MULTI 1 - Data set: COUNTRIES

INTRODUZIONE

In questo dataset sono 12 variabili su 38 osservazioni:

1. REGION: regione della country
2. AREA: area della country (Km^2)
3. IRRIGATED: area di campi irrigati (Km^2)
4. POPULATION: popolazione in milioni di persone
5. UNDER.14: % di popolazione con meno di 14 anni
6. LIFE.EXPECTANCY: speranza di vita alla nascita in anni
7. LITERACY.RATE: tasso di alfabetismo
8. UNEMPLOYMENT: tasso di disoccupazione
9. ISPS/MILLION: numero di ISPs per milione di persone
10. TVs/PERSON: numero di televisioni per persona
11. RAILWAYS: lunghezza in km della rete ferroviaria
12. AIRPORTS: numero di aeroporti

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))
}
```

```r
#-- import dei dati
ABSOLUTE_PATH <- "C:\\Users\\sbarberis\\Dropbox\\MODELLI STATISTICI"
d <- read.csv(paste0(ABSOLUTE_PATH,"\\esercizi (5)  copia\\1.mult\\countries.txt"),sep="\t")

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("Life.expectancy","Unemployment","Literacy.Rate","ISPs.million","Irrigated","Under.14")

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

Table 1: Table continues below

X	Region	Area	Irrigated	Population	Under.14
Argentina	South.America	2,766,890	17,000	37.4	26.5
Australia	Oceania	7,686,850	21,070	19.4	20.6
Bangladesh	Asia	144,000	31,000	131.3	35
Brazil	South.America	8,511,965	28,000	174.5	28.6
Bolivia	South.America	1,098,580	1,750	8.3	38.5
Cameroon	Africa	475,440	210	15.8	42.4

Table 2: Table continues below

Life.expectancy	Literacy.Rate	Unemployment	ISPs.million	Tvs.person
75.2	96.2	15	0.88	0.21
79.87	100	6.4	13.61	5.36
60.54	56	35.2	0.08	0.01
63.24	83.3	7.1	0.29	0.21
64.1	83.1	11.4	1.08	0.11
54.6	63.4	30	0.06	0.03

Railways	Airports
33,744	1,359
33,819	411
2,745	18
30,539	3,264
3,691	1,093
1,104	49

STATISTICHE DESCRITTIVE

Si vuole studiare la dipendenza delle variabili “life_expectancy” e “Unemployment” da “ISPs_million”, “irrigated”, “Under_14”, “Literacy_Rate”. Si propongono dapprima le statistiche descrittive, a seguire le matrici di correlazione tra variabili dipendenti, tra variabili esplicative e tra variabili dipendenti.

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

Table 4: Table continues below

Life.expectancy	Unemployment	Literacy.Rate	ISPs.million
Min. :37.10 1st Qu.:60.78 Median :71.70 Mean :67.58 3rd Qu.:77.45 Max. :80.80	Min. : 1.80 1st Qu.: 5.55 Median : 9.75 Mean :15.10 3rd Qu.:20.00 Max. :50.00	Min. : 38.00 1st Qu.: 78.30 Median : 87.00 Mean : 82.85 3rd Qu.: 97.75 Max. :100.00	Min. : 0.000 1st Qu.: 0.230 Median : 0.920 Mean : 3.831 3rd Qu.: 2.172 Max. :29.130

Irrigated	Under.14
Min. : 10.0 1st Qu.: 632.5 Median : 5212.0 Mean : 33385.1 3rd Qu.: 25592.5 Max. :498720.0	Min. :14.17 1st Qu.:18.90 Median :29.45 Mean :29.22 3rd Qu.:38.17 Max. :47.40

