# MULTI 2 - Data set: VITAMINA

## INTRODUZIONE

Per 2224 individui è stata studiata la quanità di energia consumata. Le variabili a disposizione sono le seguenti:

1. PERSON: identificativo dell'individuo
2. WT: peso in Kg
3. HT: altezza in cm
4. SEX: 1 maschio, 2 femmina
5. AGE: età
6. BMR: basal metabolic rate
7. E_BMR: energy per BMR
8. ENERGI: energy content (kj)
9. AVIT: vitamina A (RE)
10. RETINOL: retinol (mg)
11. BETACAR: beta carotene (mg)
12. DVIT: vitamina D (mg)
13. EVIT: vitamina E (alphaTE)
14. B1VIT: vitamina BA (mg)
15. B2VIT: vitamina B2 (mg)
16. B6VIT: vitamina B6 (mg)
17. FOLACIN: folacin (mg)
18. B12VUIR: vitamina B12 (mg)
19. CVIT: vitamina C (mg)

Analisi proposte:

1. Statistiche descrittive
2. Regressione Multivariata

```r
#-- R CODE
library(car)
library(sjstats)
library(plotrix)
library(sjPlot)
library(sjmisc)
library(lme4)
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"
d <- read.csv(paste0(ABSOLUTE_PATH,"\\esercizi (5)  copia\\2.mult\\vitamina.txt"),sep=" ")

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("bmr","E_bmr","wt","ht","Cvit")

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

Table 1: Table continues below

| id | person | wt | ht | sex | age | bmr | E_bmr | energi | Avit | retinol |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 49 | 170 | 2 | 26 | 5,246 | 2.19 | 11,481 | 1,755 | 873.9 |
| 2 | 2 | 55 | 169 | 1 | 20 | 6,351 | 2.88 | 18,310 | 1,209 | 1,085 |
| 3 | 3 | 73 | 168 | 1 | 60 | 7,327 | 1.33 | 9,746 | 816 | 588.9 |
| 4 | 4 | 71 | 173 | 1 | 24 | 7,377 | 2.07 | 15,258 | 3,384 | 1,325 |
| 5 | 5 | 69 | 178 | 1 | 38 | 7,145 | 1.61 | 11,471 | 1,981 | 1,034 |
| 6 | 6 | 66 | 174 | 2 | 27 | 6,247 | 1.52 | 9,525 | 1,176 | 781.7 |

| betacar | Dvit | Evit | B1vit | B2vit | niacin | B6vit | folacin | B12vit | Cvit |
