

# La matrice dei dati

## Contents

Una matrice $7 \times 2$	1
Correlazione lineare	4
Dati Animals	5

## Una matrice $7 \times 2$

Si costruisca la seguente matrice di dati  $X$  di dimensione  $n = 7$  e  $p = 2$ :

$$X_{7 \times 2} = \begin{bmatrix} 3 & 5 \\ 4 & 5.5 \\ 2 & 4 \\ 6 & 7 \\ 8 & 10 \\ 2 & 5 \\ 5 & 7.5 \end{bmatrix}$$

```
rm(list=ls()) # pulizia del workspace
X <- matrix(
  c(3,4,2,6,8,2,5,
    5,5.5,4,7,10,5,7.5),
  nrow=7,ncol=2,
  byrow=FALSE)
n <- nrow(X)
p <- ncol(X)
colnames(X)<-c("x1","x2")
rownames(X)<-paste("u",1:n, sep="")
X
```

```
      x1  x2
u1    3  5.0
u2    4  5.5
u3    2  4.0
u4    6  7.0
u5    8 10.0
u6    2  5.0
u7    5  7.5
```

Per ottenere la matrice trasposta

```
t(X)
```

```
      u1  u2 u3 u4 u5 u6  u7
x1    3 4.0  2  6  8  2 5.0
x2    5 5.5  4  7 10  5 7.5
```

Per calcolare le statistiche di sintesi per le due variabili (Min, Max, Primo e Terzo quartile, Mediana e Media):

```
summary(X)
```

	x1	x2
Min.	:2.000	Min. : 4.000
1st Qu.	:2.500	1st Qu.: 5.000
Median	:4.000	Median : 5.500
Mean	:4.286	Mean : 6.286
3rd Qu.	:5.500	3rd Qu.: 7.250
Max.	:8.000	Max. :10.000

Per calcolare la media e varianza per la prima variabile:

```
mean(X[,1])
```

```
[1] 4.285714
```

```
((n-1)/n) * var(X[,1])
```

```
[1] 4.204082
```

Perchè moltiplichiamo la varianza per  $(n-1)/n$ ? Guardare l'help: `?var`

Per calcolare il vettore delle medie:

```
apply(X,MARGIN=2,FUN="mean")
```

	x1	x2
4.285714	6.285714	

Per calcolare la matrice di varianze/covarianze  $S$

```
S = ((n-1)/n)*var(X)
```

```
S
```