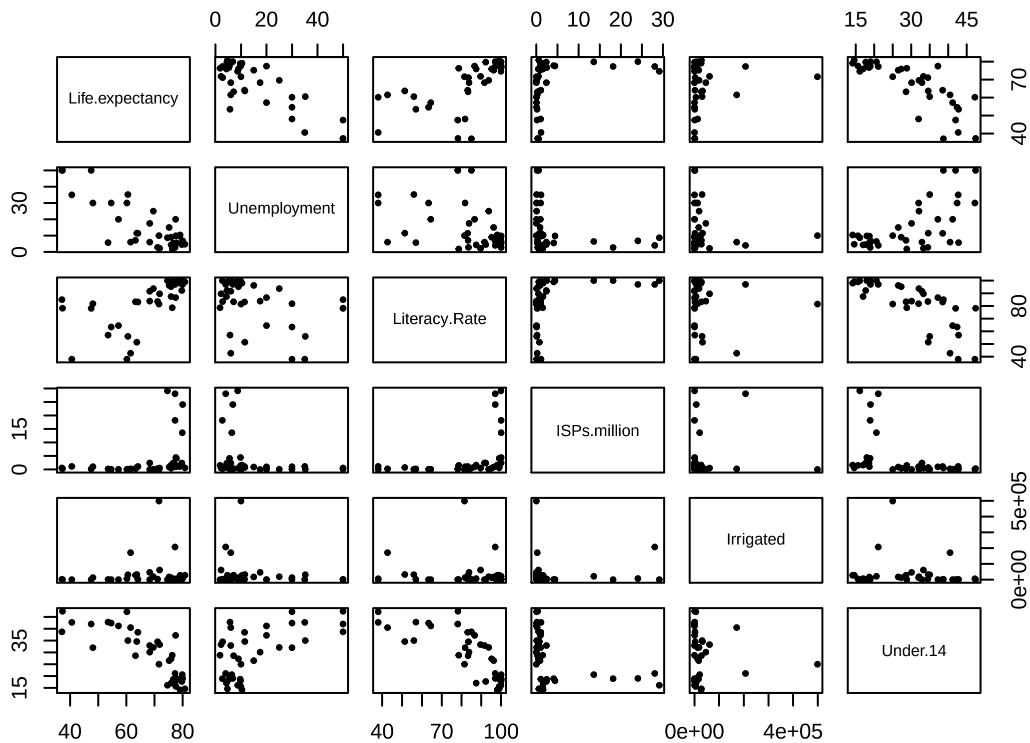
```r
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

Table 6: Table continues below

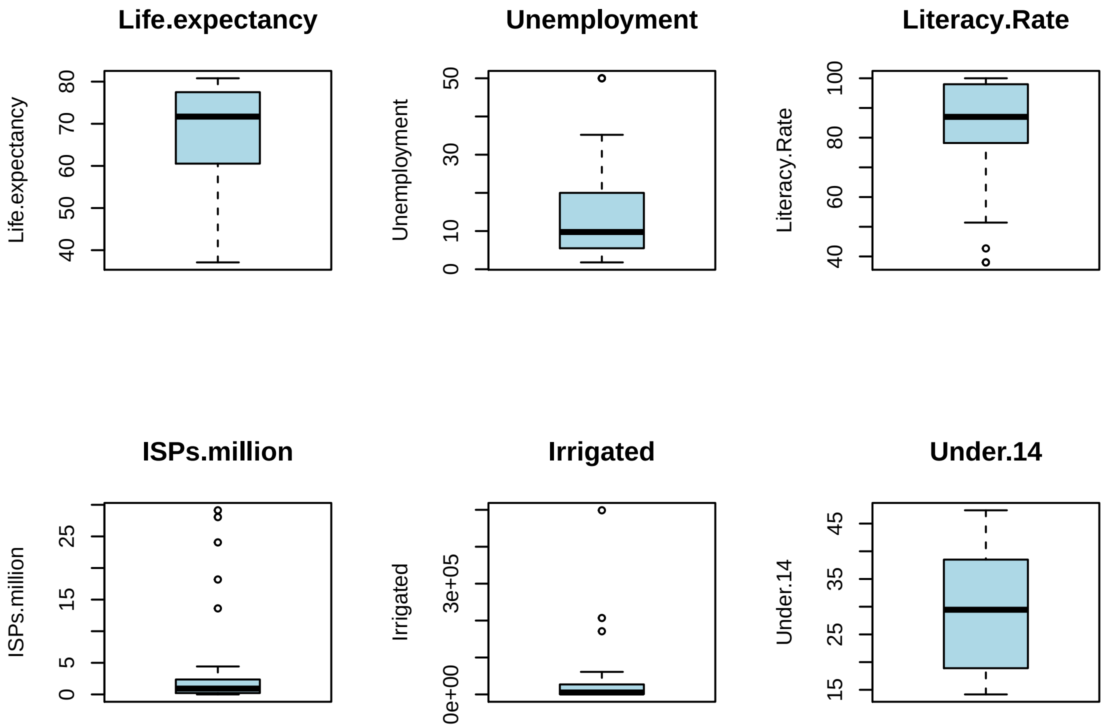
	Life.expectancy	Unemployment	Literacy.Rate
Life.expectancy	1	-0.8153	0.6385
Unemployment	-0.8153	1	-0.3976
Literacy.Rate	0.6385	-0.3976	1
ISPs.million	0.3473	-0.2932	0.378
Irrigated	0.09763	-0.1579	-0.06472
Under.14	-0.8089	0.6309	-0.7723

	ISPs.million	Irrigated	Under.14
Life.expectancy	0.3473	0.09763	-0.8089
Unemployment	-0.2932	-0.1579	0.6309
Literacy.Rate	0.378	-0.06472	-0.7723
ISPs.million	1	0.06517	-0.4452
Irrigated	0.06517	1	-0.07298
Under.14	-0.4452	-0.07298	1

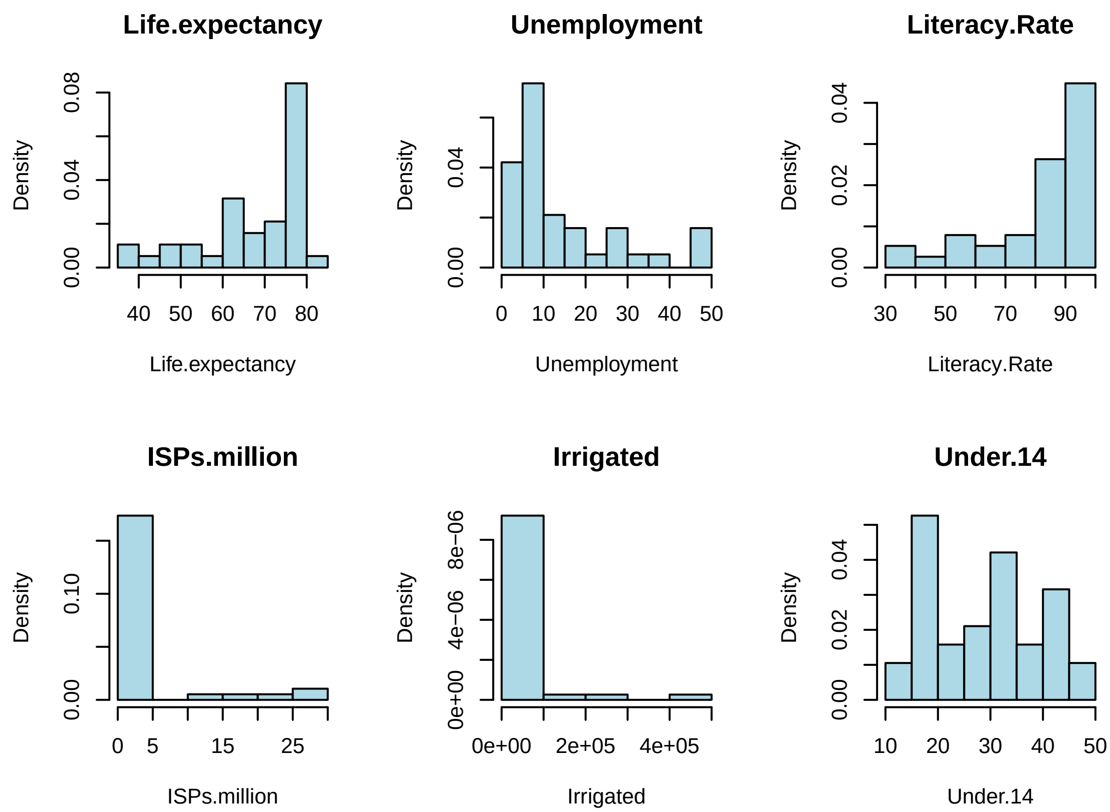
```r
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