|---|---|---|---|---|---|---|---|---|---|
| 5,280 | 4 | 7.65 | 1.65 | 3.16 | 29.04 | 1.58 | 395.5 | 6.84 | 70.72 |
| 683.9 | 1.87 | 11.25 | 2.19 | 3.02 | 38.5 | 1.68 | 348.4 | 6.57 | 27.63 |
| 1,365 | 1.4 | 8.59 | 1.45 | 2.04 | 33.99 | 1.52 | 299.1 | 3.72 | 57.69 |
| 12,359 | 2.9 | 21.08 | 2.29 | 3.7 | 38.52 | 2.26 | 591.3 | 10.02 | 148.7 |
| 5,700 | 2.26 | 6.01 | 1.93 | 2.58 | 35.22 | 1.63 | 344.9 | 7.43 | 61.47 |
| 2,352 | 2.24 | 6.75 | 1.38 | 2.14 | 22.62 | 1.02 | 207.9 | 4.34 | 42.12 |

## STATISTICHE DESCRITTIVE

Come variabili dipendenti si usa “bmr” e “Ebmr”; come variabili esplicative si usa “wt”, “ht”, “Cvit.”

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

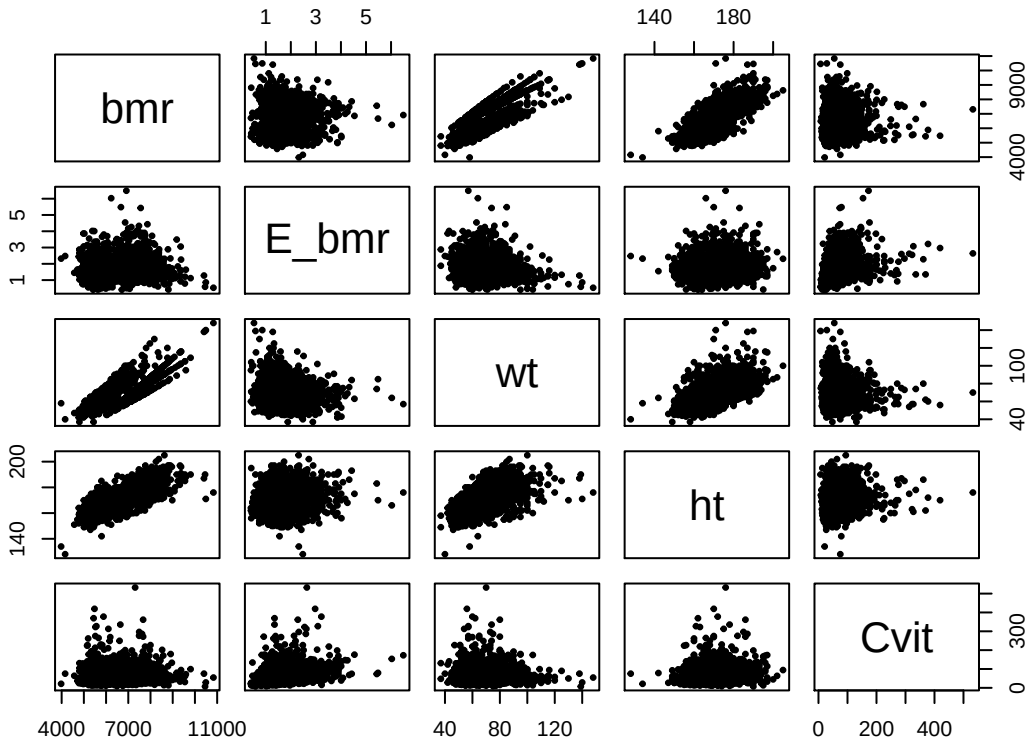
| bmr | E_bmr | wt | ht | Cvit |
|---|---|---|---|---|
| Min. : 3978 | Min. :0.410 | Min. : 37.00 | Min. :128 | Min. : 7.21 |

| bmr | E_bmr | wt | ht | Cvit |
|---|---|---|---|---|
| 1st Qu.: 5636 | 1st Qu.:1.340 | 1st Qu.: 60.00 | 1st Qu.:164 | 1st Qu.: 39.67 |
| Median : 6350 | Median :1.650 | Median : 68.00 | Median :170 | Median : 55.17 |
| Mean : 6542 | Mean :1.745 | Mean : 69.02 | Mean :171 | Mean : 65.10 |
| 3rd Qu.: 7412 | 3rd Qu.:2.050 | 3rd Qu.: 77.00 | 3rd Qu.:178 | 3rd Qu.: 77.21 |
| Max. :10834 | Max. :6.490 | Max. :148.00 | Max. :205 | Max. :532.49 |

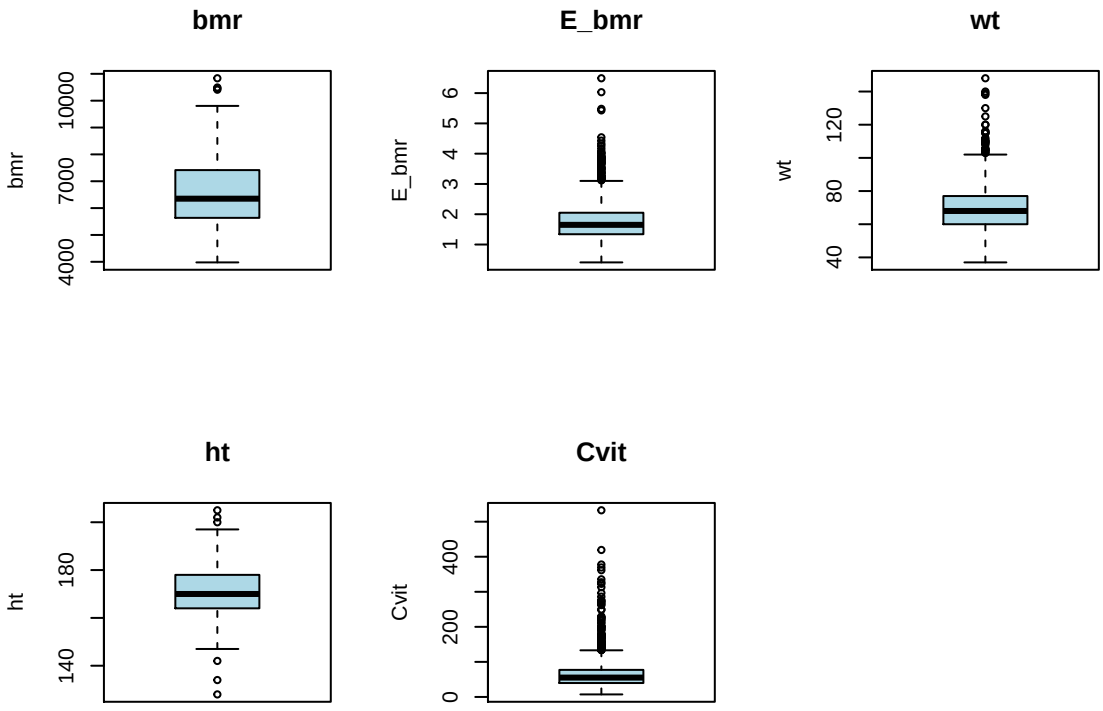
```
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

| | bmr | E_bmr | wt | ht | Cvit |
|---|---|---|---|---|---|
| **bmr** | 1 | 0.06035 | 0.825 | 0.7761 | 0.07362 |
| **E_bmr** | 0.06035 | 1 | -0.1279 | 0.1334 | 0.2993 |
| **wt** | 0.825 | -0.1279 | 1 | 0.5771 | 0.002028 |
| **ht** | 0.7761 | 0.1334 | 0.5771 | 1 | 0.1292 |
| **Cvit** | 0.07362 | 0.2993 | 0.002028 | 0.1292 | 1 |

