Correlation

Come si calcola *r* ?

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

dove

$$\sigma_{XY} = \frac{\sum_i (x_i - \mu_X)(y_i - \mu_Y)}{N - 1}$$

è la **covarianza**
(il numeratore prende il nome di **codevianza**)

$$\sigma_{XY} \in (-\infty, +\infty)$$

$\sigma_{XY} = 0$ assenza di relazione lineare

$\sigma_{XY} > 0$ relazione diretta

$\sigma_{XY} < 0$ relazione inversa

Si dimostra (disuguaglianza di Cauchy-Schwarz) che:

$$-\sigma_X \sigma_Y \leq \sigma_{XY} \leq +\sigma_X \sigma_Y$$

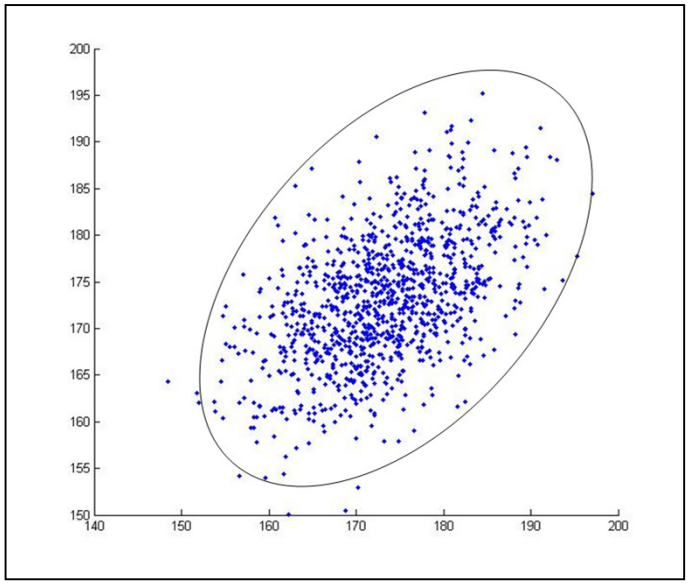
Normalizzando:

$$-1 \leq \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = r \leq 1$$

Aspetti della correlazione

- La correlazione non distingue fra variabili esplicative e di risposta
- La correlazione richiede che entrambe le variabili siano quantitative
- La correlazione non cambia se si cambiano le unità di misura di una o di entrambe le variabili
- Il segno di *r* indica la direzione dell'associazione
- *r* è sempre compreso fra –1 e +1
- La correlazione misura solo la forza di una relazione lineare fra due variabili
- La correlazione non è una misura robusta

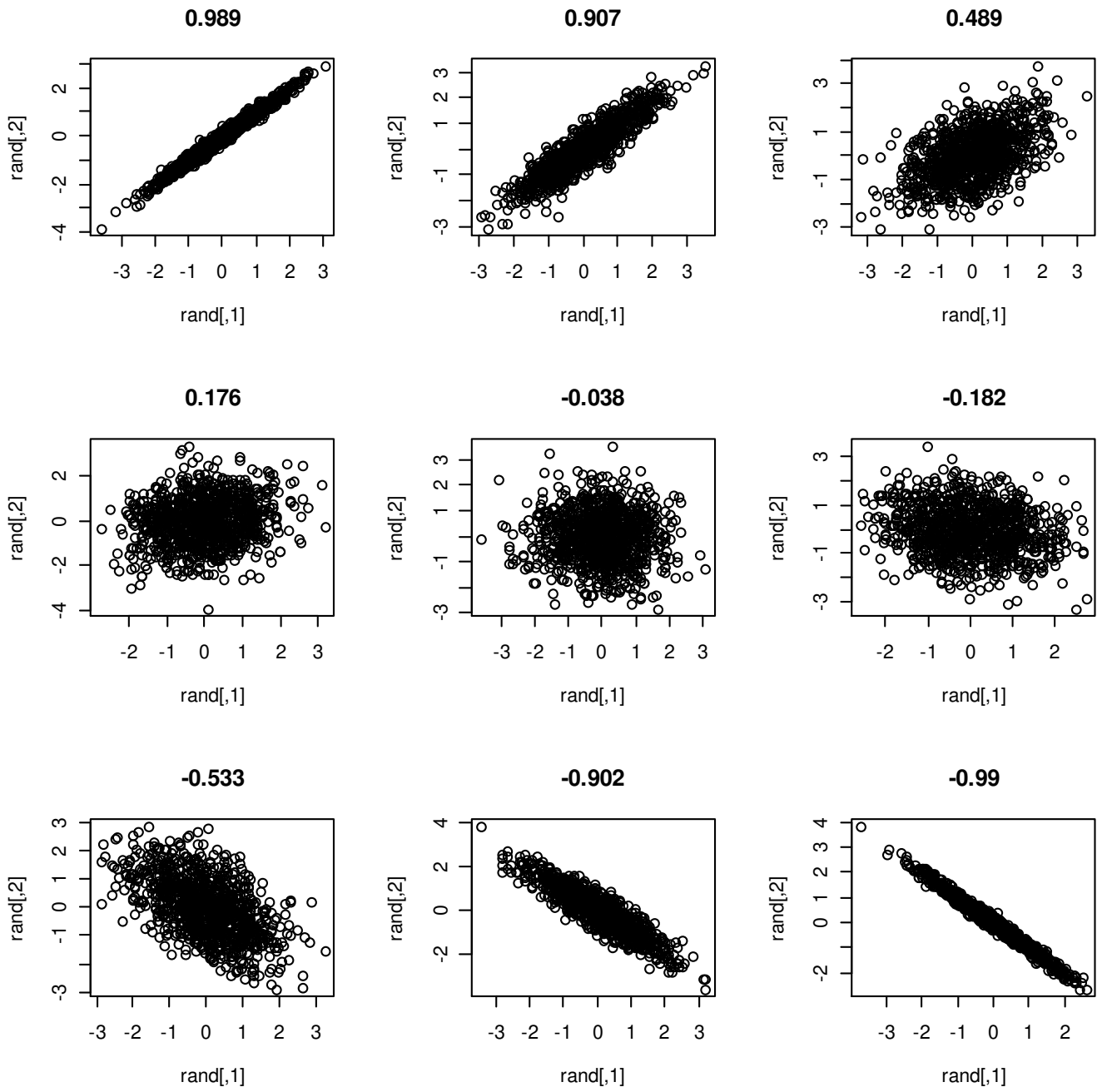
r è una misura molto utile in tutti i casi in cui lo *scatter* abbia una forma "ovale":



