statistica descrittiva bivariata

descrivere la distribuzione bivariata

due variabili categoriali

due variabili numeriche

una categoriale e una numerica

due maniere di descriverla

con grafici o tabelle

con indici numerici riassuntivi (statistiche)

descrivere l’<u>associazione</u>

imparare a valutarla da un grafico

misurarla con indici numerici

prevedere

usare l’associazione per costruire un modello del rapporto fra le due variabili

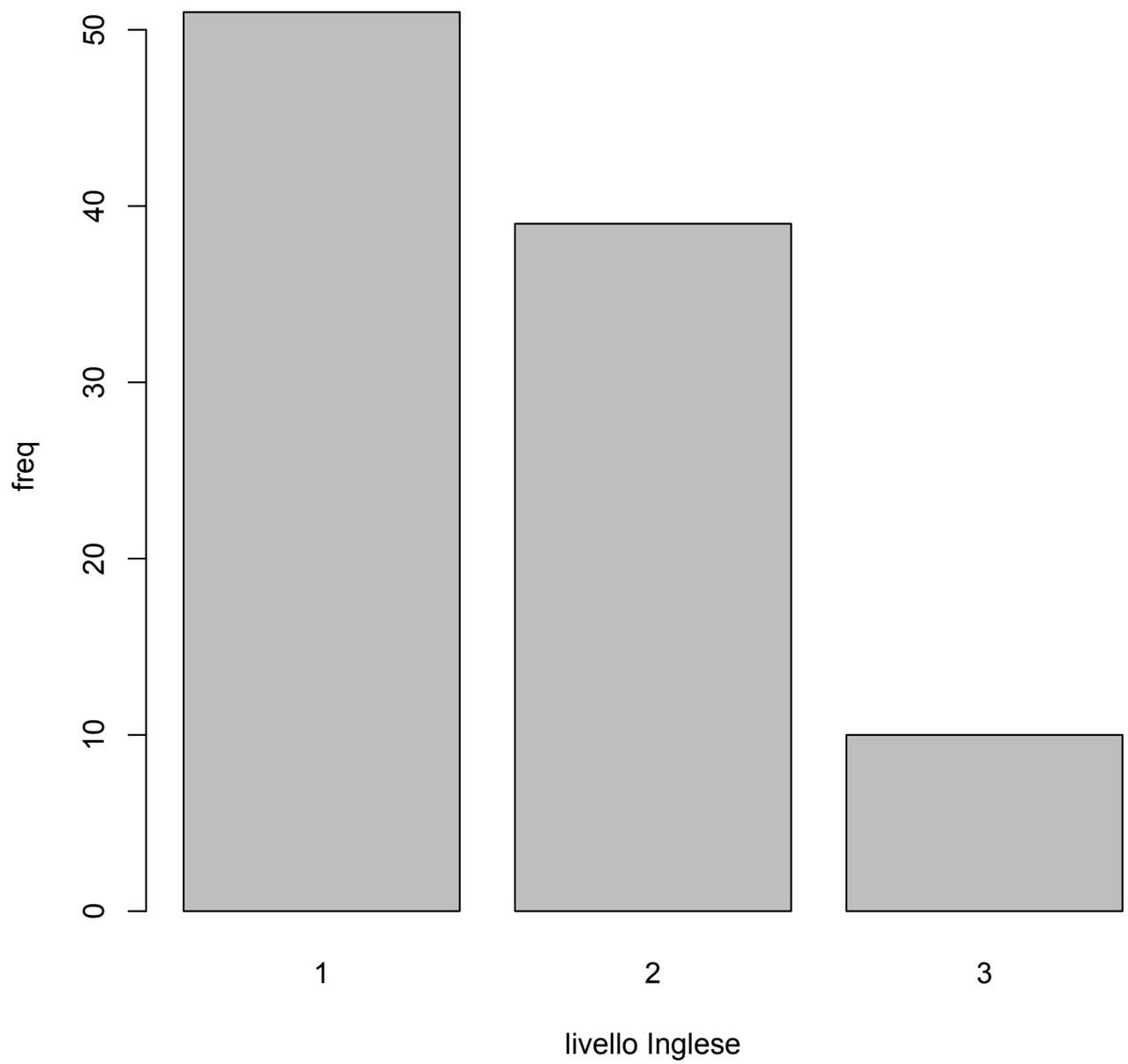
usare il modello per fare previsioni (e stimarne l’incertezza)