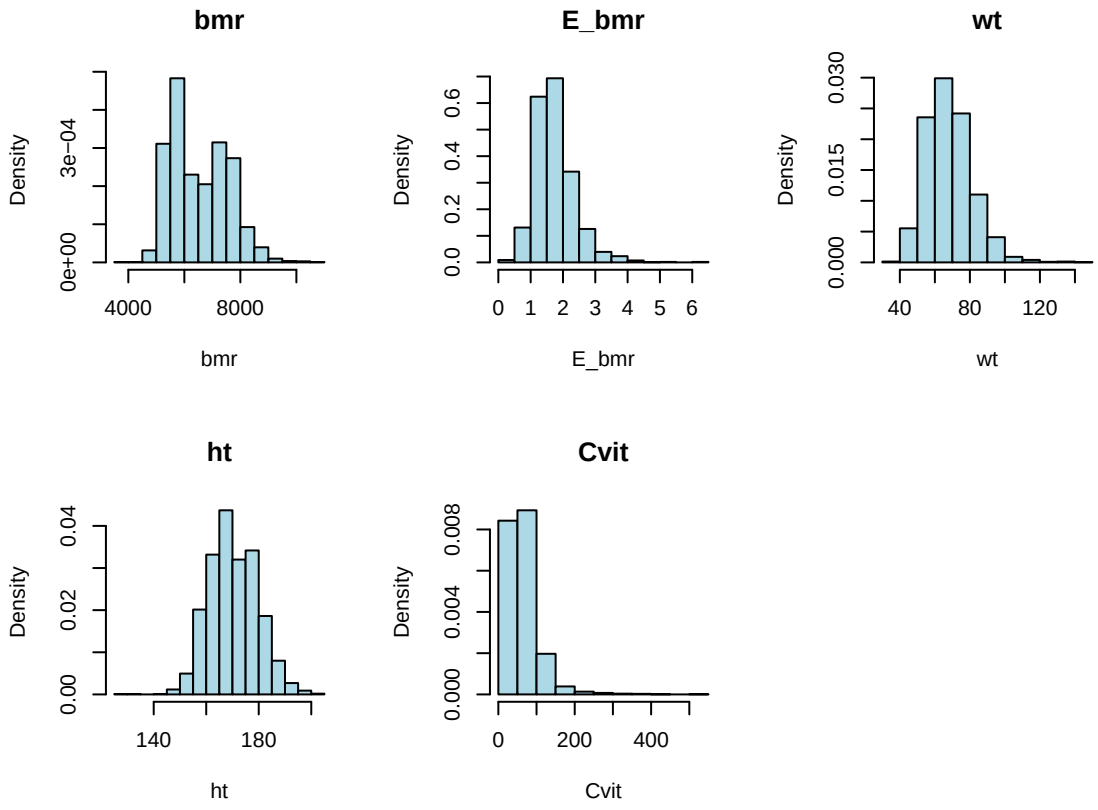
```
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```
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)
}
```




