Es1

Niccolò Puccinelli

2022-05-10

```
data <- read.csv("antrop.txt", sep="\t")
data <- data[,-1]
View(data)
```

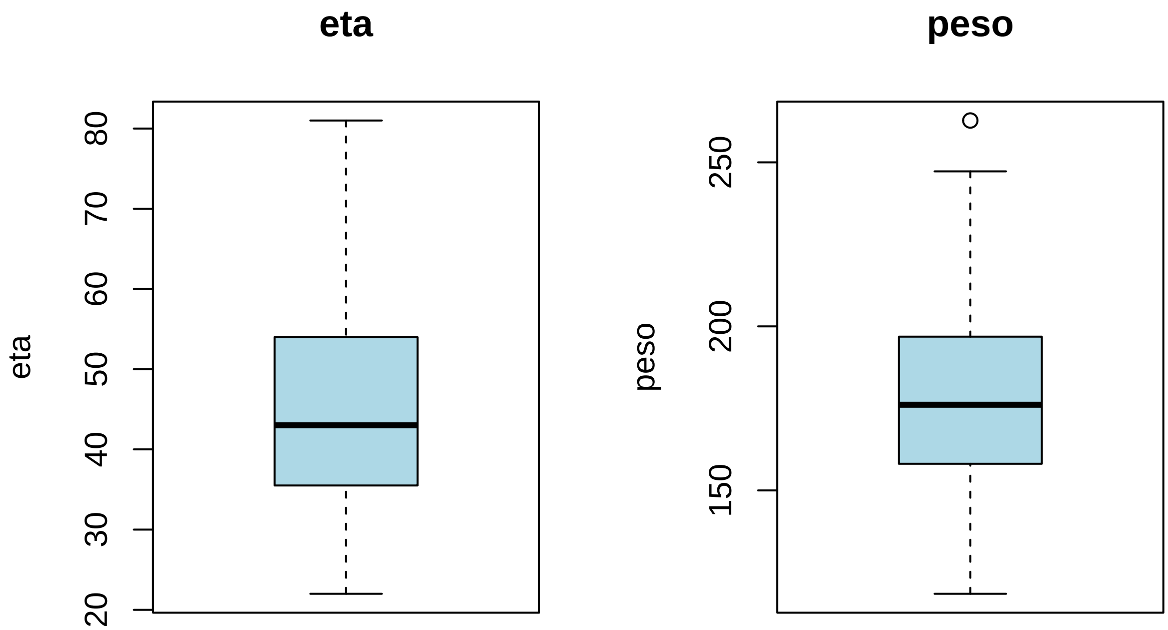
Il dataset presenta solo variabili numeriche non ordinate. Occupiamoci anzitutto delle statistiche descrittive.

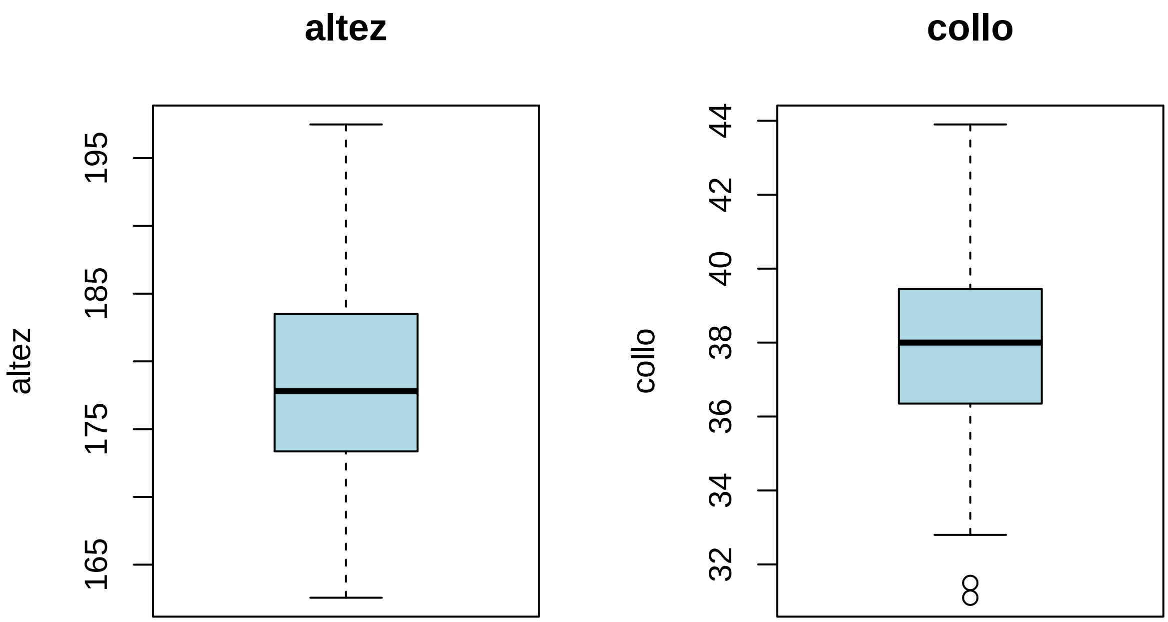
Statistiche descrittive

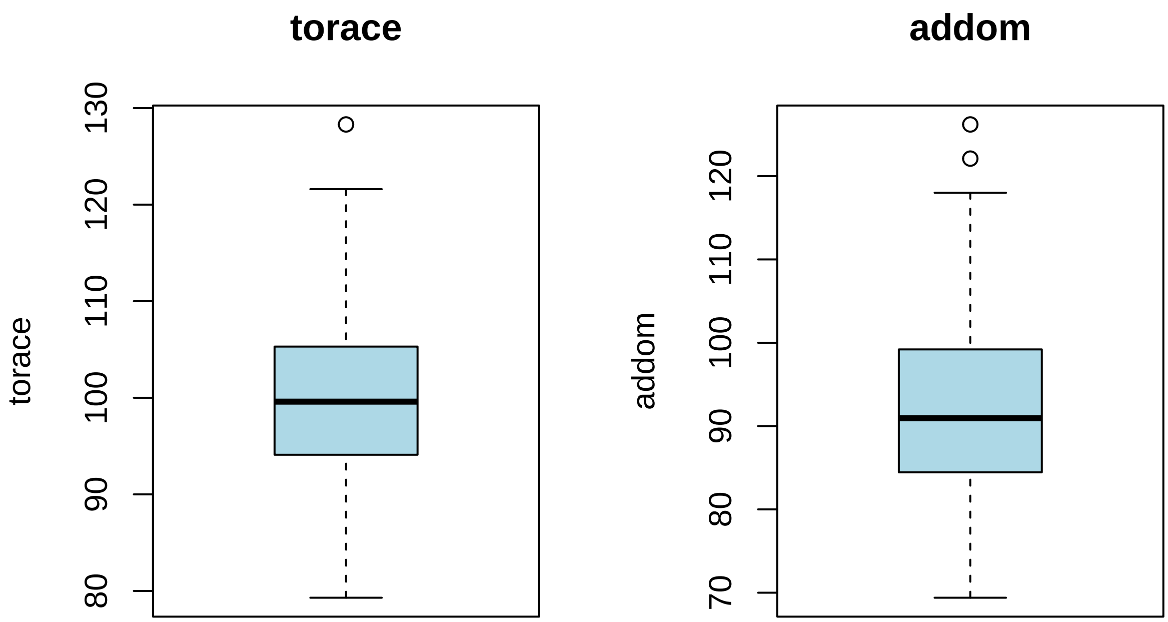
```
var = c("eta", "peso", "altez", "collo", "torace", "addom", "anca", "coscia",
        "ginocch", "caviglia", "bicipite", "avanbr", "polso")
summary(data[, var])
```

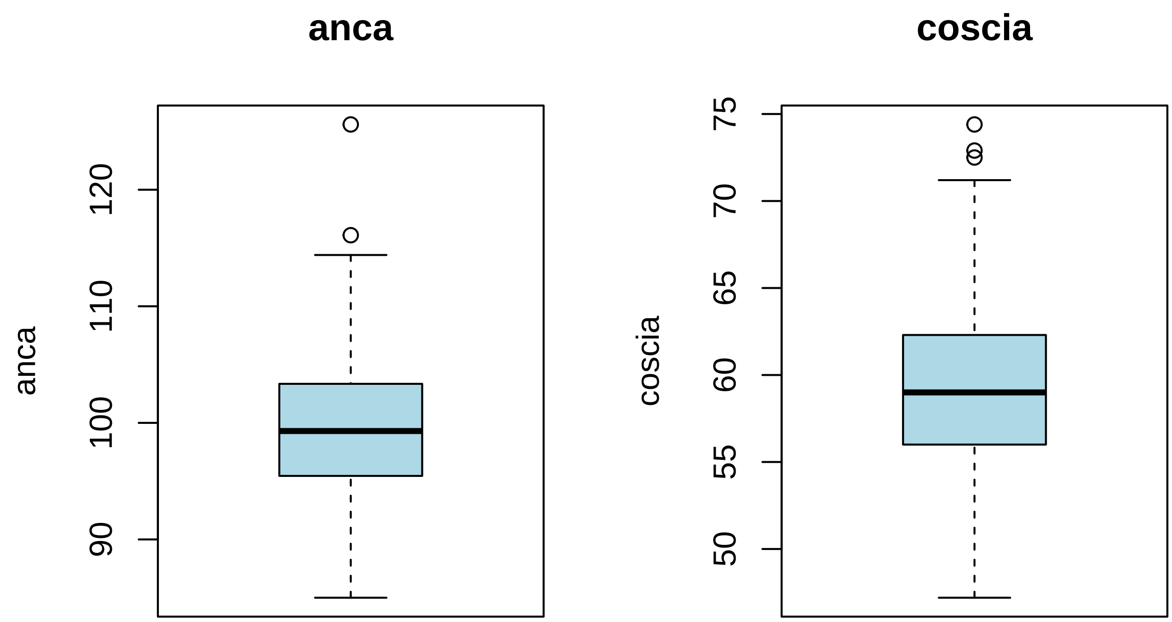
```
##       eta             peso           altez           collo      
##  Min.   :22.00   Min.   :118.5   Min.   :162.6   Min.   :31.10  
##  1st Qu.:35.75   1st Qu.:158.2   1st Qu.:173.4   1st Qu.:36.38  
##  Median :43.00   Median :176.1   Median :177.8   Median :38.00  
##  Mean   :44.85   Mean   :178.1   Mean   :178.6   Mean   :37.95  
##  3rd Qu.:54.00   3rd Qu.:196.8   3rd Qu.:183.5   3rd Qu.:39.42  
##  Max.   :81.00   Max.   :262.8   Max.   :197.5   Max.   :43.90  
##      torace           addom             anca            coscia     
##  Min.   : 79.30   Min.   : 69.40   Min.   : 85.00   Min.   :47.20  
##  1st Qu.: 94.15   1st Qu.: 84.47   1st Qu.: 95.47   1st Qu.:56.00  
##  Median : 99.60   Median : 90.95   Median : 99.30   Median :59.00  
##  Mean   :100.67   Mean   : 92.31   Mean   : 99.66   Mean   :59.27  
##  3rd Qu.:105.30   3rd Qu.: 99.20   3rd Qu.:103.28   3rd Qu.:62.30  
##  Max.   :128.30   Max.   :126.20   Max.   :125.60   Max.   :74.40  
##     ginocch         caviglia        bicipite         avanbr     
##  Min.   :33.00   Min.   :19.10   Min.   :24.80   Min.   :21.00  
##  1st Qu.:36.90   1st Qu.:22.00   1st Qu.:30.20   1st Qu.:27.30  
##  Median :38.45   Median :22.80   Median :32.00   Median :28.75  
##  Mean   :38.54   Mean   :22.99   Mean   :32.22   Mean   :28.67  
##  3rd Qu.:39.90   3rd Qu.:24.00   3rd Qu.:34.33   3rd Qu.:30.00  
##  Max.   :46.00   Max.   :27.00   Max.   :39.10   Max.   :34.90  
##      polso      
##  Min.   :15.80  
##  1st Qu.:17.60  
##  Median :18.30  
##  Mean   :18.22  
##  3rd Qu.:18.80  
##  Max.   :21.40  
```

