La matrice dei dati X

Analisi Esplorativa

Aldo Solari

I dati

I dati possono essere rappresentati con una tabella $n \times p$

- n osservazioni o unità statistiche: individui, aziende, etc.
- p variabili o misurazioni o caratteristiche: altezza, sesso, etc.

	Variabile 1	$\cdots$	Variabile j	$\cdots$	Variabile p
Unità statistica 1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1p}
Unità statistica 2	x_{21}	$\cdots$	x_{2j}	$\cdots$	x_{2p}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Unità statistica i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{ip}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Unità statistica n	x_{n1}	$\cdots$	x_{nj}	$\cdots$	x_{np}

- $n =$ numerosità dei dati
- $p =$ dimensionalità dei dati

Esempio

$n = 10$ individui e $p = 5$ variabili:

	sesso	figli	occhi	salute	peso
1	Maschio	0	Azzurri	Molto Buona	68.04
2	Maschio	1	Neri	Molto Buona	72.57
3	Maschio	0	Marroni	Media	61.23
4	Maschio	0	Neri	Cattiva	63.50
5	Maschio	1	Azzurri	Buona	49.90
6	Femmina	0	Marroni	Buona	49.90
7	Femmina	2	Azzurri	Molto Buona	54.43
8	Femmina	0	Marroni	Media	54.43
9	Femmina	0	Neri	Media	47.63
10	Femmina	1	Neri	Buona	45.36

Outline

1. Tipologia di variabili
2. Valori mancanti
3. Valori anomali
4. Matrice dei dati X
5. Diagramma di dispersione

Tipologia di variabili

Le variabili si suddividono in due tipologie:

Qualitative

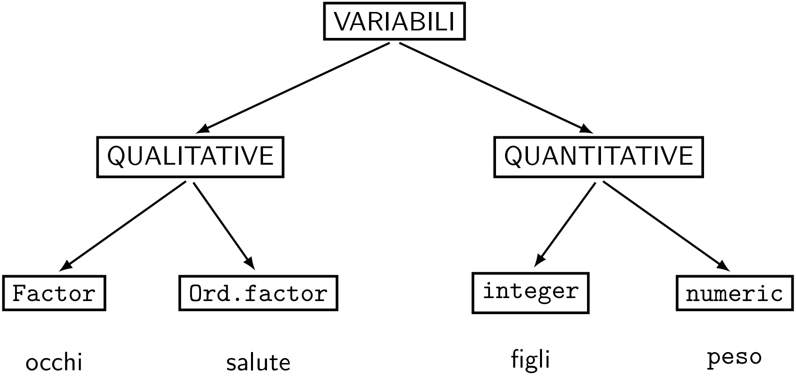
- nominali (`in R: Factor`), se non esiste nessun ordinamento naturale tra le modalità ;
- ordinali (`in R: Ord.factor`), se esiste un ordinamento naturale tra le modalità .

Quantitative

- discrete (`in R: integer`), quando sono esprimibili da numeri interi
- continue (`in R: numeric`), quando sono esprimibili da numeri reali

Variabili **Dicotomiche**: quando le modalità sono solamente due

Tipologia di variabili



Outline

Valori mancanti (*missing values*)

	sesso	figli	IQ	occhi	salute	peso
1	Maschio	0	120	Azzurri	Molto Buona	68.04
2	Maschio	1		Neri	Molto Buona	72.57
3	Maschio	0		Marroni	Media	61.23
4	Maschio	0	150	Neri	Cattiva	63.50
5	Maschio	1	92	Azzurri	Buona	49.90
6	Femmina	0	130	Marroni	Buona	49.90
7	Femmina			Azzurri	Molto Buona	54.43
8	Femmina	0		Marroni	Media	54.43
9	Femmina	0	84	Neri	Media	47.63
10	Femmina	1	70	Neri	Buona	45.36

