GLS 5 - Data set: NAZIONI

INTRODUZIONE

Nel dataset in oggetto sono riportati i risultati di un'indagine effettuata nel 1995 su 66 nazioni riguardanti alcuni fra gli aspetti socio-demografici prevalenti. Le variabili presenti nel dataset sono le seguenti:

1. DENSITA': densità di popolazione
2. URBANA: percentuale di popolazione residente nelle città
3. VITAFEM: speranza di vita alla nascita delle donne
4. VITAMAS: speranza di vita alla anascita dei maschi
5. ALFABET: percentuale di alfabetizzati sul totale della popolazione
6. PIL: prodotto interno lordo pro-capite
7. RELIG: religione prevalente nella nazione (1=cattolica, 2=ortodossa, 3=protestante)

Analisi proposte:

1. Statistiche descrittive
2. Regressione
3. Gestione dell'autocorrelazione

```r
#-- R CODE

library(Hmisc)
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,"\\F. Esercizi(22) copia\\1.Error-GLS copy(8)\\5.Error-GLS\\nazioni.
d$pil <- as.numeric(gsub(",","",paste(d$pil))) #-- trasformo pil in variabile numerica

#-- vettore di variabili numeriche presenti nei dati
```

```r
VAR_NUMERIC <- c("densita","urbana","vitafem","vitamas","alfabet","pil")

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

nazione	densita	urbana	vitafem	vitamas	alfabet	pil	relig
Argentina	12	86	75	68	95	3,408	1
Armenia	126	68	75	68	98	5,000	2