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




```r
par(mfrow=c(2,3))
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```



Non esistono correlazioni particolarmente forti che facciano pensare a collinearità o legami di dipendenza lineare perfetta. Si propongano ora le regressioni uni variate cominciando ora la variabile dipendente "life_expentancy".

ESERCIZIO 1

```r
#-- R CODE
mod1 <- lm(Life.expectancy ~ ISPs.million + Irrigated + Under.14 + Literacy.Rate, d)
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	92.61	14.62	6.333	3.634e-07
ISPs.million	-0.03255	0.1861	-0.1749	0.8622
Irrigated	6.513e-06	1.493e-05	0.4363	0.6655
Under.14	-0.9566	0.2081	-4.598	6.006e-05
Literacy.Rate	0.03402	0.1125	0.3024	0.7642

Table 9: Fitting linear model: Life.expectancy ~ ISPs.million + Irrigated + Under.14 + Literacy.Rate

Observations	Residual Std. Error	R^2	Adjusted R^2
38	7.867	0.657	0.6155

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

Table 10: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ISPs.million	1	718.1	718.1	11.6	0.001748
Irrigated	1	33.63	33.63	0.5435	0.4662
Under.14	1	3,155	3,155	50.98	3.522e-08
Literacy.Rate	1	5.66	5.66	0.09145	0.7642
Residuals	33	2,042	61.89	NA	NA

L'aspettativa alla nascita dipende solo da "Under 14" che è l'unica variabile significativa e il fitting è elevato ($R^2 = 0.6570$). Si verifica dai grafici e dal test di White che gli errori sono omoschedastici. Dal Q-Q plot e dalla distribuzione dei residui la distribuzione appare normale

