MULTI 3 - Data set: NAZIONI

INTRODUZIONE

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

1. DENSITA': densità di popolazione (abitanti per Kmq)
2. URBANA: percentuale di popolazione residente nelle città
3. VITAFEM: speranza di vita alla nascita delle donne
4. VITAMAS: speranza di vita alla nascita 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 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\\3.mult\\nazioni.csv"),sep=";")
```

```r
#d$relig <- factor(d$relig,1:3,c("catt","ortod","prot"))
d$dummy_cat <- ifelse(d$relig==1,1,0)
d$dummy_ort <- ifelse(d$relig==2,1,0)
d$dummy_prot <- ifelse(d$relig==3,1,0)

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("densita","urbana","alfabet","pil")

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

Table 1: Table continues below

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

dummy_cat	dummy_ort	dummy_prot
1	0	0
0	1	0
0	0	1
1	0	0
0	0	1
1	0	0

STATISTICHE DESCRITTIVE

Le variabili dipendenti sono “vitamas” e “vitafem”, le altre variabili sono esplicative.

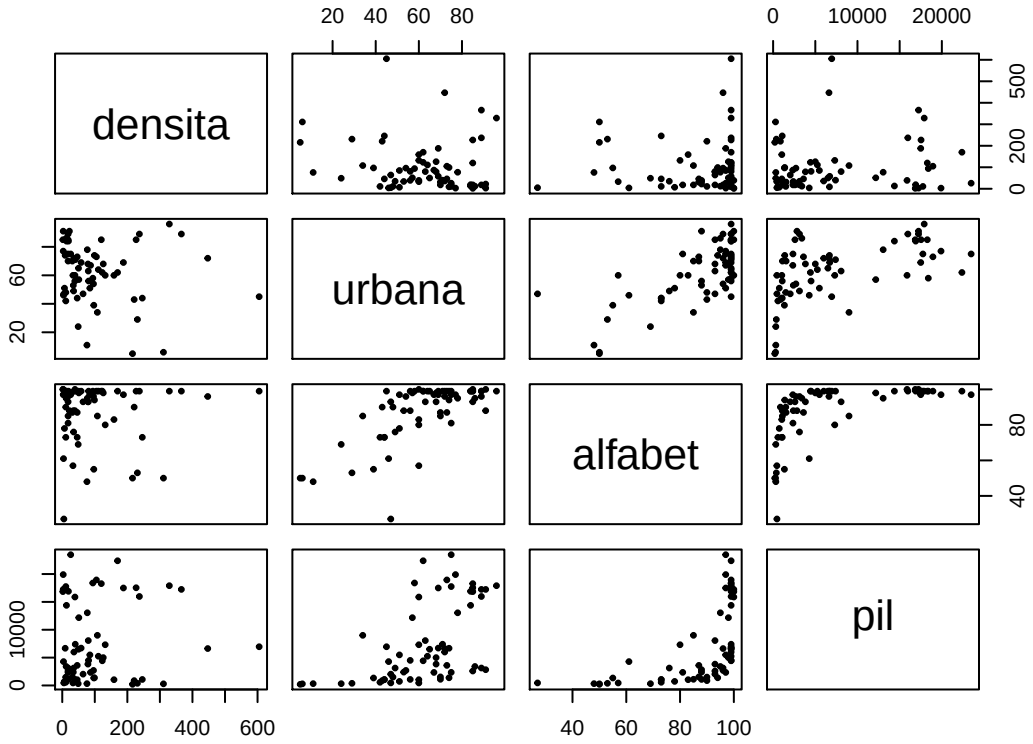
```r
#-- R CODE
pander(summary(d[,VAR_NUMERIC]),big.mark=",") #-- statistiche descrittive
```

densita	urbana	alfabet	pil
Min. : 2.00	Min. : 5.00	Min. : 27.00	Min. : 208
1st Qu.: 19.75	1st Qu.:49.50	1st Qu.: 83.50	1st Qu.: 1412
Median : 61.00	Median :64.50	Median : 95.50	Median : 4464
Mean :100.15	Mean :62.18	Mean : 87.58	Mean : 7303
3rd Qu.:122.25	3rd Qu.:75.00	3rd Qu.: 99.00	3rd Qu.:14048
Max. :605.00	Max. :96.00	Max. :100.00	Max. :23474