NA

In `R`, i valori mancanti vengono codificati con `NA` (*Not Available*)

	sesso	figli	IQ	occhi	salute	peso
1	Maschio	0	120	Azzurri	Molto Buona	68.04
2	Maschio	1	`NA`	Neri	Molto Buona	72.57
3	Maschio	0	`NA`	Marroni	Media	61.23
4	Maschio	0	150	Neri	Cattiva	63.50
5	Maschio	1	92	Azzurri	Buona	49.90
6	Femmina	0	130	Marroni	Buona	49.90
7	Femmina	`NA`	`NA`	Azzurri	Molto Buona	54.43
8	Femmina	0	`NA`	Marroni	Media	54.43
9	Femmina	0	84	Neri	Media	47.63
10	Femmina	1	70	Neri	Buona	45.36

Problema: le tecniche di analisi multivariata che andremo a considerare prevedono osservazioni con tutti i valori presenti.

Esclusione di variabili incomplete

	sesso	~~figli~~	~~IQ~~	occhi	salute	peso
1	Maschio	~~0~~	~~120~~	Azzurri	Molto Buona	68.04
2	Maschio	~~1~~	~~NA~~	Neri	Molto Buona	72.57
3	Maschio	~~0~~	~~NA~~	Marroni	Media	61.23
4	Maschio	~~0~~	~~150~~	Neri	Cattiva	63.50
5	Maschio	~~1~~	~~92~~	Azzurri	Buona	49.90
6	Femmina	~~0~~	~~130~~	Marroni	Buona	49.90
7	Femmina	~~NA~~	~~NA~~	Azzurri	Molto Buona	54.43
8	Femmina	~~0~~	~~NA~~	Marroni	Media	54.43
9	Femmina	~~0~~	~~84~~	Neri	Media	47.63
10	Femmina	~~1~~	~~70~~	Neri	Buona	45.36

Diminuisce la dimensionalità p dei nostri dati. Però le variabili escluse potrebbero essere proprio quelle di interesse per l'analisi

Esclusione di osservazioni incomplete

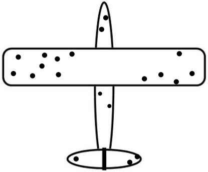
	sesso	figli	IQ	occhi	salute	peso
1	Maschio	0	120	Azzurri	Molto Buona	68.04
~~2~~	~~Maschio~~	~~1~~	~~NA~~	~~Neri~~	~~Molto Buona~~	~~72.57~~
~~3~~	~~Maschio~~	~~0~~	~~NA~~	~~Marroni~~	~~Media~~	~~61.23~~
4	Maschio	0	150	Neri	Cattiva	63.50
5	Maschio	1	92	Azzurri	Buona	49.90
6	Femmina	0	130	Marroni	Buona	49.90
~~7~~	~~Femmina~~	~~NA~~	~~NA~~	~~Azzurri~~	~~Molto Buona~~	~~54.43~~
~~8~~	~~Femmina~~	~~0~~	~~NA~~	~~Marroni~~	~~Media~~	~~54.43~~
9	Femmina	0	84	Neri	Media	47.63
10	Femmina	1	70	Neri	Buona	45.36

Diminuisce la numerosità n dei nostri dati. Vi vengono in mente altri potenziali problemi?

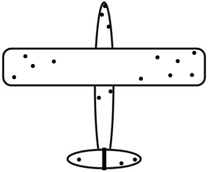
WWII

- Consideriamo come scenario la II Guerra Mondiale
- Vogliamo corazzare di più i nostri aerei, per proteggerli dai proiettili del nemico
- Le corazze sono pesanti, quindi non possiamo metterle dappertutto
- Per un aereo abbiamo 6 zone: (A) testa, (B) ali, (C) centro, (D) corpo anteriore, (E) corpo posteriore, (F) coda
- Possiamo mettere le corazze in sole 2 zone
- Dove le mettiamo?
- Guardiamo i dati degli aerei.