A volte *r* può essere una misura fuorviante e non opportuna

i) quando ci sono *outlier*;

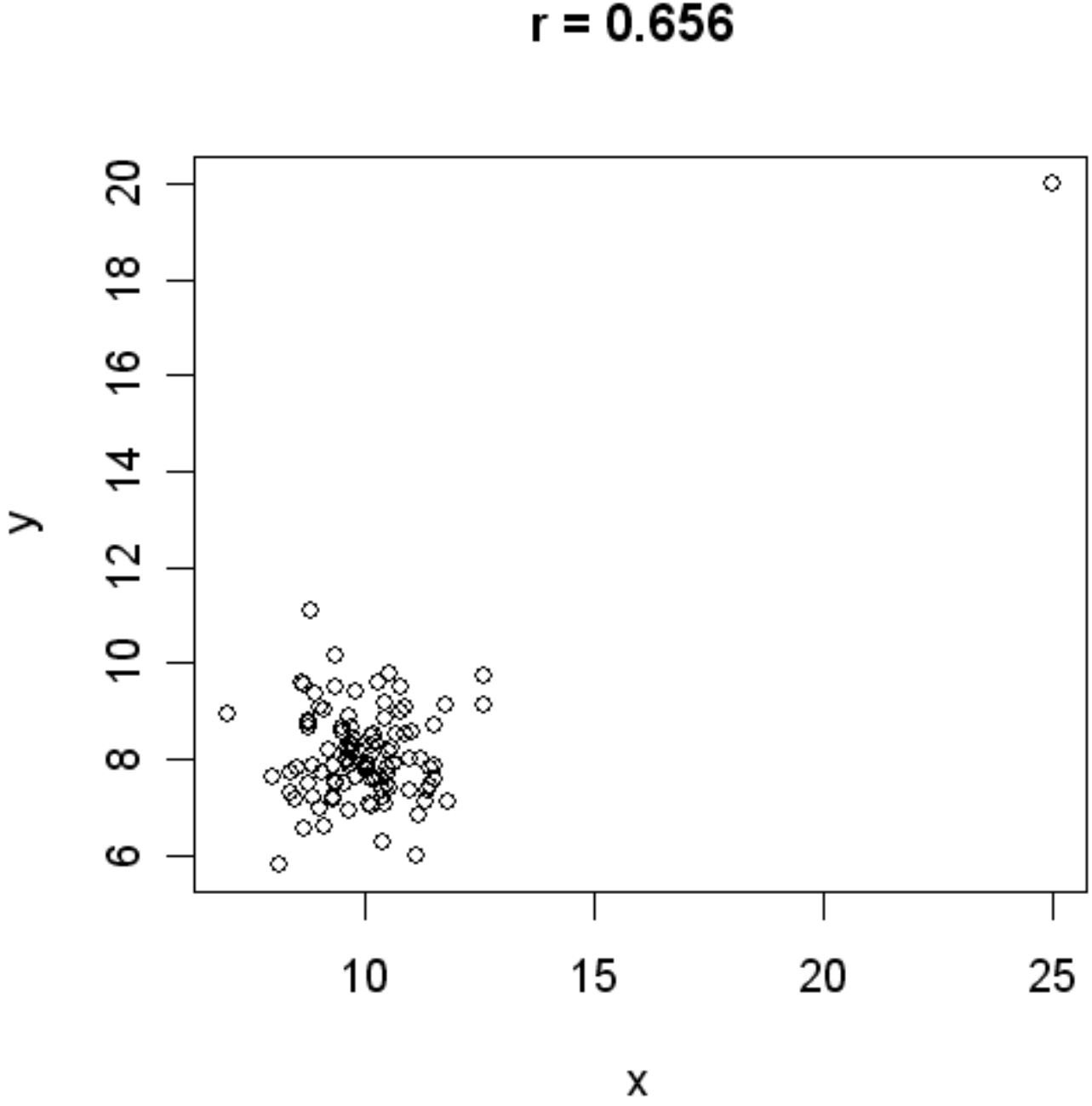
ii) quando c'è una relazione **non** lineare.