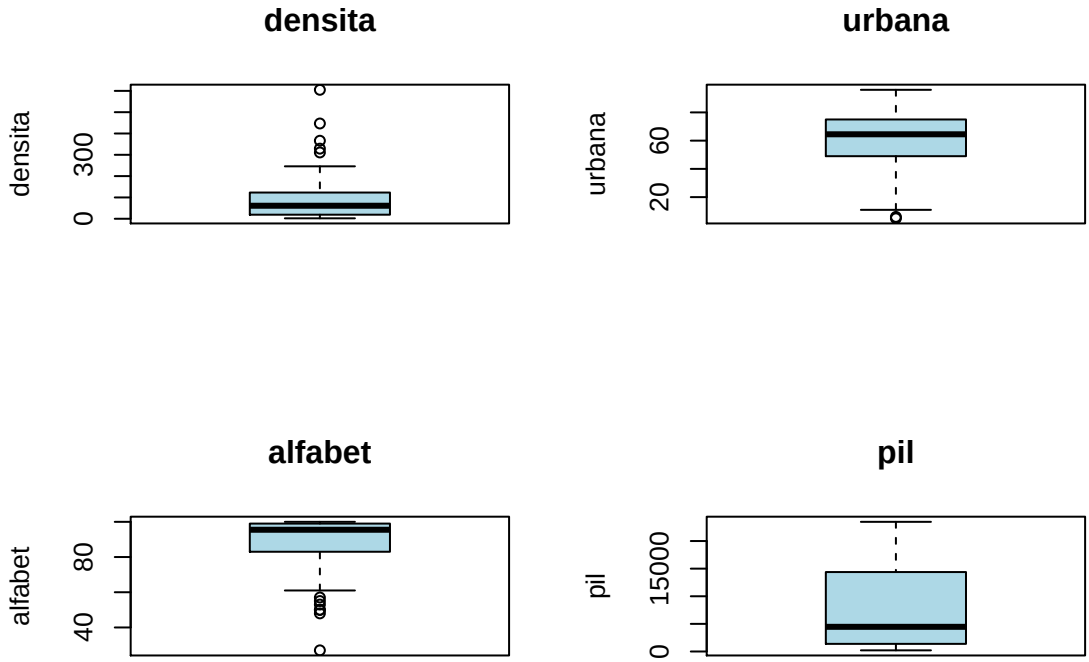
```r
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

	densita	urbana	alfabet	pil
densita	1	-0.1501	0.02142	0.09363
urbana	-0.1501	1	0.7054	0.54
alfabet	0.02142	0.7054	1	0.5629
pil	0.09363	0.54	0.5629	1

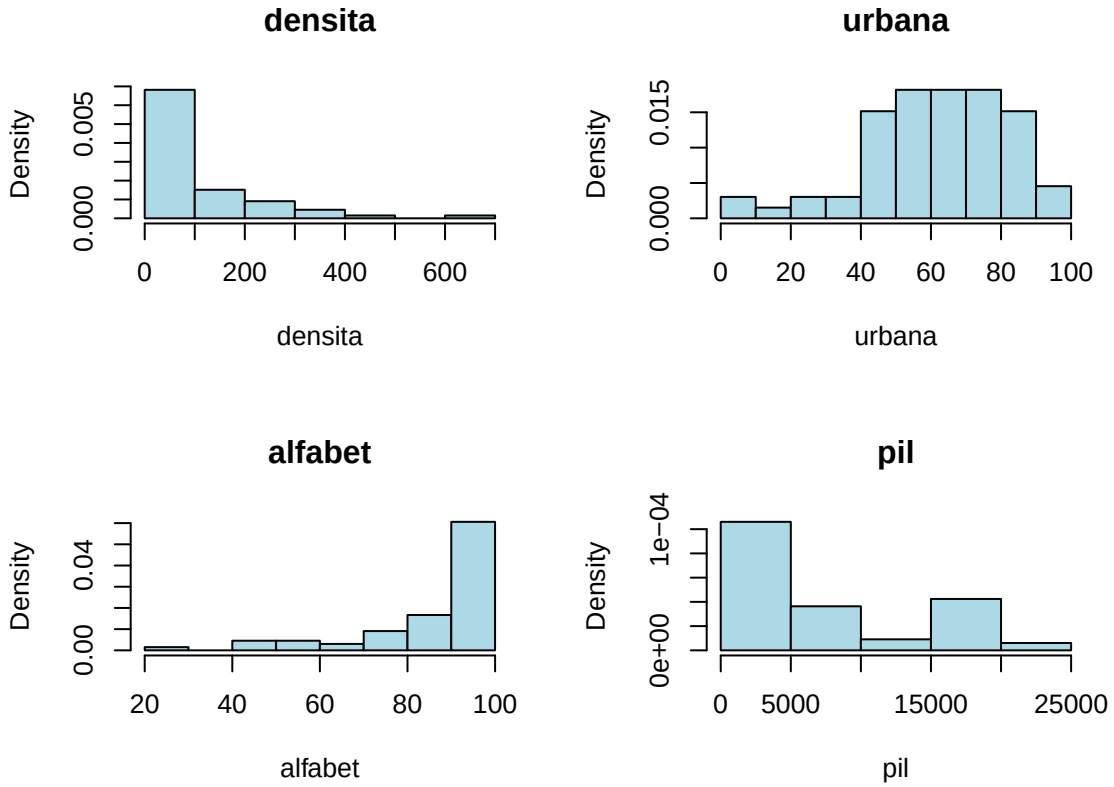
```r
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```




```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)
}
```



