LINEAR 6 - Data set: ANTRO

INTRODUZIONE

Il dataset è costituito da alcune misure antropometriche rilevate su 248 uomini.

1. ETA': età in anni compiuti
2. PESO: peso rilevato in libbre
3. ALTEZ: altezza (cm)
4. COLLO: circonferenza del collo (cm)
5. TORACE: circonferenza toracica (cm)
6. ADDOM: circonferenza addominale (cm)
7. ANCA: circonferenza dell'anca (cm)
8. COSCIA: circonferenza della coscia (cm)
9. GINOCCH: circonferenza del ginocchio (cm)
10. CAVIGLIA: circonferenza della caviglia (cm)
11. BICIPITE: circonferenza del bicipite in estensione (cm)
12. AVANBR: circonferenza dell'avambraccio (cm)
13. POLSO: circonferenza del polso (cm)

Variabile dipendente: DEATHS

Analisi proposte:

1. Statistiche descrittive
2. Regressione lineare e polinomiale

```r
#-- R CODE

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"
```

```r
d <- read.csv(paste0(ABSOLUTE_PATH,"\\F. Esercizi(22) copia\\3.lin(5)\\6.linear\\ANTROP.TXT"),sep="\t")

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- names(d)[-1] #-- tutte le variabili tranne la prima

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

Table 1: Table continues below

id_sogg	eta	peso	altez	collo	torace	addom	anca	coscia
1	23	154.2	172.1	36.2	93.1	85.2	94.5	59
2	22	173.2	183.5	38.5	93.6	83	98.7	58.7
3	22	154	168.3	34	95.8	87.9	99.2	59.6
4	26	184.8	183.5	37.4	101.8	86.4	101.2	60.1
5	24	184.2	181	34.4	97.3	100	101.9	63.2
6	24	210.2	189.9	39	104.5	94.4	107.8	66

ginocch	caviglia	bicipite	avanbr	polso
37.3	21.9	32	27.4	17.1
37.3	23.4	30.5	28.9	18.2
38.9	24	28.8	25.2	16.6
37.3	22.8	32.4	29.4	18.2
42.2	24	32.2	27.7	17.7
42	25.6	35.7	30.6	18.8

STATISTICHE DESCRITTIVE