noi ci limiteremo al modello lineare

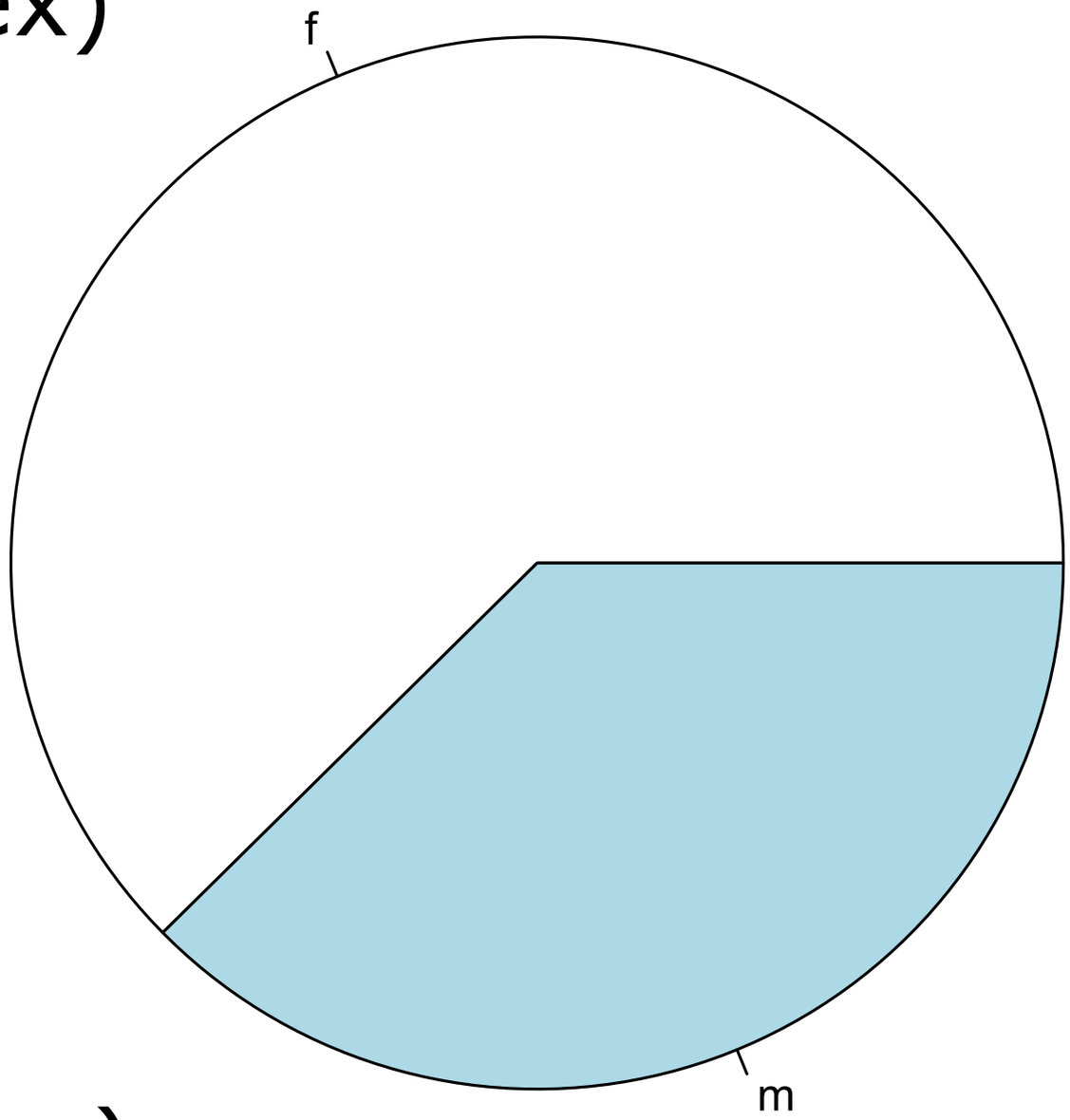
tabelle di contingenza

```
> df <- read.table("dati completi.txt",
header = TRUE)
> head(df)

  OvsR Sex HAND       RVF     LIKF   Ts  SPAF English compito
1    O   f    dx -3.9873143 1.500000 190   1.0       1       c
2    O   f    dx -0.4353741 4.166667 413   5.0       2      cd
3    R   f    dx  3.6857530 3.500000 491   2.5       1      cd
4    O   m    dx -1.4842264 5.666667 277   6.0       1       c
5    R   m    dx -1.3211927 6.000000 480   7.0       2      cd
6    R   m    dx -0.2262290 4.166667 265   3.0       2       c

> tE <- table(df$English)
> tE

 1  2  3
51 39 10
```

```
> barplot(tE, xlab = "livello Inglese",
ylab = "freq")
```



```
> tS <- table(df$Sex)
> tS

 f  m
78 47
```




```
> pie(tS)
```

```
> n <- length(df$Sex)
> n
[1] 125
```

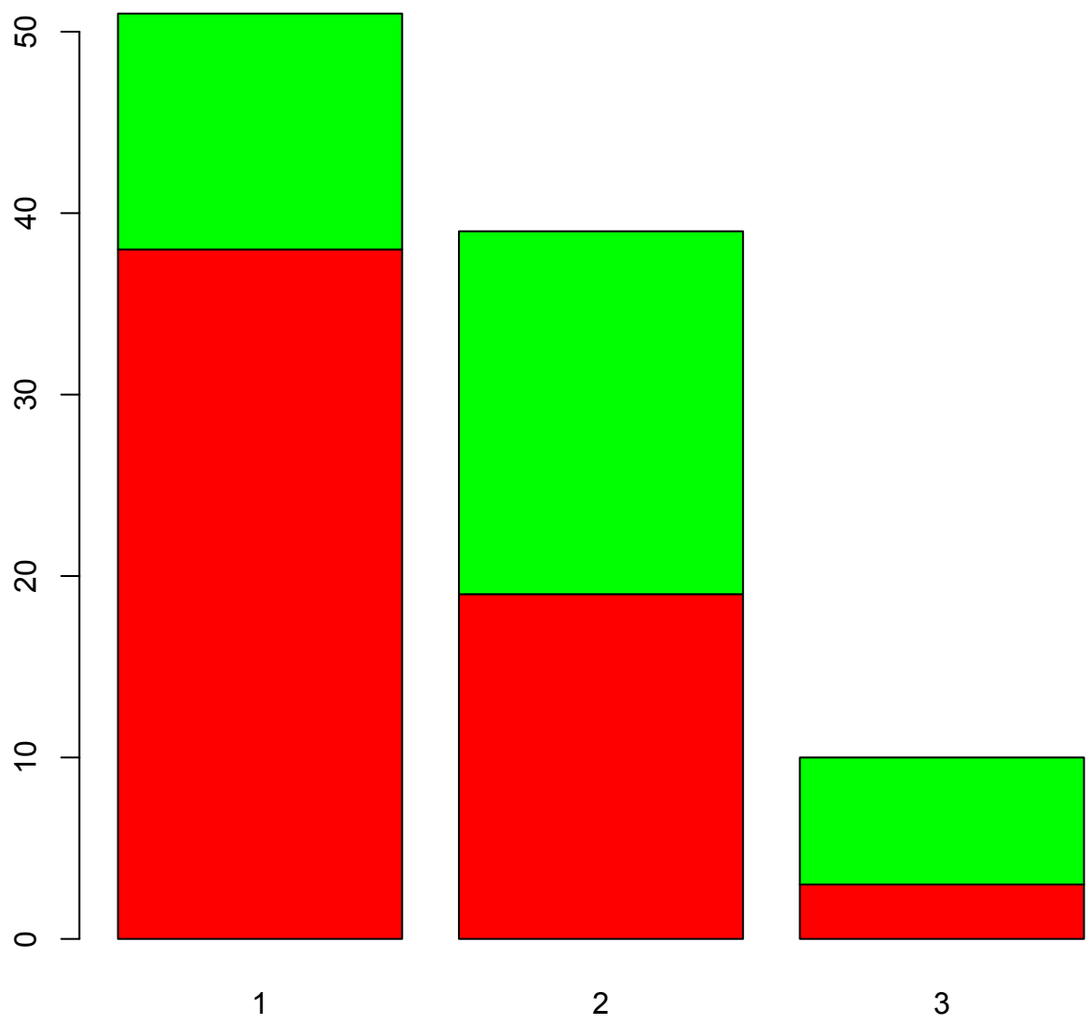
```
> table(df$Sex, df$English)

    1  2  3
f  38 19  3
m 13 20  7

> tc <- table(df$Sex, df$English)
> sum(tc)

[1] 100
```

```
> length(df$English)
[1] 125
> df$English
  [1]  1  2  1  1  2  2  1  1  1  1  1  1  1  2  2  1  2  2  2  2  2  1  3  1  1  2  3
 [28]  1  3  2  2  1  2  3  2  1  1  2  2  1  2  1  2  1  1  2  1  2  1  2  3  1  2  1
 [55]  2  2  1  2  1  2  1  1  1  2  2  2  2  1  1  3  2  1  1  2  3  1  1  1  1  1  3
 [82]  2  3  3  2  1  1  1  1  1  2  2  1  2  1  1  1  1  1  2 NA NA NA NA NA NA NA N
[109] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

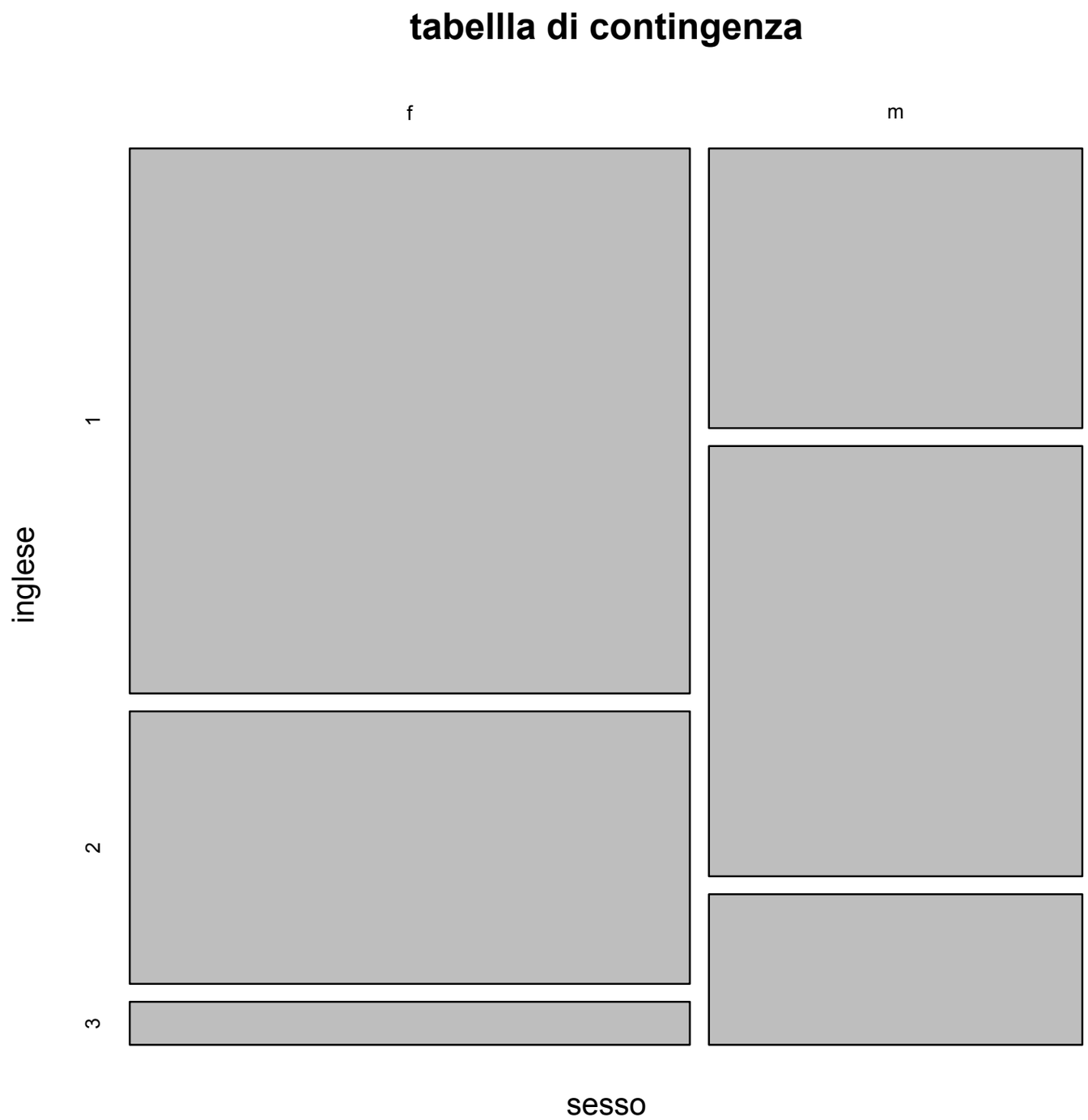
```
> tc