	x1	x2
x1	4.204082	3.704082
x2	3.704082	3.561224

Per calcolare la matrice di correlazione  $R$

```
R = cor(X)
```

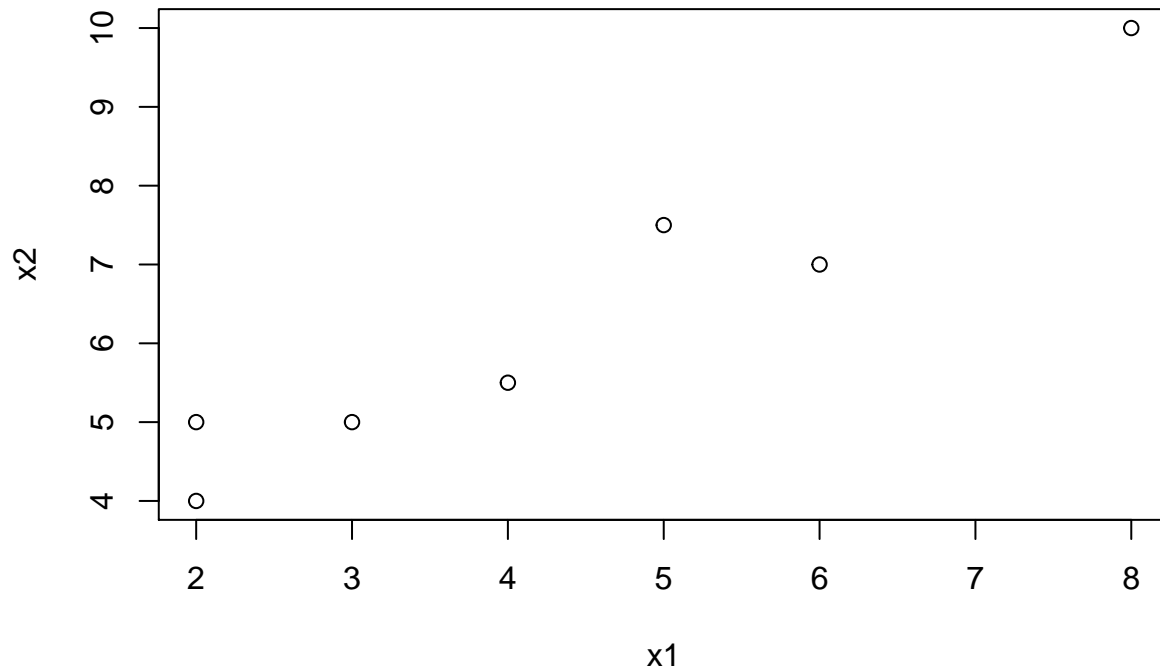
```
R
```

	x1	x2
x1	1.0000000	0.9572939
x2	0.9572939	1.0000000

Perchè in questo caso non moltiplichiamo per  $(n-1)/n$ ?

Per costruire il diagramma di dispersione:

```
plot(X)
```



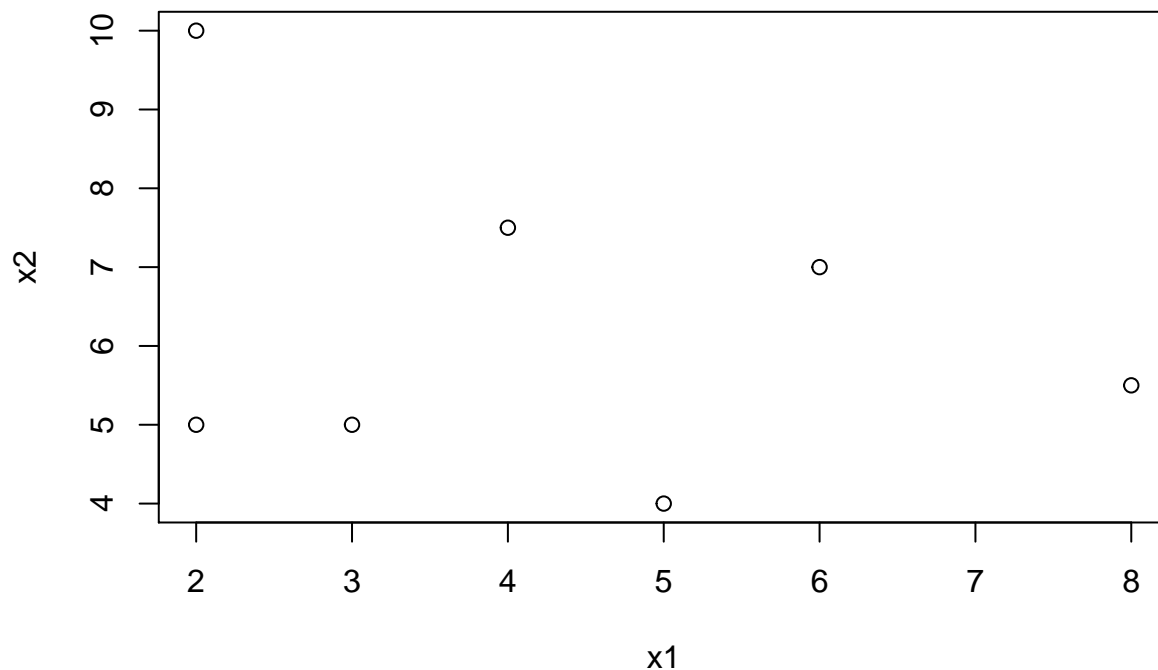
Per cambiare l'ordine dei valori della variabile  $x_1$  in maniera causale (permutazione), si può utilizzare il comando `sample()`. Per riproducibilità dei risultati, impostare all'inizio il seme generatore dei numeri casuali `set.seed(123)`.

```
set.seed(123)
X[, "x1"] <- sample(X[, "x1"])
X
```

```
  x1  x2
u1  2  5.0
u2  8  5.5
u3  5  4.0
u4  6  7.0
u5  2 10.0
u6  3  5.0
u7  4  7.5
```