I dati: dove sono i buchi da proiettile

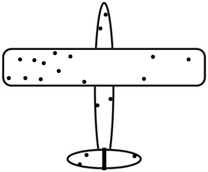


(a)



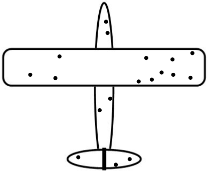
(b)

(c)



(d)

(e)



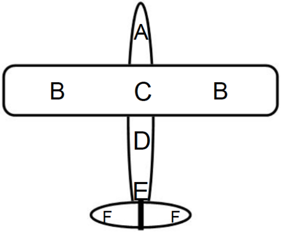
(f)

Matrice dei dati

	Testa	Ali	Centro	Corpo ant.	Corpo post.	Coda
a	2.1	2	0	1.8	0.1	1.9
b	1.9	2.1	0.2	1.7	0	1.8
c	?	?	?	?	?	?
d	2	2.2	0	1.9	0.1	2
e	?	?	?	?	?	?
f	2	2.1	0.2	1.8	0	1.9
Media	2	2.1	0.1	1.8	0.05	1.9

- Variabili espresse in termini di
  Densità di proiettili = Numero di proiettili / Area della zona
- Media calcolata escludendo i valori mancanti

Dove mettiamo le 2 protezioni?



(A) Testa

(B) Ali

(C) Centro

(D) Corpo anteriore

(E) Corpo posteriore

(F) Coda

L'opinione di uno statistico

The armor doesn't go where the bullet holes are.
It goes where the bullet holes aren't.

Abraham Wald

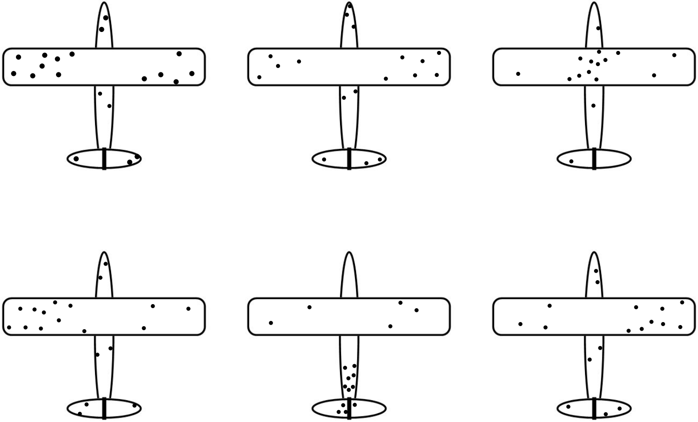
Perchè?

The observed holes showed where the planes were strongest;
that's where the planes could be shot and still survive the flight home.

The missing holes showed where the planes were weaker;
that's where the planes that didn't make it back were hit.

Pensate di far visita ad un ospedale militare durante una guerra:
vi aspettate di osservare più feriti alle gambe o alla testa?

Gli aerei mancanti (non a caso)



Valori mancanti (completamente) a caso

Si parla di valori mancanti (completamente) a caso se i valori mancanti sono un campione casuale dei $n \times p$ valori possibili.

In tale situazione non ci sono problemi se escludiamo le osservazioni che presentano almeno un valore mancante (tranne il fatto che diminuisce la numerosità n)

Outline

Valori anomali (*outliers*)

Ogni insieme di valori ha un massimo e un minimo, però può capitare di osservare uno o più valori veramente anomali (*outliers*)

Valore anomalo (*outlier*)

E' un valore che si discosta dal baricentro della distribuzione più di quanto possa essere giustificato dalla variabilità dei dati.

Perchè sono valori anomali?

Ci possono essere diverse spiegazioni, ad esempio:

Errore di rilevazione

e.g. per la variabile altezza, ho imputato 18.4 m invece di 1.84 m

Elevata variabilità intrinseca del fenomeno (code pesanti)

e.g. pensate alla variabile reddito

Valori provenienti da una distribuzione diversa (contaminazione)

e.g. pensate a misurare la temperatura degli oggetti in cucina, inclusi il forno e il tostapane accesi

Come si individuano i valori anomali?

