La matrice dei dati X

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

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

```
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)
colnames(X)<-c("x1","x2")
n <- nrow(X)
p <- ncol(X)
X
```

```
##      x1   x2
## [1,]  3  5.0
## [2,]  4  5.5
## [3,]  2  4.0
## [4,]  6  7.0
## [5,]  8 10.0
## [6,]  2  5.0
## [7,]  5  7.5
```

Per ottenere la matrice trasposta

```
t(X)
```

```
##    [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## 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
```

```r
( ((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:

```r
( apply(X,MARGIN=2,FUNC="mean") )
```

```
##       x1       x2
## 4.285714 6.285714
```

Per calcolare la matrice di varianze/covarianze S

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

```
##          x1       x2
## x1 4.204082 3.704082
## x2 3.704082 3.561224
```

Per calcolare la matrice di correlazione R

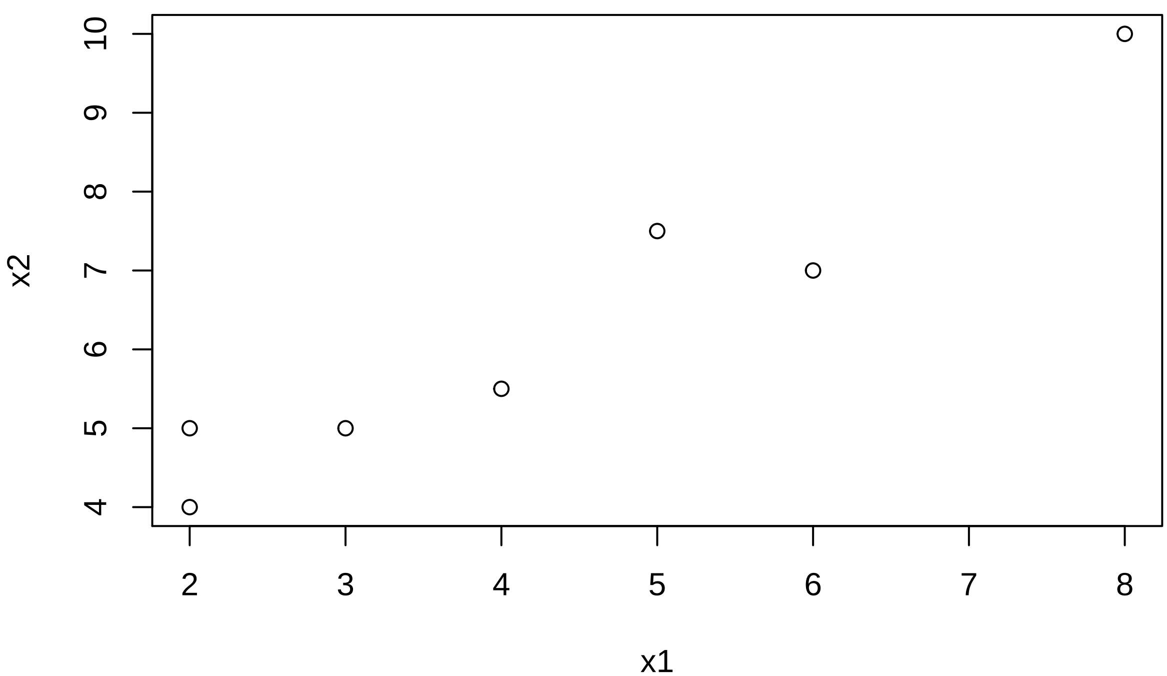
```r
( R = cor(X) )
```

```
##           x1        x2
## x1 1.0000000 0.9572939
## x2 0.9572939 1.0000000
```

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

Per costruire il diagramma di dispersione:

```r
plot(X)
```



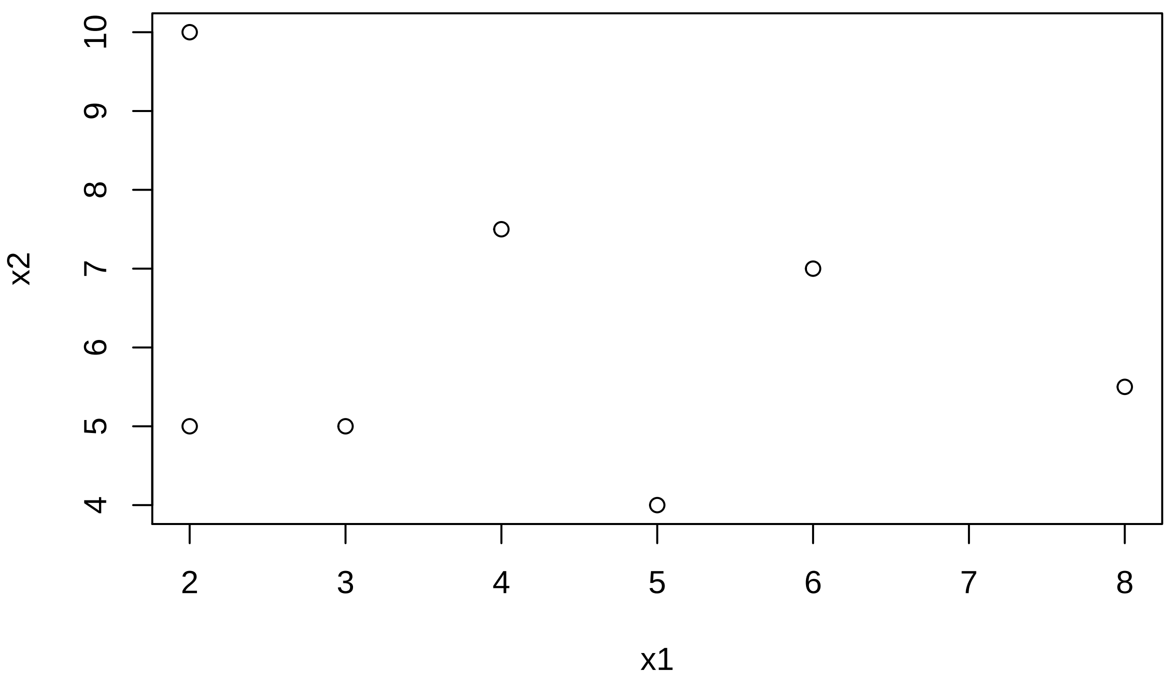
Per cambiare l'ordine (permutare) dei valori della variabile x_1 in maniera causale, si può utilizzare il comando `sample()`. Per riproducibilità dei risultati, impostare all'inizio il seme generatore dei numeri casuali `set.seed(123)`.

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

```
##      x1   x2
## [1,]  2  5.0
## [2,]  8  5.5
## [3,]  5  4.0
## [4,]  6  7.0
## [5,]  2 10.0
## [6,]  3  5.0
## [7,]  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.

```r
plot(X)
```




```r
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
```