0.989
0.907
0.489
0.176
-0.038
-0.182
-0.533
-0.902
-0.99
rand[,1]
rand[,2]

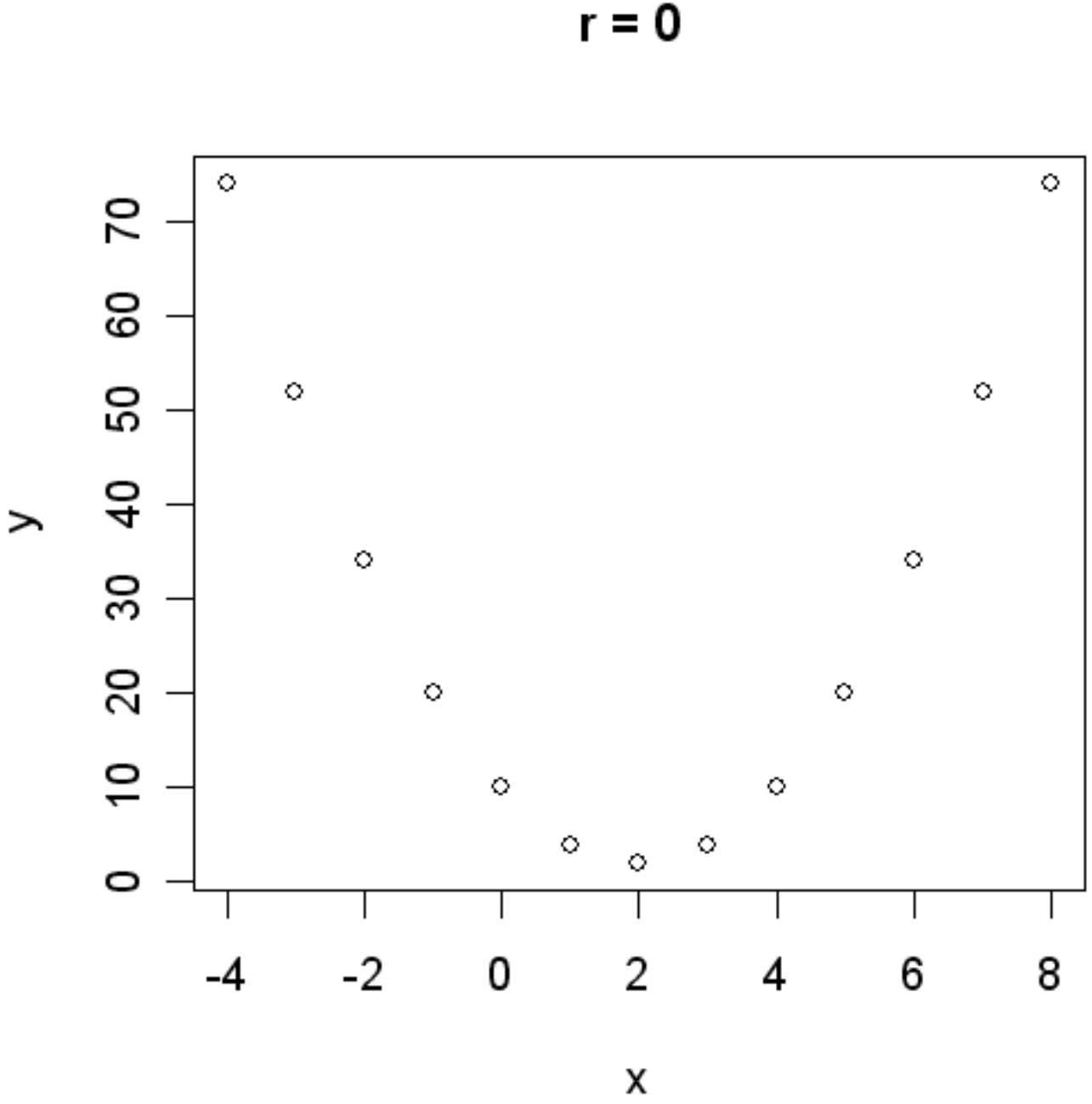
```
> x <- rnorm(100)+10; y <- rnorm(100)+8; r(x,y)
Coefficiente di correlazione lineare: 0.00183
...

> x <- c(x,25); y <- c(y,20); r(x,y)
Coefficiente di correlazione lineare: 0.6561239
…

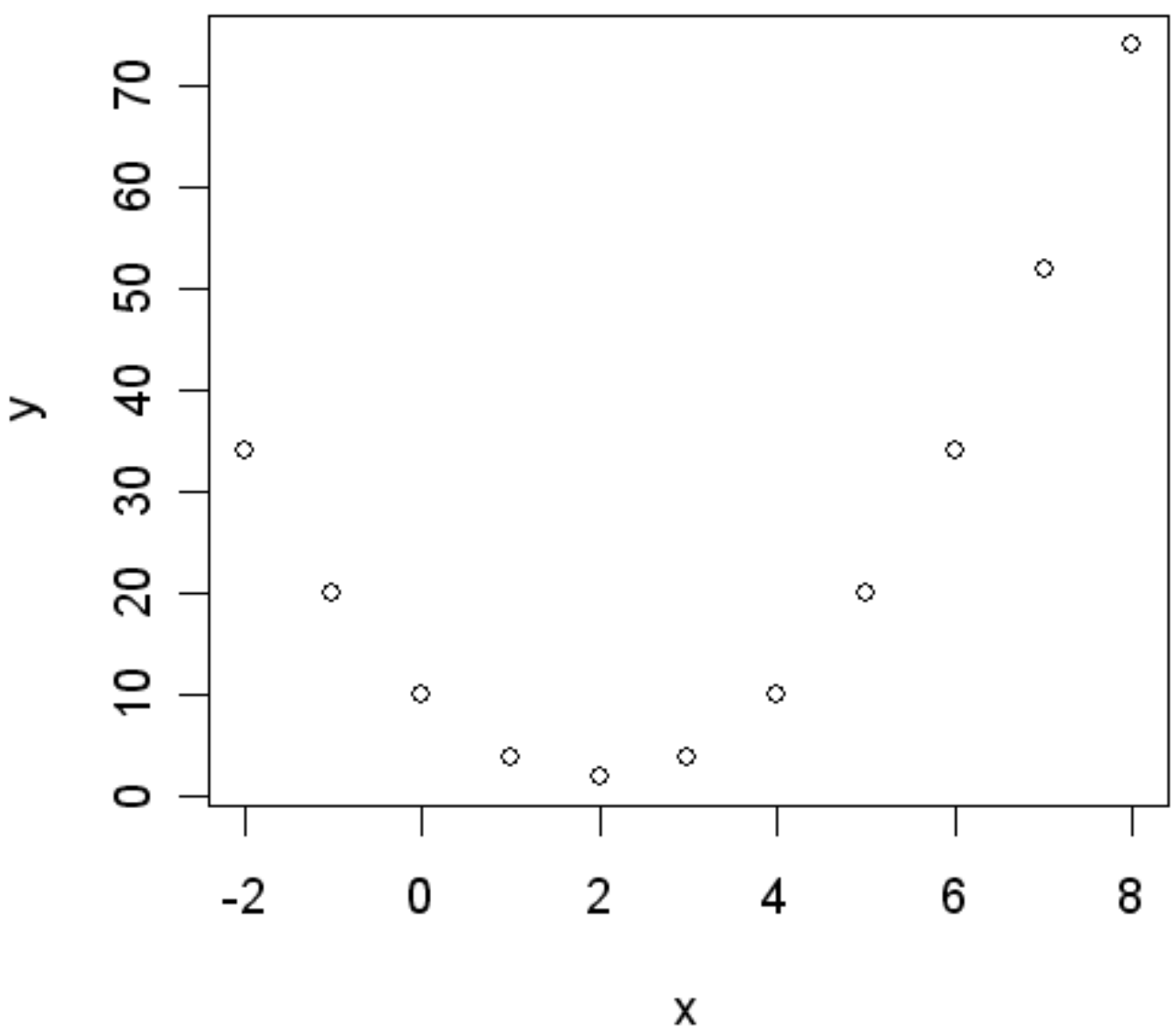
> plot(x,y,
    main=paste("r =",round(cor(x,y),3),"\n"))
```


r = 0.656
y
6
8
10
12
14
16
18
20
10
15
20
25
x

```
> x <- c(-4:8); y <- 2*x**2-8*x+10
> cbind(x,y)
       x  y
 [1,] -4 74
 [2,] -3 52
 [3,] -2 34
 [4,] -1 20
 [5,]  0 10
 [6,]  1  4
 [7,]  2  2
 [8,]  3  4
 [9,]  4 10
[10,]  5 20
[11,]  6 34
[12,]  7 52
[13,]  8 74
> cor(x,y)
[1] 0
```