Metodi basati sull'esplorazione grafica:

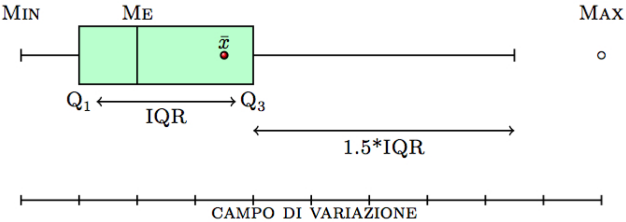
Per una singola variabile

- Diagramma a scatola con baffi (*boxplot*)

Per due variabili

- Diagramma di dispersione
- *Bagplot*

Diagramma a scatola con baffi (***boxplot***)



- Me, Q_1 e Q_3 sono la mediana, il primo e il terzo quartile
- $\mathrm{IQR} = Q_3 - Q_1$ è il *range* interquartile
- Il baffo a sinistra è il valore massimo tra Min e $Q_1 - 1.5 \cdot \mathrm{IQR}$
- Il baffo a destra è il valore minimo tra Max e $Q_3 + 1.5 \cdot \mathrm{IQR}$

Boxplot e valori anomali

Il diagramma a scatola e baffi (*boxplot*) identifica un valore anomalo (indicandolo con $\circ$) con la seguente regola:

Un valore x_i, $i = 1, \ldots, n$ è anomalo se:

- $x_i < Q_1 - 1.5 \cdot \mathrm{IQR}$ oppure se
- $x_i > Q_3 + 1.5 \cdot \mathrm{IQR}$

Comando `boxplot()` con `R`

- Se la numerosità campionaria n è un numero dispari, la descrizione coincide con quella delle slides precedenti;
- Se invece la numerosità campionaria n è un numero pari, i valori di Q_1 e Q_3 che calcola il comando `boxplot()` potrebbero essere leggermente diversi dal primo e il terzo quartile
- Potete utilizzare il comando `boxplot.stats()` per ottenere i 5 valori che compongono il *boxplot* (Min, baffo sx, Me, baffo dx, Max)

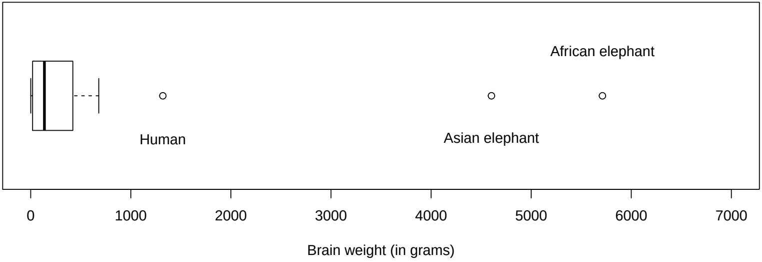
Dati Animals

- `Animals` è un *dataset* presente nella libreria `MASS`
- Per una descrizione del *dataset*, digitare `?Animals`
- *Average brain and body weights for 28 species of land animals*
- `body` : body weight in kg
- `brain` : brain weight in g
- $n = 28$ osservazioni misurate su $p = 2$ variabili