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

```
##           x1        x2
## x1  4.204082 -1.081633
## x2 -1.081633  3.561224
```

```r
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.

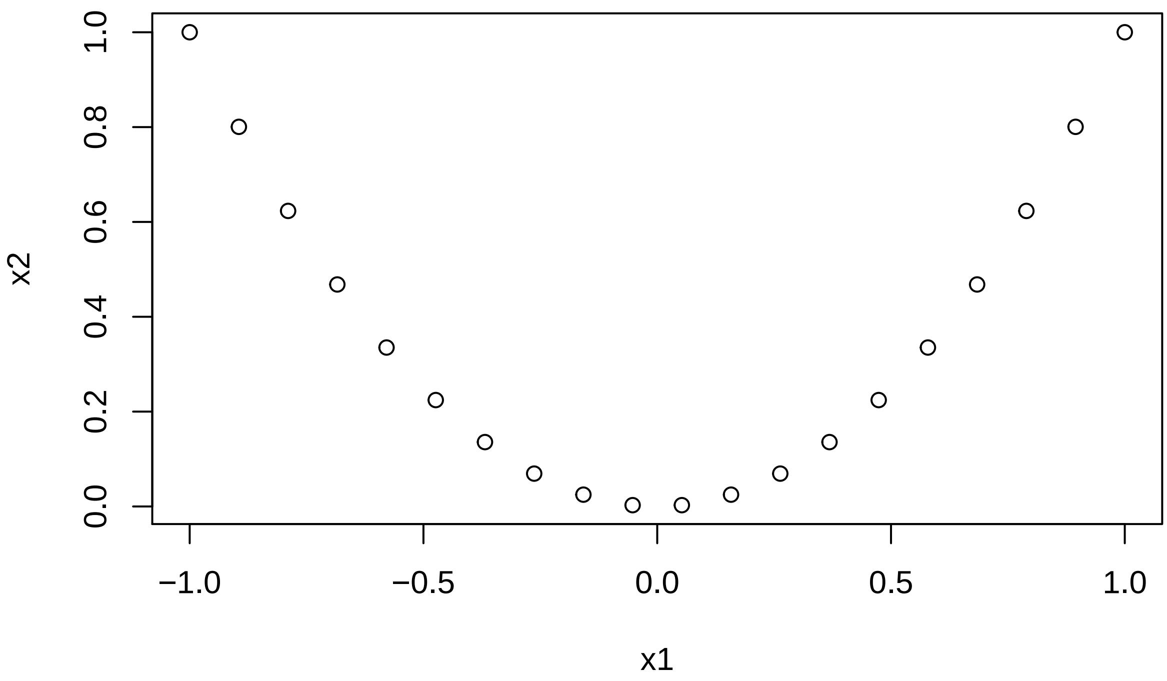
Relazione quadratica e correlazione

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

$$x_{1i} = -1 + 2\frac{(i-1)}{(n-1)}$$
$$x_{2i} = x_{1i}^2, \quad i = 1, \ldots, n$$

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.

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




```r
cor(X)
```

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

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.

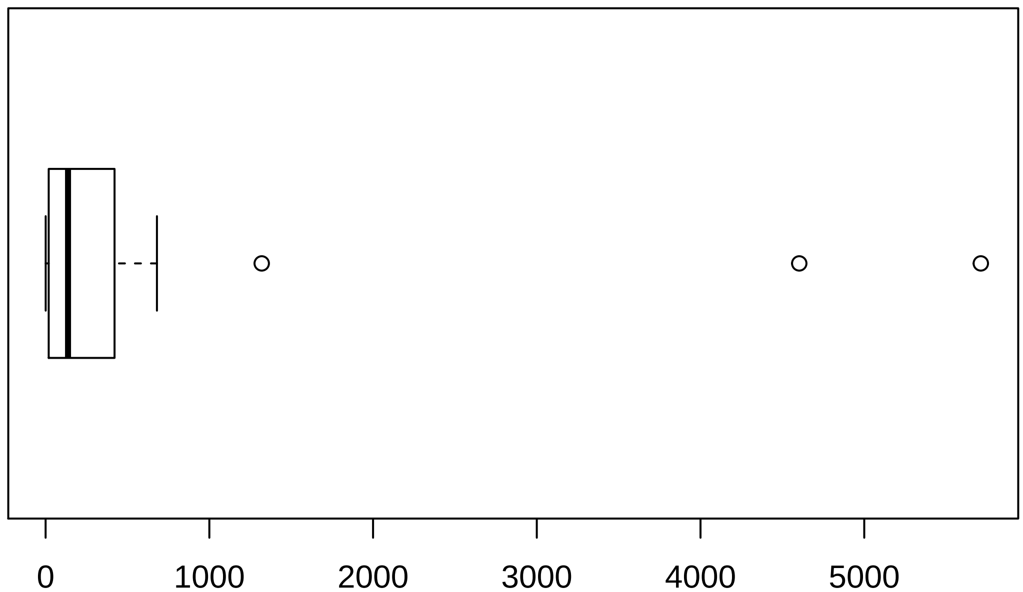
```
row.names(Animals)
```