Se calcoliamo le statistiche di sintesi, la matrice di varianze/covarianze e di correlazione, e costruiamo il diagramma di dispersione, notiamo che le due distribuzioni marginali sono le stesse, ma la covarianza  $s_{12}$  e il coefficiente di correlazione  $r_{12}$  cambiano.

```
plot(X)
```



```
summary(X)
```

x1	x2
Min. :2.000	Min. : 4.000
1st Qu.:2.500	1st Qu.: 5.000
Median :4.000	Median : 5.500
Mean :4.286	Mean : 6.286
3rd Qu.:5.500	3rd Qu.: 7.250
Max. :8.000	Max. :10.000

```
((n-1)/n)*var(X)
```

	x1	x2
x1	4.204082	-1.081633
x2	-1.081633	3.561224

```
cor(X)
```

	x1	x2
x1	1.0000000	-0.2795404
x2	-0.2795404	1.0000000

Quindi dalle due distribuzioni marginali non si può ricavare alcuna informazione sulla correlazione tra le due variabili  $x_1$  e  $x_2$ .

## Correlazione lineare

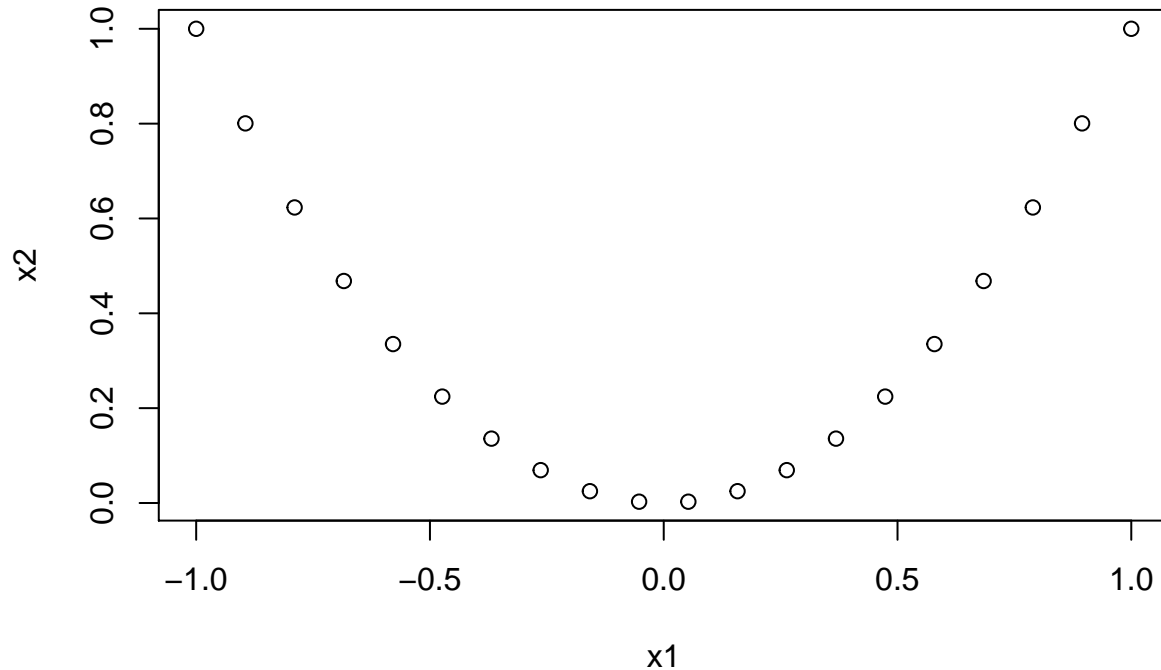
Si consideri la seguente relazione quadratica tra le variabili  $x_1$  e  $x_2$ :