Australia	2	85	80	74	100	16,848	3
Austria	94	58	79	73	99	18,396	1
Barbados	605	45	78	73	99	6,950	3
Belgio	329	96	79	73	99	17,912	1

STATISTICHE DESCRITTIVE

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

Table 2: Table continues below

densita	urbana	vitafem	vitamas
Min. : 2.00	Min. : 5.00	Min. :43.00	Min. :41.00
1st Qu.: 19.75	1st Qu.:49.50	1st Qu.:70.00	1st Qu.:64.00
Median : 61.00	Median :64.50	Median :76.00	Median :69.00
Mean :100.15	Mean :62.18	Mean :72.74	Mean :66.58
3rd Qu.:122.25	3rd Qu.:75.00	3rd Qu.:79.00	3rd Qu.:73.00
Max. :605.00	Max. :96.00	Max. :82.00	Max. :76.00

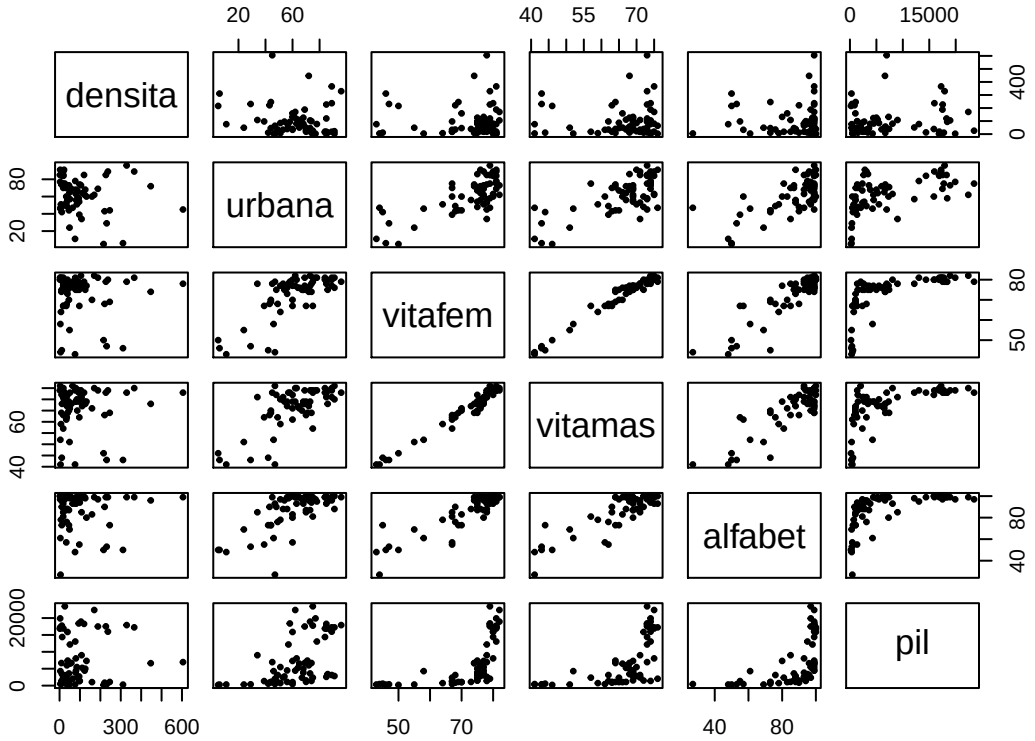
alfabet	pil
Min. : 27.00	Min. : 208
1st Qu.: 83.50	1st Qu.: 1412
Median : 95.50	Median : 4464
Mean : 87.58	Mean : 7303
3rd Qu.: 99.00	3rd Qu.:14048
Max. :100.00	Max. :23474

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

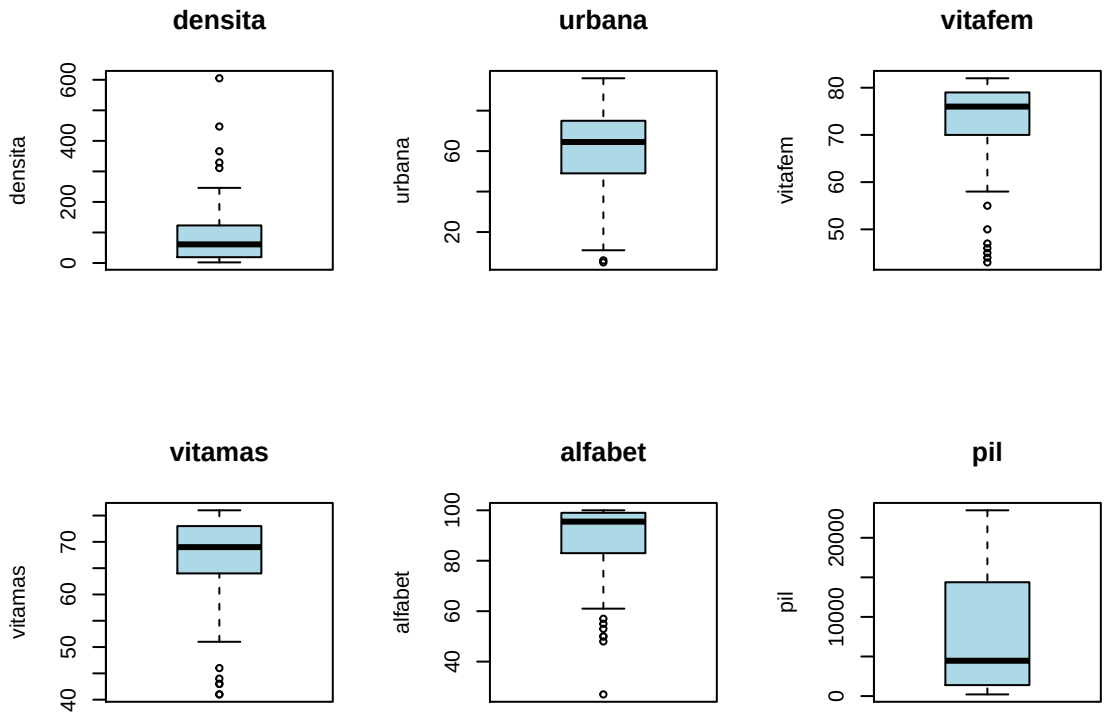
	densita	urbana	vitafem	vitamas	alfabet	pil
densita	1	-0.1501	-0.01275	0.01848	0.02142	0.09363
urbana	-0.1501	1	0.7317	0.7043	0.7054	0.54
vitafem	-0.01275	0.7317	1	0.9836	0.8874	0.601
vitamas	0.01848	0.7043	0.9836	1	0.8628	0.6039
alfabet	0.02142	0.7054	0.8874	0.8628	1	0.5629

	densita	urbana	vitafem	vitamas	alfabet	pil
pil	0.09363	0.54	0.601	0.6039	0.5629	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)
}
```