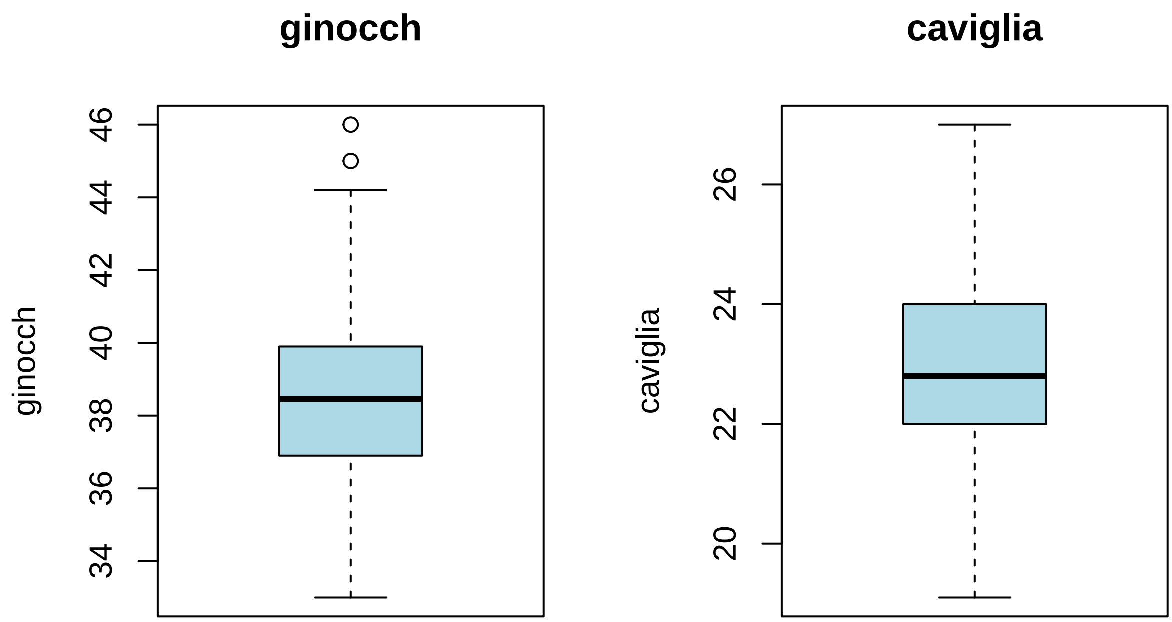
```r
# Box-plot
par(mfrow=c(1,2))
for(i in var){
  boxplot(data[,i],main=i,col="lightblue",ylab=i)
}
```

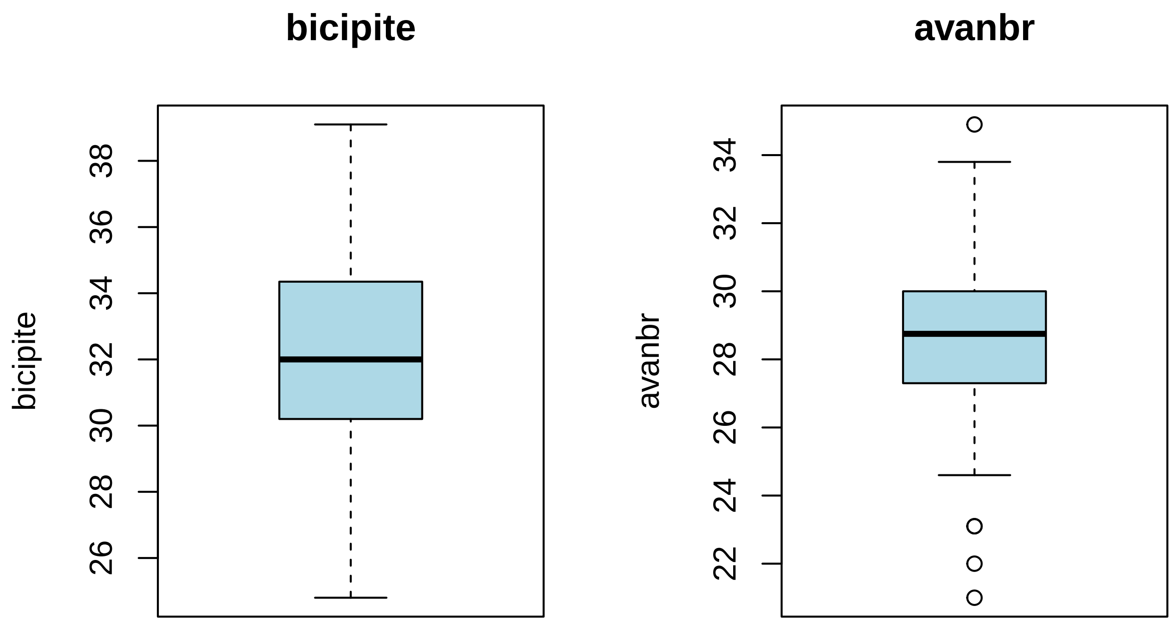


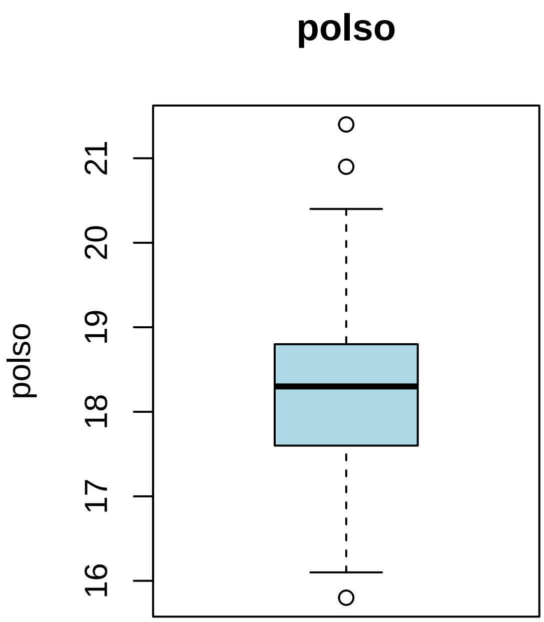

altez
195
185
175
165
altez
collo
44
42
40
38
36
34
32
collo


torace
addom
torace
addom
130
120
110
100
90
80
70


anca
120
110
100
90
anca
coscia
75
70
65
60
55
50
coscia


ginocch
ginocch
46
44
42
40
38
36
34
caviglia
caviglia
26
24
22
20

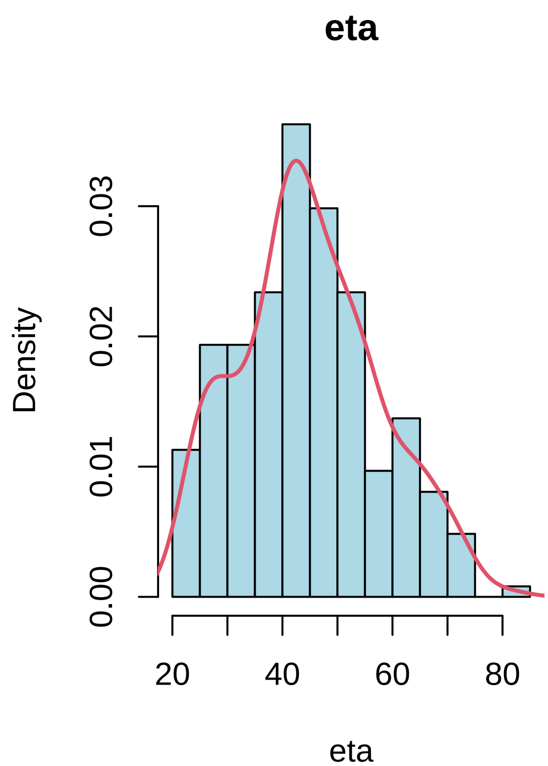

bicipite
bicipite
26
28
30
32
34
36
38
avanbr
avanbr
22
24
26
28
30
32
34



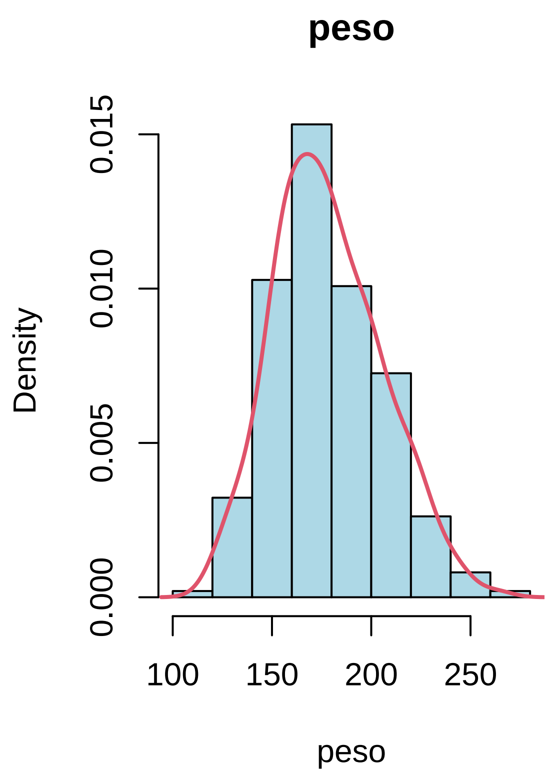
Dai box-plot notiamo che per diverse variabili la media si discosta dalla mediana (per lo più età e altezza), indicando una possibile non-normalità. Inoltre possiamo già vedere i primi outlier, presenti nelle variabili peso, collo, torace, addome, anca, coscia, ginocchio, avambraccio e polso.

Andiamo a vedere simmetria e possibile non-normalità anche con degli istogrammi.

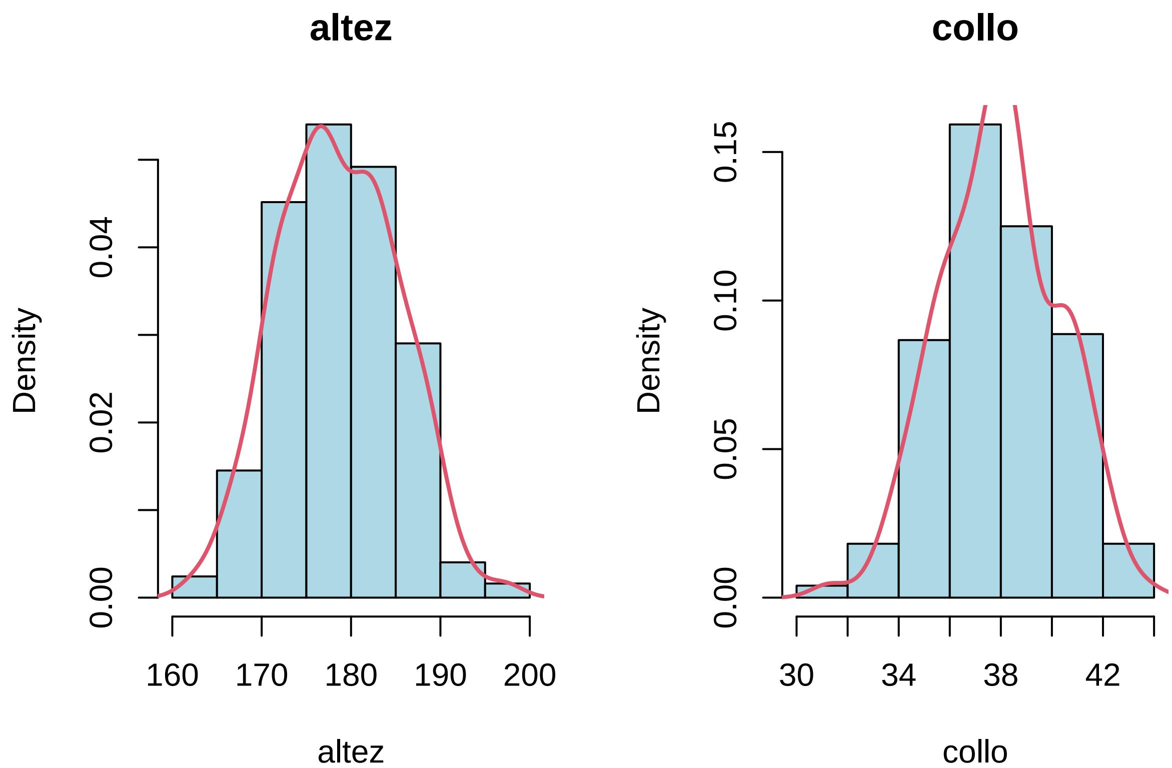
```
# Istogrammi
par(mfrow=c(1,2))
for(i in var){
  hist(data[,i],main=i,col="lightblue",xlab=i,freq=F,prob=TRUE)
  lines(density(data[,i]), col=2, lwd=2)
}
```