Si propongono la matrice di correlazione tra le variabili e alcune descrittive di base.

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

Table 3: Table continues below

eta	peso	altez	collo	torace
Min. :22.00	Min. :118.5	Min. :162.6	Min. :31.10	Min. : 79.30
1st Qu.:35.75	1st Qu.:158.2	1st Qu.:173.4	1st Qu.:36.38	1st Qu.: 94.15
Median :43.00	Median :176.1	Median :177.8	Median :38.00	Median : 99.60
Mean :44.85	Mean :178.1	Mean :178.6	Mean :37.95	Mean :100.67
3rd Qu.:54.00	3rd Qu.:196.8	3rd Qu.:183.5	3rd Qu.:39.42	3rd Qu.:105.30
Max. :81.00	Max. :262.8	Max. :197.5	Max. :43.90	Max. :128.30

Table 4: Table continues below

addom	anca	coscia	ginocch
Min. : 69.40	Min. : 85.00	Min. :47.20	Min. :33.00

addom	anca	coscia	ginocch
1st Qu.: 84.47	1st Qu.: 95.47	1st Qu.:56.00	1st Qu.:36.90
Median : 90.95	Median : 99.30	Median :59.00	Median :38.45
Mean : 92.31	Mean : 99.66	Mean :59.27	Mean :38.54
3rd Qu.: 99.20	3rd Qu.:103.28	3rd Qu.:62.30	3rd Qu.:39.90
Max. :126.20	Max. :125.60	Max. :74.40	Max. :46.00

caviglia	bicipite	avanbr	polso
Min. :19.10	Min. :24.80	Min. :21.00	Min. :15.80
1st Qu.:22.00	1st Qu.:30.20	1st Qu.:27.30	1st Qu.:17.60
Median :22.80	Median :32.00	Median :28.75	Median :18.30
Mean :22.99	Mean :32.22	Mean :28.67	Mean :18.22
3rd Qu.:24.00	3rd Qu.:34.33	3rd Qu.:30.00	3rd Qu.:18.80
Max. :27.00	Max. :39.10	Max. :34.90	Max. :21.40

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

Table 6: Table continues below

	eta	peso	altez	collo	torace	addom
eta	1	-0.01269	-0.2363	0.1257	0.1848	0.2452
peso	-0.01269	1	0.5136	0.8099	0.8914	0.8742
altez	-0.2363	0.5136	1	0.3224	0.2241	0.1886
collo	0.1257	0.8099	0.3224	1	0.7691	0.7293
torace	0.1848	0.8914	0.2241	0.7691	1	0.9103