Dati Animals

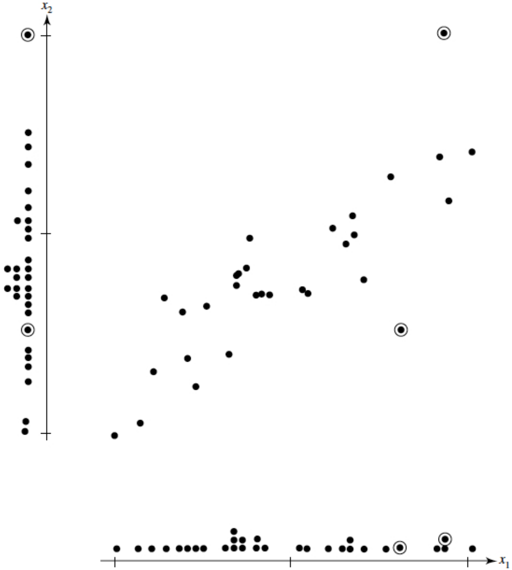
	body	brain
Mountain beaver	1.35	8.10
Cow	465.00	423.00
Grey wolf	36.33	119.50
Goat	27.66	115.00
Guinea pig	1.04	5.50
Dipliodocus	11700.00	50.00
Asian elephant	2547.00	4603.00
Donkey	187.10	419.00
Horse	521.00	655.00
Potar monkey	10.00	115.00
Cat	3.30	25.60
Giraffe	529.00	680.00
Gorilla	207.00	406.00
Human	62.00	1320.00
African elephant	6654.00	5712.00
Triceratops	9400.00	70.00
Rhesus monkey	6.80	179.00
Kangaroo	35.00	56.00
Golden hamster	0.12	1.00
Mouse	0.02	0.40
Rabbit	2.50	12.10
Sheep	55.50	175.00
Jaguar	100.00	157.00
Chimpanzee	52.16	440.00
Rat	0.28	1.90
Brachiosaurus	87000.00	154.50
Mole	0.12	3.00
Pig	192.00	180.00

Boxplot dati Animals

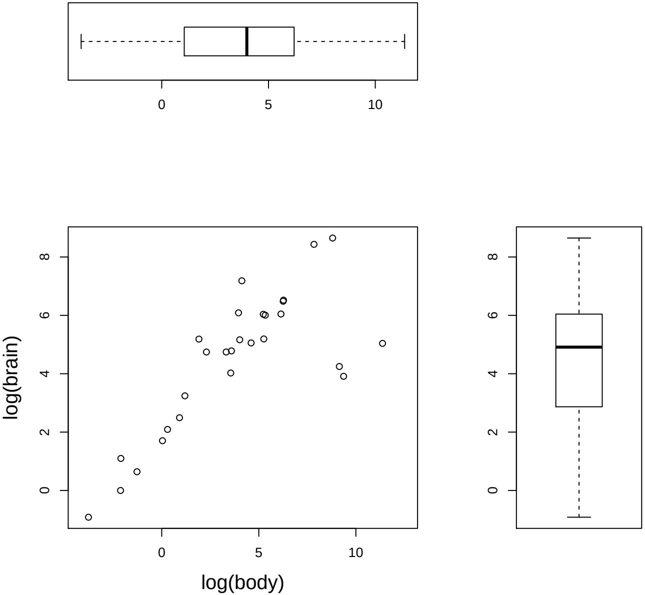



```
library("MASS")
boxplot(Animals$brain)
boxplot.stats(Animals$brain)
$stats
[1]   0.40  18.85 137.00 421.00 680.00
$out
[1] 4603 1320 5712
```