$$x_{1i} = -1 + 2 \frac{(i-1)}{(n-1)}$$

$$x_{2i} = x_{1i}^2, \quad i = 1, \dots, n$$

Generare i dati come descritto sopra per  $n = 20$ , costruire il diagramma di dispersione e calcolare la matrice di correlazione, e commentare i risultati ottenuti.

```
n <- 20
x1 <- -1 + 2* ((1:n) - 1)/(n-1)
x2 <- x1^2
X <- cbind(x1,x2)
plot(x1,x2)
```



```
cor(X)

      x1      x2
x1  1.00000e+00 -8.77515e-17
x2 -8.77515e-17  1.00000e+00

round(cor(X),1) # arrotondo al primo decimale

      x1 x2
x1  1  0
x2  0  1
```

Come si vede, sebbene ci sia una dipendenza perfettamente quadratica tra le variabili  $x_1$  e  $x_2$ , il coefficiente di correlazione  $r_{12} \approx 0$ , perchè misura solo la dipendenza lineare (ovvero, la correlazione) tra le due variabili

## Dati Animals

Caricare il data set **Animals**, presente nella libreria MASS, che si carica R con il comando `library("MASS")`:

```
rm(list=ls())
library("MASS")
data(Animals)
?Animals
```

Dalla descrizione ottenuta con il comando `?Animals`, vediamo che si tratta di *Average brain and body weights*

for 28 species of land animals.

E' un data.frame con  $n = 28$  osservazioni misurate su  $p = 2$  variabili:

- **body** body weight in kg.
- **brain** brain weight in g.