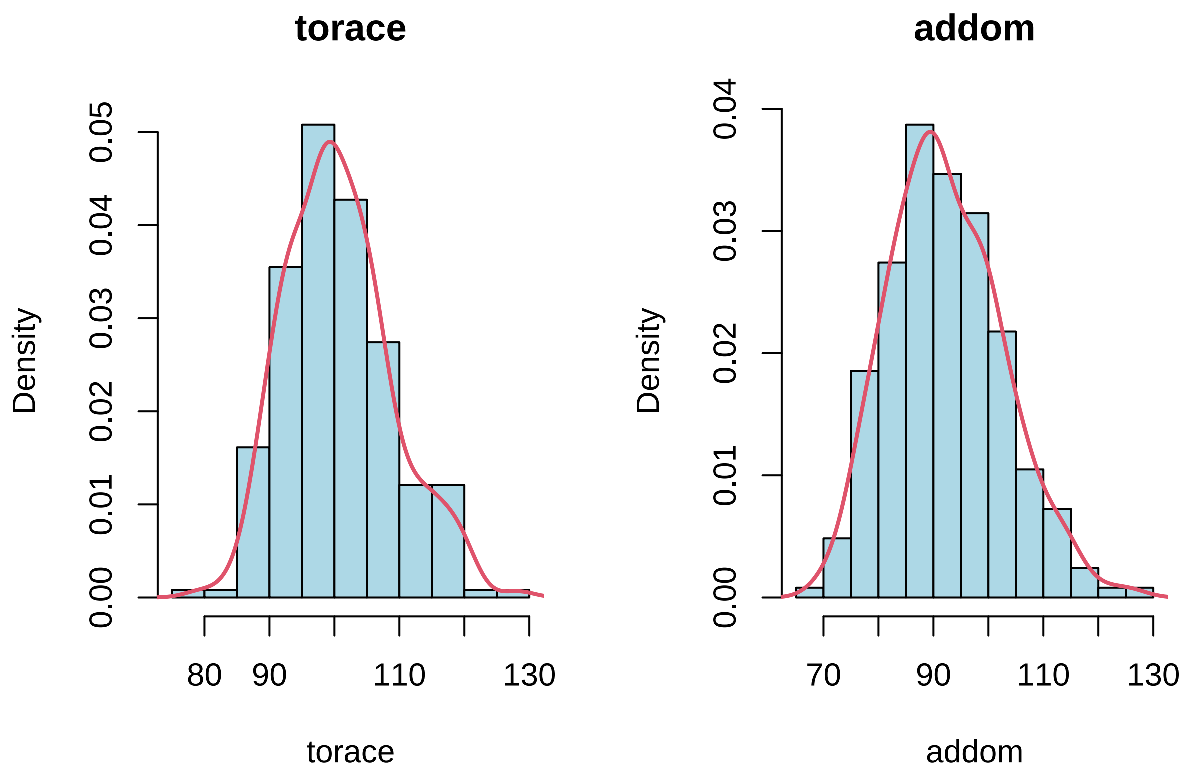


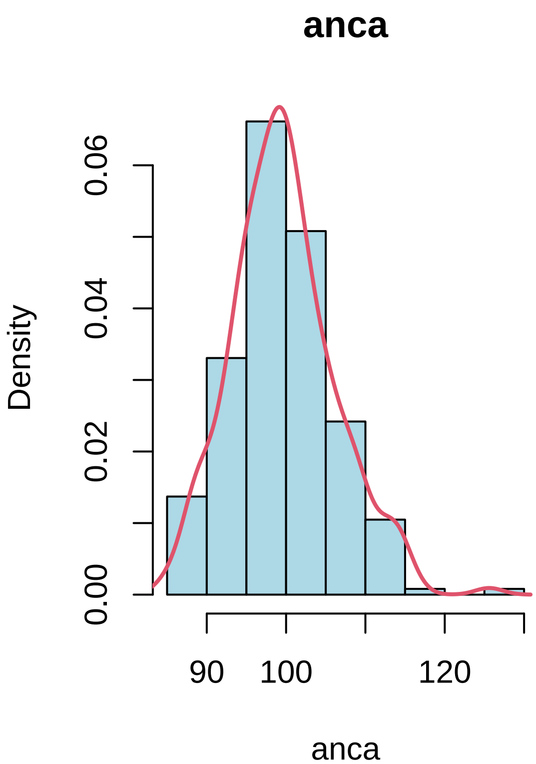
eta
Density
0.00
0.01
0.02
0.03
20
40
60
80
eta

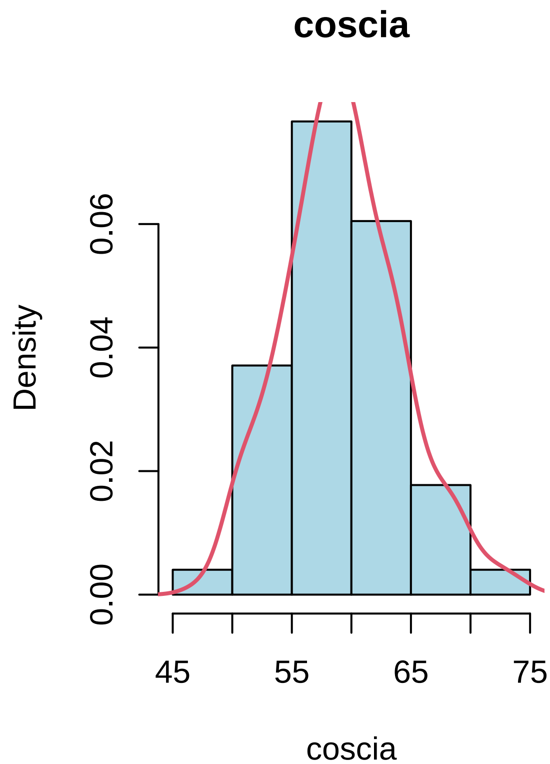


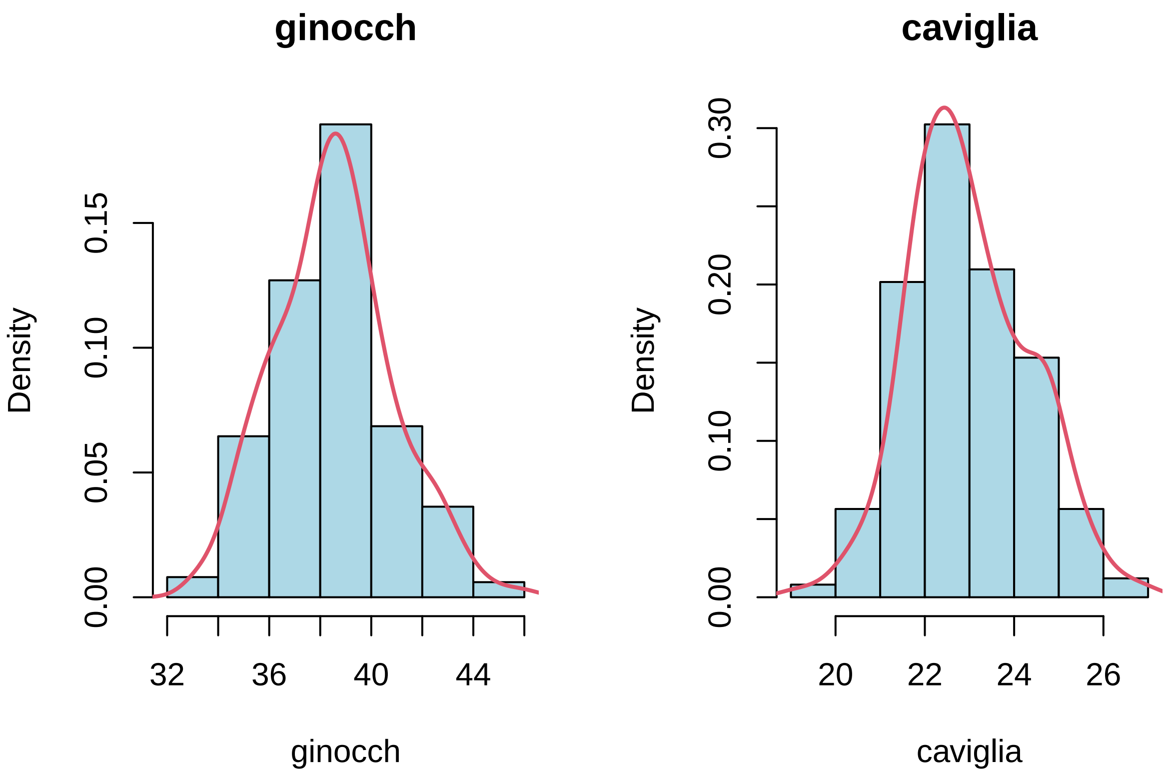
peso
Density
0.000
0.005
0.010
0.015
100
150
200
250
peso

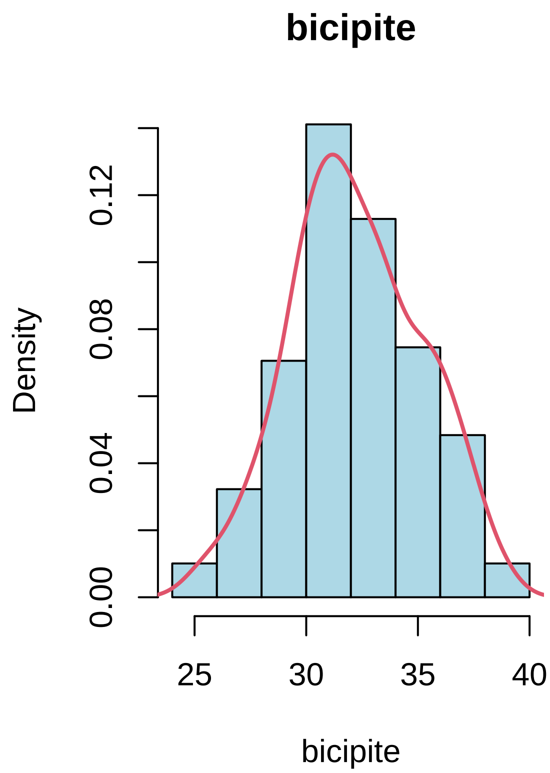

altez
Density
0.00
0.02
0.04
160
170
180
190
200
altez
collo
Density
0.00
0.05
0.10
0.15
30
34
38
42
collo

