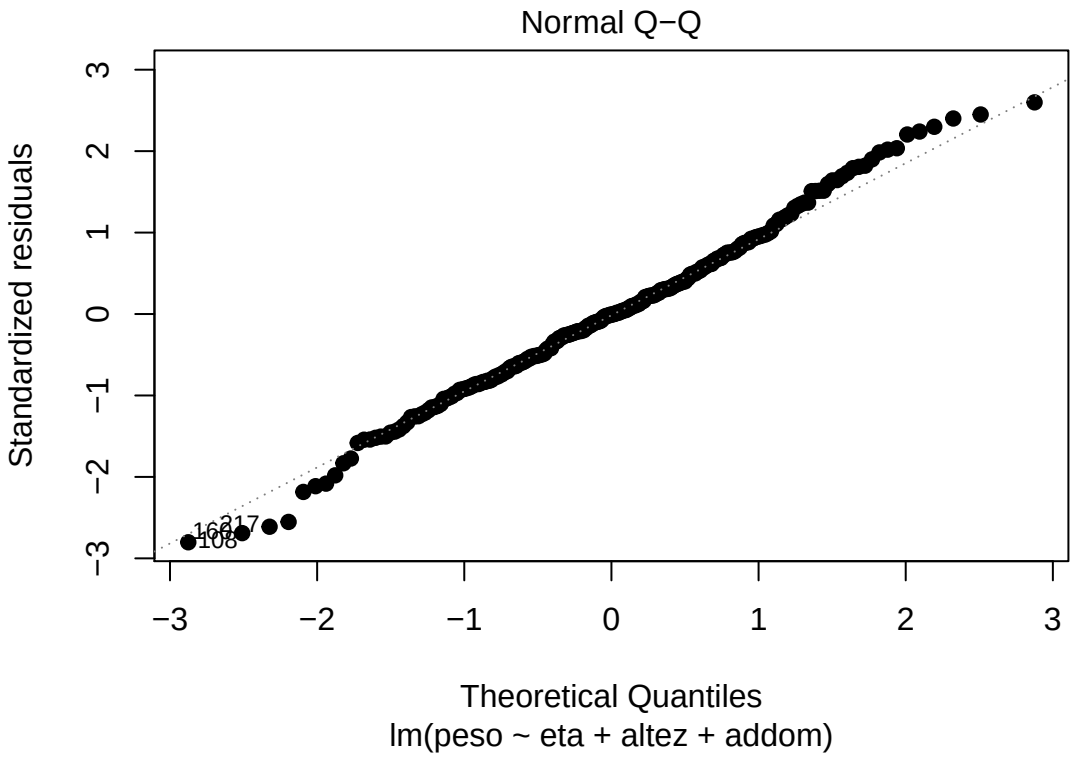
| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 0.3765 | 0 * * * | two-sided |

Possiamo accettare l'ipotesi di normalità dei residui.