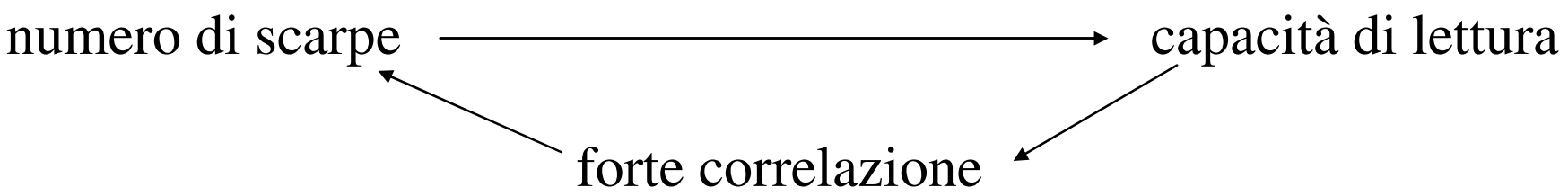

r = 0
y
0
10
20
30
40
50
60
70
-4
-2
0
2
4
6
8
x

```
> x <- x[-c(1,2)]; y <- y[-c(1,2)]
> cbind(x,y)
       x  y
 [1,] -2 34
 [2,] -1 20
 [3,]  0 10
 [4,]  1  4
 [5,]  2  2
 [6,]  3  4
 [7,]  4 10
 [8,]  5 20
 [9,]  6 34
[10,]  7 52
[11,]  8 74
> cor(x,y)
[1] 0.5822225
```


r = 0.582
y
0
10
20
30
40
50
60
70
-2
0
2
4
6
8
x

Correlazione e causalità

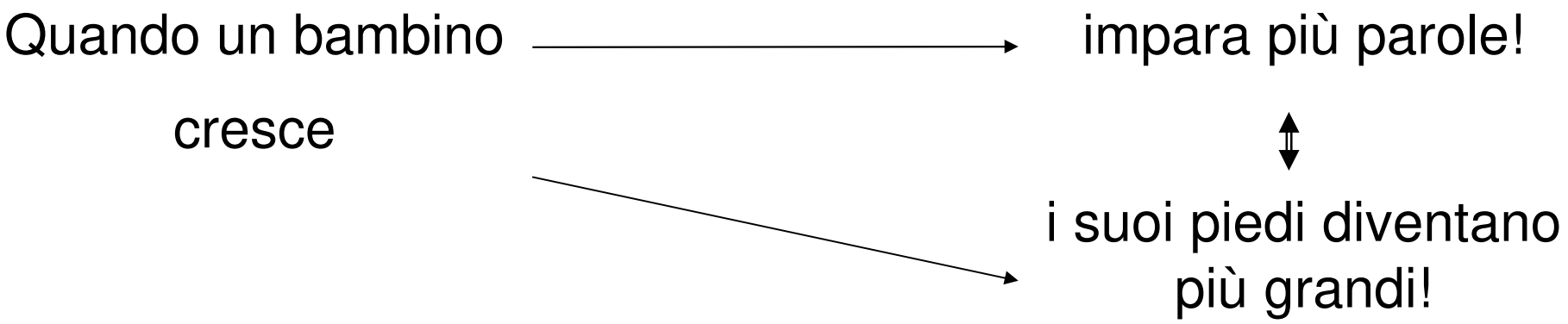
Bambini:



La capacità di imparare nuove lettere può rendere più grande il piede?

NO!!

Esiste un fattore di disturbo: l'età!



r non protegge dai fattori di disturbo

r misura la forza ed i verso della relazione lineare tra *X* e *Y*

$\neq$

nessi di causalità

Correlazioni spurie e variabili nascoste

- http://www.tylervigen.com/spurious-correlations
- Numero di cicogne e numero di bambini (https://onlinelibrary.wiley.com/doi/epdf/10.1111/1467-9639.00013)

Durante i mesi invernali, nelle case in cui è presente un neonato la temperatura viene mantenuta più alta della norma, passando indicativamente dai 16 ai 20 gradi centigradi. Soprattutto nei periodi più rigidi, le cicogne sono attratte dal maggior calore emesso dai camini e nidificano più facilmente su di essi o vi si soffermano più a lungo.

È semplice suddividere un'ampia area rurale in zone con una popolazione equivalente e contare per ognuna il numero dei camini con cicogne e quello dei bambini neonati.