addom	0.2452	0.8742	0.1886	0.7293	0.9103	1
anca	-0.05476	0.9327	0.3968	0.7073	0.825	0.8608
coscia	-0.2132	0.8528	0.3502	0.669	0.7082	0.7373
ginocch	0.01988	0.8427	0.5143	0.6481	0.6975	0.7106
caviglia	-0.1593	0.7248	0.4805	0.5456	0.5588	0.5222
bicipite	-0.04456	0.7856	0.3202	0.7093	0.707	0.6568
avanbr	-0.08449	0.6837	0.3246	0.6615	0.5995	0.5297
polso	0.2203	0.7253	0.3982	0.7317	0.6446	0.6029

Table 7: Table continues below

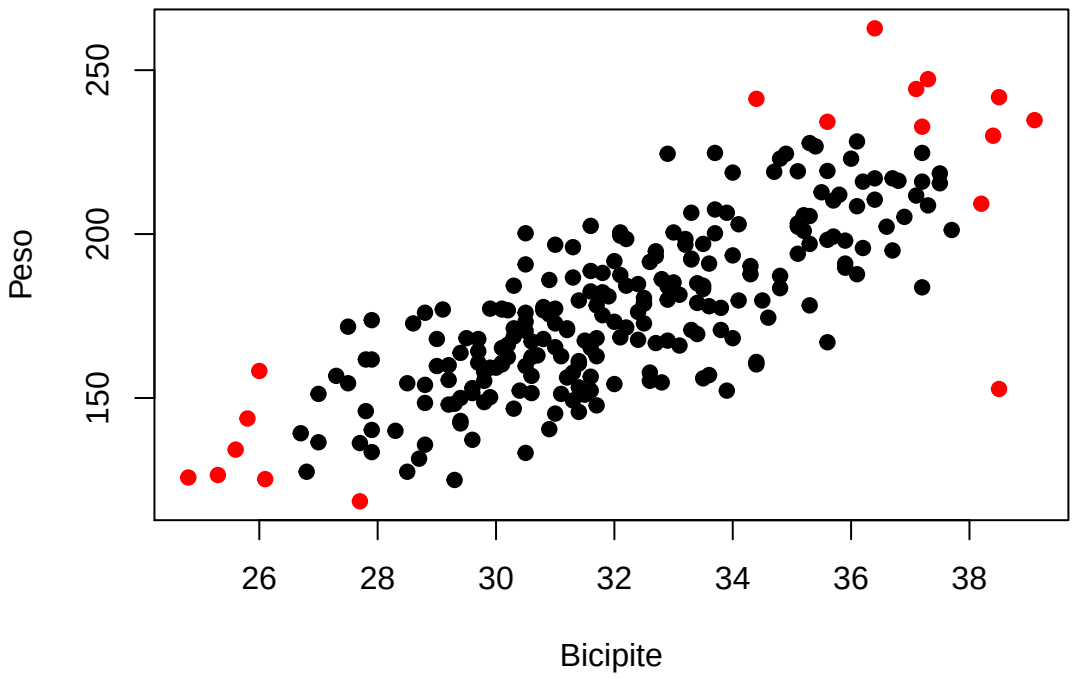
	anca	coscia	ginocch	caviglia	bicipite	avanbr
eta	-0.05476	-0.2132	0.01988	-0.1593	-0.04456	-0.08449
peso	0.9327	0.8528	0.8427	0.7248	0.7856	0.6837
altez	0.3968	0.3502	0.5143	0.4805	0.3202	0.3246
collo	0.7073	0.669	0.6481	0.5456	0.7093	0.6615
torace	0.825	0.7082	0.6975	0.5588	0.707	0.5995
addom	0.8608	0.7373	0.7106	0.5222	0.6568	0.5297
anca	1	0.8814	0.8091	0.6593	0.7222	0.6032
coscia	0.8814	1	0.7781	0.6635	0.7459	0.6036
ginocch	0.8091	0.7781	1	0.7293	0.6544	0.5787
caviglia	0.6593	0.6635	0.7293	1	0.5484	0.5607
bicipite	0.7222	0.7459	0.6544	0.5484	1	0.7021
avanbr	0.6032	0.6036	0.5787	0.5607	0.7021	1

	anca	coscia	ginocch	caviglia	bicipite	avanbr
polso	0.6267	0.545	0.6558	0.6662	0.6137	0.5993

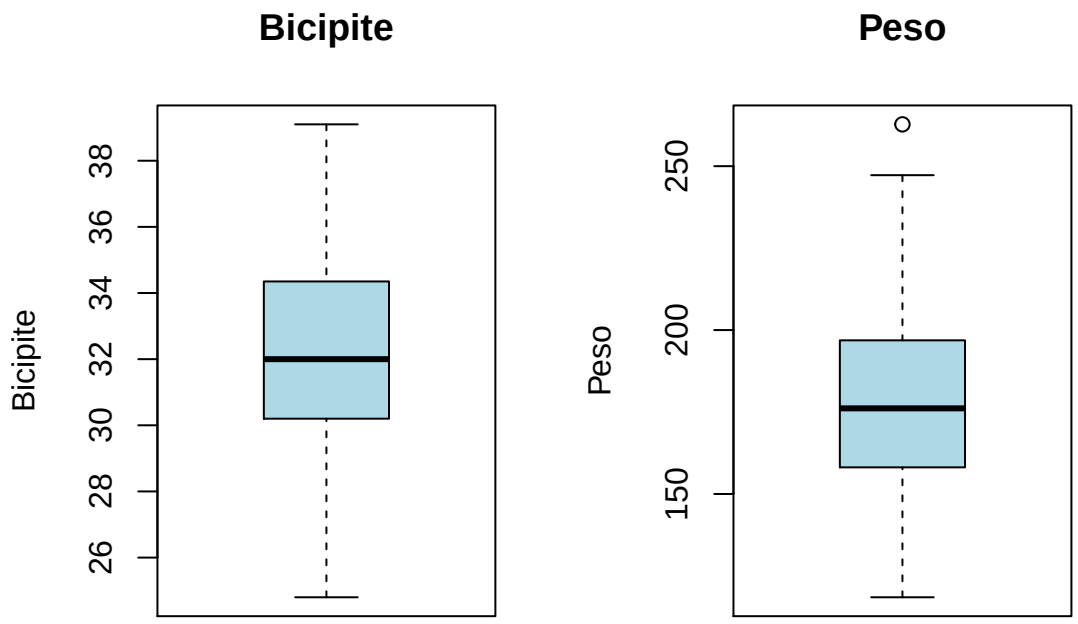
	polso
eta	0.2203
peso	0.7253
altez	0.3982
collo	0.7317
torace	0.6446
addom	0.6029
anca	0.6267
coscia	0.545
ginocch	0.6558
caviglia	0.6662
bicipite	0.6137
avanbr	0.5993
polso	1

Si decide di studiare il nesso lineare tra peso e circonferenza del bicipite. Si propongono quindi innanzitutto il grafico a dispersione inerente le due variabili, i box plot, i quantili e le osservazioni estreme.

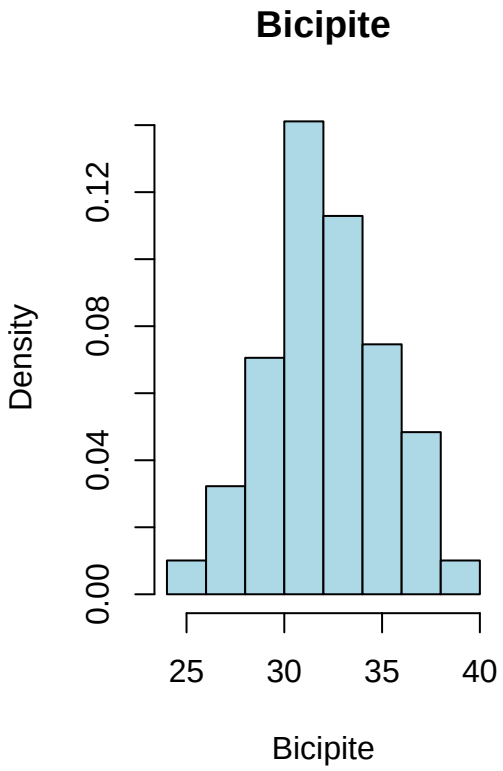
```
#-- R CODE