## ESERCIZIO 1

Non appaiono correlazioni di particolare valore. Si propone innanzitutto la regressione multipla di "bmr" sulle 3 variabili esplicative.

```
#-- R CODE
mod1 <- lm(bmr ~ wt + ht + Cvit, d)
pander(summary(mod1),big.mark=",")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | -5,088 | 186.6 | -27.26 | 2.015e-141 |
| **wt** | 44 | 0.869 | 50.63 | 0 |
| **ht** | 50.11 | 1.266 | 39.58 | 9.744e-260 |
| **Cvit** | 0.3634 | 0.2278 | 1.595 | 0.1109 |

Table 6: Fitting linear model: bmr ~ wt + ht + Cvit

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 2224 | 440.7 | 0.8158 | 0.8155 |

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

Table 7: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **wt** | 1 | 1.593e+09 | 1.593e+09 | 8,202 | 0 |
| **ht** | 1 | 3.16e+08 | 3.16e+08 | 1,627 | 2.661e-267 |
| **Cvit** | 1 | 494,204 | 494,204 | 2.544 | 0.1109 |
| **Residuals** | 2,220 | 431,250,153 | 194,257 | NA | NA |

Il fitting è molto elevato e "wt" e "ht" sono significative. Tuttavia gli errori sono eteroschedatici come si vede dal grafici residui-predetti e residui-variabile esplicativa ht e dal test di White.

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

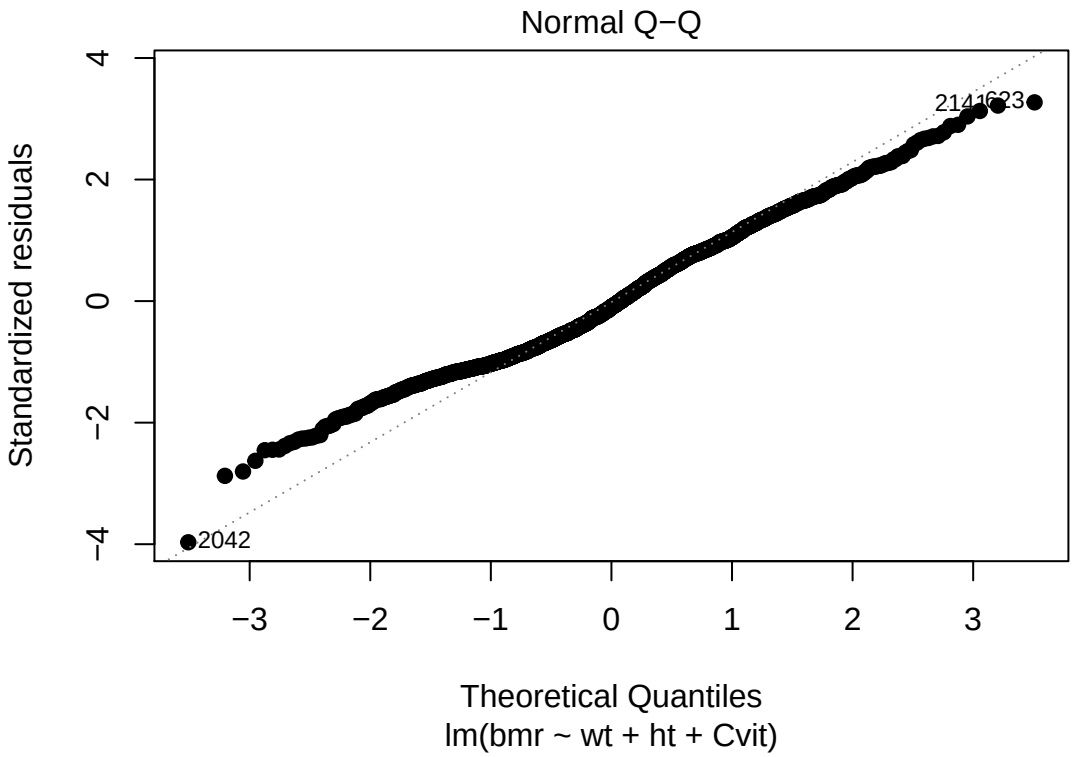
| Test.statistic | P.value |
|---|---|
| 39.15 | 3.154e-09 |

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

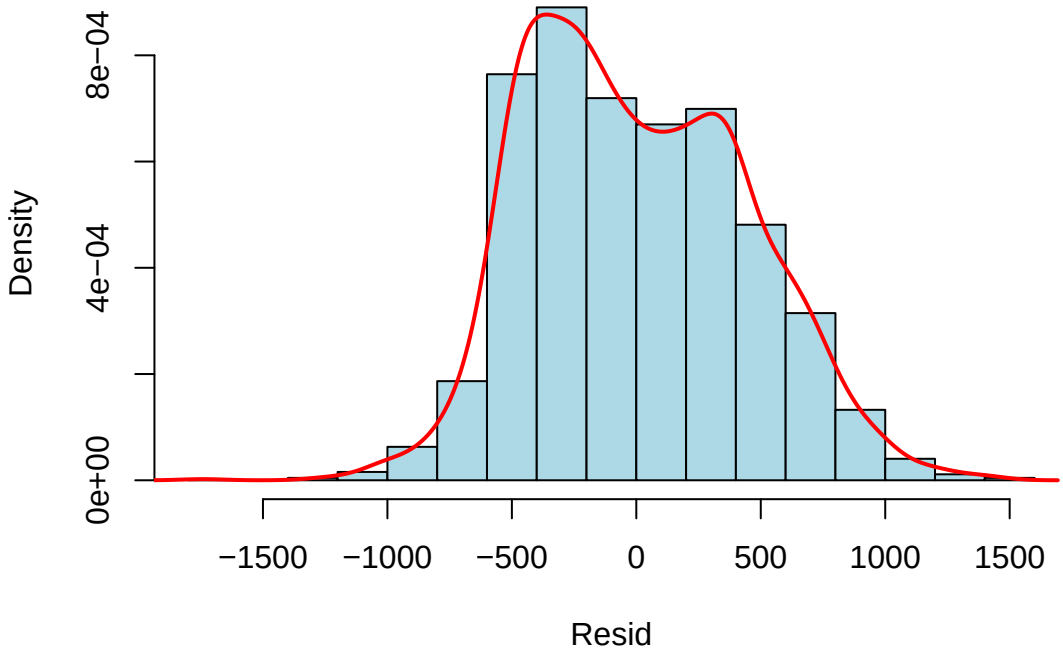
Table 9: Durbin-Watson test: `mod1`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 1.945 | 0.09694 | true autocorrelation is greater than 0 |

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

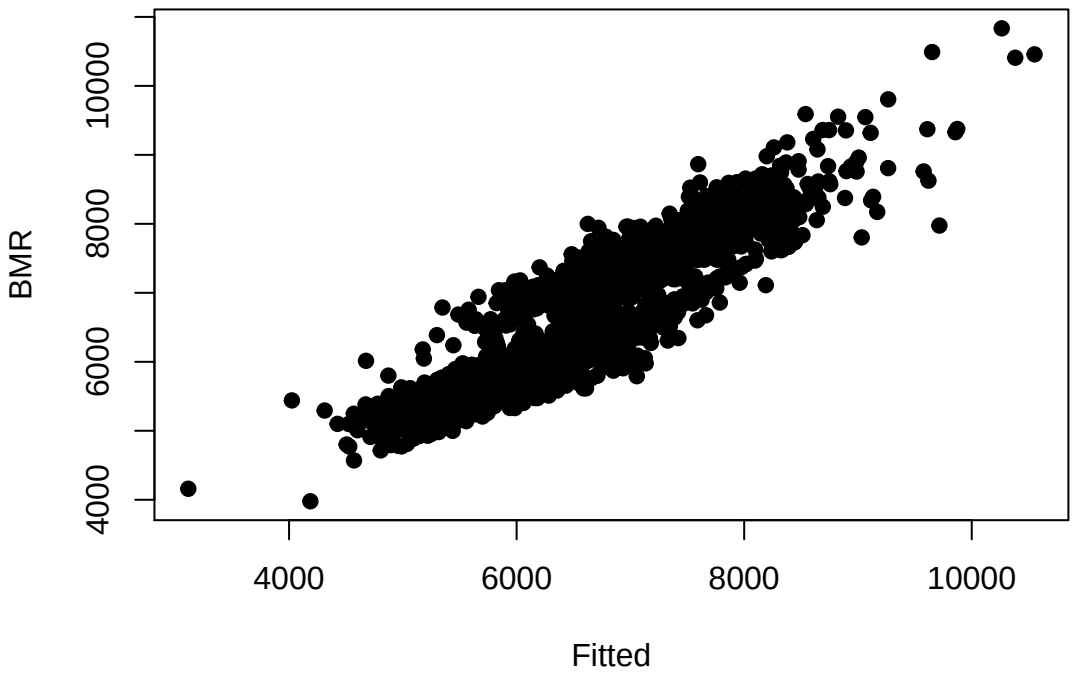