   1  2  3
f 38 19  3
m 13 20  7

> totr <- apply(tc, 1, sum)
> tc <- cbind(tc, totr)
> tc

   1  2  3 totr
f 38 19  3   60
m 13 20  7   40
```

```
> totc <- apply(tc, 2, sum)
> tc <- rbind(tc, totc)
> tc

     1  2  3 totr
f   38 19  3   60
m   13 20  7   40
totc 51 39 10  100
```

```
> tc <- table(df$Sex, df$English)
> barplot(tc, col = c("red", "green"))
```



```
> mosaicplot(tc, main = "tabellla di contingenza", xlab = "sesso", ylab = "inglese")
```



associazione

definizioni

associazione = NON indipendenza

indipendenza = la probabilità di A <u>non</u> cambia se sappiamo che B

probabilità = rapporto fra casi favorevoli e casi possibili

esempio 1

lancio un dado onesto due volte, se al primo lancio esce 2, la probabilità che esca 2 al secondo lancio rimane la stessa (1/6)

ogni lancio è un evento indipendente, il dado non ha "memoria"

esempio 2

estraggo a caso una carta da un mazzo ben mescolato di carte da Poker

la probabilità che sia l'asso di picche (o qualsiasi altra carta) è 1/52)

esempio 2

estraggo a caso una seconda carta dal mazzo

la probabilità che sia l'asso di cuori (o qualsiasi altra carta eccetto l'asso di picche) è adesso 1/51

gli eventi non sono indipendenti

probabilità

regola generale:

p(A e B) = p(A) p(B/A)

definizione di indipendenza:

A e B sono indipendenti SE E SOLO SE p(B/A) = p(B), per cui p(A e B) = p(A) p(B)

tabella di contingenza

inglese

sesso	1	2	3	totr
f	38	19	3	60
m	13	20	7	40
totc	51	39	10	100

p(m) = 40/100

p (3) = 10/100

se sesso e inglese sono indipendenti, allora p(m e 3) = p(m) p(3)

dato empirico

inglese

sesso		1	2	3	totr
	f	38	19	3	60
	m	13	20	7	40
	totc	51	39	10	100

p(m) p(3) = 0.4 x 0.1 = 0.04

il dato empirico è che

p(m e 3) = 7/100 = 0.07

in generale

inglese

sesso	1	2	3	totr
f	38	19	3	60
m	13	20	7	40
totc	51	39	10	100

se sesso e inglese sono indipendenti,per ogni cella della tabella di contingenza la probabilità osservata dovrebbe essere uguale al prodotto delle probabilità marginali

inglese

sesso	1	2	3	totr
f	38	19	3	60
m	13	20	7	40
totc	51	39	10	100

p(f e 1) = p(f) p(1) = 60/100 x 51/100

p(m e 1) = p(m) p(1) = 40/100 x 51/100

p(f e 2) = p(f) p(2) = 60/100 x 39/100

p(m e 2) = p(m) p(2) = 40/100 x 39/100

p(f e 3) = p(f) p(3) = 60/100 x 10/100

p(m e 3) = p(m) p(3) = 40/100 x 10/100

frequenze attese se indipendenti

inglese

sess	1	2	3	totr
f	38	19	3	60
m	13	20	7	40
totc	51	39	10	100

freq(f e 1) = p(f e 1) x totale

freq(m e 1) = p(m e 1) x totale

freq(f e 2) = ...