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

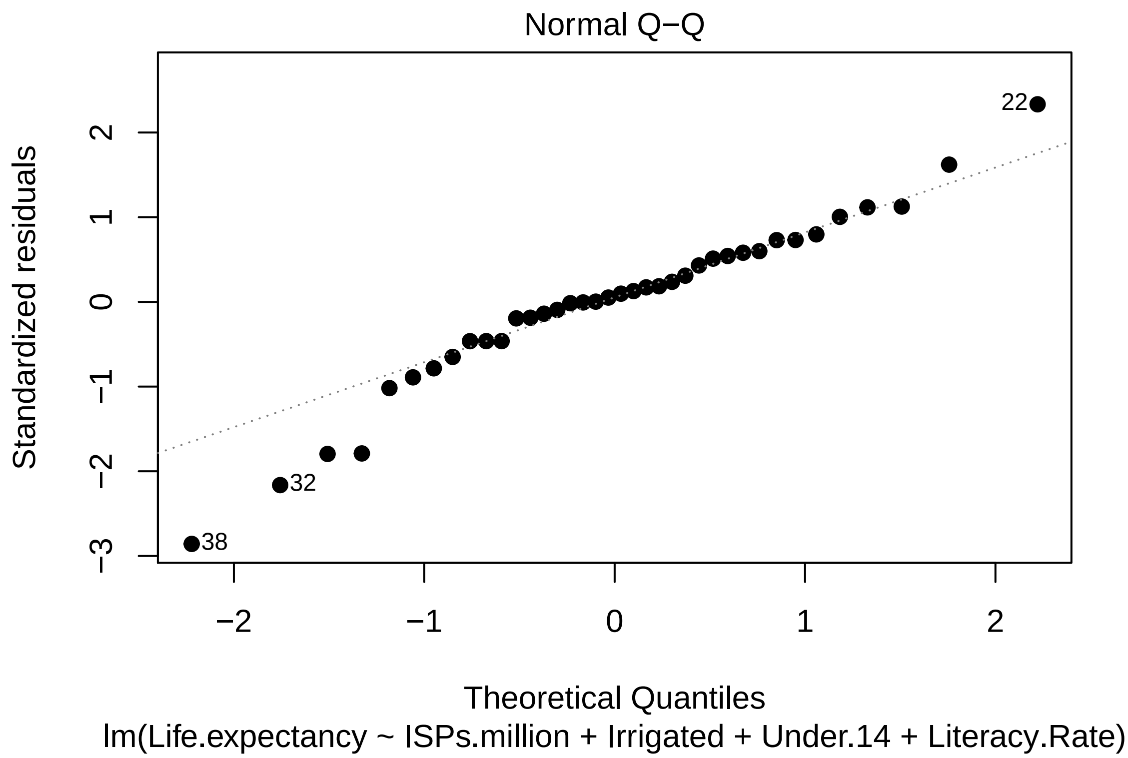
Test.statistic	P.value
6.533	0.03813

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

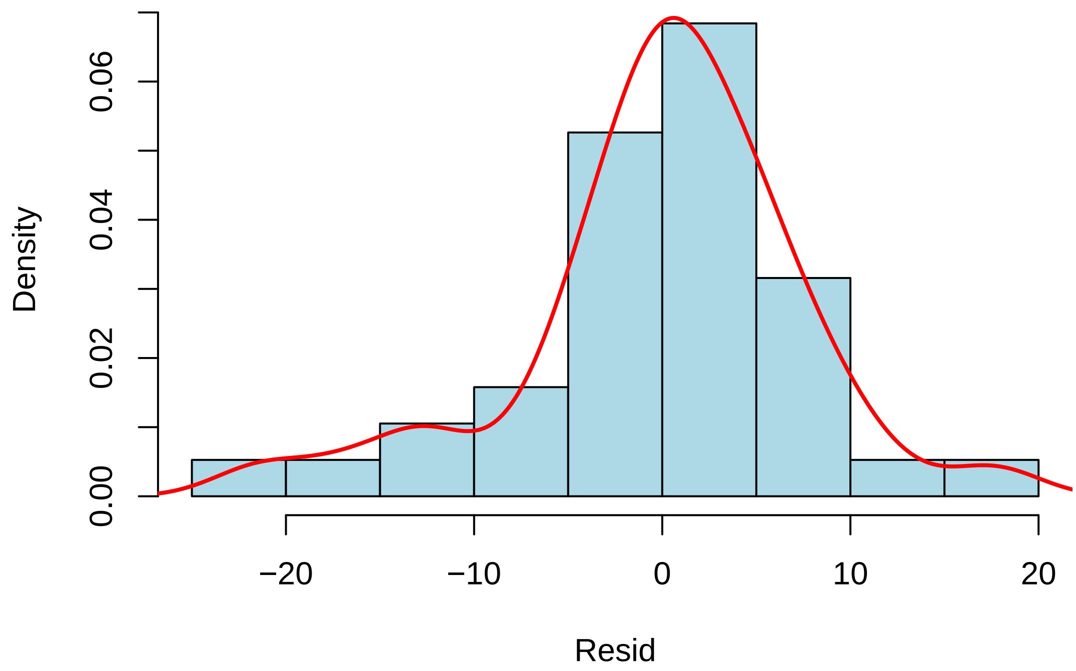
Table 12: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.848	0.3211	true autocorrelation is greater than 0

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

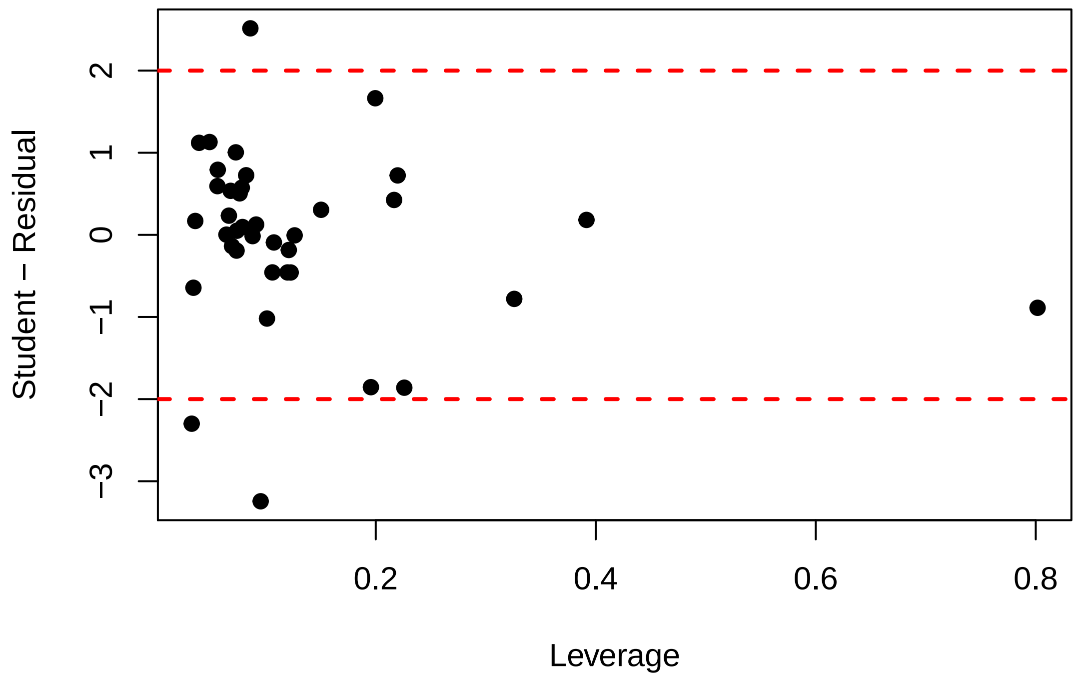



```r
hist(resid(mod1),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod1)),col=2,lwd=2)
```