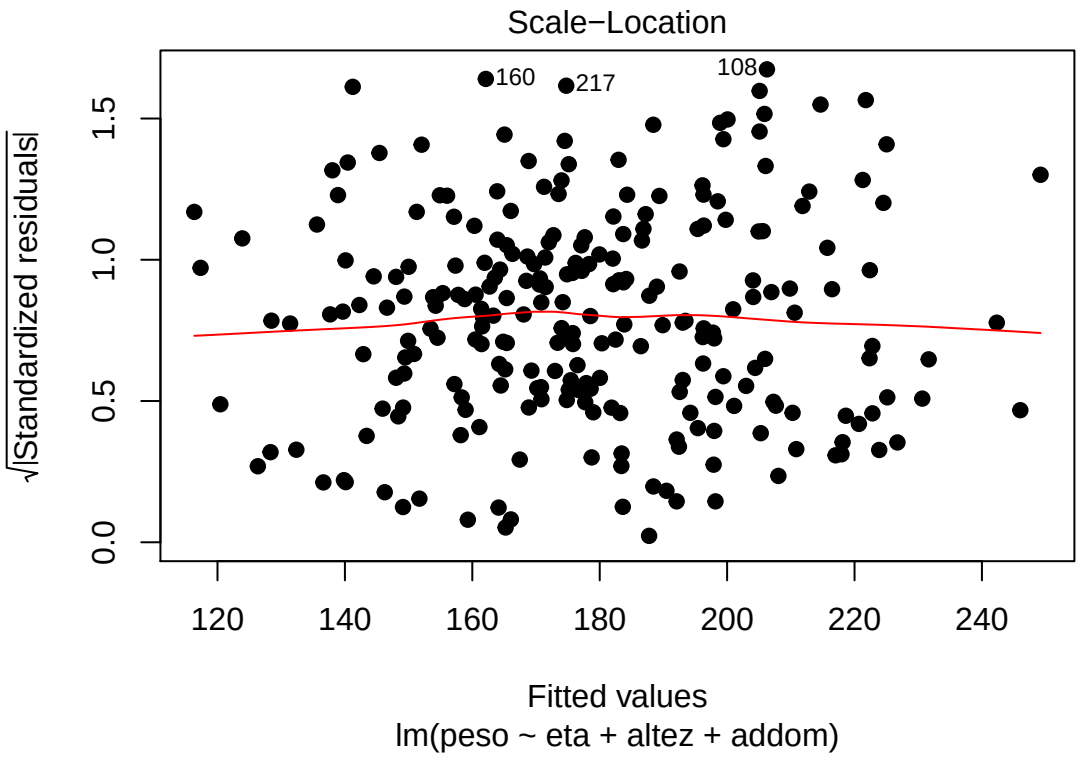
```
#-- R CODE
plot(mod1,which=1,pch=19)
```