```
##  [1] "Mountain beaver"  "Cow"              "Grey wolf"
##  [4] "Goat"             "Guinea pig"       "Dipliodocus"
##  [7] "Asian elephant"   "Donkey"           "Horse"
## [10] "Potar monkey"     "Cat"              "Giraffe"
## [13] "Gorilla"          "Human"            "African elephant"
## [16] "Triceratops"      "Rhesus monkey"    "Kangaroo"
## [19] "Golden hamster"   "Mouse"            "Rabbit"
## [22] "Sheep"            "Jaguar"           "Chimpanzee"
## [25] "Rat"              "Brachiosaurus"    "Mole"
## [28] "Pig"
```

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

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 ∘), 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] )
```

```
## [1] 0.4
```

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

```
## [1] 680
```

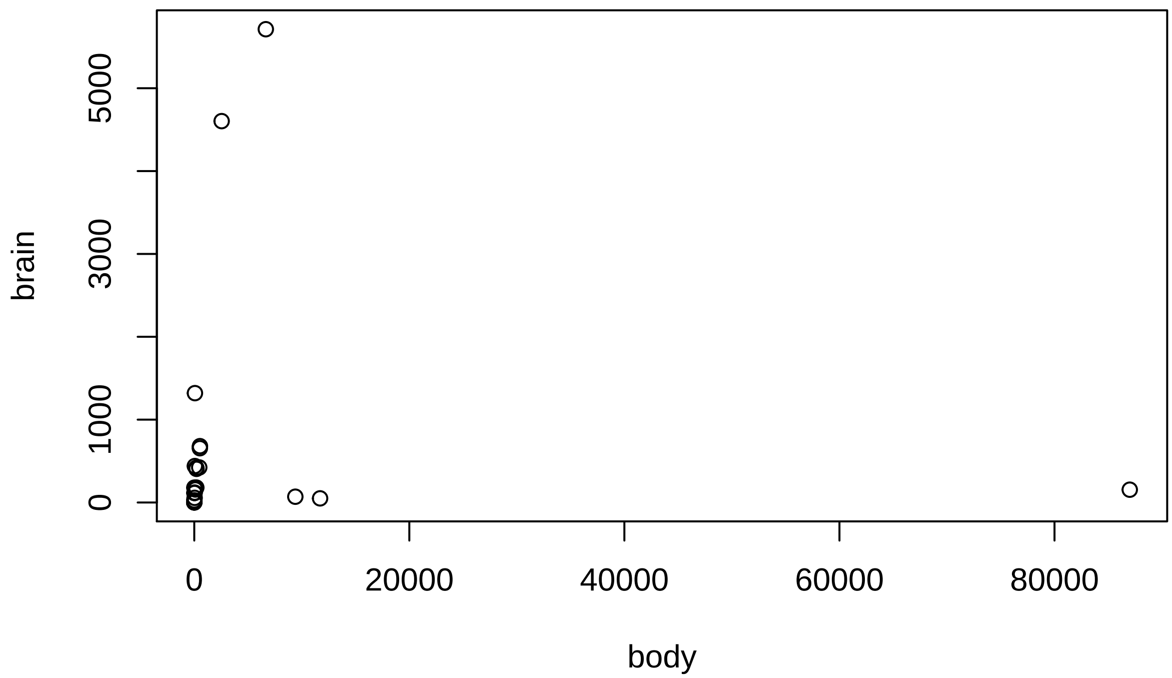
3. Ricavare i nomi delle specie corrispondenti agli *outliers*:

```
outs <- boxplot.stats(Animals$brain)$out
which.outs <- which( Animals$brain %in% outs )
( names.outs <- 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:

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



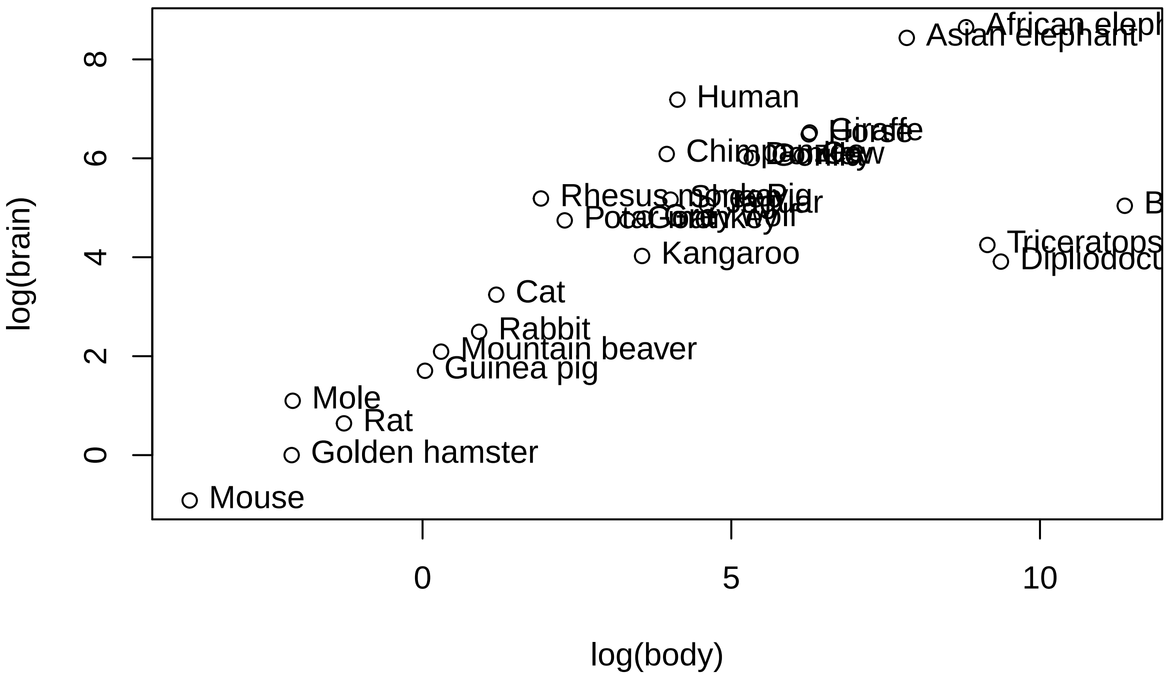