ESERCIZIO 1

Si propongano ora le regressioni multiple con "vitamas" e "vitafem" variabili dipendenti.

```
#-- R CODE
mod1 <- lm(vitamas ~ densita + urbana + alfabet + pil + dummy_ort + dummy_prot, d)
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.29	3.352	8.139	3.18e-11
densita	0.001103	0.004899	0.2252	0.8226
urbana	0.07802	0.04058	1.922	0.05939
alfabet	0.3862	0.0524	7.371	6.36e-10
pil	0.0002248	0.000106	2.121	0.03818
dummy_ort	-2.747	1.79	-1.535	0.1301
dummy_prot	-3.123	1.376	-2.27	0.02689

Table 6: Fitting linear model: vitamas ~ densita + urbana + alfabet + pil + dummy_ort + dummy_prot

Observations	Residual Std. Error	R^2	Adjusted R^2
66	4.345	0.7984	0.7778

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

Table 7: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
densita	1	1.886	1.886	0.09992	0.753
urbana	1	2,825	2,825	149.6	7.969e-18
alfabet	1	1,391	1,391	73.68	5.678e-12
pil	1	66.6	66.6	3.528	0.0653
dummy_ort	1	27.91	27.91	1.478	0.2289
dummy_prot	1	97.27	97.27	5.152	0.02689
Residuals	59	1,114	18.88	NA	NA

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

Test.statistic	P.value
16.18	0.0003064

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

Table 9: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.653	0.07424	true autocorrelation is greater than 0

```
#-- R CODE
mod2 <- lm(vitafem ~ densita + urbana + alfabet + pil + dummy_ort + dummy_prot, d)
pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.57	3.316	8.614	5.045e-12
densita	-0.0008131	0.004846	-0.1678	0.8673
urbana	0.09874	0.04015	2.459	0.01687
alfabet	0.4287	0.05184	8.27	1.918e-11
pil	0.0002326	0.0001049	2.218	0.03042
dummy_ort	-1.737	1.771	-0.9807	0.3307
dummy_prot	-3.54	1.361	-2.601	0.01174

Table 11: Fitting linear model: vitafem ~ densita + urbana + alfabet + pil + dummy_ort + dummy_prot

Observations	Residual Std. Error	R^2	Adjusted R^2
66	4.299	0.8375	0.821

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

Table 12: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
densita	1	1.091	1.091	0.05903	0.8089
urbana	1	3,655	3,655	197.8	1.678e-20
alfabet	1	1,776	1,776	96.12	5.366e-14
pil	1	54.03	54.03	2.924	0.09251
dummy_ort	1	6.798	6.798	0.3679	0.5465
dummy_prot	1	125	125	6.764	0.01174
Residuals	59	1,090	18.48	NA	NA

```
pander(white.test(mod2),big.mark=",")
```

Test.statistic	P.value
15.37	0.0004602

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

Table 14: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
1.725	0.1252	true autocorrelation is greater than 0

In entrambe le regressioni il fitting è molto elevato. Si passi ora al modello multivariato e all'analisi dei test multivariati.

```
#-- R CODE
mod3 <- lm(cbind(vitamas,vitafem) ~ densita + urbana + alfabet + pil + dummy_ort + dummy_prot, d)

library(ppcor)
```

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

```
pander(pcor.test(d$vitamas,d$vitafem,d[,c("densita","urbana","alfabet","pil","dummy_ort","dummy_prot")])
```