Esiste una fortissima correlazione fra "vitafem" e "vitamas" che fa presagire una collinearità fra le due variabili. Effettuiamo ora una regressione di "pil" su "urbana", "vitamas", "vitafem", "alfabet".

REGRESSIONE

```
#-- R CODE
mod1 <- lm(pil ~ urbana + vitamas + vitafem + alfabet, d) #-- stima modello lineare semplice
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-19,325	5,645	-3.423	0.001111
urbana	75.85	51.38	1.476	0.145
vitamas	404.1	425	0.9507	0.3455
vitafem	-123.1	430.9	-0.2856	0.7762
alfabet	45.24	94.14	0.4806	0.6325

Table 6: Fitting linear model: pil ~ urbana + vitamas + vitafem + alfabet

Observations	Residual Std. Error	R^2	Adjusted R^2
66	5623	0.3933	0.3535

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

Table 7: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
urbana	1	927,255,670	927,255,670	29.32	1.095e-06
vitamas	1	315,409,111	315,409,111	9.974	0.002469
vitafem	1	411,646	411,646	0.01302	0.9095
alfabet	1	7,304,548	7,304,548	0.231	0.6325
Residuals	61	1.929e+09	31,623,430	NA	NA

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

Test.statistic	P.value
11.78	0.00277

```r
pander(dwtest(mod1),big.mark=",") #-- Durbin-Whatson test
```

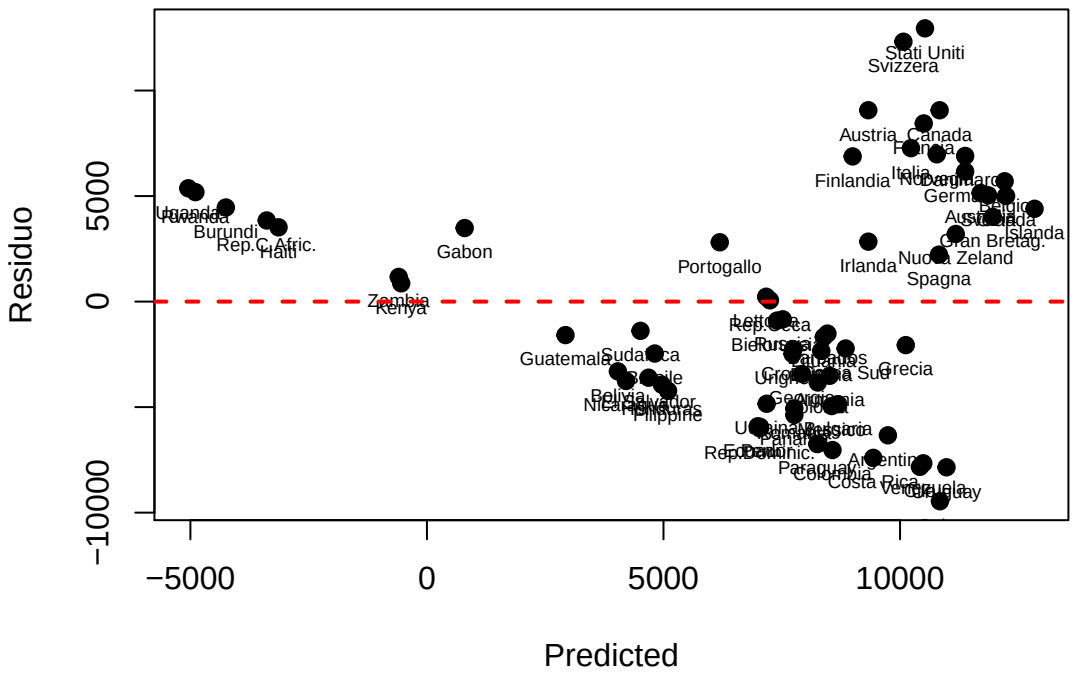
Table 9: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.659	0.07578	true autocorrelation is greater than 0

Il modello è significativo ma non interpreta molto bene la variabile dipendente e infatti solo l'intercetta risulta significativa.

Si sarebbe portati a cambiar modello, ma prima occorre verificare la diagnostica. Tralasciando la verifica di multicollinearità che pur sarebbe molto opportuna consideriamo la sfericità degli errori. Analizziamo dapprima l'eteroschedasticità degli errori.