```
hist(resid(mod1),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1)),col=2,lwd=2)
```

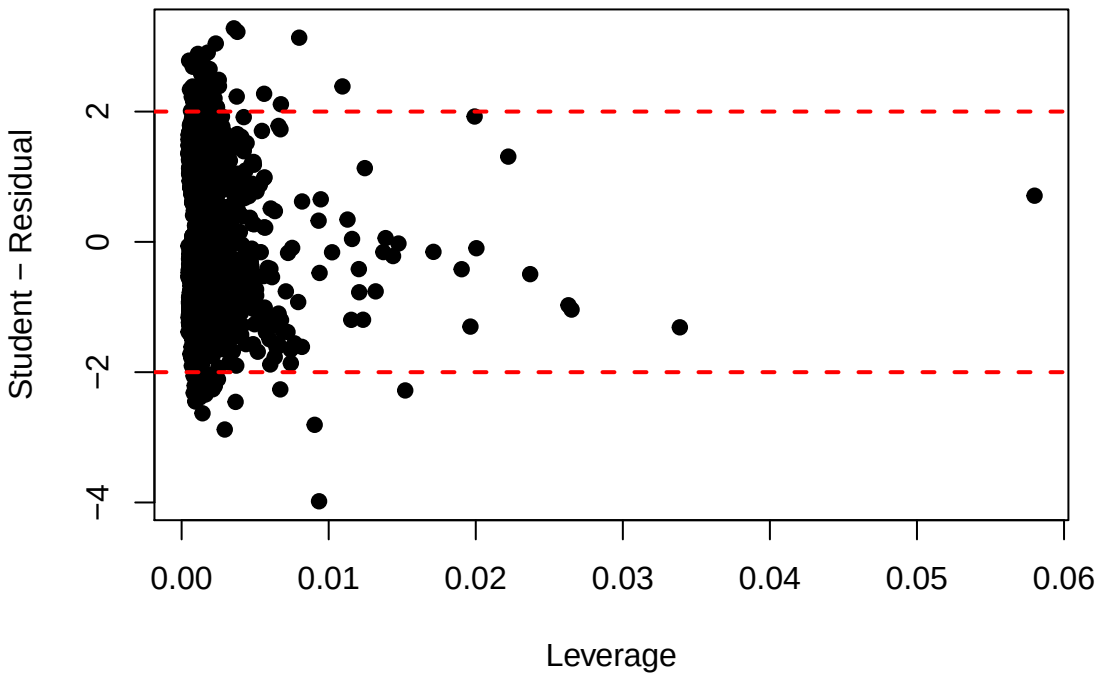



Esistono anche diversi outlier come si vede dai grafici inerenti le misure che analizzano tali outlier. La distribuzione dei residui appare comunque normale.

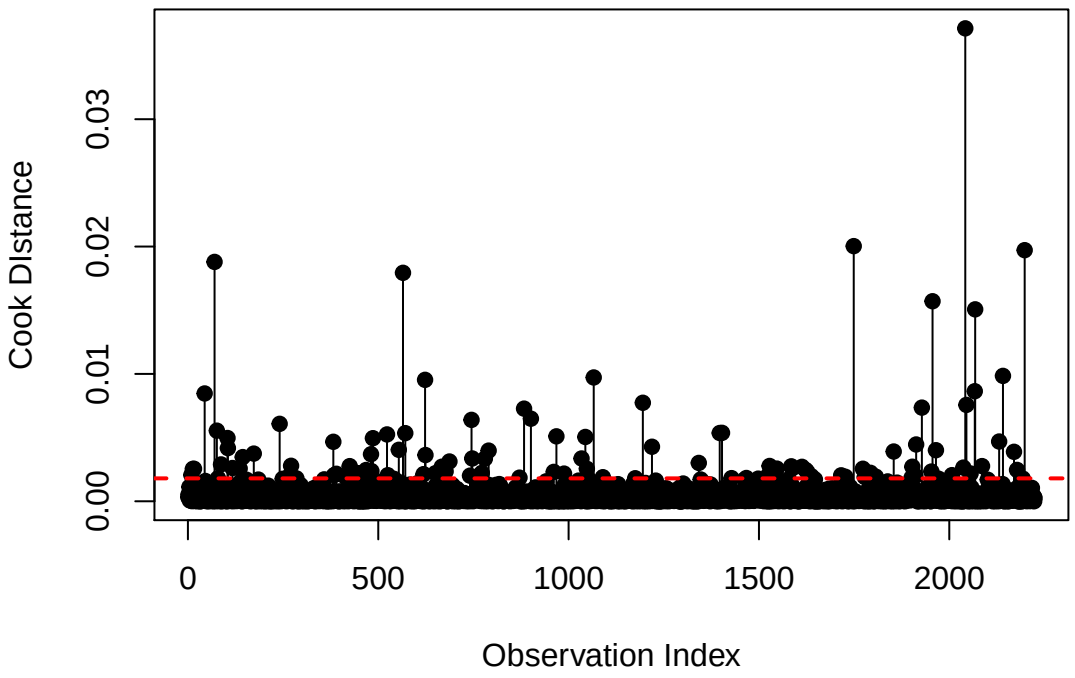
```
#-- R CODE
plot(fitted(mod1),d$bmr,pch=19,xlab="Fitted",ylab="BMR")
```