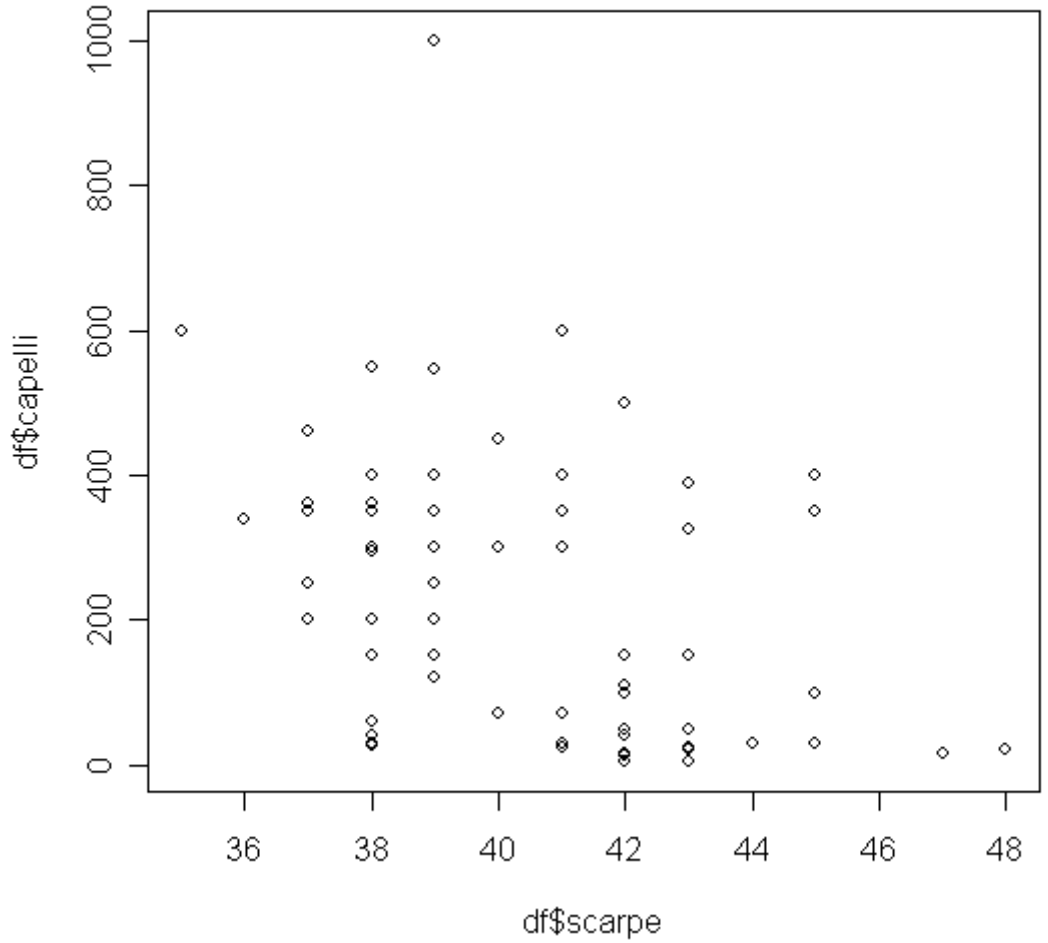
Calcolare la correlazione fra queste due variabili sembra implicare una relazione di causa-effetto tra presenza di cicogne e nascite di bambini che porterebbe a concludere che i bambini sono portati dalle cicogne.