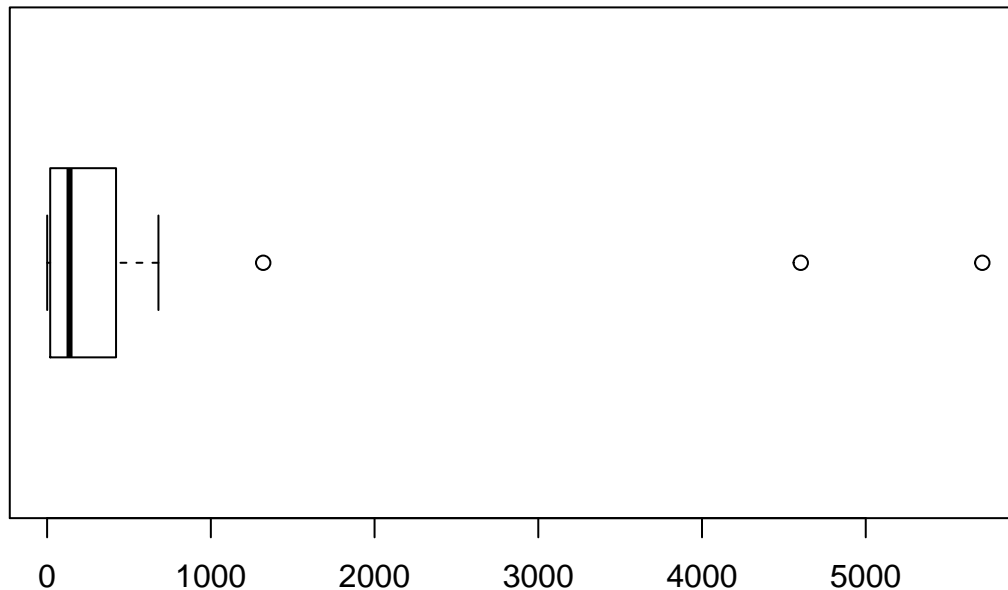
```
row.names(Animals)
```

[1] "Mountain beaver"	"Cow"	"Grey wolf"	"Goat"	"Guinea pig"	"Diplodocus"
[12] "Giraffe"	"Gorilla"	"Human"	"African elephant"	"Triceratops"	"Rhino"
[23] "Jaguar"	"Chimpanzee"	"Rat"	"Brachiosaurus"	"Mole"	"Pig"

Notare che sono presenti alcune specie estinte, come il Brachiosaurus, il Triceratops e il Diplodocus.

1. Si verifichi graficamente la presenza di valori anomali per la variabile **brain**, utilizzando il *boxplot*, e commentare:

```
with(Animals,
boxplot(brain, horizontal=TRUE)
)
```



Per la variabile **brain**, il *boxplot* identifica 3 valori anomali (evidenziandoli con  $\circ$ ), perchè risultano superiori al baffo di destra.

2. Ricavare i valori corrispondenti al baffo sinistro e al baffo destro del *boxplot* utilizzando il comando `boxplot.stats()`:

```
boxplot.stats(Animals$brain)
```

```
$stats
[1] 0.40 18.85 137.00 421.00 680.00
```

```
$n
[1] 28
```

```
$conf
[1] 16.92125 257.07875
```

```
$out
```

```
[1] 4603 1320 5712
```

```
baffo.sx <- boxplot.stats(Animals$brain)$stats[1]
baffo.sx
```

```
[1] 0.4
```

```
baffo.dx <- boxplot.stats(Animals$brain)$stats[5]
baffo.dx
```

```
[1] 680
```

3. Ricavare i nomi delle specie corrispondenti agli *outliers*:

```
outs <- boxplot.stats(Animals$brain)$out
outsTRUE <- Animals$brain %in% outs # restituisce un vettore logico
outsTRUE
```

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FA
which.outs <- which(outsTRUE) # indici delle osservazioni anomale
which.outs
```

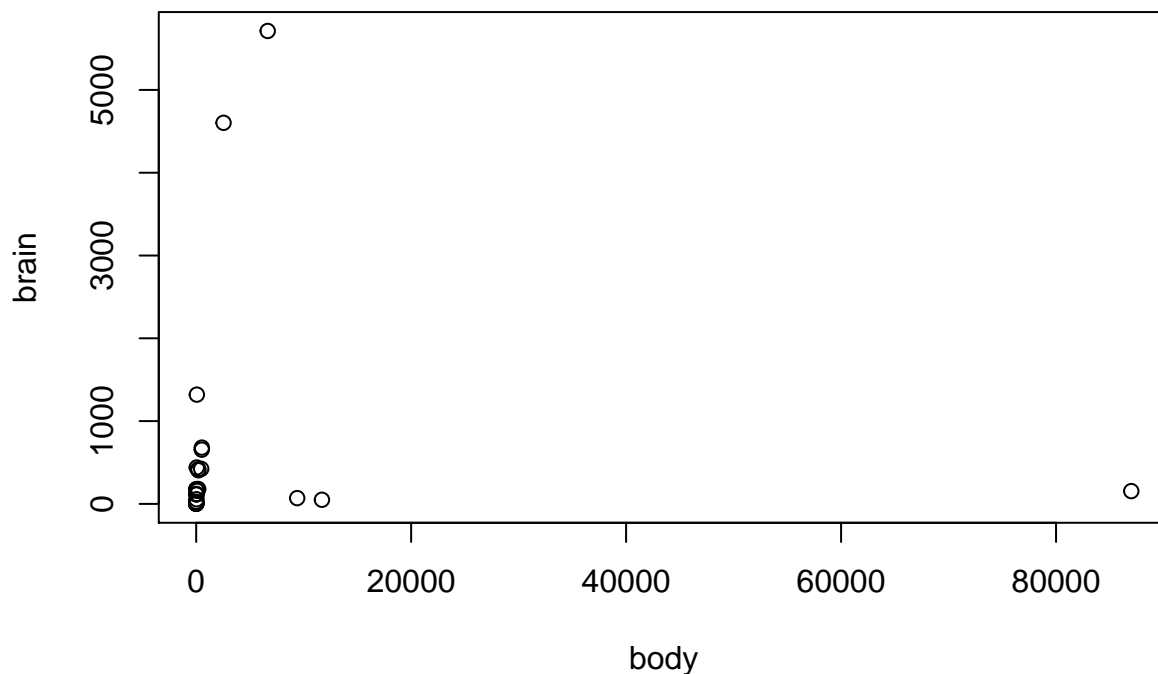
```
[1] 7 14 15
```