```
> df <- read.table("~/Desktop/dati
completi.txt", header = TRUE)
> tc <- table(df$Sex, df$English)
> tc

    1  2  3
 f 38 19  3
 m 13 20  7

> tot <- sum(tc)
> totr <- apply(tc, 1, sum)
> tc <- cbind(tc, totr)
> totc <- apply(tc, 2, sum)
> tc <- rbind(tc, totc)
```

```
> tc

      1    2   3  totr
f    38   19   3    60
m    13   20   7    40
totc 51   39  10   100

> faf1 <- (tc[3, 1]/tot) * (tc[1, 4]/tot) * tot
> faf1
[1] 30.6

> faf1 <- (tc[3, 1] * tc[1, 4]) / tot
> faf1
[1] 30.6
```

```
> tc

      1   2  3  totr
f    38  19  3   60
m    13  20  7   40
totc 51  39 10  100

> faf1 <- (tc[3, 1] * tc[1, 4]) / tot
> faf2 <- (tc[3, 2] * tc[1, 4]) / tot
> faf3 <- (tc[3, 3] * tc[1, 4]) / tot
> fam1 <- (tc[3, 1] * tc[2, 4]) / tot
> fam2 <- (tc[3, 2] * tc[2, 4]) / tot
> fam3 <- (tc[3, 3] * tc[2, 4]) / tot
```

misurare l’associazione (fra categorie)

idea di base

associazione = NON indipendenza

confrontare le frequenze osservate con le frequenze attese se A e B sono indipendenti

maggiore la differenza, più forte l'associazione

chi-quadrato

somma delle differenze (osservate - attese), al quadrato, espresse come proporzione delle frequenze attese:

$$\frac{(\text{fr. osservate} - \text{fr. attese})^2}{\text{fr. attese}}$$

```
> v <- c(faf1, faf2, faf3, fam1, fam2, fam3)
> fa <- matrix(v, nrow = 2, byrow = TRUE)
> fa
     [,1]   [,2]  [,3]
[1,] 30.6 23.4    6
[2,] 20.4 15.6    4
```

```
> fo <- table(df$Sex, df$English)
> fo
     1  2  3
  f 38 19  3
  m 13 20  7
```

```
> fa
     [,1]   [,2]  [,3]
[1,] 30.6 23.4    6
[2,] 20.4 15.6    4

> fo
     1  2  3
  f 38 19  3
  m 13 20  7

> chiq <- sum( (fo - fa)^2/ fa)
> chiq
[1] 10.29223
```

```
> fo
     1  2   3
  f  38 19  3
  m 13 20  7

> chisq.test(fo)

        Pearson's Chi-squared test

data:  fo
X-squared = 10.2922, df = 2, p-value =
0.005822
```

chi-quadrato

non è mai negativo

è tanto più grande quanto maggiore la differenza con le frequenze attese

non è facilmente interpretabile come indice di associazione

V di Cramér

radice quadrata di chi-quadrato, diviso per n x (k - 1)

n = totale delle freq osservate

k = numero di righe o di colonne, se diversi usare il più piccolo

V di Cramér

non è mai negativo

varia da 0 (assenza di associazione) a 1 (associazione perfetta)

con k = 2, è detto anche phi di Cramér, phi = sqrt(chisq/N)

```
> chiq
[1] 10.29223

> tot
[1] 100

> V <- sqrt(chiq/tot)

> V
[1] 0.3208151
```

```
> mm <- matrix(c(35, 12, 11, 30, 16, 60, 80,
20, 15), nrow = 3,  byrow = TRUE)
> str(mm)
 num [1:3, 1:3] 35 30 80 12 16 20 11 60 15
> dim(mm)
[1] 3 3
> min(dim(mm))
[1] 3
```

```
> k <- min(dim(mm))
> n <- sum(mm)
> chisq <- chisq.test(mm)
>  V <- sqrt(chisq$statistic[[1]]/(n * (k -1)))
> V
[1] 0.3197222
```

```
> str(chisq)
List of 9
 $ statistic: Named num 57
  ..- attr(*, "names")= chr "X-squared"
 $ parameter: Named int 4
  ..- attr(*, "names")= chr "df"
 $ p.value  : num 1.21e-11
 $ method   : chr "Pearson's Chi-squared test"
 $ data.name: chr "chisq"
 $ observed : num [1:3, 1:3] 35 30 80 12 16
20 11 60 15
 $ expected : num [1:3, 1:3] 30.14 55.09
59.77 9.98 18.24 ...
 $ residuals: num [1:3, 1:3] 0.885 -3.38 2.617
0.64 -0.524 ...
 $ stdres   : num [1:3, 1:3] 1.434 -6.194
4.926 0.79 -0.731 ...
 - attr(*, "class")= chr "htest"
```

```
> chisq$observed
     [,1] [,2] [,3]
[1,]   35   12   11
[2,]   30   16   60
[3,]   80   20   15