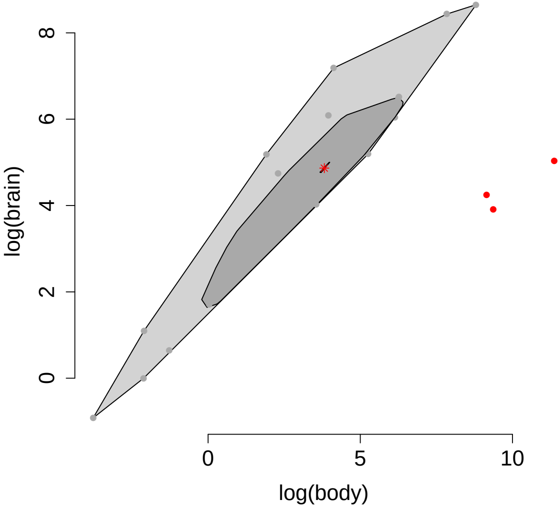
Outlier bivariato



Boxplot dati Animals

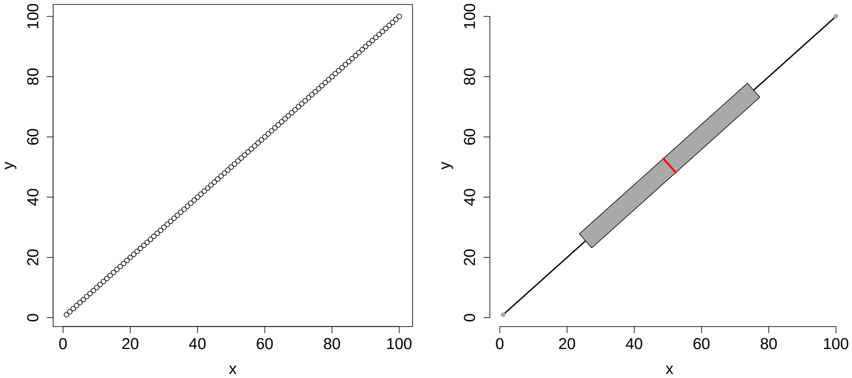


Bagplot = boxplot bivariato



50% delle osservazioni all'interno del *bag* (area grigio scuro)

Bagplot per dati unidimensionali



Outline

Matrice X

$$\underset{n\times p}{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

Medie e varianze

- Media per la j-sima variabile

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \qquad j = 1, \ldots, p$$

- Varianza per la j-sima variabile

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2, \qquad j = 1, \ldots, p$$

Covarianze e correlazioni

- Covarianza tra la j-sima e la k-sima variabile

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \qquad j = 1, \ldots, p, \ k = 1, \ldots, p$$

  Si noti che $s_{jk} = s_{kj}$ e che $s_{jj} = s_j^2$

- Correlazione tra la j-sima e la k-sima variabile

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}, \qquad j = 1, \ldots, p, \ k = 1, \ldots, p$$

  Si noti che $-1 \leq r_{jk} \leq 1$

Vettore delle medie

$$\underset{p\times 1}{\bar{x}} = \begin{bmatrix} \bar{x}_1 \\ \cdots \\ \bar{x}_j \\ \cdots \\ \bar{x}_p \end{bmatrix}$$

Matrice di varianze/covarianze

$$\underset{p\times p}{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1j} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2j} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ s_{j1} & s_{j2} & \cdots & s_{jj} & \cdots & s_{jp} \\ \vdots & \vdots & \ldots & \vdots & \ddots & \ldots \\ s_{p1} & s_{p2} & \cdots & s_{pj} & \cdots & s_{pp} \end{bmatrix}$$

Matrice di correlazione

$$\underset{p\times p}{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1j} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2j} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ r_{j1} & r_{j2} & \cdots & 1 & \cdots & r_{jp} \\ \vdots & \vdots & \ldots & \vdots & \ddots & \ldots \\ r_{p1} & r_{p2} & \cdots & r_{pj} & \cdots & 1 \end{bmatrix}$$

Esempio

Variabile 1 (prezzo in Dollari per libro):	42	52	48	58
Variabile 2 (numero di libri venduti):	4	5	4	3

Matrice X

$$\underset{4\times 2}{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

Vettore delle medie

$$\underset{2\times 1}{\bar{x}} = \left[\begin{array}{c} \bar{x}_1 \\ \bar{x}_2 \end{array} \right] = \left[\begin{array}{c} 50 \\ 4 \end{array} \right]$$

Matrice di varianze/covarianze

$$\underset{2\times 2}{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \begin{bmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{bmatrix}$$

Matrice di correlazione

$$\underset{2\times 2}{R} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} = \begin{bmatrix} 1 & -0.36 \\ -0.36 & 1 \end{bmatrix}$$