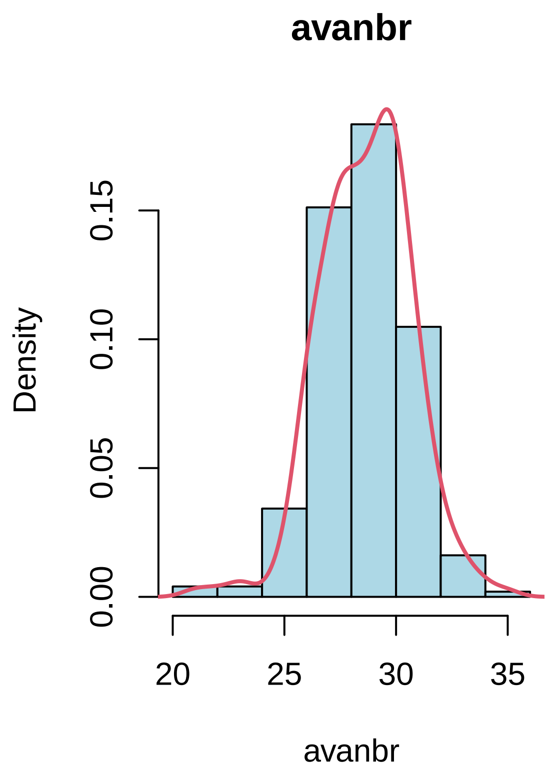

torace
Density
0.00
0.01
0.02
0.03
0.04
0.05
80
90
110
130
torace
addom
Density
0.00
0.01
0.02
0.03
0.04
70
90
110
130
addom

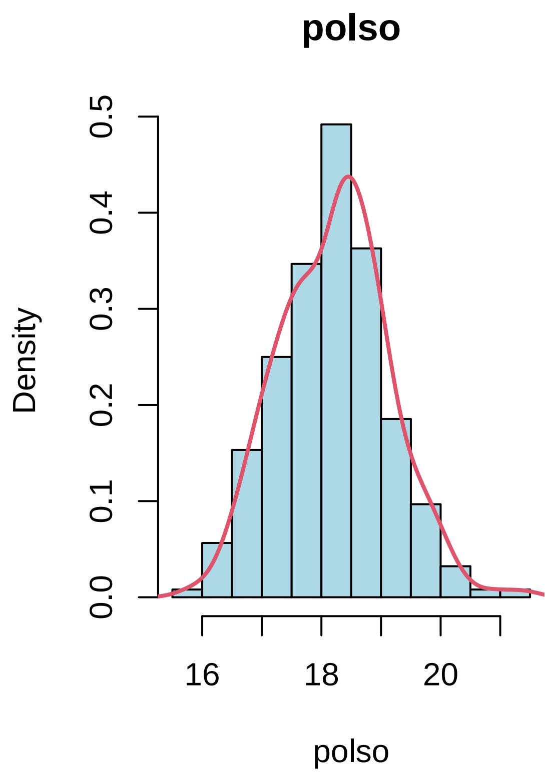

anca
Density
0.00
0.02
0.04
0.06
90
100
120
anca


coscia
Density
0.00
0.02
0.04
0.06
45
55
65
75
coscia


ginocch
Density
0.00
0.05
0.10
0.15
32
36
40
44
ginocch
caviglia
Density
0.00
0.10
0.20
0.30
20
22
24
26
caviglia


bicipite
Density
0.00
0.04
0.08
0.12
25
30
35
40
bicipite

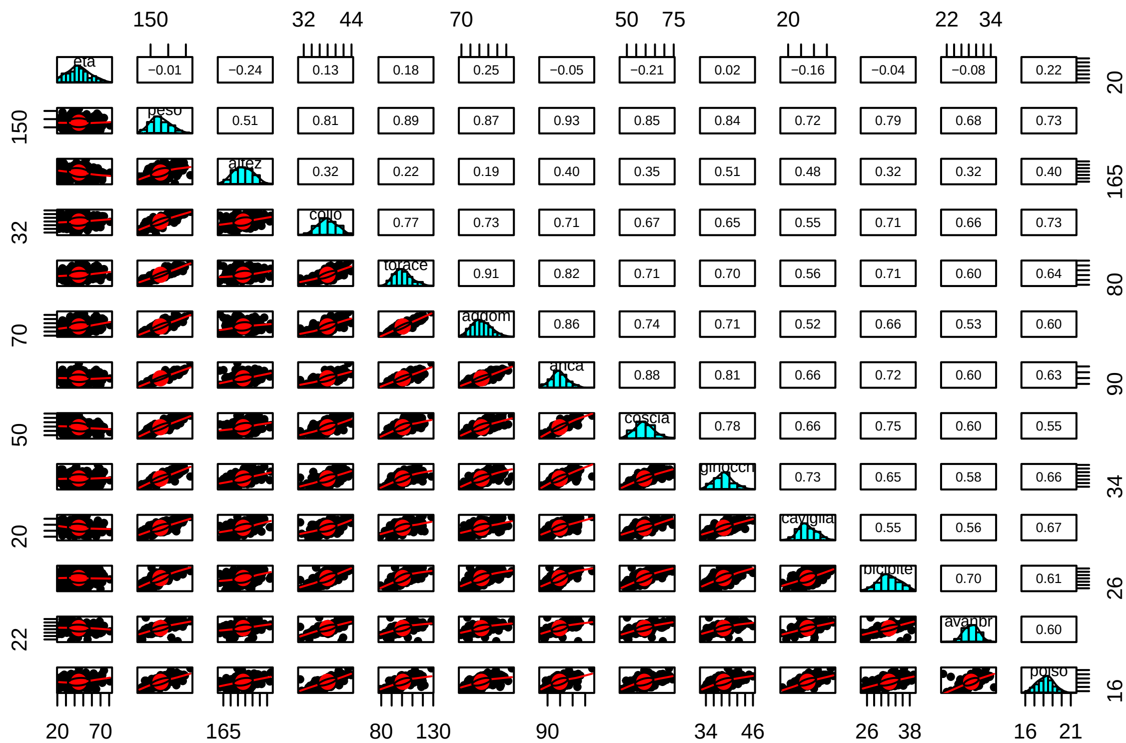

avanbr
Density
0.00
0.05
0.10
0.15
20
25
30
35
avanbr



Alcune variabili presentano un'asimmetria positiva (e.g. età, peso, torace, addome, anca, caviglia), altre negativa (e.g. collo, avambraccio, polso), indicando una possibile non-normalità, da verificare con gli appositi grafici e test.

Andiamo a studiare le correlazioni.

```
# Studio delle correlazioni
pairs.panels(data)
```



```
cor(data)
```

```
##                  eta        peso     altez     collo    torace     addom
## eta       1.00000000 -0.01269094 -0.2362936 0.1256832 0.1847965 0.2452463
## peso     -0.01269094  1.00000000  0.5135600 0.8099153 0.8913646 0.8741842
## altez    -0.23629355  0.51356002  1.0000000 0.3224054 0.2240882 0.1886275
## collo     0.12568319  0.80991528  0.3224054 1.0000000 0.7691450 0.7293274
## torace    0.18479651  0.89136456  0.2240882 0.7691450 1.0000000 0.9102506
## addom     0.24524634  0.87418418  0.1886275 0.7293274 0.9102506 1.0000000
## anca     -0.05475672  0.93268793  0.3968323 0.7072583 0.8249990 0.8608247
## coscia   -0.21320509  0.85281577  0.3502220 0.6689794 0.7082294 0.7372990
## ginocch   0.01988356  0.84267934  0.5142569 0.6480645 0.6975475 0.7106229
## caviglia -0.15928620  0.72484213  0.4804845 0.5455551 0.5588112 0.5221625
## bicipite -0.04455524  0.78558121  0.3201975 0.7092980 0.7070395 0.6567847
## avanbr   -0.08448871  0.68370728  0.3246422 0.6614849 0.5995031 0.5296607
## polso     0.22030070  0.72528679  0.3981916 0.7316718 0.6445865 0.6028574
##                 anca     coscia    ginocch   caviglia    bicipite      avanbr
## eta      -0.05475672 -0.2132051 0.01988356 -0.1592862 -0.04455524 -0.08448871
## peso      0.93268793  0.8528158 0.84267934  0.7248421  0.78558121  0.68370728
## altez     0.39683229  0.3502220 0.51425689  0.4804845  0.32019750  0.32464225
## collo     0.70725827  0.6689794 0.64806451  0.5455551  0.70929800  0.66148492
## torace    0.82499902  0.7082294 0.69754751  0.5588112  0.70703948  0.59950310
## addom     0.86082474  0.7372990 0.71062293  0.5221625  0.65678474  0.52966068
## anca      1.00000000  0.8814257 0.80910516  0.6593495  0.72221413  0.60318884
## coscia    0.88142565  1.0000000 0.77810920  0.6635202  0.74589642  0.60358993
## ginocch   0.80910516  0.7781092 1.00000000  0.7293425  0.65436933  0.57873467
```

```
## caviglia  0.65934955  0.6635202 0.72934248  1.0000000  0.54841146  0.56068107
## bicipite  0.72221413  0.7458964 0.65436933  0.5484115  1.00000000  0.70206128
## avanbr    0.60318884  0.6035899 0.57873467  0.5606811  0.70206128  1.00000000
## polso     0.62674915  0.5450004 0.65583443  0.6661933  0.61366767  0.59925545
##              polso
## eta      0.2203007
## peso     0.7252868
## altez    0.3981916
## collo    0.7316718
## torace   0.6445865
## addom    0.6028574
## anca     0.6267491
## coscia   0.5450004
## ginocch  0.6558344
## caviglia 0.6661933
## bicipite 0.6136677
## avanbr   0.5992555
## polso    1.0000000
```

Notiamo, in generale, diverse correlazioni abbastanza elevate tra le variabili. Consideriamo da ora in avanti esclusivamente le variabili di nostro interesse (i.e. bicipite e peso), che presentano un indice di correlazione pari allo 0.79. Tuttavia, essendo questo un task di regressione univariata, non andremo ad occuparci della multicollinearità.

Regressione di peso su bicipite

Effetttuiamo la regressione e commentiamone i risultati.

```
m1<-lm(peso~bicipite,data)
summary(m1)
```

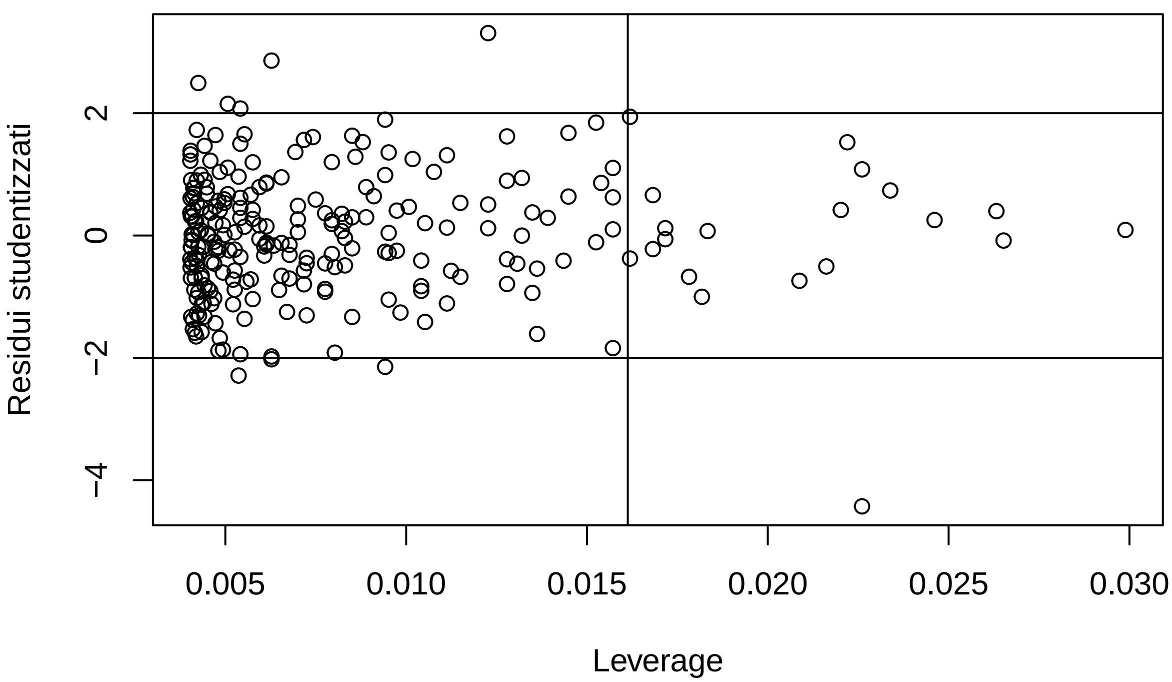
```
##
## Call:
## lm(formula = peso ~ bicipite, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.019 -11.099   0.163  10.467  54.237
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -55.9264    11.8015  -4.739 3.64e-06 ***
## bicipite      7.2648     0.3648  19.913  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.82 on 246 degrees of freedom
## Multiple R-squared:  0.6171, Adjusted R-squared:  0.6156
## F-statistic: 396.5 on 1 and 246 DF,  p-value: < 2.2e-16
```

La variabile bicipite risulta significativa e viene respinta l'ipotesi nulla H0 del test F. Il modello è dunque significativo. L'R quadro è buono: il modello riesce a spiegare circa il 62% della variabilità totale ed è molto simile all'R quadro aggiustato, che tiene conto della significatività delle variabili del modello.