> chisq$expected
       [,1]      [,2]        [,3]
[1,] 30.14337  9.978495 17.87814
[2,] 55.08961 18.236559 32.67384
[3,] 59.76703 19.784946 35.44803
```

diagramma di dispersione

(scattergram)

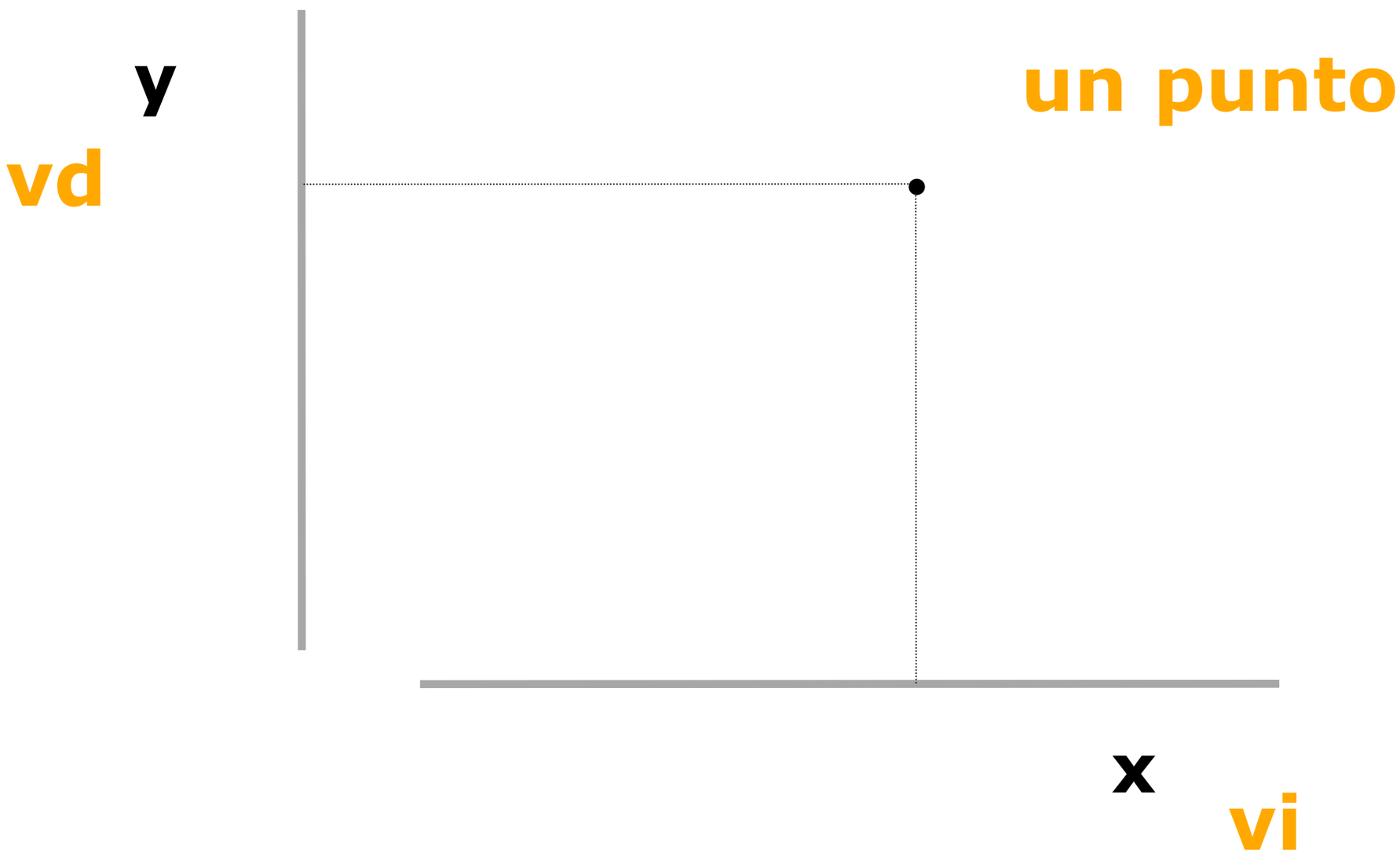
scattergram

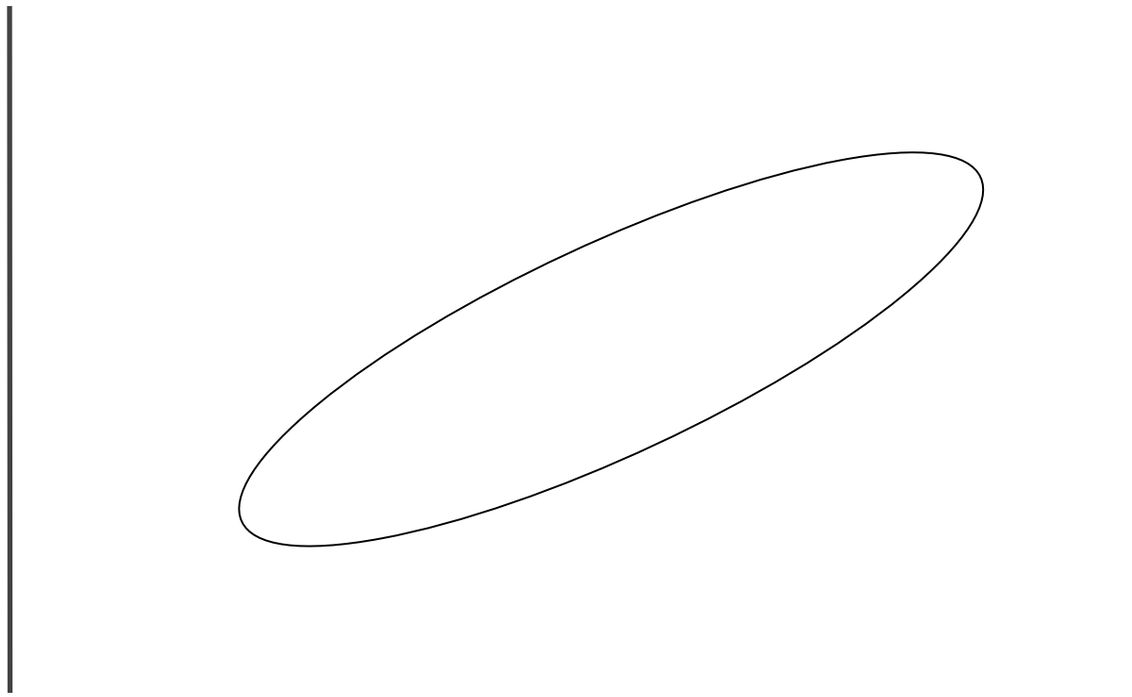