d$EXTREME <- 1
d$EXTREME[c(FIND_EXTREME_OBSERVARION(d$bicipite),FIND_EXTREME_OBSERVARION(d$peso))] <- 2

#-- Evidenzio in rosso le osservazioni estreme (superiori ed inferiori)
plot(d$bicipite,d$peso,pch=19,xlab="Bicipite",ylab="Peso",col=d$EXTREME)
```

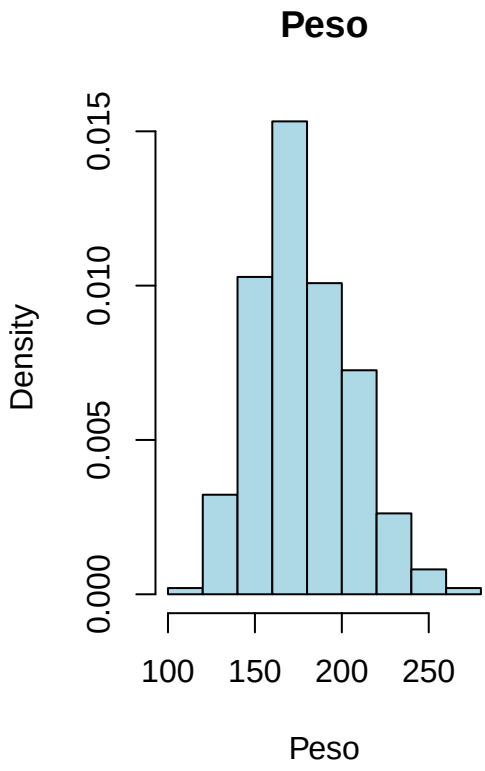



```
par(mfrow=c(1,2))
boxplot(d[,"bicipite"],main="Bicipite",col="lightblue",ylab="Bicipite",freq=F)
boxplot(d[,"peso"],main="Peso",col="lightblue",ylab="Peso",freq=F)
```




```
par(mfrow=c(1,2))
hist(d[,"bicipite"],main="Bicipite",col="lightblue",xlab="Bicipite",freq=F)
hist(d[,"peso"],main="Peso",col="lightblue",xlab="Peso",freq=F)
```





REGRESSIONE

A questa prima vista non si vedono particolari aspetti anomali delle due distribuzioni. Si propone prima il legame lineare tra le due variabili.

```
#-- R CODE
mod1 <- lm(peso~bicipite,d)
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-55.93	11.8	-4.739	3.637e-06
bicipite	7.265	0.3648	19.91	3.342e-53

Table 10: Fitting linear model: peso ~ bicipite

Observations	Residual Std. Error	R^2	Adjusted R^2
248	16.82	0.6171	0.6156

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

Table 11: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bicipite	1	112,233	112,233	396.5	3.342e-53
Residuals	246	69,627	283	NA	NA

```
pander(white.test(mod1),big.mark=",") #-- White test (per dettagli ?bptest)
```