```
plot(hatvalues(mod1),rstudent(mod1),pch=19,xlab="Leverage",ylab="Student - Residual")
abline(h=2,col=2,lty=2,lwd=2)
abline(h=-2,col=2,lty=2,lwd=2)
```




```
plot(cooks.distance(mod1),pch=19,xlab="Observation Index",ylab="Cook DIstance",type="h")
points(cooks.distance(mod1),pch=19)
abline(h=4/nrow(d),col=2,lty=2,lwd=2)
```




Si consideri allora la seconda regressione multipla di "E_bmr" su "wt", "ht", "Cwit."

```
#-- R CODE
mod2 <- lm(E_bmr ~ wt + ht + Cvit, d)
pander(summary(mod2),big.mark=",")
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | -0.6096 | 0.2412 | -2.528 | 0.01154 |
| **wt** | -0.01299 | 0.001123 | -11.57 | 4.359e-30 |
| **ht** | 0.0175 | 0.001636 | 10.7 | 4.472e-26 |
| **Cvit** | 0.003958 | 0.0002944 | 13.44 | 1.156e-39 |

Table 11: Fitting linear model: E_bmr ~ wt + ht + Cvit

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 2224 | 0.5696 | 0.1499 | 0.1488 |

```
pander(anova(mod2),big.mark=",")
```

Table 12: Analysis of Variance Table

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **wt** | 1 | 13.87 | 13.87 | 42.75 | 7.686e-11 |
| **ht** | 1 | 54.54 | 54.54 | 168.1 | 4.132e-37 |
| **Cvit** | 1 | 58.63 | 58.63 | 180.7 | 1.156e-39 |
| **Residuals** | 2,220 | 720.3 | 0.3245 | NA | NA |

Nonostante le 3 variabili esplicative siano significative il fitting è molto scadente. Inoltre gli errori sono eteroschedastici ed esistono anche diversi outlier che risultano essere tuttavia incorrelati.

```
#-- R CODE
pander(white.test(mod2),big.mark=",")
```