visualizza la distribuzione bivariata

ogni punto è una coppia y, x (variabile dip. e indip.)

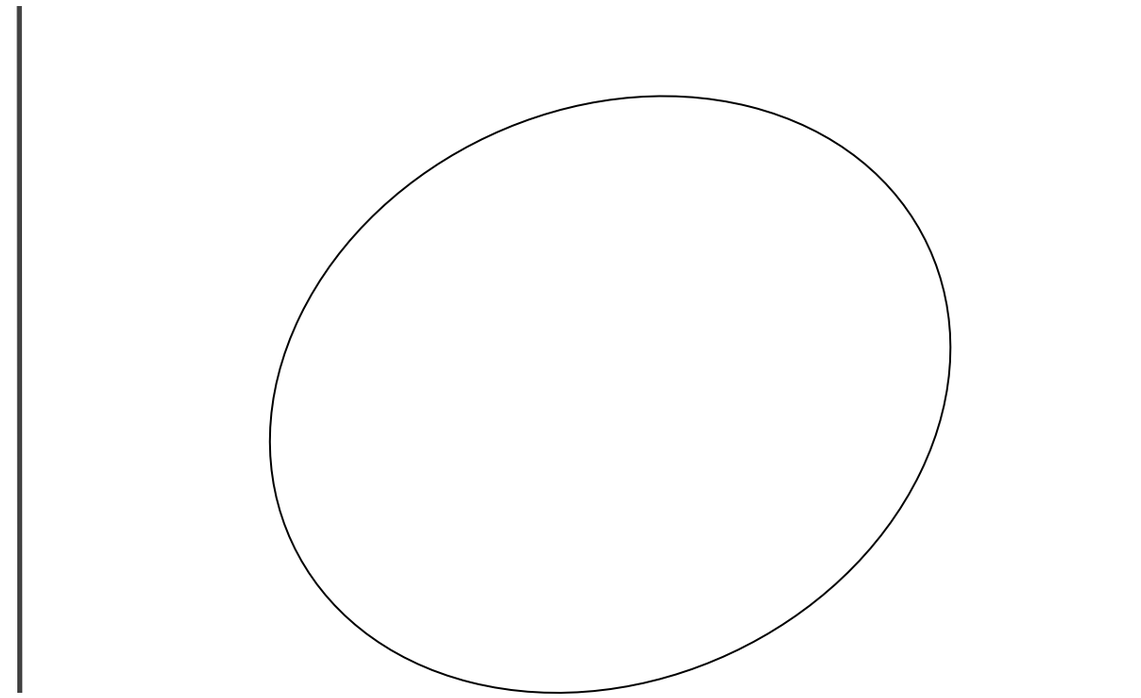
non ha propriamente un'origine (y = 0, x = 0)

scattergram

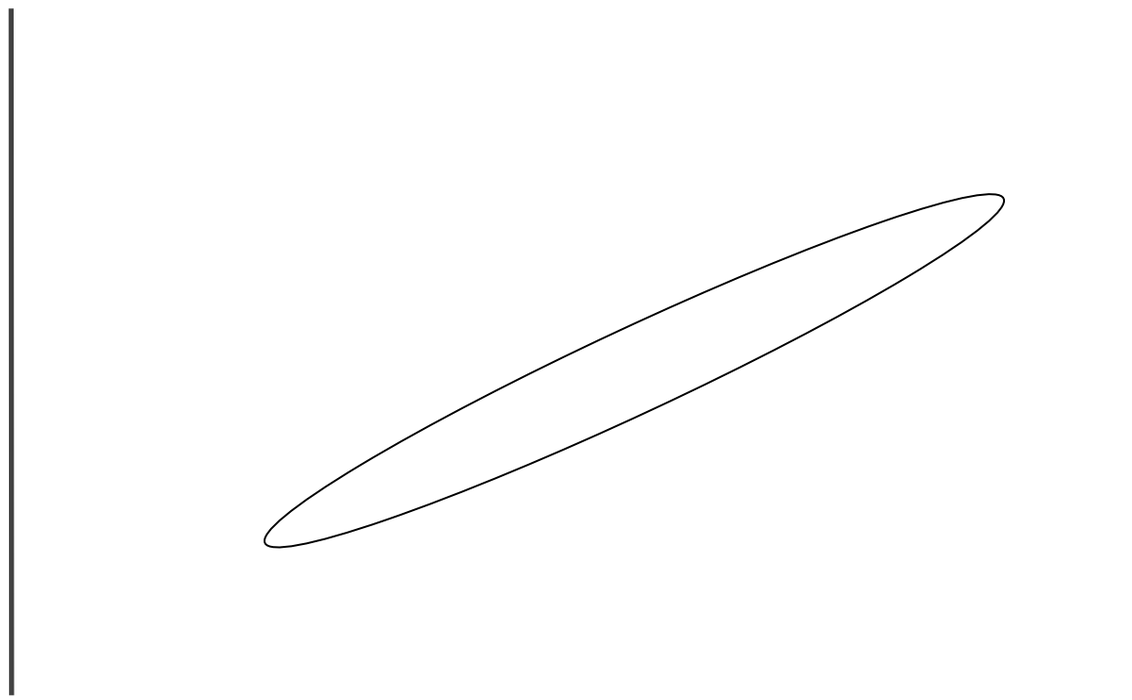




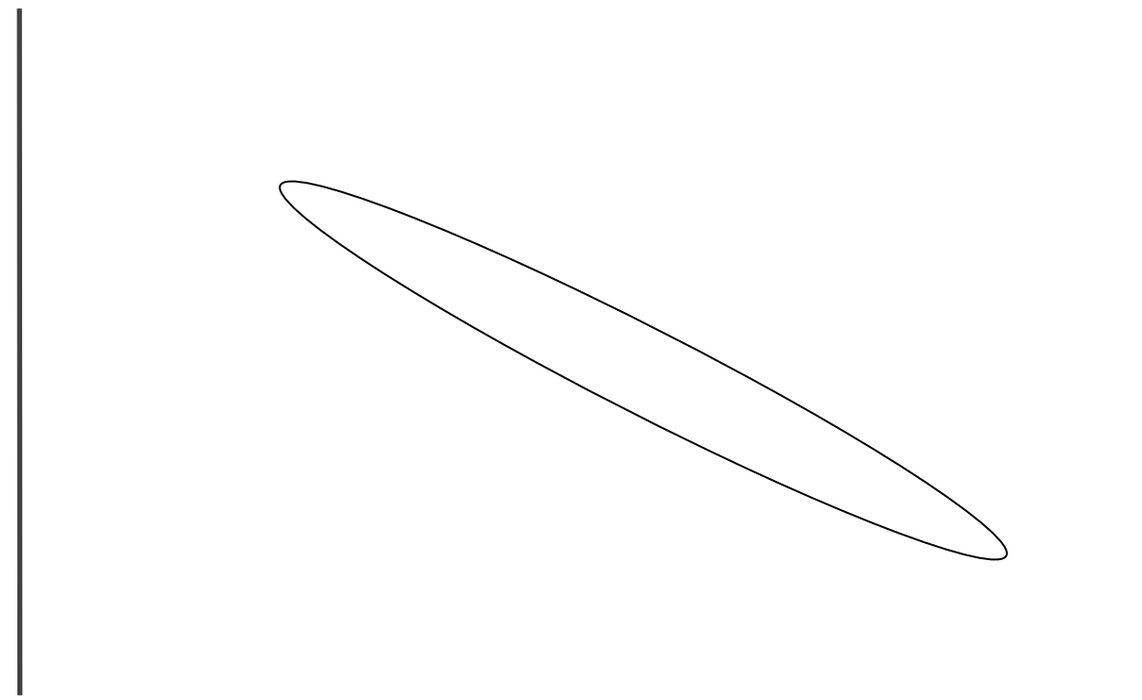
associazione forte



associazione debole