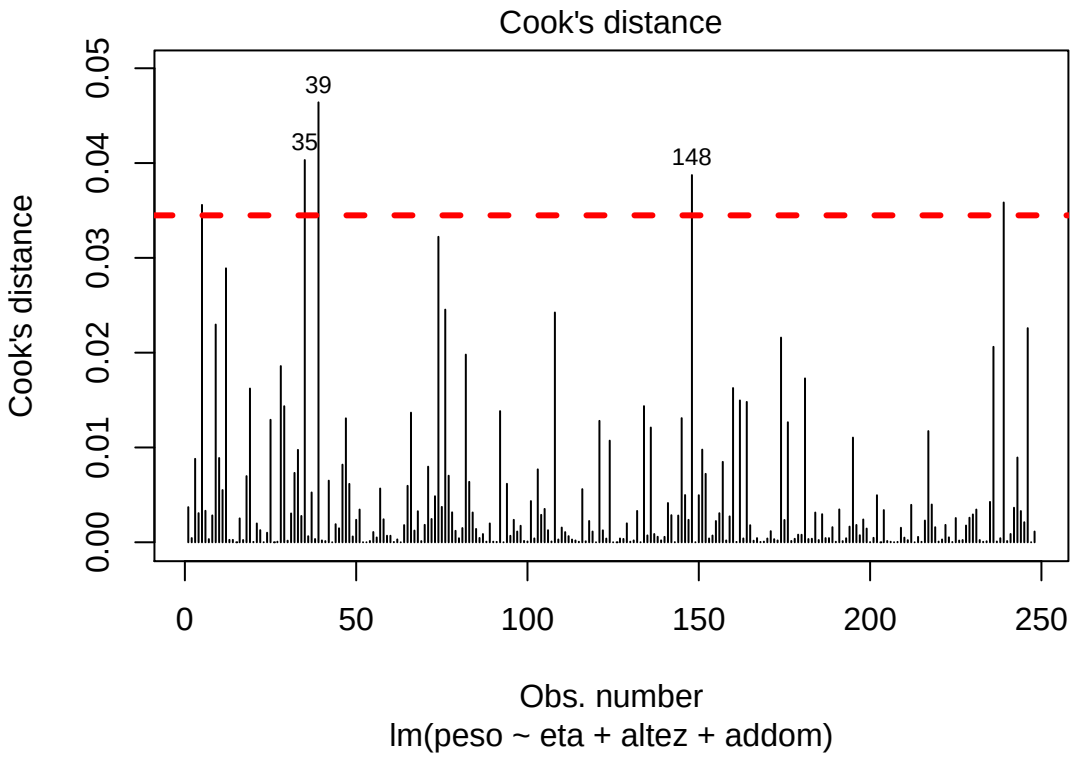

```
plot(mod1,which=2,pch=19)
```



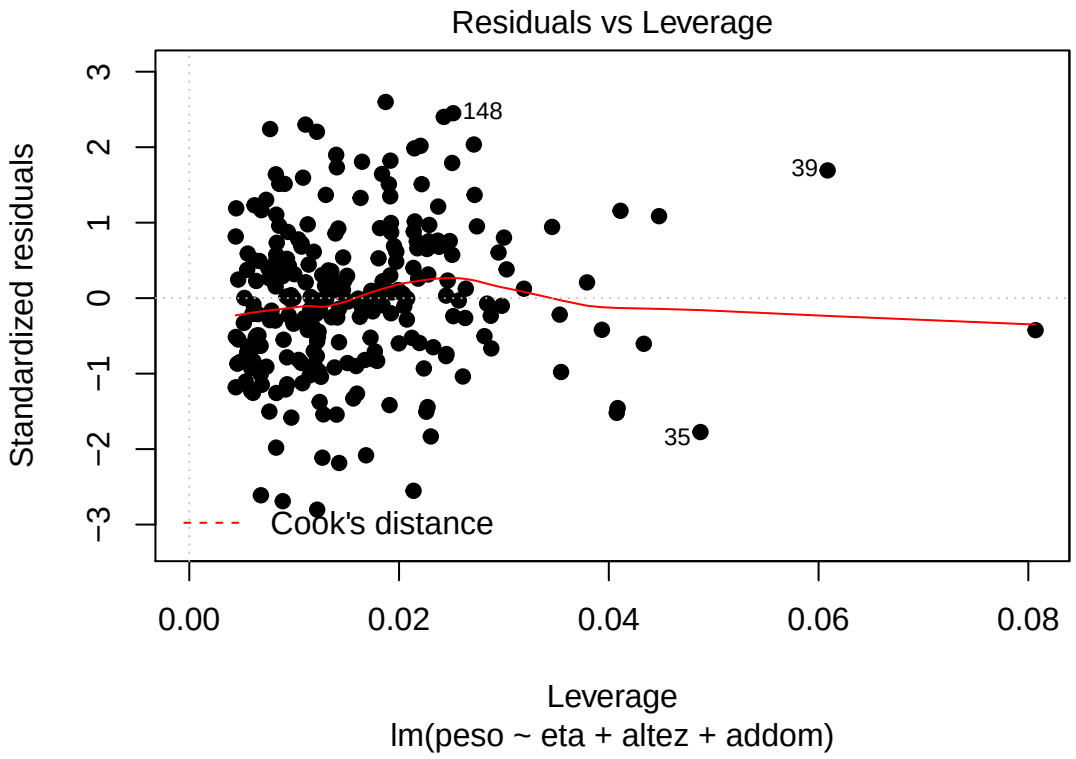

```
plot(mod1,which=3,pch=19)
```



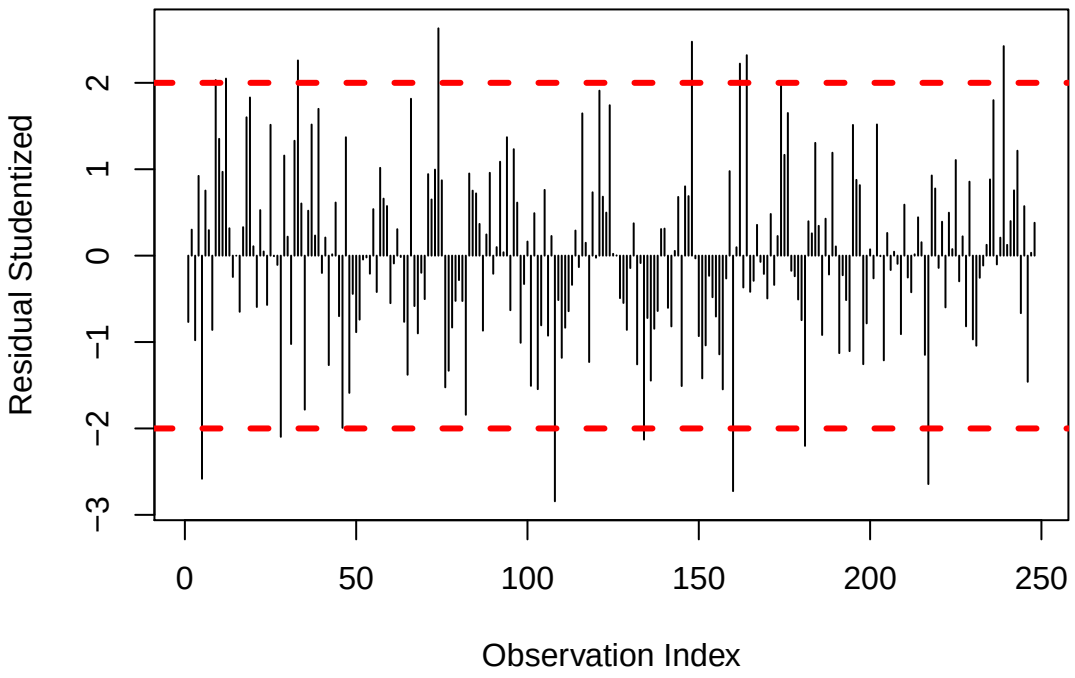

```
plot(mod1,which=4,pch=19)
abline(h=2*4/nrow(d1),col=2,lwd=3,lty=2)
```




```
plot(mod1,which=5,pch=19)
```




```r
#-- R CODE
plot(1:nrow(d),rstudent(mod1),pch=19,xlab="Observation Index",ylab="Residual Studentized",type="h")
abline(h=2,lwd=3,lty=2,col=2)
abline(h=-2,lwd=3,lty=2,col=2)
```