```
rownames(Animals)[which.outs]
```

```
[1] "Asian elephant" "Human" "African elephant"
```

4. Costruire il diagramma di dispersione del peso del cervello in funzione del peso del corpo, calcolare la matrice di correlazione, e commentare il risultato ottenuto:

```
plot(brain~body, Animals)
```



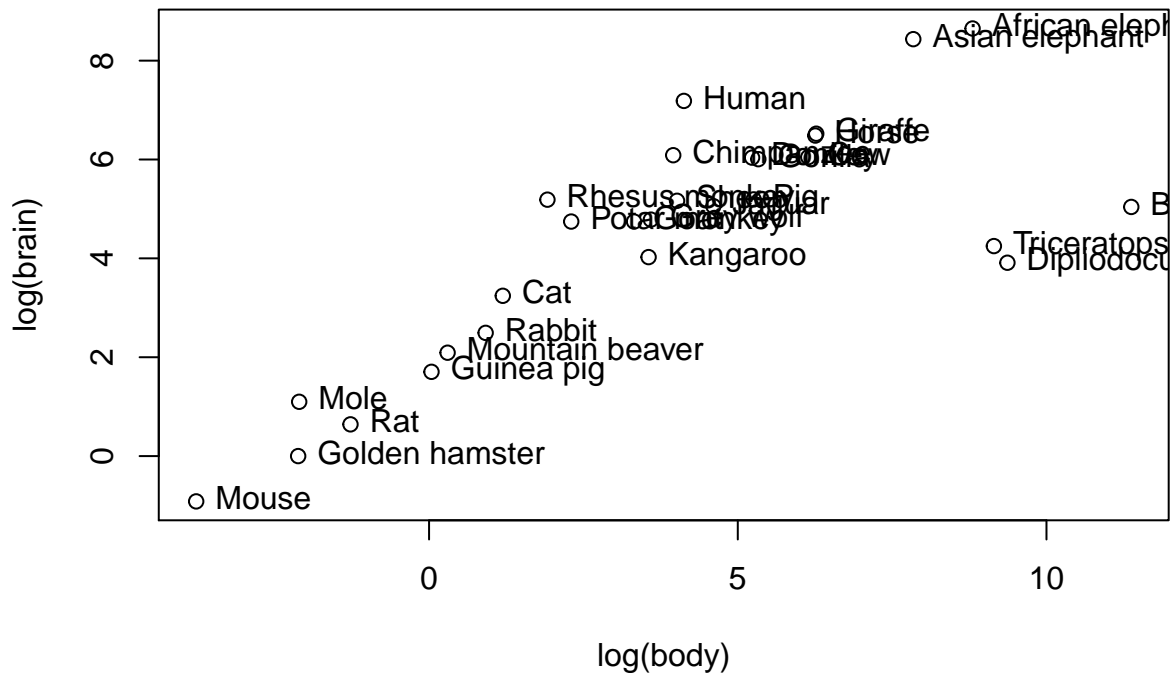
```
cor(Animals)
```

```
      body      brain
body  1.000000000 -0.005341163
brain -0.005341163  1.000000000
```

La correlazione è quasi nulla, e quindi c'è poca dipendenza lineare tra il peso del cervello e il peso del corpo delle specie considerate.

5. Trasformare entrambe le variabili con la trasformazione logaritmica (che è una trasformazione non lineare), costruire il diagramma di dispersione e calcolare la matrice di correlazione, commentando i risultati ottenuti.