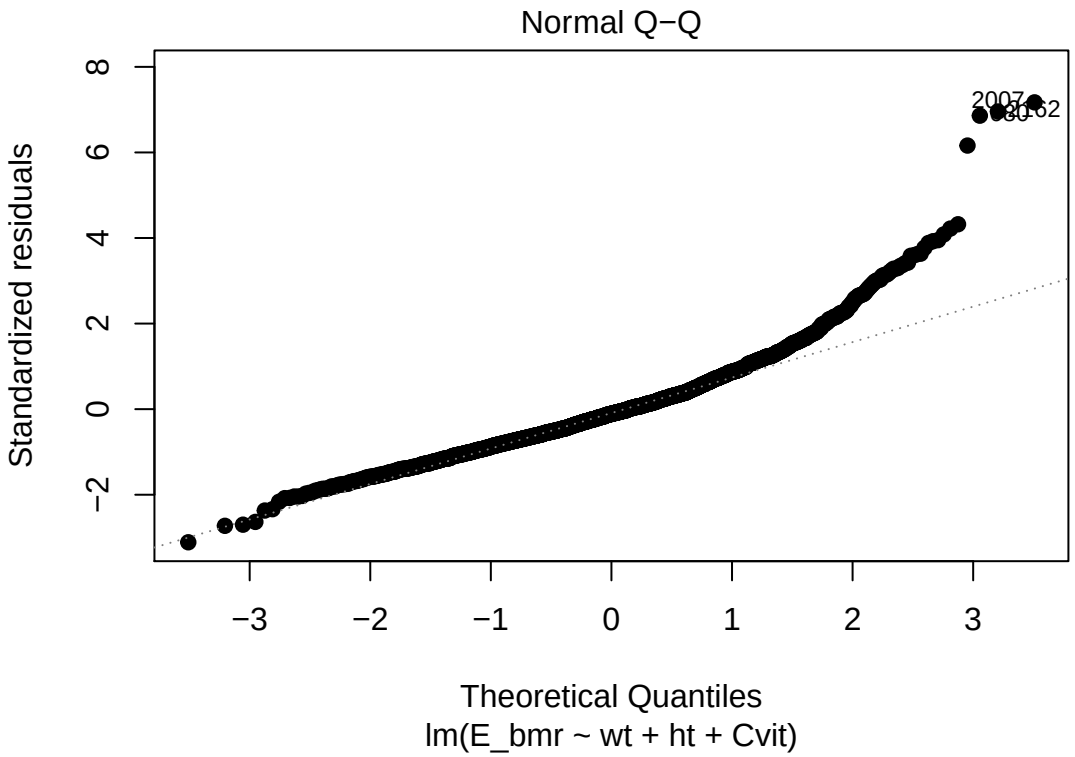
| Test.statistic | P.value |
|---|---|
| 55.88 | 7.33e-13 |

```
pander(dwtest(mod2),big.mark=",")
```

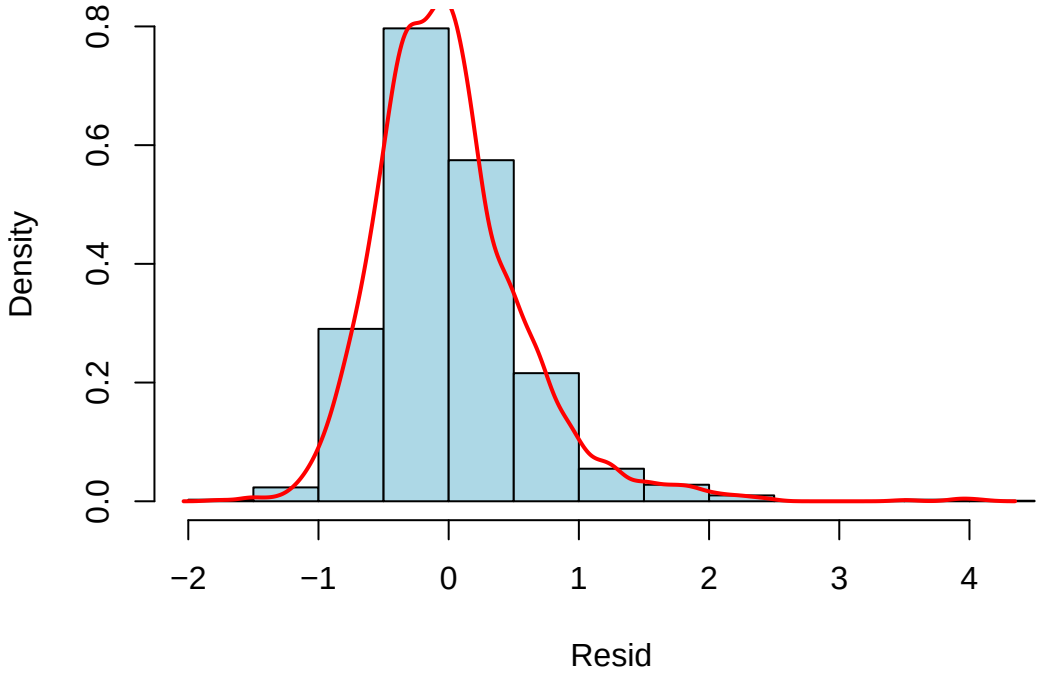
Table 14: Durbin-Watson test: `mod2`

| Test statistic | P value | Alternative hypothesis |
|---|---|---|
| 1.956 | 0.1473 | true autocorrelation is greater than 0 |

```
plot(mod2,which=2,pch=19)
```




```
hist(resid(mod2),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod2)),col=2,lwd=2)
```




Prima di proseguire si dovrebbe a questo punto eliminare l'eteroschedasticità degli errori dividendo variabile dipendente, variabili esplicative, errori medesimi per lo scarto quadratico degli errori stessi. Inoltre si dovrebbero eliminare gli outlier individuati nelle due regressioni multiple. Si prosegue invece senza operare queste trasformazioni per potere mostrare, a puri scopi didattici, il nesso tra regressione multipla OLS e regressione multivariata OLS.