Outlier

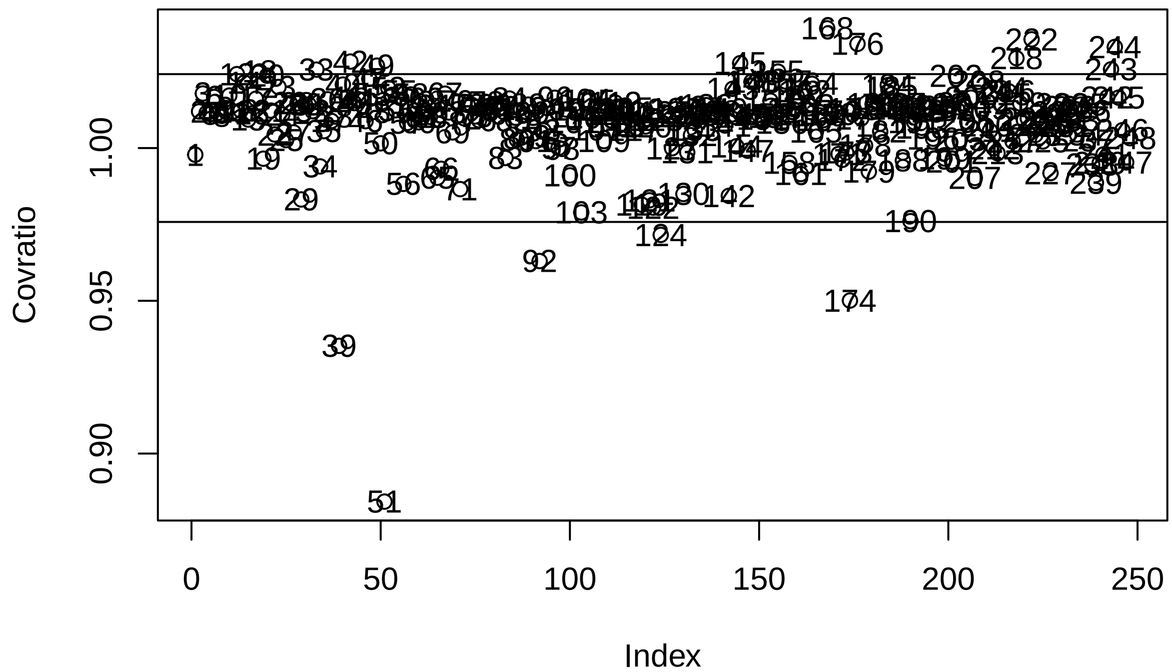
Verifichiamo la presenza o meno di outlier tramite il grafico residui studentizzati vs. leverage.

```
plot(hatvalues(m1), rstudent(m1), ylab="Residui studentizzati", xlab="Leverage")
# Soglie: 2, -2, 2k/n, oltre la quale si considerano outlier
# k=#coefficienti
# n=#osservazioni
abline(h=2)
abline(h=-2)
k=length(coef(m1))
n=nrow(data)
abline(v=2*k/n)
```

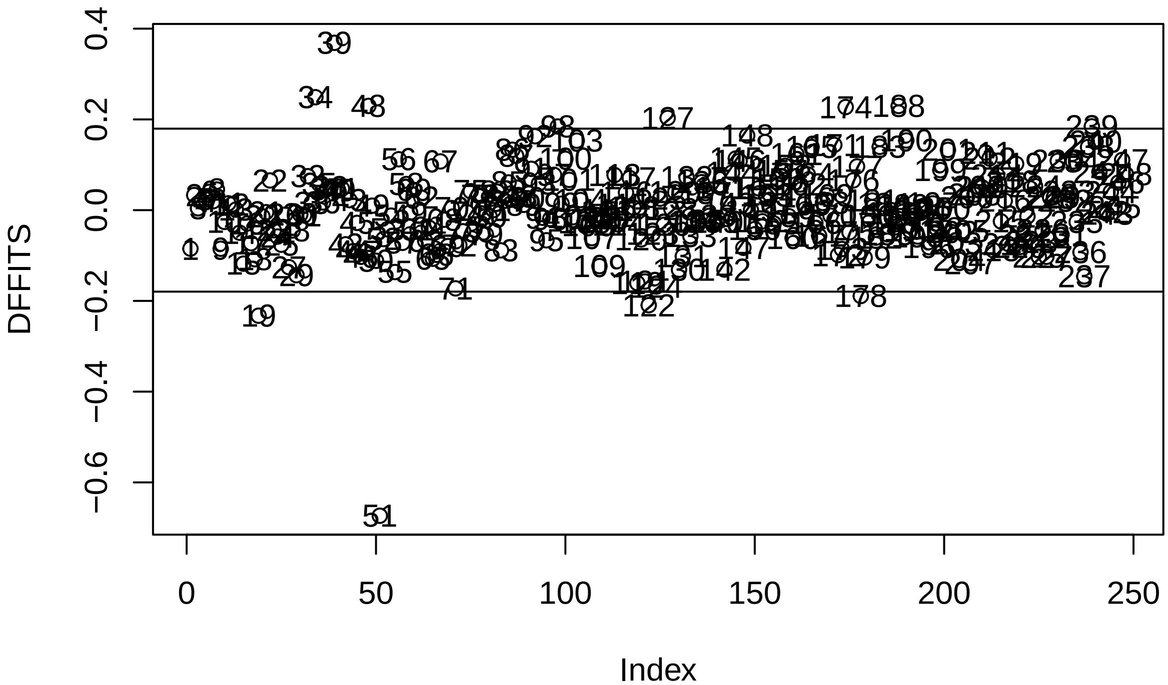


Notiamo la presenza di diversi outlier, i.e. i punti fuori dalla zona delimitata da v, 2 e -2. Utilizziamo i test.

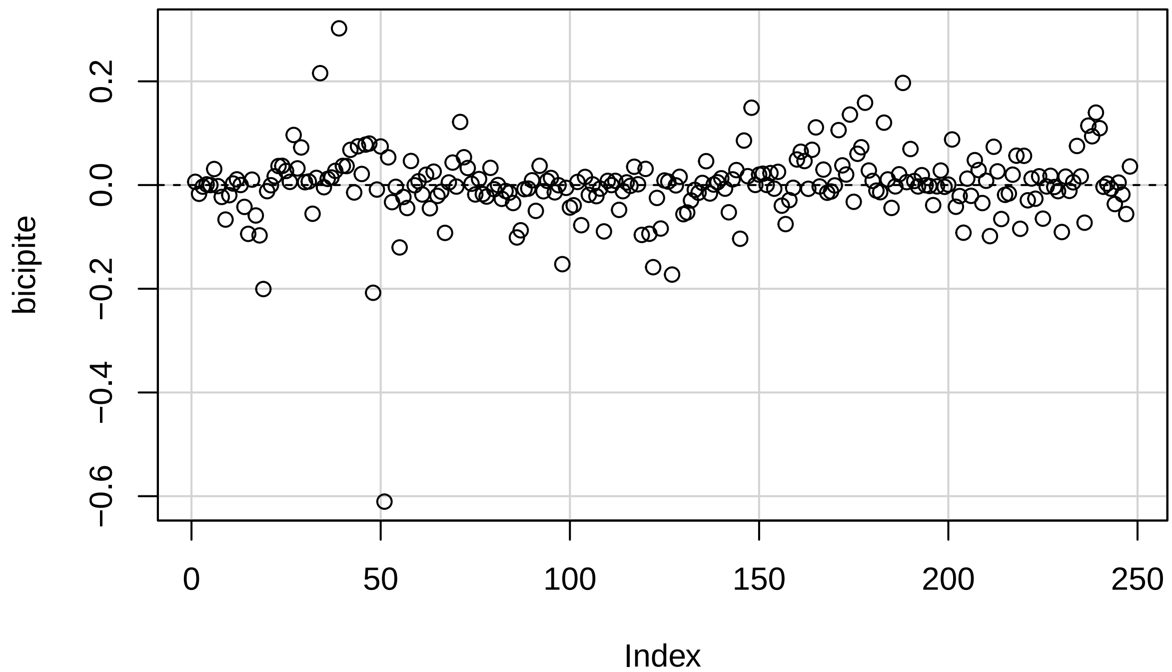
```
# COVRATIO
plot(covratio(m1), ylab="Covratio")
abline(h=1+3*k/n)
abline(h=1-3*k/n)
text(covratio(m1))
```



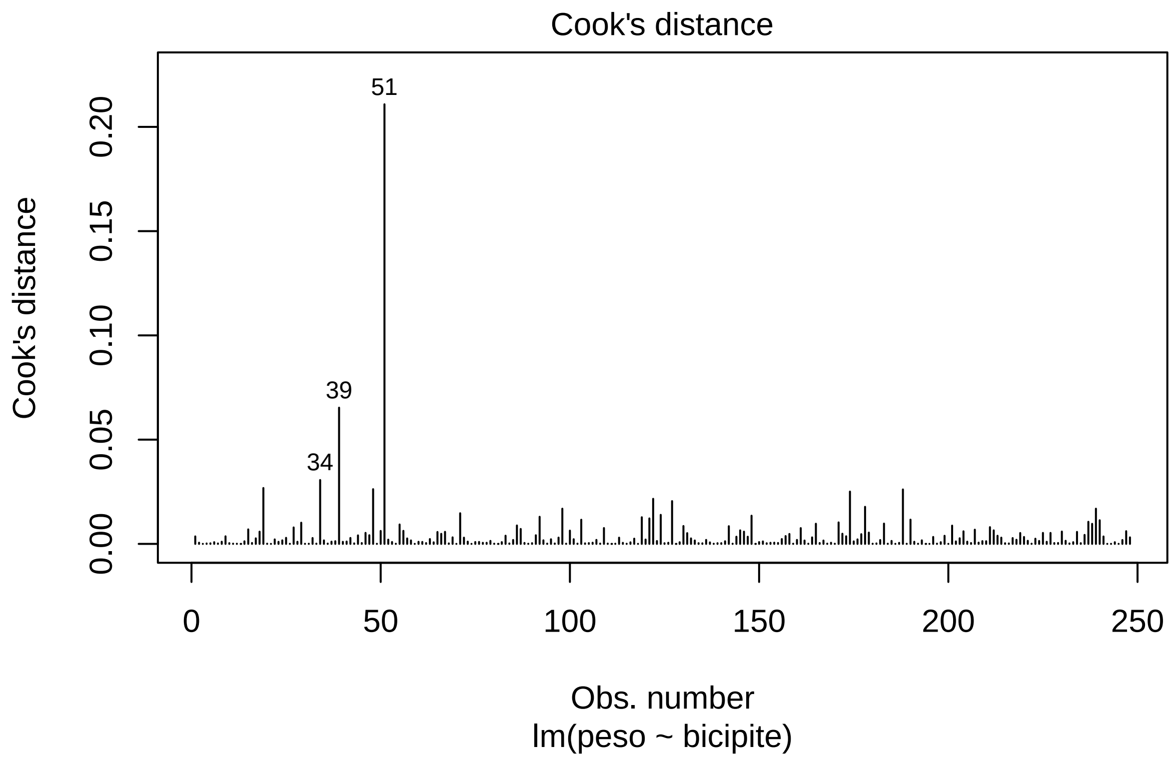

```
# DFITTS
plot(dffits(m1), ylab="DFFITS")
text(dffits(m1))
abline(h=2*sqrt(k/n))
abline(h=-2*sqrt(k/n))
```




```r
# DFBETAS
dfbetasPlots(m1)
```




```
# COOK
plot(m1, which=4)
```