```r
#-- R CODE
plot(fitted(mod1),resid(mod1),pch=19,xlab="Predicted",ylab="Residuo",type="p",col=1,lwd=2)
text(fitted(mod1),resid(mod1),d$nazione,pos=1,cex=.6)
abline(h=0,lwd=2,lty=2,col=2)
```



La rappresentazione grafica dei residui ben lontana da una forma rettangolare e il test di White mostrano con chiarezza che l'ipotesi di omoschedasticità degli errori va respinta. Inoltre si vede la presenza di outlier (Stati Uniti e Svizzera). Per ciò che concerne la non correlazione degli errori si vede dal test di Durbin-Watson che è respinta l'ipotesi di non correlazione dei residui.

Per risolvere il problema si propone un metodo di stima FGLS.

```
#-- R CODE
mod2 <- lm(resid(mod1)^2 ~ urbana + vitamas + vitafem + alfabet, d)
weight <- 1/fitted(mod2)

mod3 <- lm(pil ~ urbana + vitamas + vitafem + alfabet, d,weights=weight)
pander(summary(mod3),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-13,079	3,263	-4.009	0.0001688
urbana	26.99	36.69	0.7356	0.4648
vitamas	627.6	274.1	2.289	0.02553
vitafem	-358.5	297.2	-1.206	0.2324
alfabet	34.3	70.98	0.4833	0.6306

Table 11: Fitting linear model: pil ~ urbana + vitamas + vitafem + alfabet

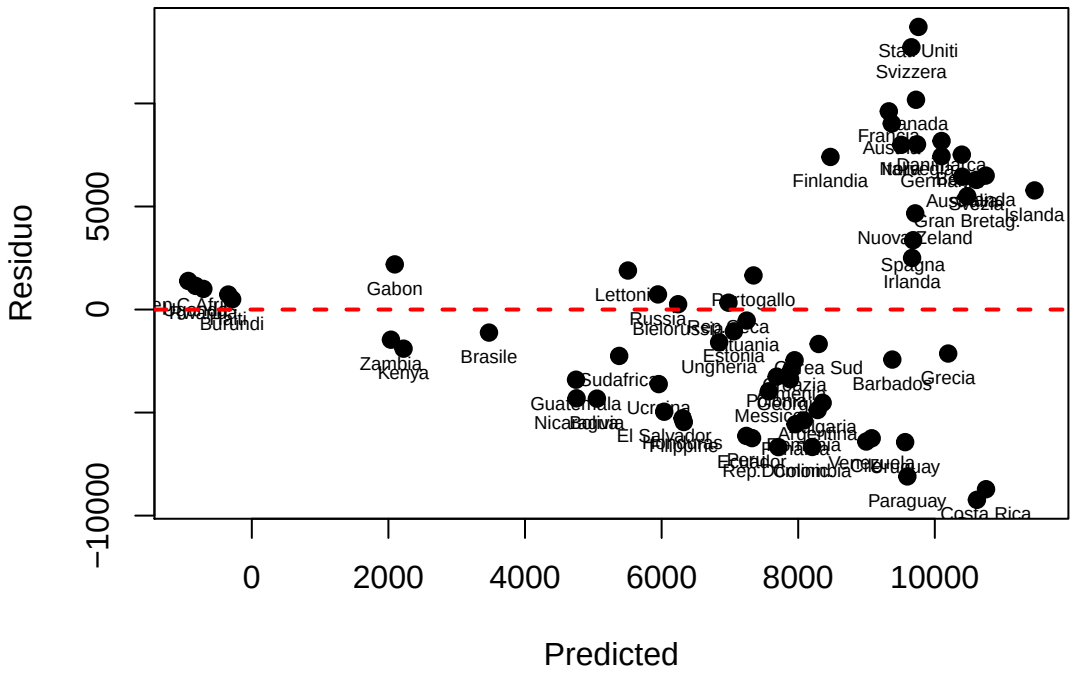
Observations	Residual Std. Error	R^2	Adjusted R^2
66	0.9919	0.5146	0.4828