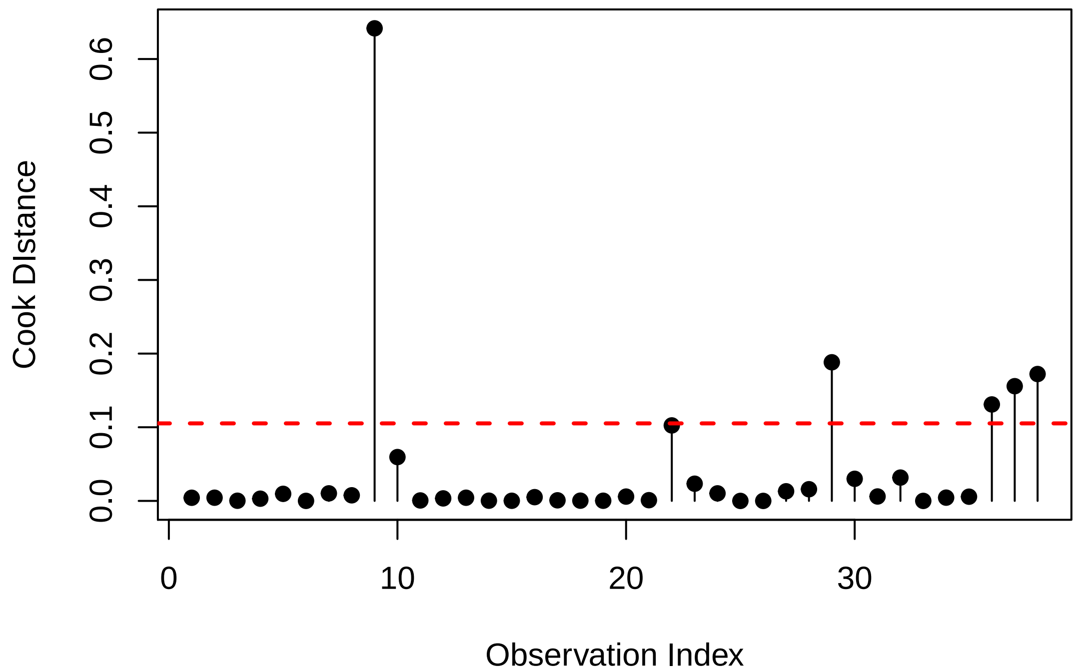


Si osserva qualche outlier che andrebbe eliminato:

```
#-- R CODE
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 passa ora alla regressione multipla dove la variabile dipendente è "unemployment." Anche in questo caso l'unica variabile significativa rimane "under 14" con un discreto fitting. Gli errori sono anche in questo caso omoschedastici e gli errori sono anche non correlati con distribuzione normale.

```
#-- R CODE
mod2 <- lm(Unemployment ~ ISPs.million + Irrigated + Under.14 + Literacy.Rate, d)
pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.25	20.91	-1.303	0.2015
ISPs.million	-0.03646	0.266	-0.1371	0.8918
Irrigated	-1.388e-05	2.134e-05	-0.6505	0.5199
Under.14	1.045	0.2974	3.515	0.001301
Literacy.Rate	0.1497	0.1608	0.931	0.3586

Table 14: Fitting linear model: Unemployment ~ ISPs.million + Irrigated + Under.14 + Literacy.Rate

Observations	Residual Std. Error	R^2	Adjusted R^2
38	11.25	0.4257	0.3561

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

Table 15: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ISPs.million	1	624.8	624.8	4.94	0.03321
Irrigated	1	140.6	140.6	1.112	0.2994
Under.14	1	2,220	2,220	17.55	0.0001962
Literacy.Rate	1	109.6	109.6	0.8667	0.3586
Residuals	33	4,174	126.5	NA	NA