Si osserva tuttavia che vi sono osservazioni influenti come anche dai grafici dei residui studentizzati, della distanza di Cook e dei quantili dei residui.

## REGRESSIONE - Esempio 3

Si propone un terzo esempio in cui la variabile dipendente è il "peso" e le variabili esplicative sono "circonferenza torace", "collo", "addome", "coscia" Si considerano innanzitutto le analisi descrittive.

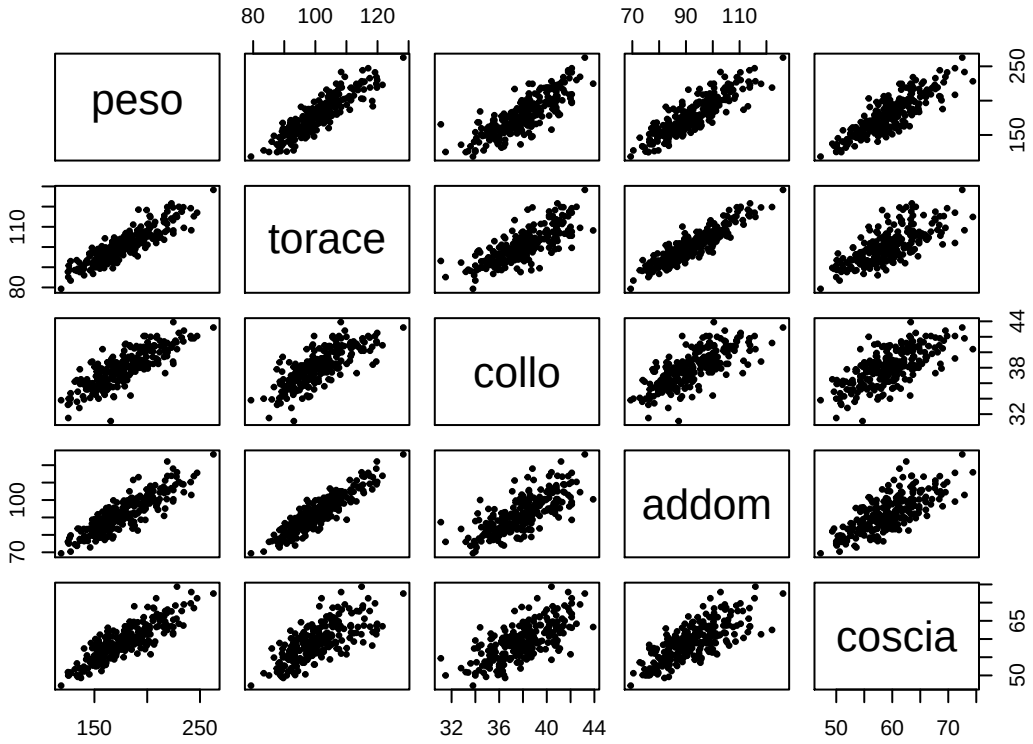
```
#-- R CODE
VAR_NUMERIC <- c("peso","torace","collo","addom","coscia")
pander(summary(d[,VAR_NUMERIC]),big.mark=",") #-- statistiche descrittive
```