I test ci identificano diversi outlier. Prendiamo come test per identificare gli outlier la distanza di Cook e rimuoviamo i punti 34, 39 e 51 dal dataset.

```
data2 = data[-c(34, 39, 51), ]
```

Andiamo a ristimare il modello.

```
m1<-lm(peso~bicipite,data2)
summary(m1)
```

```
##
## Call:
## lm(formula = peso ~ bicipite, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.085 -10.919   0.227  10.353  47.269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -56.8499    11.2796   -5.04 9.09e-07 ***
## bicipite      7.2916     0.3494   20.87  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

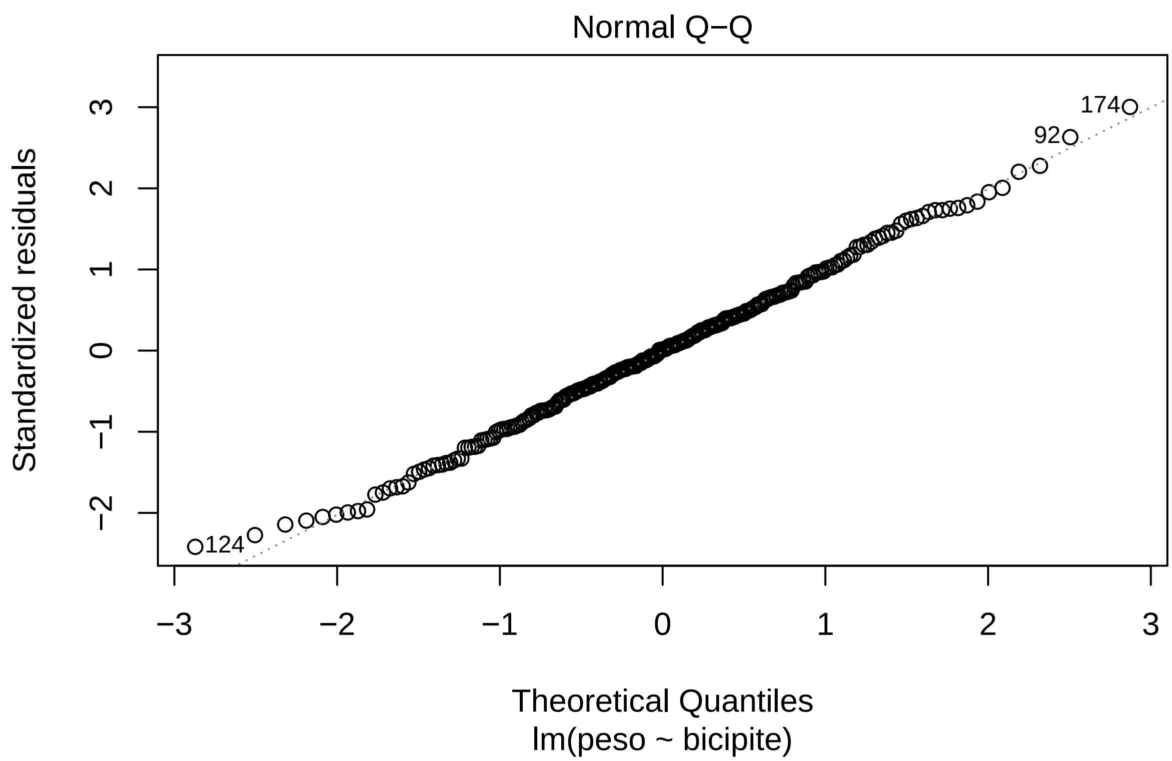
```
## Residual standard error: 15.79 on 243 degrees of freedom
## Multiple R-squared:  0.6418, Adjusted R-squared:  0.6404
## F-statistic: 435.5 on 1 and 243 DF,  p-value: < 2.2e-16
```

La variabile esplicativa e il modello rimangono significativi. Notiamo inoltre un incremento sia dell'R quadro, sia dell'R quadro aggiustato. Il modello senza outlier riesce a spiegare una porzione leggermente maggiore della variabilità.

Normalità

Adesso visualizziamo il Normal Q-Q Plot per verificare la normalità.

```
plot(m1, which=2)
```



La distribuzione, almeno da un punto di vista grafico, risulta pressoché normale, eccezion fatta per un po' di fuoriuscita dei punti sulle code. Verifichiamo la nostra ipotesi con i test di Shapiro-Wilk e di Kolmogorov-Smirnov.

```
ols_test_normality(m1)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.997          0.9236
```

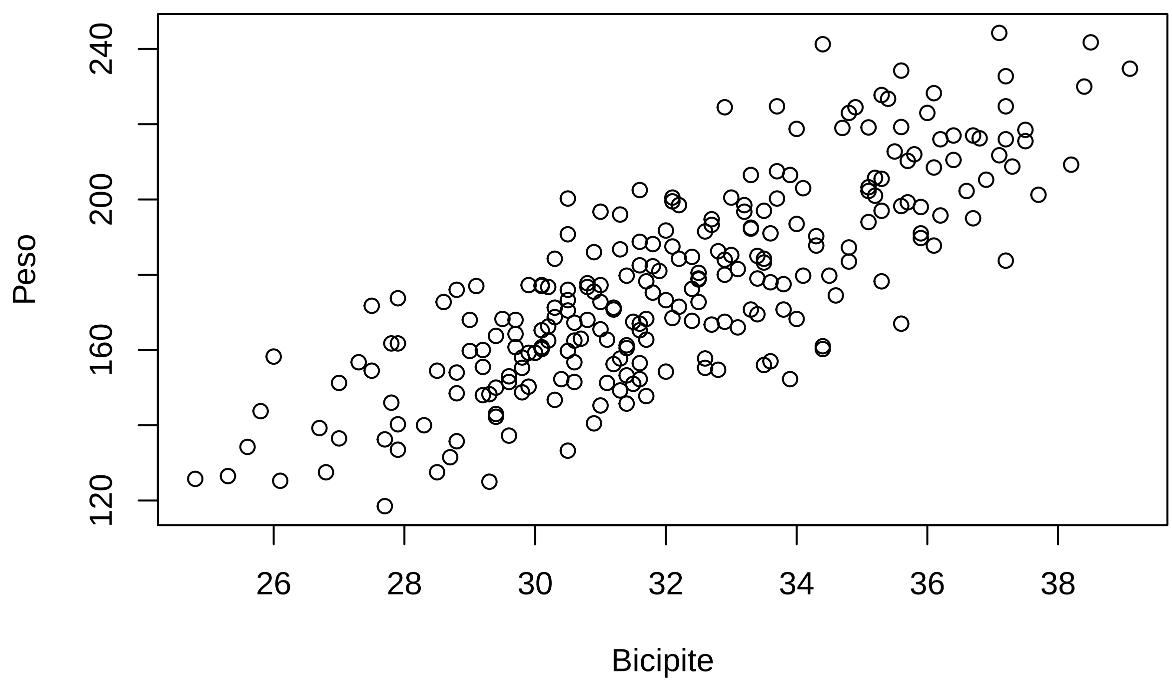
```
## Kolmogorov-Smirnov        0.018          1.0000
## Cramer-von Mises         19.2038         0.0000
## Anderson-Darling          0.1033         0.9953
## ----------------------------------------------
```

La nostra ipotesi di normalità risulta verificata.

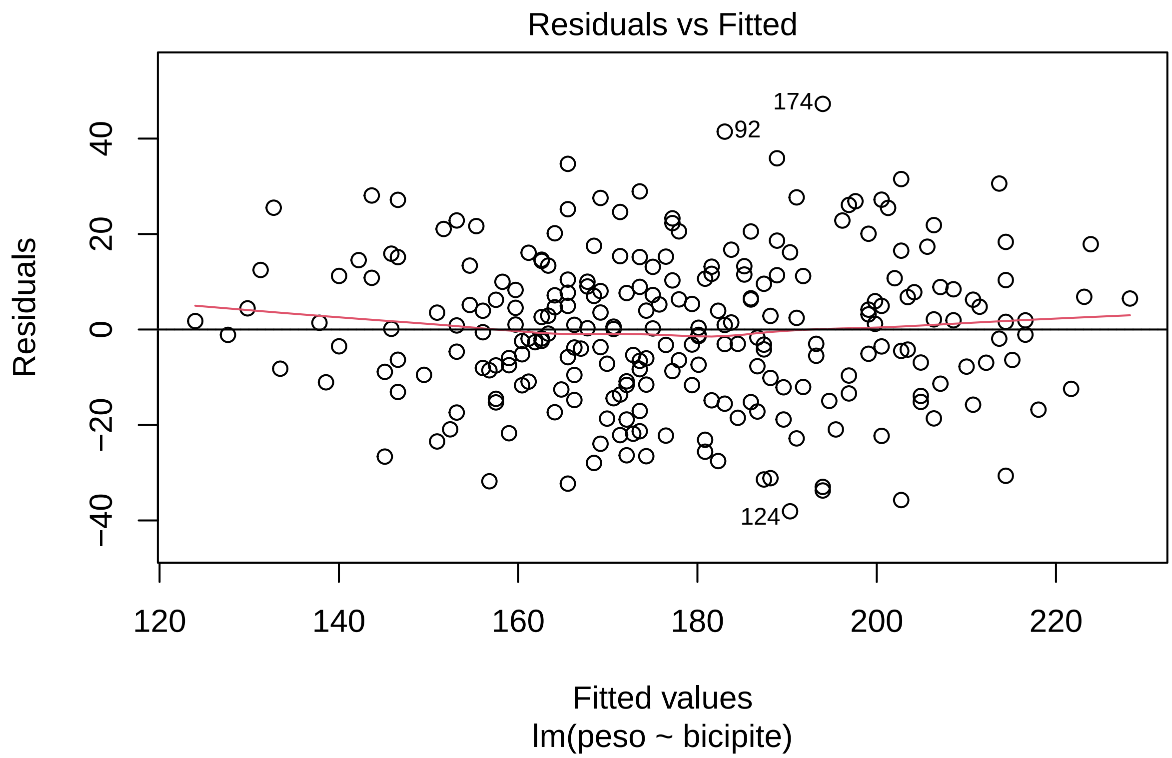
Eteroschedasticità

Verifichiamo ora se c'è omoschedasticità o meno dei residui, tramite i seguenti grafici.

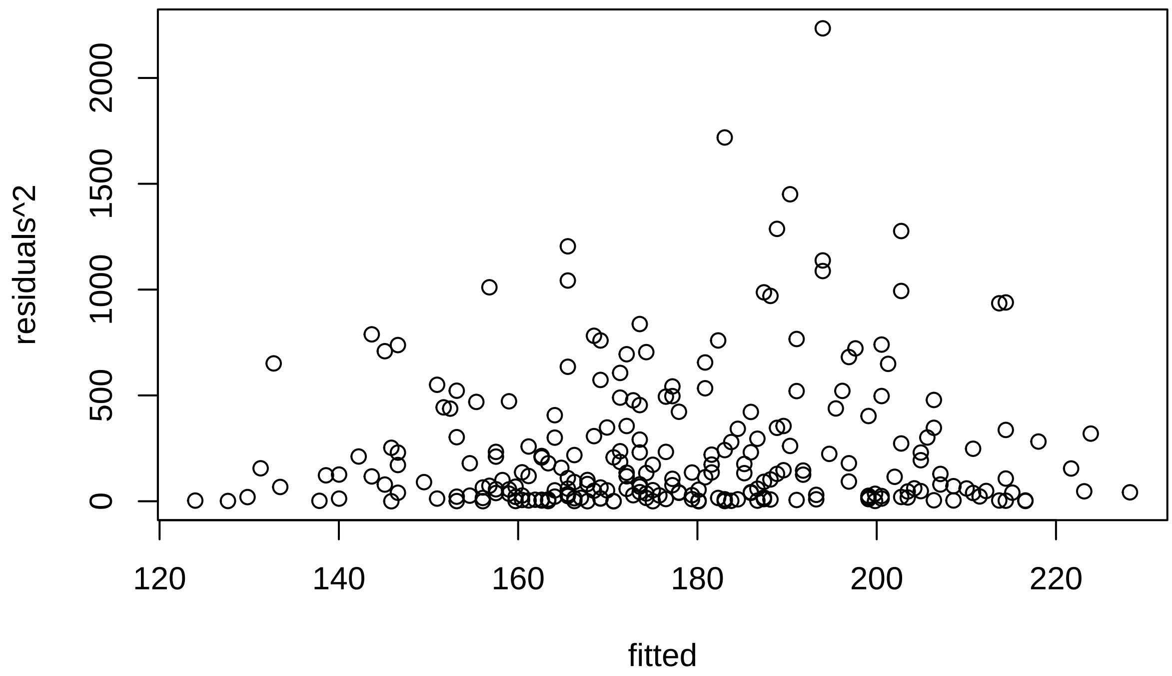
```r
# Valori osservati regressore x vs. variabile dipendente y
plot(data2$bicipite, data2$peso, xlab="Bicipite", ylab="Peso")
```



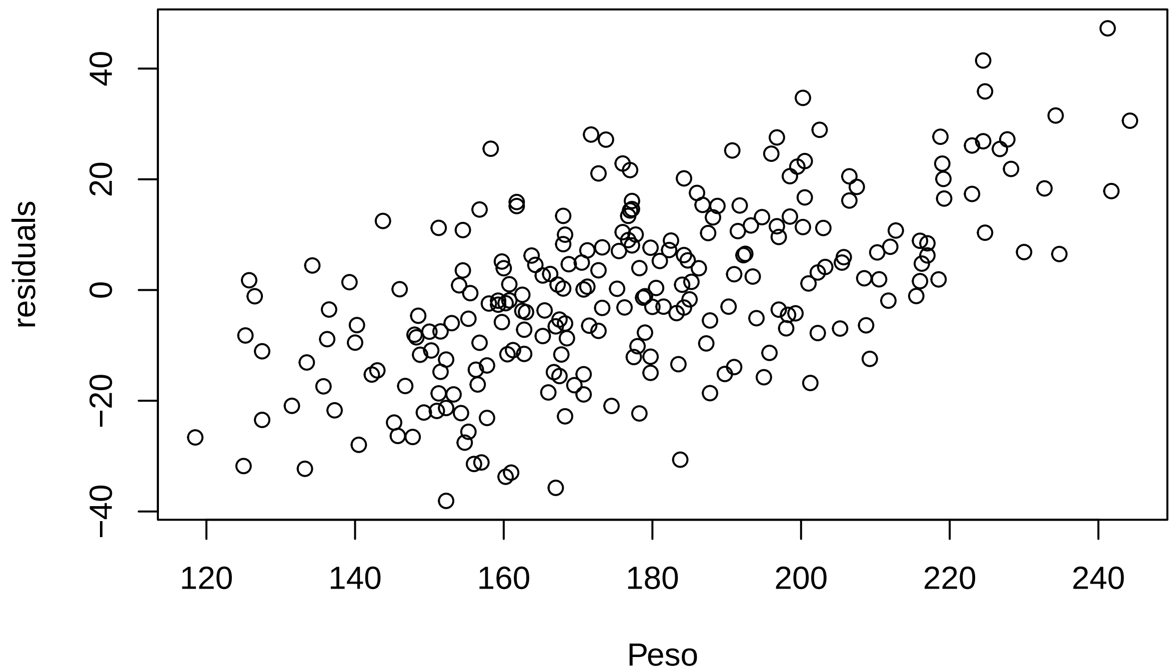