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

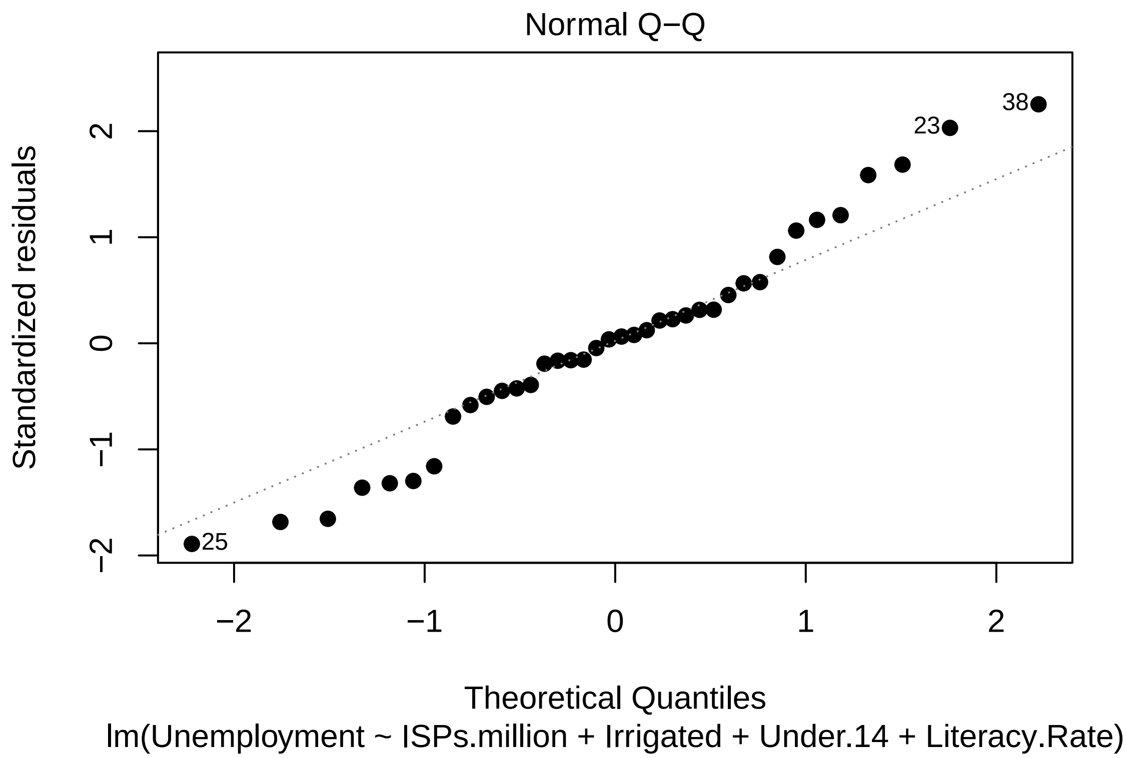
Test.statistic	P.value
12.94	0.00155

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

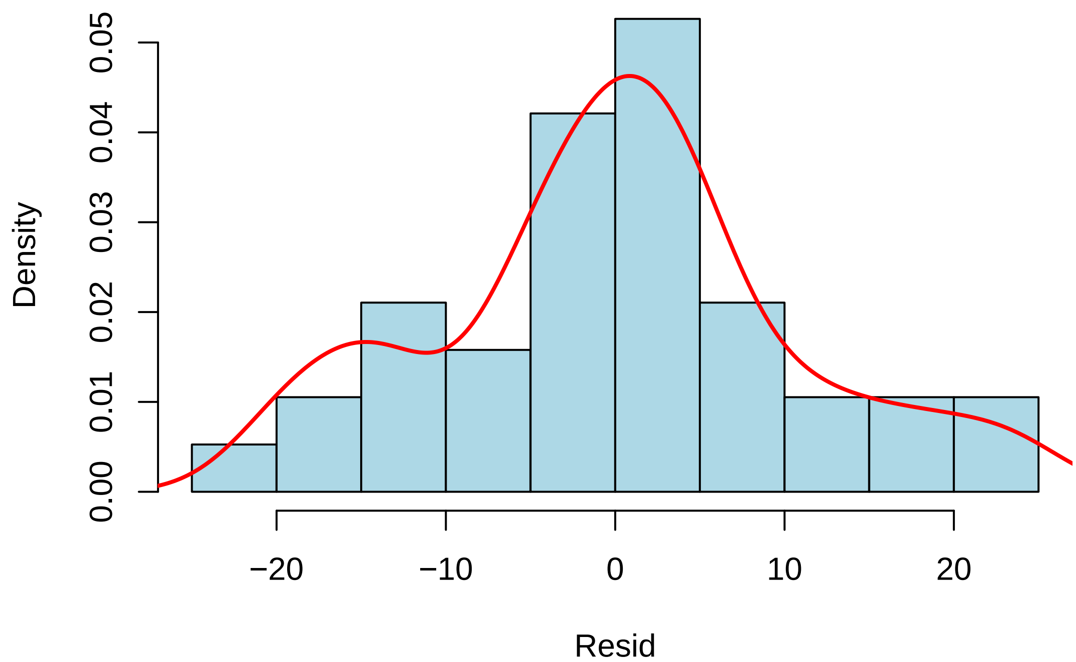
Table 17: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
1.707	0.1825	true autocorrelation is greater than 0