| peso | torace | collo | addom | coscia |
|---|---|---|---|---|
| Min. :118.5 | Min. : 79.30 | Min. :31.10 | Min. : 69.40 | Min. :47.20 |
| 1st Qu.:158.2 | 1st Qu.: 94.15 | 1st Qu.:36.38 | 1st Qu.: 84.47 | 1st Qu.:56.00 |
| Median :176.1 | Median : 99.60 | Median :38.00 | Median : 90.95 | Median :59.00 |
| Mean :178.1 | Mean :100.67 | Mean :37.95 | Mean : 92.31 | Mean :59.27 |
| 3rd Qu.:196.8 | 3rd Qu.:105.30 | 3rd Qu.:39.42 | 3rd Qu.: 99.20 | 3rd Qu.:62.30 |
| Max. :262.8 | Max. :128.30 | Max. :43.90 | Max. :126.20 | Max. :74.40 |

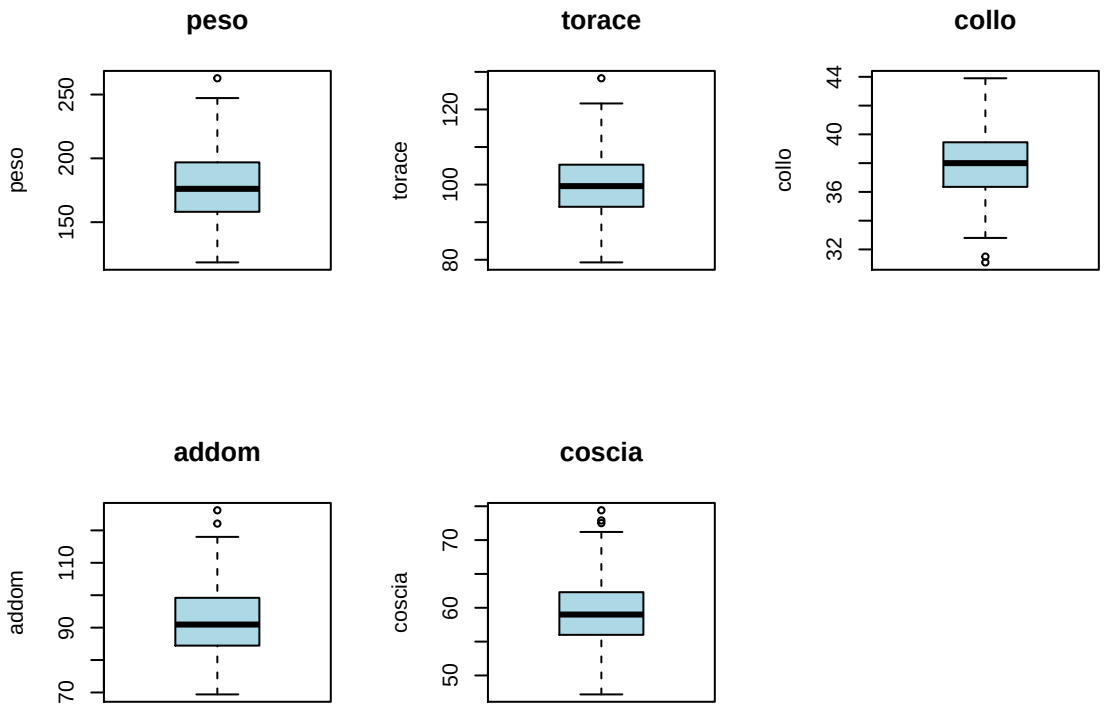
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

| | peso | torace | collo | addom | coscia |
|---|---|---|---|---|---|
| **peso** | 1 | 0.8914 | 0.8099 | 0.8742 | 0.8528 |
| **torace** | 0.8914 | 1 | 0.7691 | 0.9103 | 0.7082 |
| **collo** | 0.8099 | 0.7691 | 1 | 0.7293 | 0.669 |
| **addom** | 0.8742 | 0.9103 | 0.7293 | 1 | 0.7373 |
| **coscia** | 0.8528 | 0.7082 | 0.669 | 0.7373 | 1 |