```r
cor(Animals)
```

```
##               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.

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




```r
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.

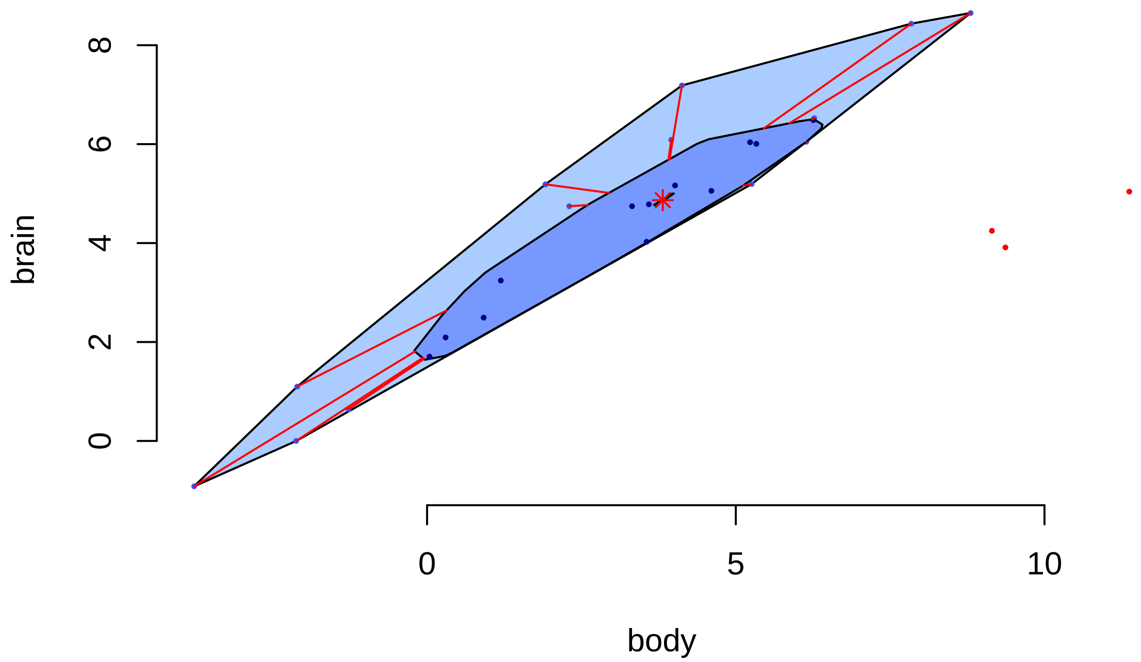
Notiamo però che ci sono 3 osservazioni che non si comportano come le altre (le specie estinte). Si tratta forse di valori anomali (*outliers*)? Potrebbe trattarsi di *outliers* bivariati, perchè non riusciamo ad identificare queste osservazioni anomale con l'utilizzo dei *boxplot*.

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