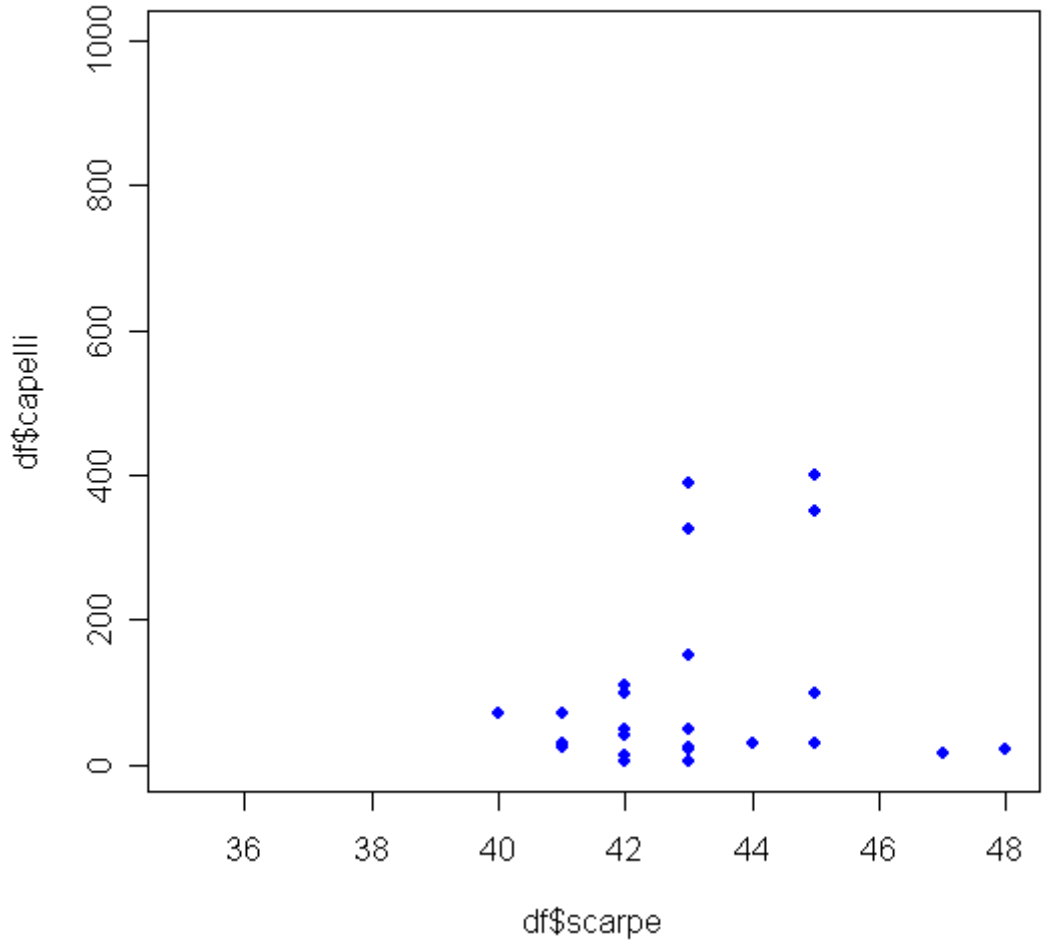
Correlazioni spurie e variabili nascoste

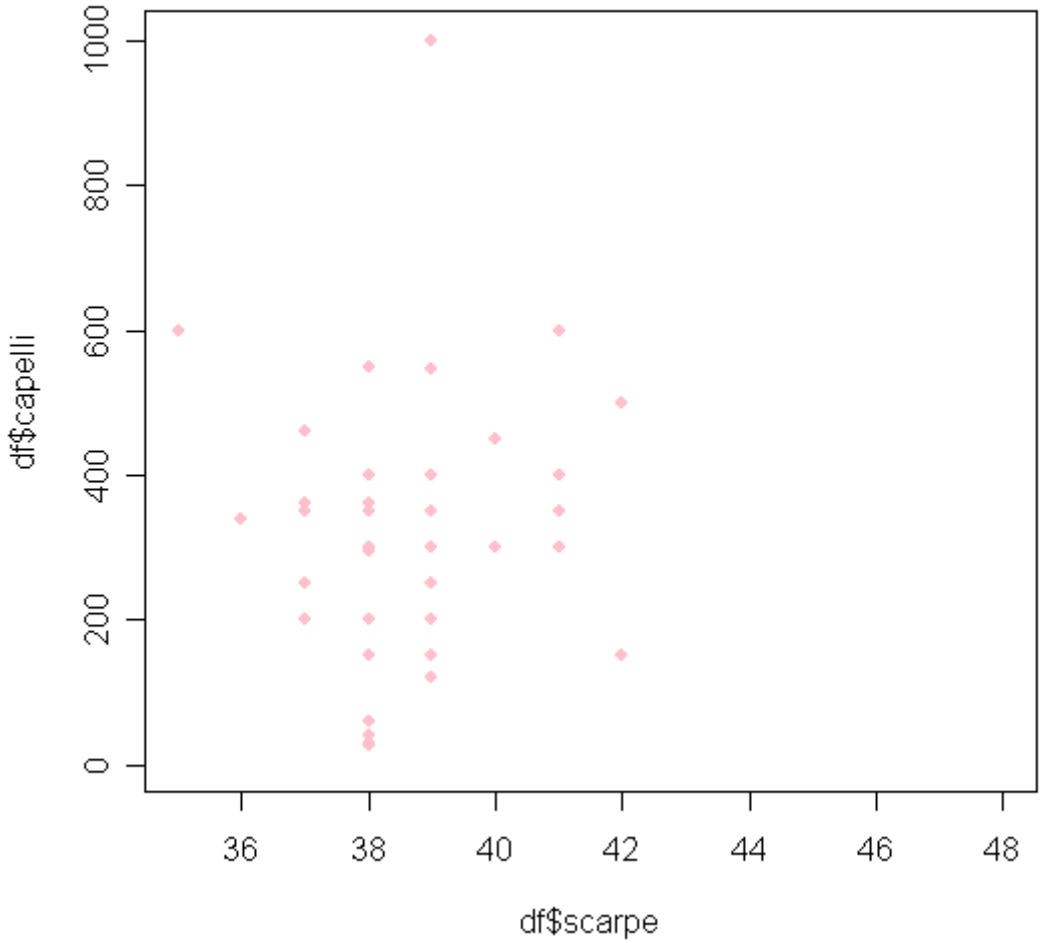
- Numero di scarpe e lunghezza dei capelli