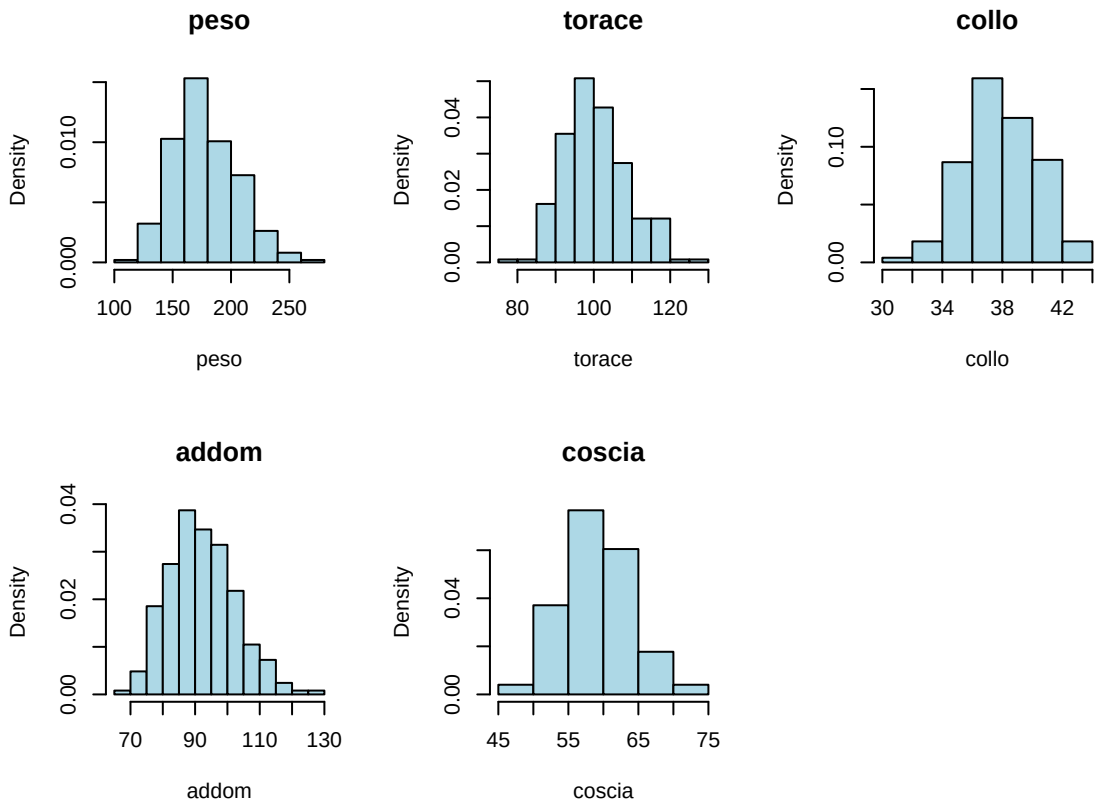
```
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```
par(mfrow=c(2,3))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
par(mfrow=c(2,3))
```




```
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```




```
#-- R CODE
mod1 <- lm(peso ~ torace + collo + addom + coscia, d)

pander(summary(mod1),big.mark=",")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | -180 | 9.6 | -18.75 | 4.185e-49 |
| **torace** | 1.212 | 0.1675 | 7.236 | 6.038e-12 |
| **collo** | 2.137 | 0.3734 | 5.722 | 3.079e-08 |
| **addom** | 0.3527 | 0.131 | 2.693 | 0.007582 |
| **coscia** | 2.065 | 0.1653 | 12.49 | 5.503e-28 |

Table 41: Fitting linear model: peso ~ torace + collo + addom + coscia

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 248 | 8.281 | 0.9084 | 0.9069 |

```
pander(anova(mod1),big.mark=",")
```