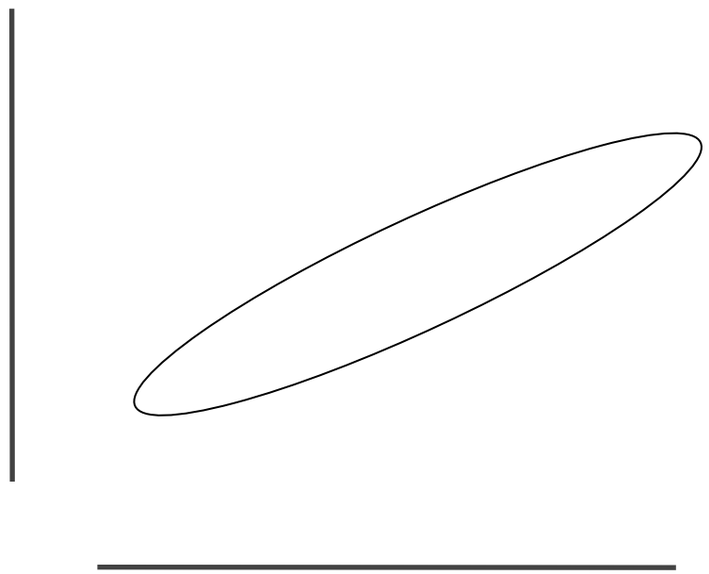


associazione positiva

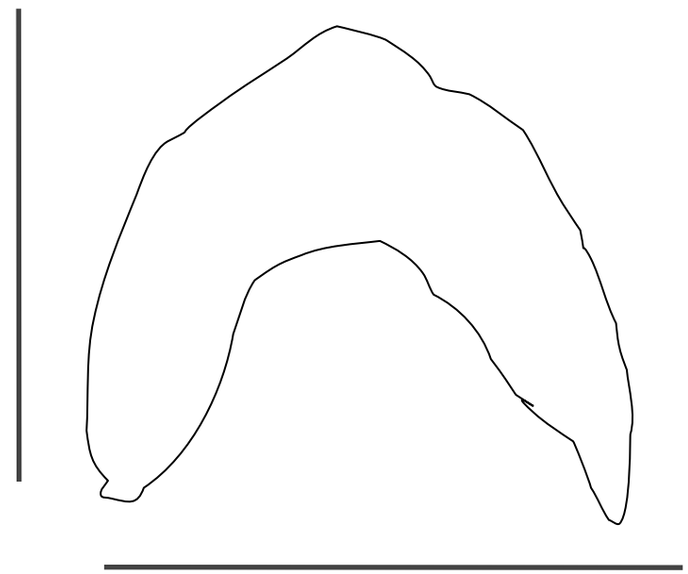


associazione negativa

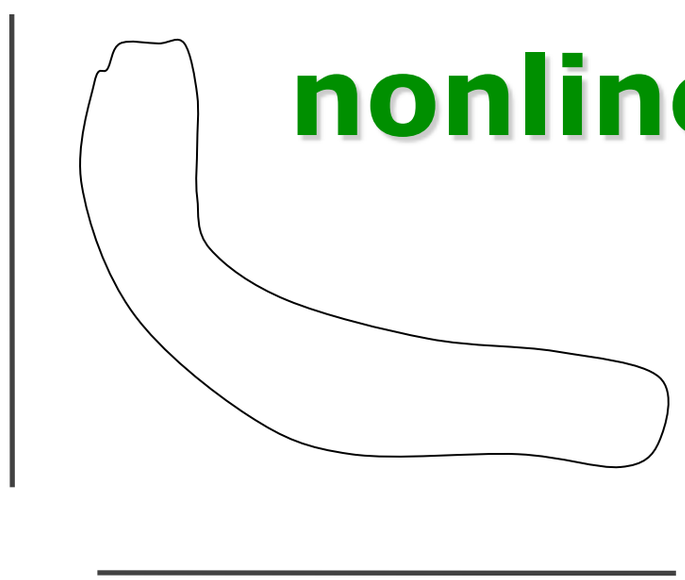
lineare

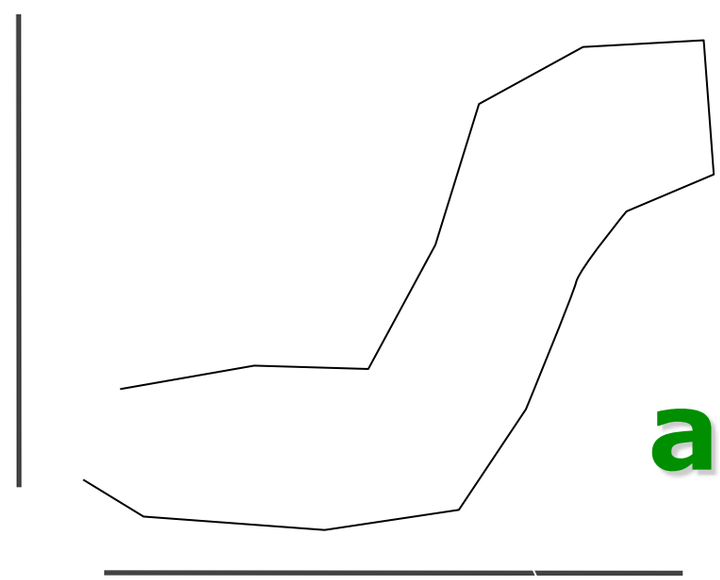


U rovesciata



nonlineare



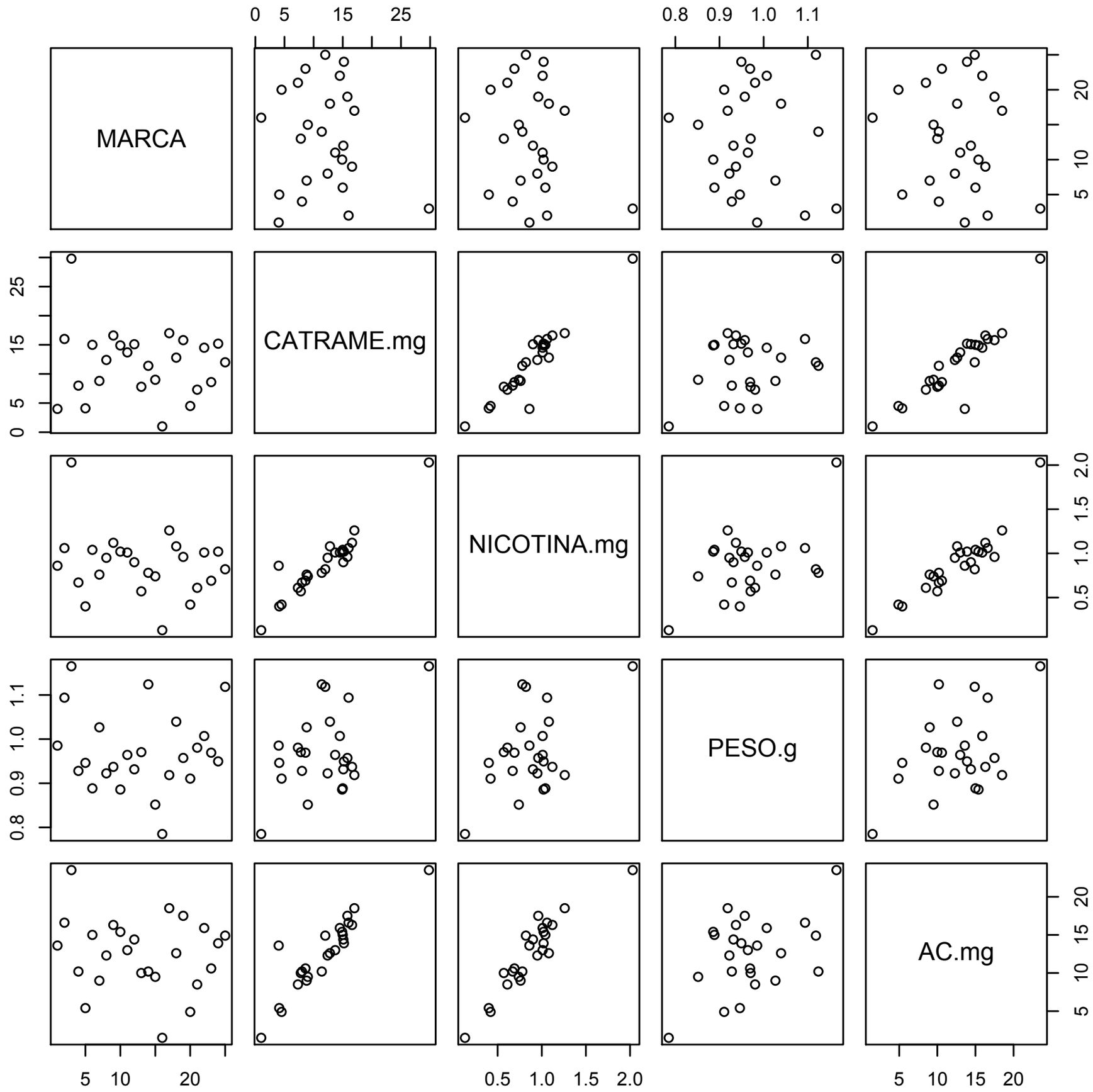


a S