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

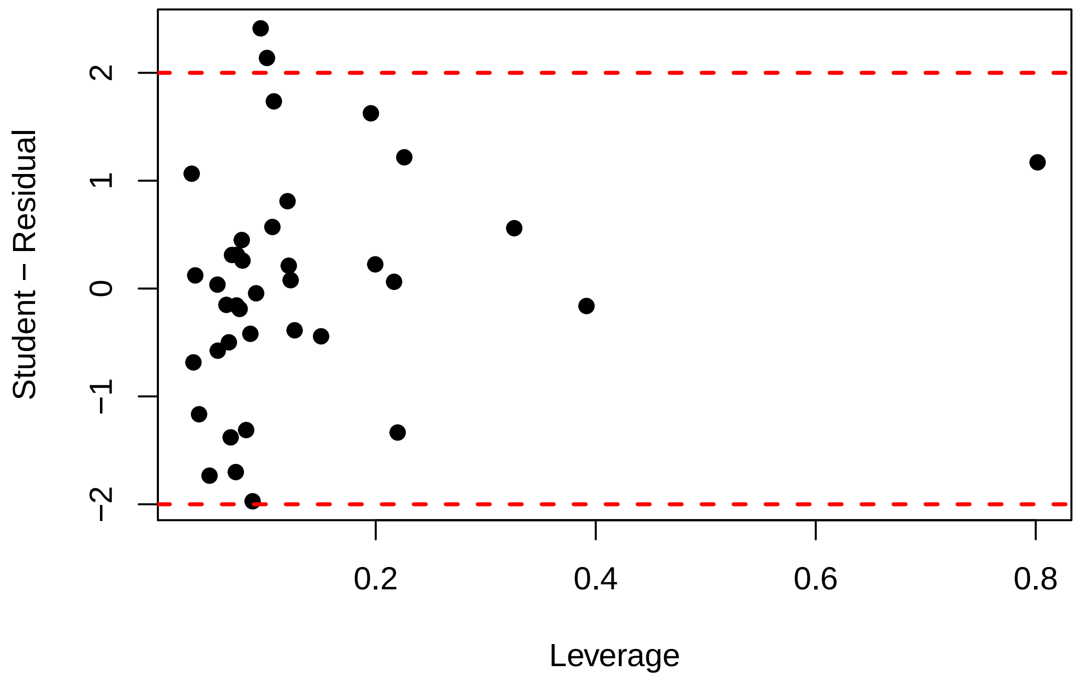



```r
hist(resid(mod2),col="lightblue",freq=F,xlab="Resid",main="")
lines(density(resid(mod2)),col=2,lwd=2)
```

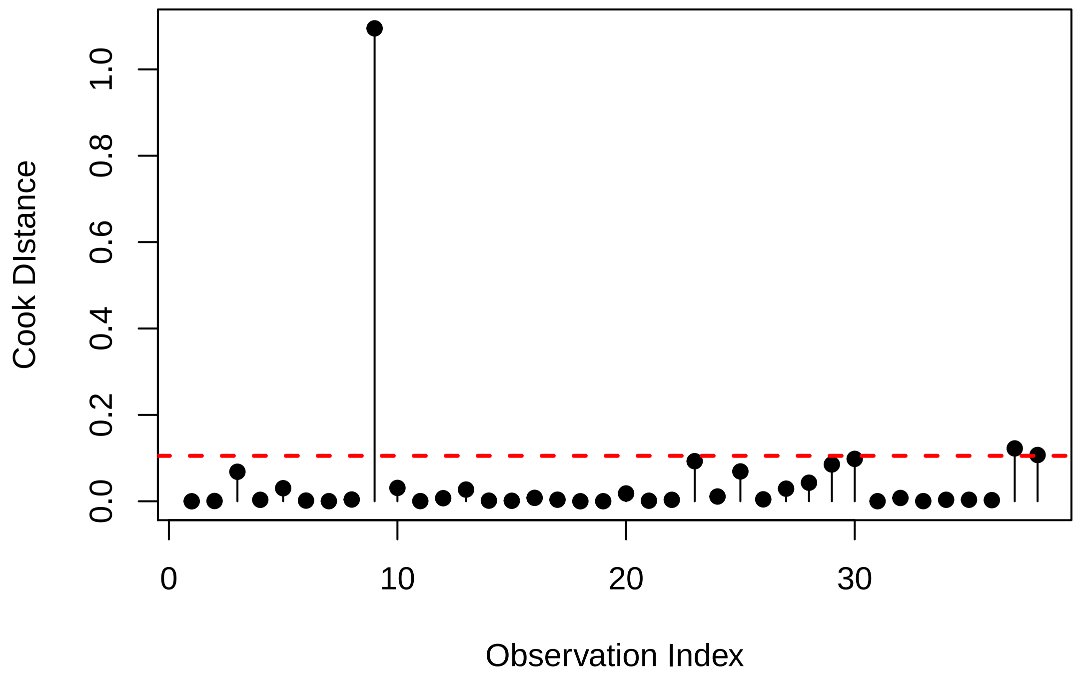


Si osserva anceh in questo caso che qualche outlier che andrebbe eliminato:

```
#-- R CODE
plot(hatvalues(mod2),rstudent(mod2),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(mod2),pch=19,xlab="Observation Index",ylab="Cook DIstance",type="h")
points(cooks.distance(mod2),pch=19)
abline(h=4/nrow(d),col=2,lty=2,lwd=2)
```