```
plot(log(brain)~log(body), Animals)
with(Animals,
      text(log(brain)~log(body), labels = row.names(Animals), pos=4)
)
```



```
cor(log(Animals))
```

```
      body    brain
body 1.0000000 0.7794935
brain 0.7794935 1.0000000
```

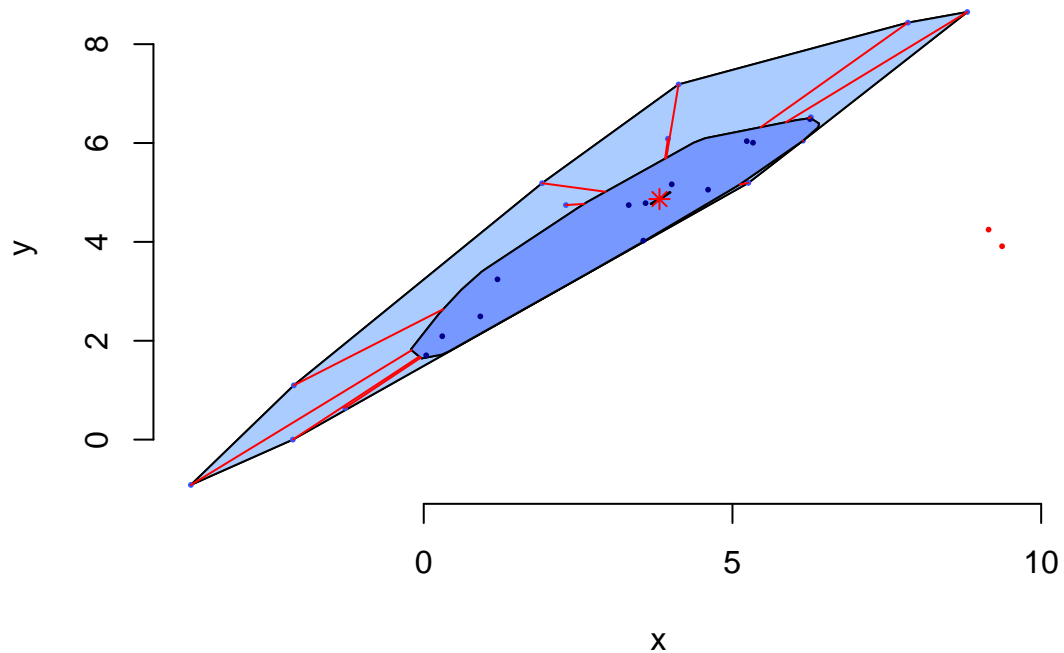
Vediamo che ora è presente una sostanziale correlazione tra il logaritmo del peso del cervello e il logaritmo del peso del corpo. Questo esempio dimostra inoltre che la matrice di correlazione non è invariante rispetto a trasformazioni non lineari.

Notiamo però che ci sono 3 osservazioni che non si comportano come le altre (le specie estinte). Si tratta forse di valori anomali (*outliers*)? Potrebbe trattarsi di *outliers* bivariati, perchè non riusciamo ad identificare queste osservazioni anomale con l'utilizzo dei *boxplot*.

6. Costruire il *bagplot* per verificare se si tratta effettivamente di *outliers* bivariati, e ricavare i valori di queste osservazioni anomale.