Outline

Dati

	u_1	u_2	u_3	u_4	u_5	u_6	u_7
x_1	3	4	2	6	8	2	5
x_2	5	5.5	4	7	10	5	7.5

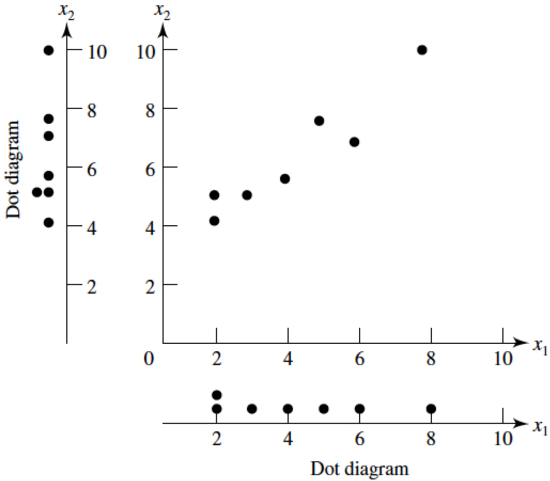
Medie: $\bar{x}_1 = 4.2$, $\bar{x}_2 = 6.2$

Varianze: $s_{11} = 4.2$, $s_{22} = 0.56$

Covarianza: $s_{12} = 3.70$

Correlazione: $r_{12} = 0.95$

Diagramma di dispersione



Dati

	u_1	u_2	u_3	u_4	u_5	u_6	u_7
x_1	5	4	6	2	2	8	3
x_2	5	5.5	4	7	10	5	7.5

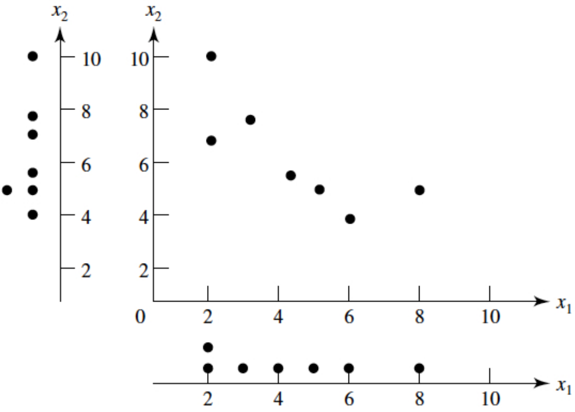
Medie: $\bar{x}_1 = 4.2$, $\bar{x}_2 = 6.2$