Table 42: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **torace** | 1 | 144,493 | 144,493 | 2,107 | 9.864e-122 |
| **collo** | 1 | 6,883 | 6,883 | 100.4 | 5.333e-20 |
| **addom** | 1 | 3,123 | 3,123 | 45.54 | 1.088e-10 |
| **coscia** | 1 | 10,698 | 10,698 | 156 | 5.503e-28 |
| **Residuals** | 243 | 16,663 | 68.57 | NA | NA |

```
pander(white.test(mod1),big.mark=",")
```

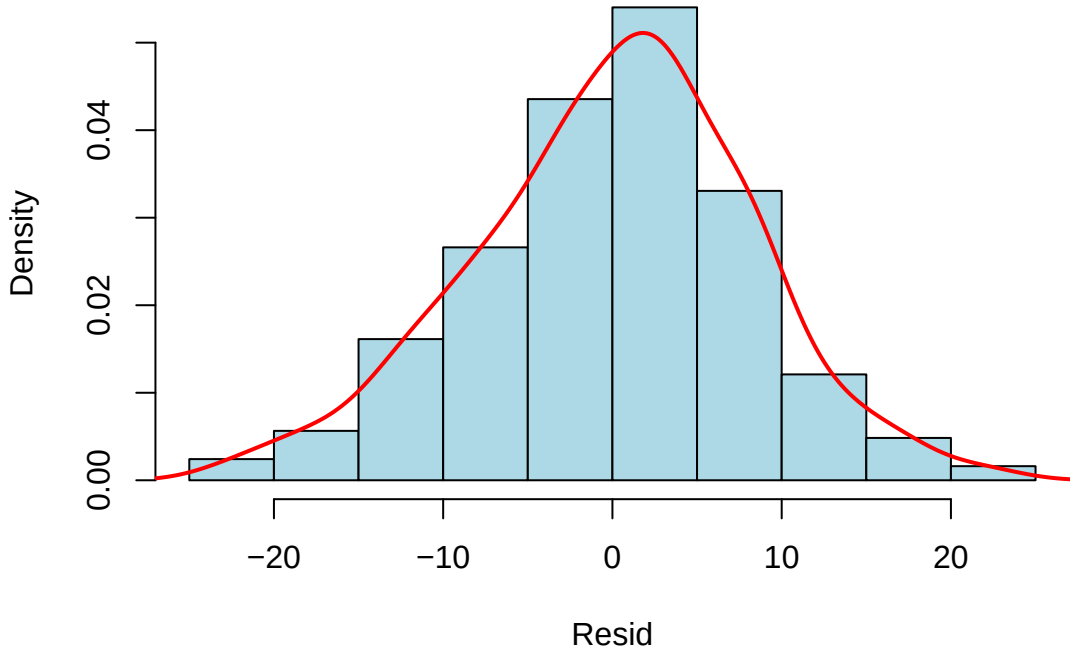
| Test.statistic | P.value |
|---|---|
| 4.156 | 0.1252 |

```
pander(dwtest(mod1),big.mark=",")
```

Table 44: Durbin-Watson test: `mod1`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 1.758 | 0.02538 * | true autocorrelation is greater than 0 |

```
#-- R CODE
hist(resid(mod1),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1)),col=2,lwd=2)
```




```
pander(shapiro.test(resid(mod1)))
```