```r
# Residui stimati vs. valori predetti
plot(m1, which=1)
abline(h=0)
```




```r
# Residui stimati al quadrato vs. valori predetti
plot(m1$fitted, (m1$residuals)^2, xlab="fitted", ylab="residuals^2")
```




```
# Residui regressore x vs. variabile dipendente y
plot(data2$peso,  m1$residuals , xlab="Peso", ylab="residuals")
```



Dai grafici si direbbe che vi è omoschedasticità. Andiamo a vedere i test di White e di Breusch-Pagan.

```
white.test(m1)
```

```
##   Test.Statistic         P
## 1       3.574388 0.1674293
```

```
ols_test_breusch_pagan(m1)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##               Data
##  --------------------------------
##  Response : peso
##  Variables: fitted values of peso
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    1.84819
##  Prob > Chi2   =    0.1739941
```

Il p-value dei test ci porta ad accettare l'ipotesi di omoschedasticità. Proviamo anche a stimare altri modelli.

Altri modelli

Linear-Log

Modello con trasformazione logaritmica della variabile esplicativa.

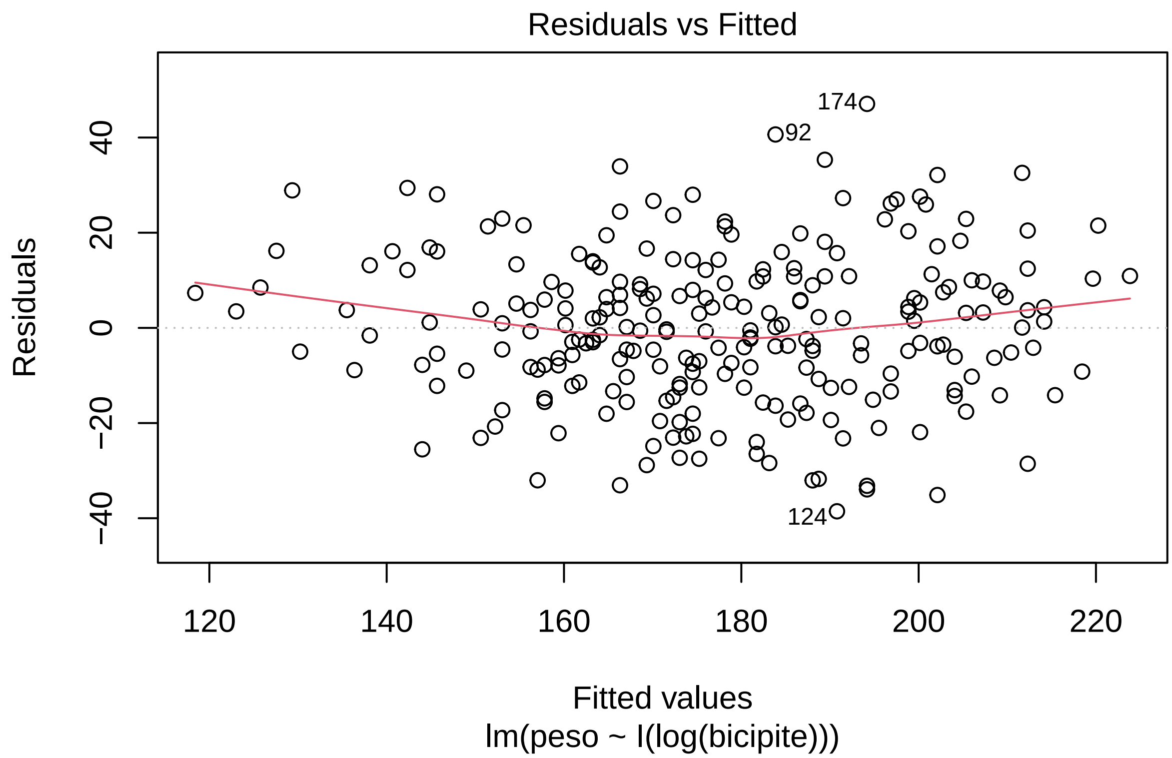
```
linlog<-lm(peso~I(log(bicipite)),data2)
summary(linlog)
```

```
##
## Call:
## lm(formula = peso ~ I(log(bicipite)), data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.540 -10.324   0.072  10.347  47.070
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -625.11      39.04  -16.01   <2e-16 ***
## I(log(bicipite))   231.56      11.26   20.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.94 on 243 degrees of freedom
## Multiple R-squared:  0.6352, Adjusted R-squared:  0.6337
## F-statistic:   423 on 1 and 243 DF,  p-value: < 2.2e-16
```

Il modello e la variabile esplicativa sono ancora significativi, tuttavia assistiamo ad un leggero calo dell'indice R quadro. Per questo modello, ad una variazione percentuale dell'1% di "bicipite" corrisponde una variazione di 2.31 di "peso".

Verifichiamo di nuovo l'ipotesi di omoschedasticità tramite il grafico residui vs. valori fitted e l'apposito test di White.

```
plot(linlog, which=1)
```




```
white.test(linlog)
```

```
##   Test.Statistic         P
## 1       2.223203 0.3290316
```

L'omoschedasticità risulta ancora verificata. Proviamo con un altro modello.

Log-Log

Modello con trasformazione logaritmica della variabile esplicativa e della variabile dipendente.