```r
library(aplpack) # bagplot richiede questo pacchetto
```

```
## Warning: running command ''/usr/bin/otool' -L '/Library/Frameworks/
## R.framework/Resources/library/tcltk/libs//tcltk.so'' had status 1
```

```r
bag <- bagplot(log(Animals))
```

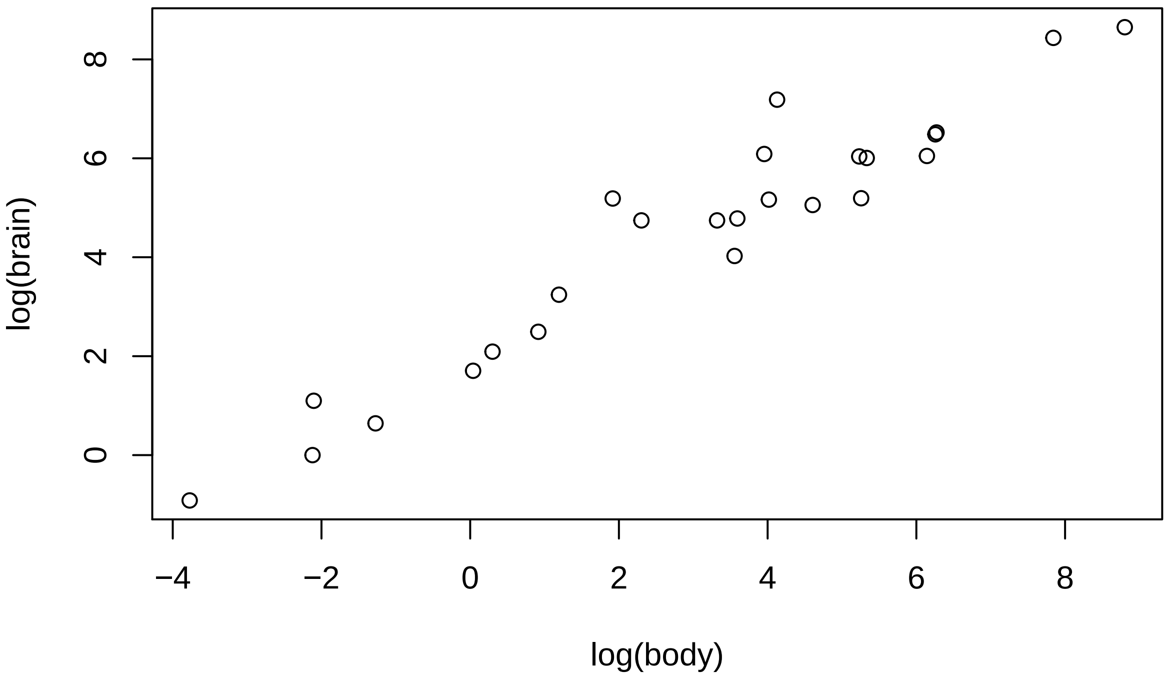



```r
# per ricavare i valori anomali
bag$pxy.outlier
```

```
##                     body    brain
## Dipliodocus     9.367344 3.912023
## Triceratops     9.148465 4.248495
## Brachiosaurus 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.

```r
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(body)` e `log(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, $r_{12} = 0.78$).