Si propone quindi la regressione multivariata delle due variabili dipendenti sulle tre variabili esplicative. Il modello multivariato con le stesse variabili esplicative sotto il profilo descrittivo è accostamento di due regressioni multiple che vengono risolte l'una indipendentemente dall' altra perciò gli $R^2$ e le stime dei parametri usando il test sono identici. Il test F conferma i risultati perché la f non è altro che il quadrato della t.

```
#-- R CODE
mod3 <- lm(cbind(E_bmr,bmr) ~ wt + ht + Cvit, d)

#-- calcolo correlazione parziale tra "Life.expectancy" e "Unemployment"
#-- al netto delle altre variabili
library(ppcor)
```

```
## Warning: package 'ppcor' was built under R version 3.4.3
```

```
pander(pcor.test(d$E_bmr,d$bmr,d[,c("wt","ht","Cvit")]))
```

```
##
## ----------------------------------------------------------
##  estimate    p.value    statistic    n     gp   Method
## ---------- ----------- ----------- ------ ---- ---------
```

```
##   0.1741    1.438e-16     8.326     2224   3    pearson
## ---------------------------------------------------------
```

```
summary(mod3)
```

```
## Response E_bmr :
##
## Call:
## lm(formula = E_bmr ~ wt + ht + Cvit, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7616 -0.3704 -0.0620  0.2664  4.0759
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.6096482  0.2411713  -2.528   0.0115 *
## wt          -0.0129909  0.0011232 -11.566   <2e-16 ***
## ht           0.0175036  0.0016363  10.697   <2e-16 ***
## Cvit         0.0039582  0.0002944  13.443   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5696 on 2220 degrees of freedom
## Multiple R-squared:  0.1499, Adjusted R-squared:  0.1488
## F-statistic: 130.5 on 3 and 2220 DF,  p-value: < 2.2e-16
##
##
## Response bmr :
##
## Call:
## lm(formula = bmr ~ wt + ht + Cvit, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1740.77  -350.15   -43.05   335.08  1438.80
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5087.6485   186.6061 -27.264   <2e-16 ***
## wt             43.9965     0.8690  50.627   <2e-16 ***
## ht             50.1144     1.2661  39.583   <2e-16 ***
## Cvit            0.3634     0.2278   1.595    0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 440.7 on 2220 degrees of freedom
## Multiple R-squared:  0.8158, Adjusted R-squared:  0.8155
## F-statistic:  3277 on 3 and 2220 DF,  p-value: < 2.2e-16
```

```
pander(manova(mod3),big.mark=",")
```

```
##
## ------------------------------------------------------------------------
##              Df     Pillai    approx F   num Df   den Df     Pr(>F)
```

```
## ---------------- ------- --------- ---------- -------- -------- ------------
##     **wt**          1      0.797     4,356       2      2,219         0
##
##     **ht**          1     0.4283     831.2       2      2,219     3.844e-270
##
##    **Cvit**         1     0.07549     90.6       2      2,219     1.504e-38
##
##  **Residuals**    2,220     NA          NA        NA       NA          NA
## ---------------------------------------------------------------------------
```

```
Anova(mod3, type="III")
```

```
##
## Type III MANOVA Tests: Pillai test statistic
##             Df test stat approx F num Df den Df    Pr(>F)
## (Intercept)  1   0.25212   374.03      2   2219 < 2.2e-16 ***
## wt           1   0.57400  1494.98      2   2219 < 2.2e-16 ***
## ht           1   0.41606   790.52      2   2219 < 2.2e-16 ***
## Cvit         1   0.07549    90.60      2   2219 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si costruiscono quindi i test multivariati inerenti le tre variabile esplicative. Poiché l'ipotesi testata è che i parametri relativi alle 3 variabili risultino nulle per entrambe le equazioni, tutti i parametri risultano significativi per tutti i test, come era prevedibile dati i risultati delle regressioni multiple.

```
#-- R CODE
summary(manova(cbind(E_bmr,bmr) ~ wt, data = d))
```

```
##             Df  Pillai approx F num Df den Df    Pr(>F)
## wt           1 0.71231   2749.6      2   2221 < 2.2e-16 ***
## Residuals 2222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(manova(cbind(E_bmr,bmr) ~ ht, data = d))
```

```
##             Df  Pillai approx F num Df den Df    Pr(>F)
## ht           1 0.60991   1736.3      2   2221 < 2.2e-16 ***
## Residuals 2222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(manova(cbind(E_bmr,bmr) ~ Cvit, data = d))
```

```
##             Df   Pillai approx F num Df den Df    Pr(>F)
## Cvit         1 0.092688   113.44      2   2221 < 2.2e-16 ***
## Residuals 2222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Infine effettuiamo un test congiunto multivariato per tutte le variabile esplicative congiuntamente. Come deve essere per quanto visto viene respinta l'ipotesi nulla di non significatività di almeno una delle variabili.

```r
#-- R CODE
summary(manova(cbind(E_bmr,bmr) ~ wt + ht + Cvit, data = d))
```

```
##             Df  Pillai approx F num Df den Df    Pr(>F)
## wt           1 0.79699   4355.7      2   2219 < 2.2e-16 ***
## ht           1 0.42829    831.2      2   2219 < 2.2e-16 ***
## Cvit         1 0.07549     90.6      2   2219 < 2.2e-16 ***
## Residuals 2220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```