Table 45: Shapiro-Wilk normality test: `resid(mod1)`

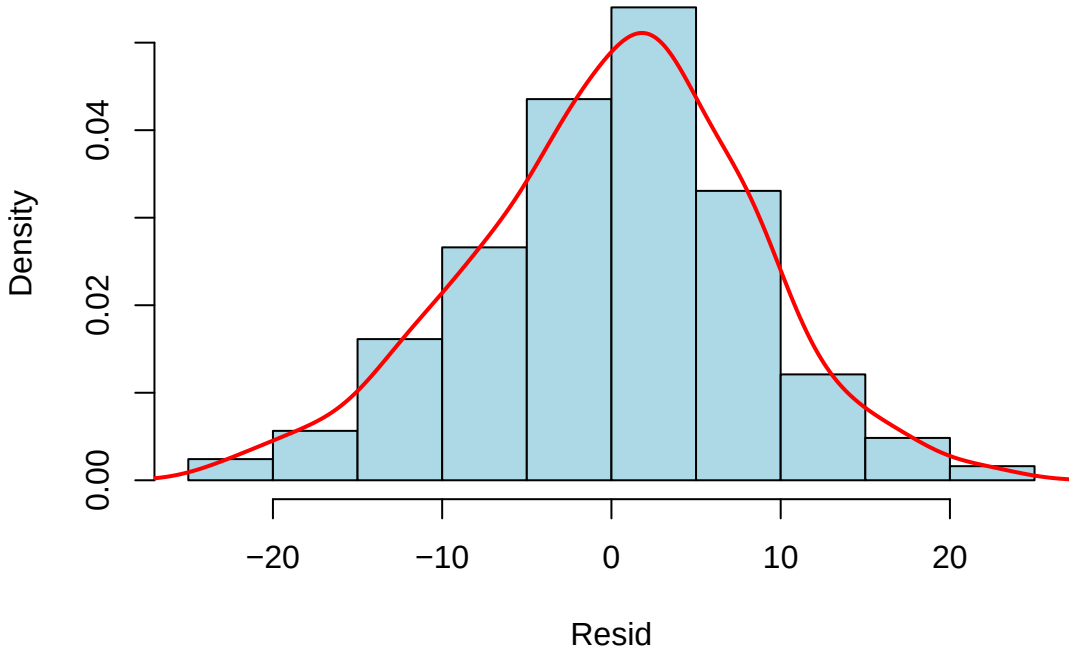
| Test statistic | P value |
|---|---|
| 0.994 | 0.4261 |

```
pander(ks.test(resid(mod1),"pnorm"))
```

Table 46: One-sample Kolmogorov-Smirnov test: `resid(mod1)`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 0.4134 | 0 * * * | two-sided |

```
#-- R CODE
hist(resid(mod1),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1)),col=2,lwd=2)
```




```
pander(shapiro.test(resid(mod1)))
```

Table 47: Shapiro-Wilk normality test: `resid(mod1)`

| Test statistic | P value |
|---|---|
| 0.994 | 0.4261 |

```
pander(ks.test(resid(mod1),"pnorm"))
```

Table 48: One-sample Kolmogorov-Smirnov test: `resid(mod1)`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 0.4134 | 0 * * * | two-sided |

```
#-- R CODE
pander(ols_eigen_cindex(mod1),big.mark=",")
```

Table 49: Table continues below

| Eigenvalue | Condition Index | intercept | torace | collo | addom |
|---|---|---|---|---|---|
| 4.989 | 1 | 0.0001201 | 3.876e-05 | 5.505e-05 | 7.48e-05 |
| 0.006611 | 27.47 | 0.2273 | 0.006842 | 0.008227 | 0.1158 |

| Eigenvalue | Condition Index | intercept | torace | collo | addom |
|---|---|---|---|---|---|
| 0.002341 | 46.17 | 0.05424 | 0.03962 | 0.01108 | 0.04392 |
| 0.001111 | 67.02 | 0.6577 | 0.0009945 | 0.8181 | 0.1508 |
| 0.000665 | 86.62 | 0.06058 | 0.9525 | 0.1625 | 0.6893 |

| coscia |
|---|
| 0.0001147 |
| 0.001077 |
| 0.9623 |
| 0.007978 |
| 0.02852 |

```
pander(ols_vif_tol(mod1),big.mark=",")
```

| Variables | Tolerance | VIF |
|---|---|---|
| torace | 0.1476 | 6.774 |
| collo | 0.3772 | 2.651 |
| addom | 0.1541 | 6.488 |
| coscia | 0.4192 | 2.385 |

Gli indici di tolleranza e varianza multifattoriale sono rispettivamente molto lontani da zero e molto al di sotto della soglia di 20. Tuttavia guardando il condition index si vede che per il 3, 4, 5 autovalore è superiore alla soglia di 30.

Si osserva poi guardando la proporzione di varianza associata ad ogni variabile che tale proporzione è molto elevata per "coscia" in corrispondenza dell'autovalore 3 (0.96), per "collo" in corrispondenza dell'autovalore 4 (0.81), per "torace" in corrispondenza dell'autovalore 5 (0.95). Il modello è quindi inficiato da multicollinearità: occorre riconcepirlo in un altro modo per superare il problema.