Varianze: $s_{11} = 4.20$, $s_{22} = 0.56$

Covarianza $s_{12} = -3.01$

Correlazione $r_{12} = -0.78$

Diagramma di dispersione



Web Apps

Guess the correlation

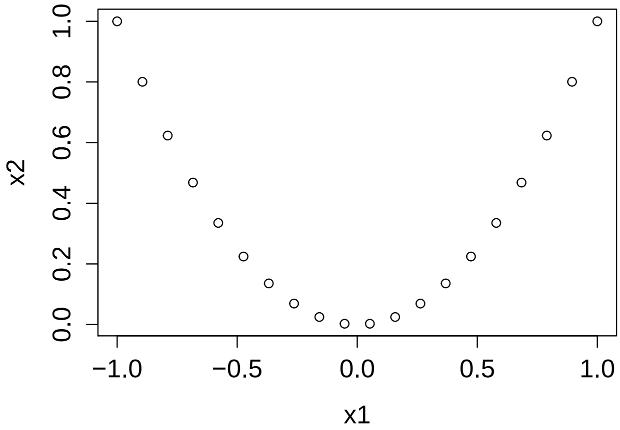
Correlation game

Relazione quadratica

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

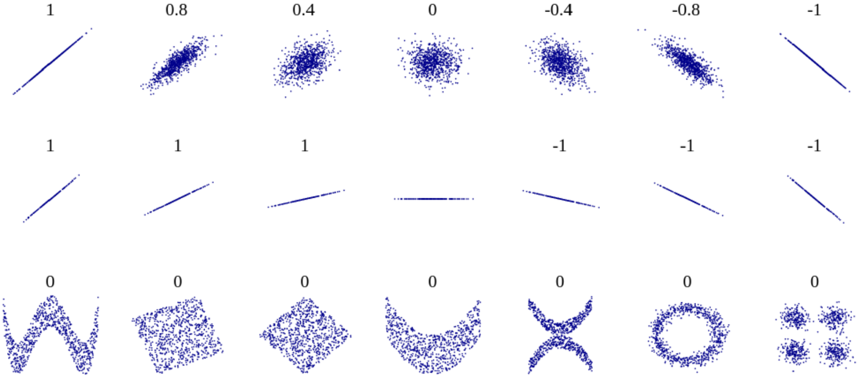
Relazione quadratica

Per $n = 20$:

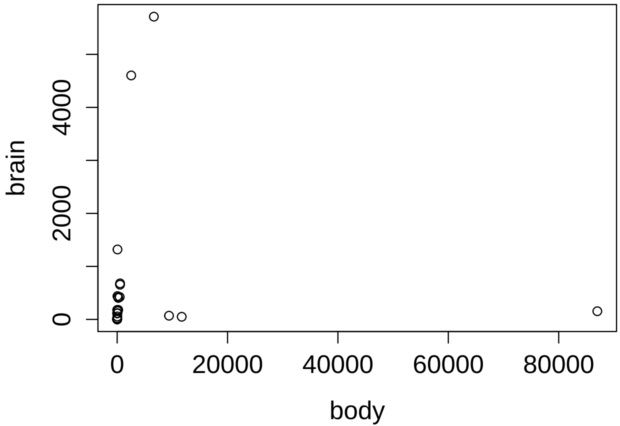


$$r_{12} \approx 0$$

Correlazione = relazione LINEARE

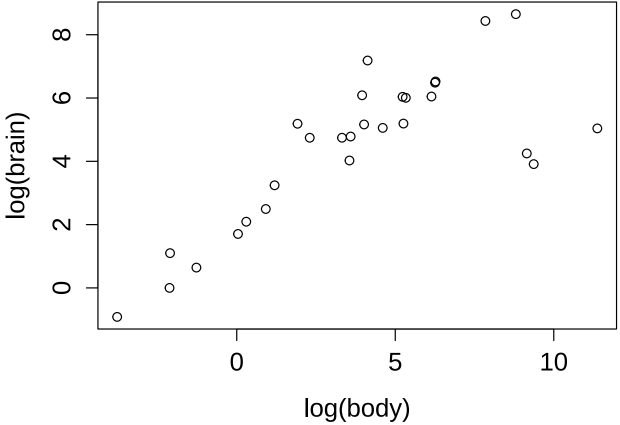


`Animals`: diagramma di dispersione



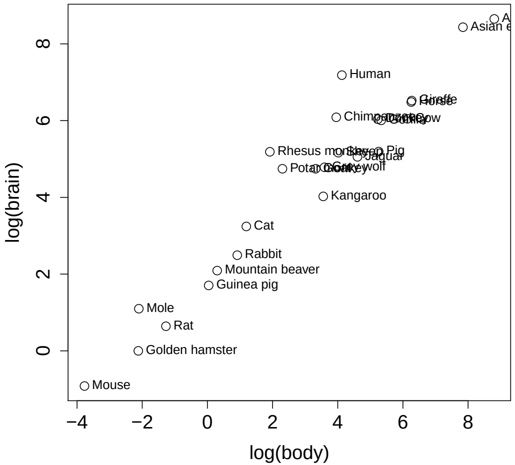
$$r_{12} = -0.0053$$

Animals: trasformazione logaritmica



$$r_{12} = 0.779$$

`Animals`: escludendo 3 osservazioni anomale



$$r_{12} = 0.932$$