```
loglog<-lm(I(log(peso))~I(log(bicipite)),data2)
summary(loglog)
```

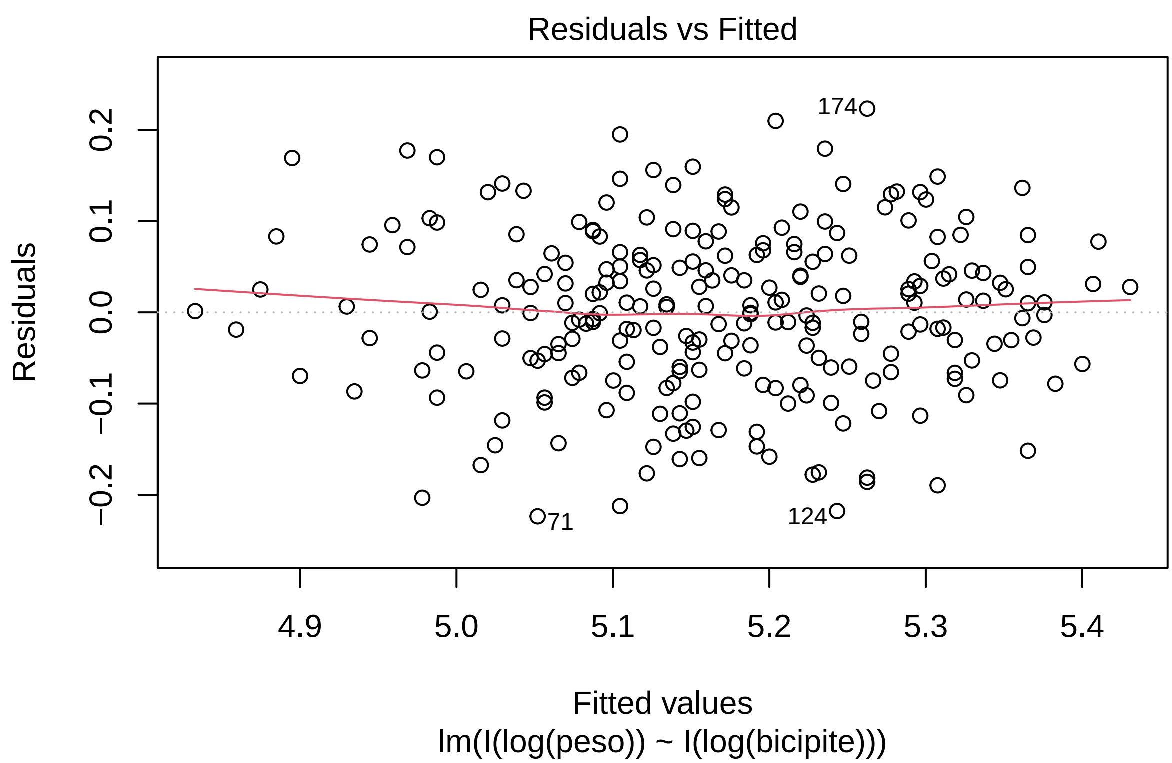
```
##
## Call:
## lm(formula = I(log(peso)) ~ I(log(bicipite)), data = data2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.223570 -0.060496  0.005825  0.062137  0.223265
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)        0.61733    0.21935   2.814  0.00529 **
## I(log(bicipite))  1.31293    0.06326  20.755  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08956 on 243 degrees of freedom
## Multiple R-squared:  0.6393, Adjusted R-squared:  0.6379
## F-statistic: 430.8 on 1 and 243 DF,  p-value: < 2.2e-16
```

Il modello e la variabile esplicativa sono significativi anche per questo modello. L'indice R quadro risulta migliore del precedente ma non del modello iniziale (per pochissimo). Per questo modello, ad una variazione percentuale dell'1.3% di "bicipite" corrisponde una variazione dell'1% di "peso".

Verifichiamo di nuovo l'ipotesi di omoschedasticità tramite il grafico residui vs. valori fitted e l'apposito test di White.

```
plot(loglog, which=1)
```




```
white.test(loglog)
```

```
##   Test.Statistic         P
## 1       1.873917 0.3918178
```

L'omoschedasticità risulta ancora verificata. Proviamo con un'ulteriore combinazione.

Log-Linear

Modello con trasformazione logaritmica della variabile dipendente.

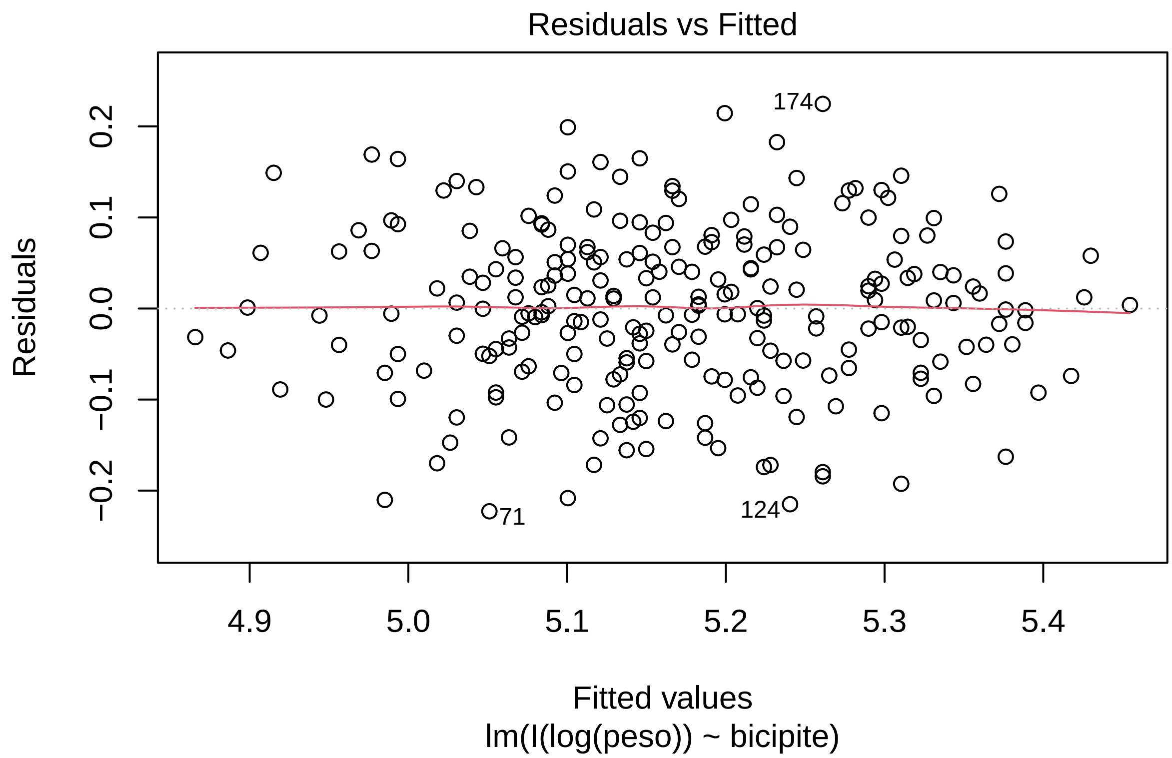
```
loglin<-lm(I(log(peso))~bicipite,data2)
summary(loglin)
```

```
##
## Call:
## lm(formula = I(log(peso)) ~ bicipite, data = data2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.222732 -0.057669  0.001034  0.061917  0.224756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.844392   0.063810   60.25   <2e-16 ***
## bicipite    0.041183   0.001977   20.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08934 on 243 degrees of freedom
## Multiple R-squared:  0.6411, Adjusted R-squared:  0.6396
## F-statistic: 434.1 on 1 and 243 DF,  p-value: < 2.2e-16
```

Il modello e la variabile esplicativa sono significativi anche per questo modello. L'indice R quadro risulta pressoché simile al modello iniziale. Per questo modello, ad una variazione percentuale del 4% di "bicipite" corrisponde una variazione unitaria di "peso".

Verifichiamo di nuovo l'ipotesi di omoschedasticità tramite il grafico residui vs. valori fitted e l'apposito test di White.

```
plot(loglin, which=1)
```




```
white.test(loglin)
```

```
##   Test.Statistic        P
## 1       2.171941 0.337574
```

Anche questo modello presenta omoschedasticità dei residui. Proviamo con un modello quadratico.

Quadratico

Modello a cui aggiungiamo la trasformazione quadratica della variabile esplicativa (dopo averla centrata per evitare la collinearità con sé stessa al quadrato).

```
data2$bicipite_c<-data2$bicipite-mean(data2$bicipite)
quadratico<-lm(peso~bicipite_c+I(bicipite_c^2),data2)
summary(quadratico)
```

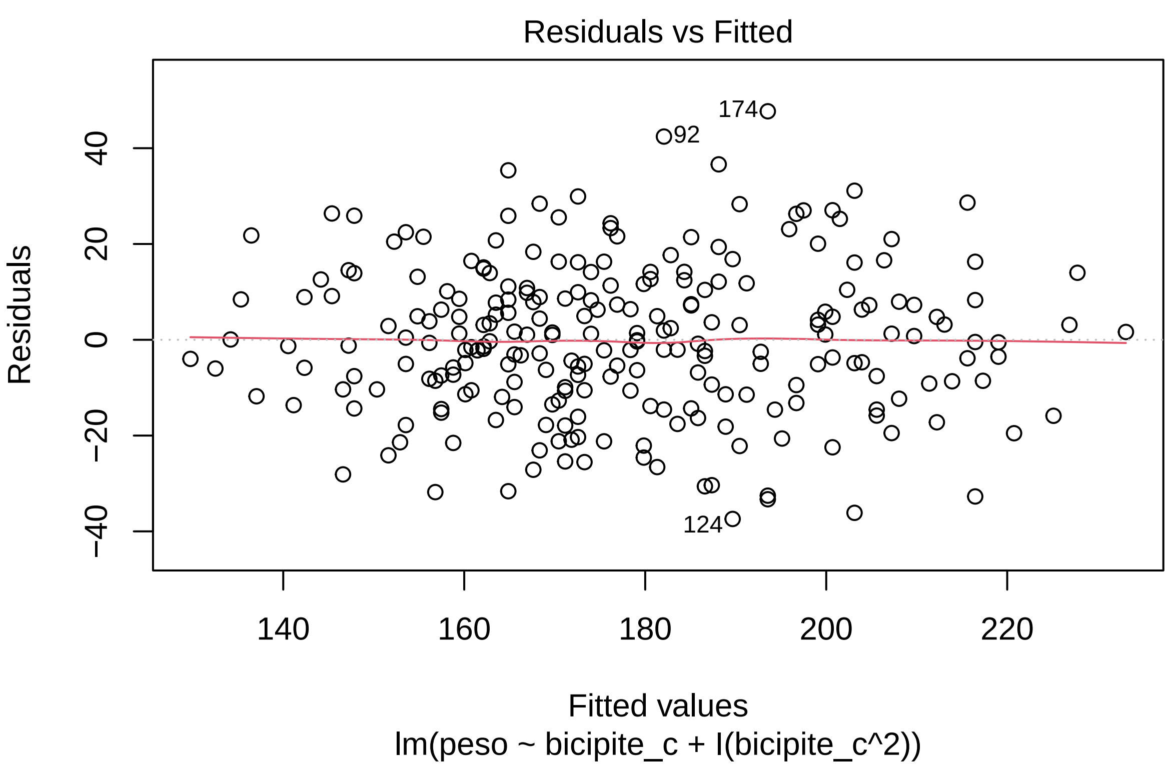
```
##
## Call:
## lm(formula = peso ~ bicipite_c + I(bicipite_c^2), data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.408 -10.541  -0.353  10.127  47.705
##
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    176.55368    1.29261 136.587   <2e-16 ***
## bicipite_c       7.27920    0.34910  20.851   <2e-16 ***
## I(bicipite_c^2)  0.12405    0.09711   1.277    0.203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.77 on 242 degrees of freedom
## Multiple R-squared:  0.6442, Adjusted R-squared:  0.6413
## F-statistic: 219.1 on 2 and 242 DF,  p-value: < 2.2e-16
```

Il modello e la variabile esplicativa "bicipite" sono significativi anche per questo modello. tuttavia, la variabile al quadrato non è significativa. L'indice R quadro risulta pressoché simile al modello iniziale.

Verifichiamo di nuovo l'ipotesi di omoschedasticità tramite il grafico residui vs. valori fitted e l'apposito test di White.

```
plot(quadratico, which = 1)
```




```
white.test(quadratico)
```

```
##   Test.Statistic          P
## 1       4.752178 0.09291325
```

Accettiamo l'ipotesi di omoschedasticità. Tuttavia, questa volta il p-value indica che siamo molto più vicini a rifiutarla rispetto ai casi precedenti. Se avessimo scelto un alfa del 10%, avremmo dovuto rifiutare l'ipotesi.

Modello finale

Il modello che si è comportato meglio in termini di eteroschedasticità (i.e. quello col p-value più alto) è il modello Log-Log. Testiamo dunque anche l'ipotesi di normalità. tramite i test di Shapiro-Wilk e di Kolmogorv-Smirnov.

```
loglog<-lm(I(log(peso))~I(log(bicipite)),data2)
ols_test_normality(loglog)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9947         0.5589
## Kolmogorov-Smirnov        0.0303         0.9778
## Cramer-von Mises         67.7982         0.0000
## Anderson-Darling          0.2366         0.7849
## -----------------------------------------------
```

Accettiamo dunque anche l'ipotesi di normalità.