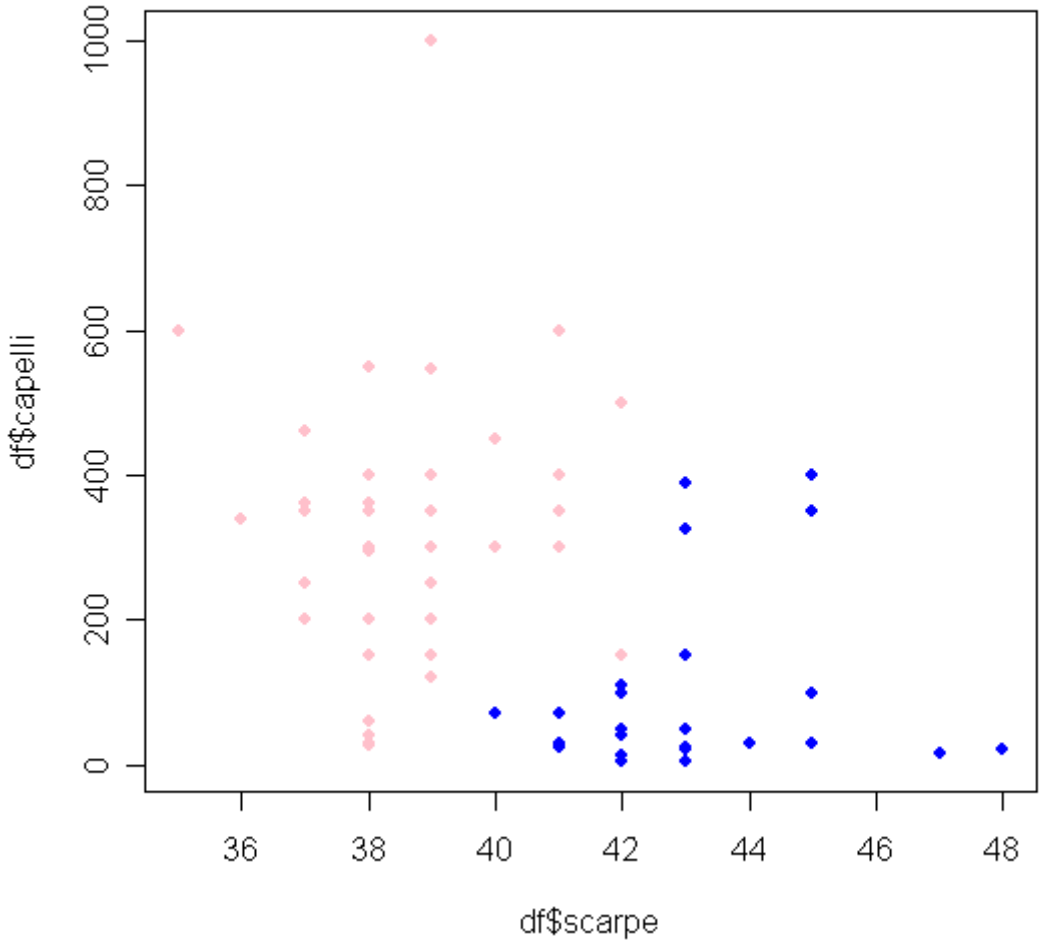
```
> r(df$scarpe,df$capelli)
Coefficiente di correlazione lineare:  -0.4061388
Numero di coppie di osservazioni    : 75

> r(df$scarpe,df$peso)
Coefficiente di correlazione lineare:  0.8048932
Numero di coppie di osservazioni    : 88
```


df$capelli
0
200
400
600
800
1000
36
38
40
42
44
46
48
df$scarpe


df$capelli
0
200
400
600
800
1000
36
38
40
42
44
46
48
df$scarpe


df$capelli
0
200
400
600
800
1000
36
38
40
42
44
46
48
df$scarpe


df$capelli
0
200
400
600
800
1000
36
38
40
42
44
46
48
df$scarpe

Inference

Significance Test

$$\frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Confidence Interval

$$T(r) = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] \qquad T(r) \sim \mathrm{N}\left(T(\rho), \frac{1}{n\text{-}3}\right)$$

$$r = \frac{e^{2T} - 1}{e^{2T} + 1}$$