Rinunciando a eliminare gli outlier (provare per esercizio) come sarebbe comunque opportuno si passa ora alla regressione multivariata.

```
#-- R CODE
mod3 <- lm(cbind(Unemployment,Life.expectancy) ~ ISPs.million + Irrigated + Under.14 + Literacy.Rate, 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$Life.expectancy,d$Unemployment,d[,c("ISPs.million","Irrigated","Under.14","Literacy.R
```

```
##
## ---------------------------------------------------------
##  estimate    p.value    statistic   n    gp   Method
## ---------- ----------- ----------- ---- ---- ---------
##  -0.6864    7.408e-06    -5.339     38   4    pearson
## ---------------------------------------------------------
```

```
summary(mod3)
```

```
## Response Unemployment :
##
## Call:
## lm(formula = Unemployment ~ ISPs.million + Irrigated + Under.14 +
```

```
##     Literacy.Rate, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3115  -5.2740   0.5191   5.1450  24.1047
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.725e+01  2.091e+01  -1.303   0.2015
## ISPs.million  -3.646e-02  2.660e-01  -0.137   0.8918
## Irrigated     -1.388e-05  2.134e-05  -0.650   0.5199
## Under.14       1.045e+00  2.974e-01   3.515   0.0013 **
## Literacy.Rate  1.497e-01  1.608e-01   0.931   0.3586
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.25 on 33 degrees of freedom
## Multiple R-squared:  0.4257, Adjusted R-squared:  0.3561
## F-statistic: 6.116 on 4 and 33 DF,  p-value: 0.0008505
##
##
## Response Life.expectancy :
##
## Call:
## lm(formula = Life.expectancy ~ ISPs.million + Irrigated + Under.14 +
##     Literacy.Rate, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3828  -3.3426   0.5599   4.3173  17.5504
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.261e+01  1.462e+01   6.333 3.63e-07 ***
## ISPs.million  -3.255e-02  1.861e-01  -0.175    0.862
## Irrigated      6.513e-06  1.493e-05   0.436    0.665
## Under.14      -9.566e-01  2.081e-01  -4.598 6.01e-05 ***
## Literacy.Rate  3.402e-02  1.125e-01   0.302    0.764
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.867 on 33 degrees of freedom
## Multiple R-squared:  0.657,  Adjusted R-squared:  0.6155
## F-statistic:  15.8 on 4 and 33 DF,  p-value: 2.542e-07
```

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

```
##
## ---------------------------------------------------------------------------
##                 Df   Pillai    approx F   num Df   den Df    Pr(>F)
## ------------------- ---- --------- ---------- -------- -------- -----------
##  **ISPs.million**    1    0.2606     5.638       2        32     0.007987
##
##    **Irrigated**     1    0.0326     0.5391      2        32      0.5885
##
```

```
##    **Under.14**      1    0.6115     25.18       2        32     2.696e-07
##
##  **Literacy.Rate**   1    0.07153    1.233       2        32       0.305
##
##    **Residuals**     33     NA          NA        NA       NA        NA
## ---------------------------------------------------------------------------
```

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

```
##
## Type III MANOVA Tests: Pillai test statistic
##              Df test stat approx F num Df den Df    Pr(>F)
## (Intercept)   1   0.63584  27.9373      2     32 9.563e-08 ***
## ISPs.million  1   0.00469   0.0755      2     32  0.927482
## Irrigated     1   0.01267   0.2053      2     32  0.815503
## Under.14      1   0.39319  10.3674      2     32  0.000338 ***
## Literacy.Rate 1   0.07153   1.2327      2     32  0.304984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il modello multivariato con le stesse variabili esplicative sotto il profilo descrittivo è l'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.

La variabile "under 14", come previsto risulta significativa. "Literacy" non risulta significativa

Si passa ora a verificare ipotesi multiple mediante il test Manova cominciando con le 2 variabili Isps e literacy che risultano congiuntamente non significative:

```r
#-- R CODE
summary(manova(cbind(Life.expectancy, Unemployment) ~ ISPs.million, data = d))
```

```
##              Df Pillai approx F num Df den Df Pr(>F)
## ISPs.million  1 0.1209   2.4067      2     35 0.1049
## Residuals    36
```

```r
Anova(lm(cbind(Life.expectancy, Unemployment) ~ ISPs.million, data = d),type="III")
```

```
##
## Type III MANOVA Tests: Pillai test statistic
##              Df test stat approx F num Df den Df Pr(>F)
## (Intercept)   1   0.98975  1690.06      2     35 <2e-16 ***
## ISPs.million  1   0.12090     2.41      2     35 0.1049
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(manova(cbind(Life.expectancy, Unemployment) ~ Irrigated, data = d))
```

```
##           Df   Pillai approx F num Df den Df Pr(>F)
## Irrigated  1 0.027818  0.50075      2     35 0.6103
## Residuals 36
```

```r
summary(manova(cbind(Life.expectancy, Unemployment) ~ Under.14, data = d))
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)
## Under.14   1 0.65678   33.487      2     35 7.457e-09 ***
## Residuals 36
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(manova(cbind(Life.expectancy, Unemployment) ~ Literacy.Rate, data = d))
```

```
##               Df  Pillai approx F num Df den Df    Pr(>F)
## Literacy.Rate  1 0.45279    14.48      2     35 2.616e-05 ***
## Residuals     36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#-- R CODE
summary(manova(cbind(Life.expectancy, Unemployment) ~ Literacy.Rate + ISPs.million, data = d))
```

```
##               Df  Pillai approx F num Df den Df    Pr(>F)
## Literacy.Rate  1 0.45542  14.2166      2     34 3.259e-05 ***
## ISPs.million   1 0.02890   0.5059      2     34    0.6075
## Residuals     35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(lm(cbind(Life.expectancy, Unemployment) ~ Literacy.Rate + ISPs.million, data = d),type="III")
```

```
##
## Type III MANOVA Tests: Pillai test statistic
##               Df test stat approx F num Df den Df    Pr(>F)
## (Intercept)    1   0.80954   72.259      2     34 5.709e-13 ***
## Literacy.Rate  1   0.39552   11.123      2     34 0.0001921 ***
## ISPs.million   1   0.02890    0.506      2     34 0.6074542
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(manova(cbind(Life.expectancy, Unemployment) ~ Irrigated + Under.14, data = d))
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)
## Irrigated  1 0.04199    0.745      2     34    0.4823
## Under.14   1 0.65494   32.267      2     34 1.394e-08 ***
## Residuals 35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```