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

Table 12: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
urbana	1	41.51	41.51	42.19	1.692e-08
vitamas	1	20.69	20.69	21.03	2.301e-05
vitafem	1	1.207	1.207	1.226	0.2725
alfabet	1	0.2298	0.2298	0.2336	0.6306
Residuals	61	60.02	0.9839	NA	NA

```
#-- R CODE
plot(fitted(mod3),resid(mod3),pch=19,xlab="Predicted",ylab="Residuo",type="p",col=1,lwd=2)
text(fitted(mod3),resid(mod3),d$nazione,pos=1,cex=.6)
abline(h=0,lwd=2,lty=2,col=2)
```



Il modello ora fitta in modo rilevante i dati e “vitamas” è significativa.

Si propone ora un modello basato su stime FGLS con errori espressi in forma esponenziale.

```
#-- R CODE
mod5 <- lm(log(resid(mod1)^2) ~ urbana + vitamas + vitafem + alfabet, d)
sd_error <- sqrt(exp(fitted(mod5)))

mod6 <- lm(I(pil/sd_error) ~ 0 + I(1/sd_error) + I(urbana/sd_error) + I(vitamas/sd_error) + I(vitafem/s
pander(summary(mod6),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
I(1/sd_error)	-16,206	4,283	-3.784	0.0003544
I(urbana/sd_error)	30.54	44.55	0.6856	0.4956
I(vitamas/sd_error)	655.4	310.3	2.112	0.03877
I(vitafem/sd_error)	-343.6	353	-0.9733	0.3342
I(alfabet/sd_error)	33.4	77.84	0.4291	0.6694

Table 14: Fitting linear model: I(pil/sd_error) ~ 0 + I(1/sd_error) + I(urbana/sd_error) + I(vitamas/sd_error) + I(vitafem/sd_error) + I(alfabet/sd_error)

Observations	Residual Std. Error	R^2	Adjusted R^2
66	1.386	0.731	0.7089

```
pander(anova(mod6),big.mark=",")
```

Table 15: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(1/sd_error)	1	233	233	121.4	3.842e-16
I(urbana/sd_error)	1	44.49	44.49	23.18	1.012e-05
I(vitamas/sd_error)	1	38.81	38.81	20.22	3.154e-05
I(vitafem/sd_error)	1	1.513	1.513	0.7881	0.3782
I(alfabet/sd_error)	1	0.3534	0.3534	0.1841	0.6694
Residuals	61	117.1	1.92	NA	NA

```
pander(white.test(mod6),big.mark=",") #-- white test
```

Test.statistic	P.value
4.693	0.0957

```
pander(dwtest(mod6),big.mark=",") #-- Durbin-Whatson test
```

Table 17: Durbin-Watson test: `mod6`

Test statistic	P value	Alternative hypothesis
1.828	0.2416	true autocorrelation is greater than 0

Il modello ora ha un fitting ancora più elevato che il precedente modello e “vitamas” rimane l’unica variabile con parametri significativi a mostrare che il “Pil” è influenzato solo dalla speranza di vita maschile tra le variabili esplicative prescelte. I test di White e Durbin Watson mostrano che i residui sono omoschedastici e incorrelati.