```
library(aplpack) # bagplot richiede l'installazione del pacchetto aplpack
bag <- bagplot(log(Animals))
```



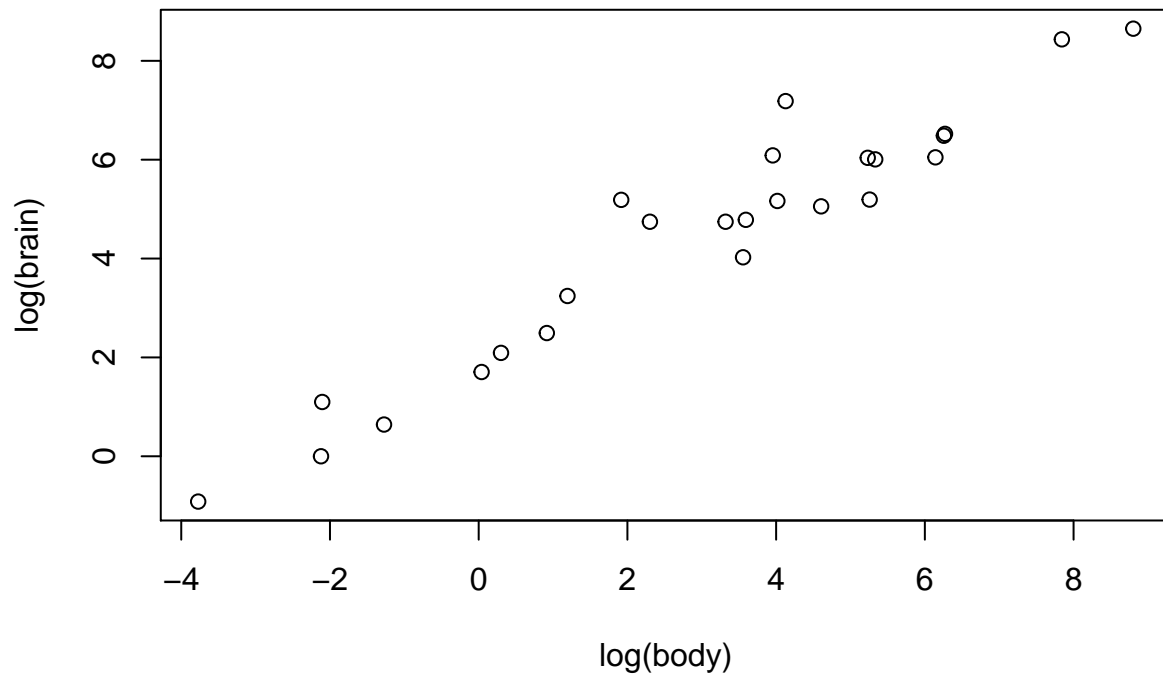


```
# per ricavare i valori anomali del bagplot
bag$pxy.outlier
```

```
      x      y
[1,] 9.367344 3.912023
[2,] 9.148465 4.248495
[3,] 11.373663 5.040194
```

7. Le unità statistiche sospette sono le specie Brachiosaurus, Triceratops e Dipliodocus. Identificare a quali righe della matrice corrispondono, costruire il diagramma di dispersione e la matrice di correlazione senza queste osservazioni, commentando il risultato.

```
which.out<-which( rownames(Animals) %in% c("Brachiosaurus", "Triceratops", "Dipliodocus"))
plot(log(brain)~log(body), Animals[-which.out,])
```



```
cor(log(Animals[-which.out,]))
```

```

      body      brain
body 1.0000000 0.9600516
brain 0.9600516 1.0000000

```

Dal diagramma di dispersione senza le osservazioni anomale si vede che le variabili  $\log(\text{body})$  e  $\log(\text{brain})$  presentano una dipendenza lineare positiva (sono correlate positivamente), quantificata come molto forte dal coefficiente di correlazione lineare  $r_{12} = 0.96$  (Si noti che senza rimuovere le osservazioni anomale,  $r_{12} = 0.78$ ).