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

```
##   0.9252    4.35e-26     18.57     66   6    pearson
## -----------------------------------------------------
```

```
summary(mod3)
```

```
## Response vitamas :
##
## Call:
## lm(formula = vitamas ~ densita + urbana + alfabet + pil + dummy_ort +
##     dummy_prot, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7743  -2.7699   0.0802   2.2627  10.0207
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.2853144  3.3522247   8.139 3.18e-11 ***
## densita      0.0011033  0.0048985   0.225   0.8226
## urbana       0.0780176  0.0405844   1.922   0.0594 .
## alfabet      0.3862275  0.0523973   7.371 6.36e-10 ***
## pil          0.0002248  0.0001060   2.121   0.0382 *
## dummy_ort   -2.7473814  1.7898836  -1.535   0.1301
## dummy_prot  -3.1233300  1.3760087  -2.270   0.0269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.345 on 59 degrees of freedom
## Multiple R-squared:  0.7984, Adjusted R-squared:  0.7778
## F-statistic: 38.93 on 6 and 59 DF,  p-value: < 2.2e-16
##
##
## Response vitafem :
##
## Call:
## lm(formula = vitafem ~ densita + urbana + alfabet + pil + dummy_ort +
##     dummy_prot, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5905  -2.1488  -0.1886   1.8190  10.7722
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5674922  3.3162840   8.614 5.05e-12 ***
## densita     -0.0008131  0.0048460  -0.168   0.8673
## urbana       0.0987418  0.0401493   2.459   0.0169 *
## alfabet      0.4286552  0.0518356   8.270 1.92e-11 ***
## pil          0.0002326  0.0001049   2.218   0.0304 *
## dummy_ort   -1.7365632  1.7706934  -0.981   0.3307
## dummy_prot  -3.5403502  1.3612558  -2.601   0.0117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.299 on 59 degrees of freedom
```

```
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.821
## F-statistic: 50.68 on 6 and 59 DF,  p-value: < 2.2e-16
```

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

```
##
## ------------------------------------------------------------------------
##               Df    Pillai    approx F   num Df   den Df    Pr(>F)
## ---------------- ---- --------- ---------- -------- -------- -----------
##   **densita**      1    0.03424    1.028       2        58      0.3641
##
##    **urbana**      1     0.774     99.32       2        58     1.858e-19
##
##   **alfabet**      1    0.6237     48.06       2        58     4.919e-13
##
##     **pil**        1    0.0565     1.737       2        58      0.1852
##
##  **dummy_ort**     1    0.05365    1.644       2        58      0.2021
##
##  **dummy_prot**    1    0.1046     3.388       2        58      0.04058
##
##  **Residuals**     59     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   0.55774   36.572      2     58 5.305e-11 ***
## densita      1   0.01722    0.508      2     58   0.60425
## urbana       1   0.10491    3.399      2     58   0.04020 *
## alfabet      1   0.53881   33.881      2     58 1.789e-10 ***
## pil          1   0.07743    2.434      2     58   0.09661 .
## dummy_ort    1   0.05898    1.818      2     58   0.17155
## dummy_prot   1   0.10462    3.388      2     58   0.04058 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#-- R CODE
summary(manova(cbind(vitamas,vitafem) ~ densita, data = d))
```

```
##           Df   Pillai approx F num Df den Df Pr(>F)
## densita    1 0.029743  0.96563      2     63 0.3863
## Residuals 64
```

```
summary(manova(cbind(vitamas,vitafem) ~ urbana, data = d))
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)
## urbana     1 0.54271   37.385      2     63 1.976e-11 ***
## Residuals 64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(manova(cbind(vitamas,vitafem) ~ alfabet, data = d))
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)
## alfabet    1 0.79069      119      2     63 < 2.2e-16 ***
## Residuals 64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(manova(cbind(vitamas,vitafem) ~ pil, data = d))
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)
## pil        1 0.36622   18.202      2     63 5.769e-07 ***
## Residuals 64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(manova(cbind(vitamas,vitafem) ~ dummy_ort, data = d))
```

```
##           Df  Pillai approx F num Df den Df  Pr(>F)
## dummy_ort  1 0.07755   2.6482      2     63 0.07865 .
## Residuals 64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(manova(cbind(vitamas,vitafem) ~ dummy_prot, data = d))
```

```
##            Df   Pillai approx F num Df den Df Pr(>F)
## dummy_prot  1 0.011389  0.36289      2     63 0.6971
## Residuals  64
```