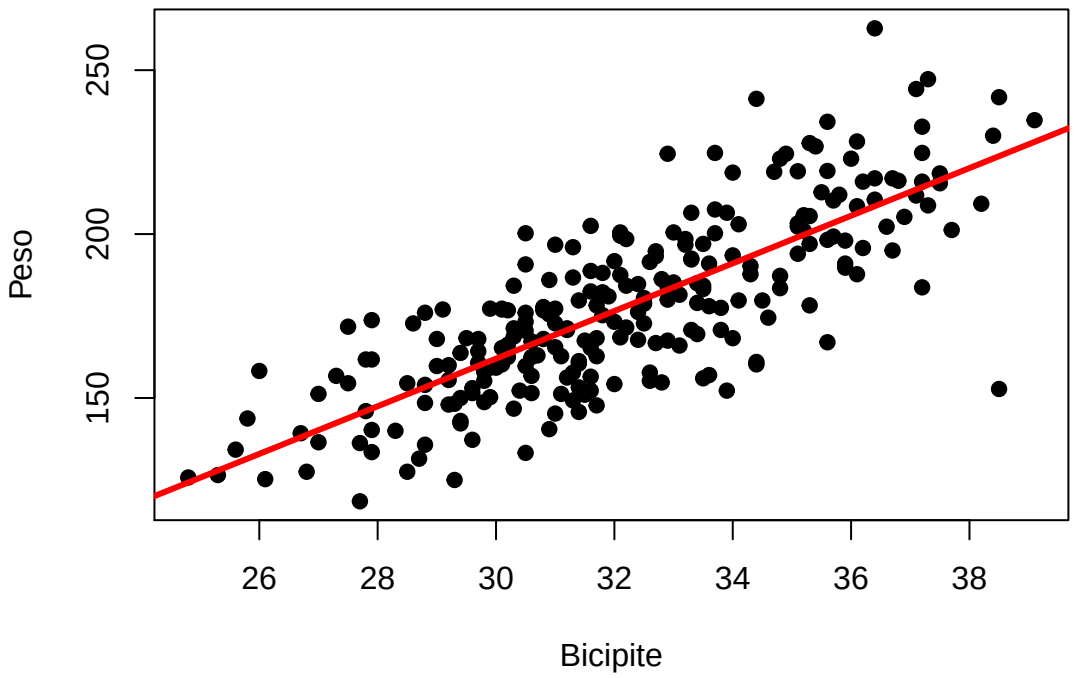
Test.statistic	P.value
10.88	0.004346

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

Table 13: Durbin-Watson test: `mod1`

Test statistic	P value	Alternative hypothesis
1.647	0.002537 * *	true autocorrelation is greater than 0

```
#-- R CODE
plot(d$bicipite,d$peso,pch=19,xlab="Bicipite",ylab="Peso")
abline(mod1,col=2,lwd=3) #-- abline del modello lineare
```



La variabile esplicativa bicipite è significativa e spiega in modo notevole peso (osservare il valore dell'R^2). Inoltre gli errori sono omoschedastici come si vede dal test di White. Si verifica ora se un modello linear-log sia preferibile al modello lineare.

```
#-- R CODE
mod2 <- lm(peso~I(log(bicipite)),d)
pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-624.8	40.77	-15.33	1.129e-37
I(log(bicipite))	231.5	11.75	19.7	1.678e-52

Table 15: Fitting linear model: peso ~ I(log(bicipite))

Observations	Residual Std. Error	R^2	Adjusted R^2
248	16.93	0.6121	0.6105

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

Table 16: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(bicipite))	1	111,316	111,316	388.2	1.678e-52
Residuals	246	70,545	286.8	NA	NA

```
pander(white.test(mod2),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
10.61	0.004972

```
pander(dwtest(mod2),big.mark=",") #-- Durbin-Whatson test
```

Table 18: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
1.66	0.003508 * *	true autocorrelation is greater than 0

Si utilizza per il confronto l'R^2 e si vede che la differenza è minima a favore del modello lineare. In ogni caso anche il modello linear-log ha errori omoschedastici. A questo punto si propone un modello log-lineare.

```
#-- R CODE
mod3 <- lm(I(log(peso))~bicipite,d)
pander(summary(mod3),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.858	0.06586	58.57	1.779e-146
bicipite	0.04076	0.002036	20.02	1.513e-53

Table 20: Fitting linear model: I(log(peso)) ~ bicipite

Observations	Residual Std. Error	R^2	Adjusted R^2
248	0.09389	0.6196	0.618

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

Table 21: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bicipite	1	3.532	3.532	400.7	1.513e-53
Residuals	246	2.169	0.008816	NA	NA

```r
pander(white.test(mod3),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
3.068	0.2157

```r
pander(dwtest(mod3),big.mark=",") #-- Durbin-Whatson test
```

Table 23: Durbin-Watson test: `mod3`

Test statistic	P value	Alternative hypothesis
1.62	0.001292 * *	true autocorrelation is greater than 0

Anche in questo caso confrontando gli R^2 il modello lineare è preferibile leggermente al modello log-lineare a sua volta leggermente migliore del modello linear-log. Il modello log-lineare ha anche esso errori omoschedastici. Si propone a questo punto il modello log-log:

```r
#-- R CODE
mod4 <- lm(I(log(peso))~I(log(bicipite)),d)
pander(summary(mod4),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6485	0.2261	2.868	0.004484
I(log(bicipite))	1.304	0.06516	20.01	1.571e-53

Table 25: Fitting linear model: I(log(peso)) ~ I(log(bicipite))

Observations	Residual Std. Error	R^2	Adjusted R^2
248	0.09391	0.6195	0.6179

```r
pander(anova(mod4),big.mark=",")
```

Table 26: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(bicipite))	1	3.532	3.532	400.5	1.571e-53
Residuals	246	2.169	0.008819	NA	NA

```r
pander(white.test(mod4),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
2.319	0.3137

```r
pander(dwtest(mod4),big.mark=",") #-- Durbin-Whatson test
```

Table 28: Durbin-Watson test: `mod4`

Test statistic	P value	Alternative hypothesis
1.634	0.001829 * *	true autocorrelation is greater than 0

Ancora una volta il modello lineare è migliore del modello log-log che ha ancora errori omoschedastici. Si sceglie quindi in definitiva il modello lineare.