```
> df <- read.table("SIGAR.txt", header = TRUE)
> head(df)
```

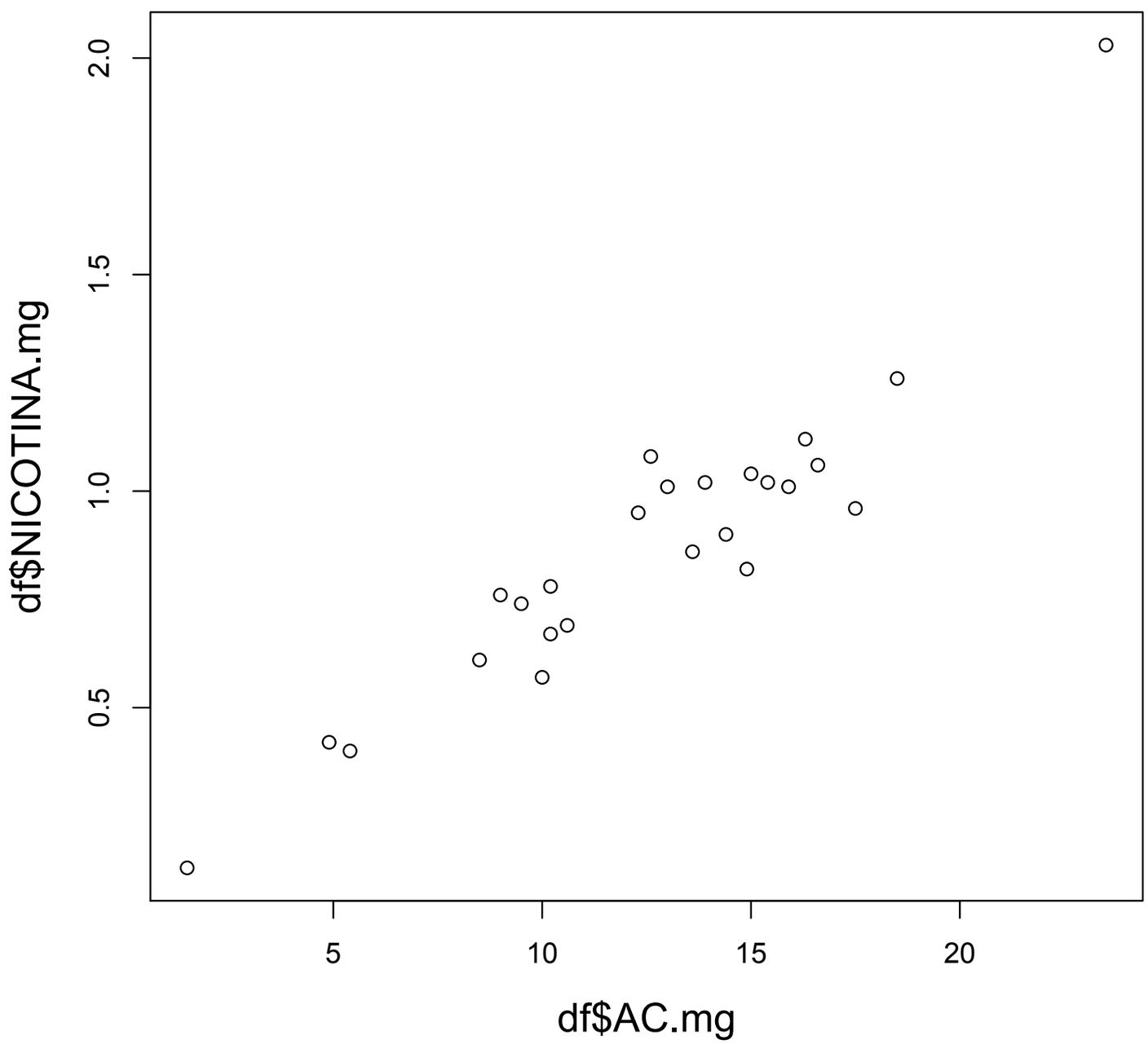
	MARCA	CATRAME.mg	NICOTINA.mg	PESO.g	AC.mg
1	Alpine	4.0	0.86	0.9853	13.6
2	Benson&Hedges	16.0	1.06	1.0938	16.6
3	BullDurham	29.8	2.03	1.1650	23.5
4	CamelLights	8.0	0.67	0.9280	10.2
5	Carlton	4.1	0.40	0.9462	5.4
6	Chesterfield	15.0	1.04	0.8885	15.0

```
>plot(df)
```

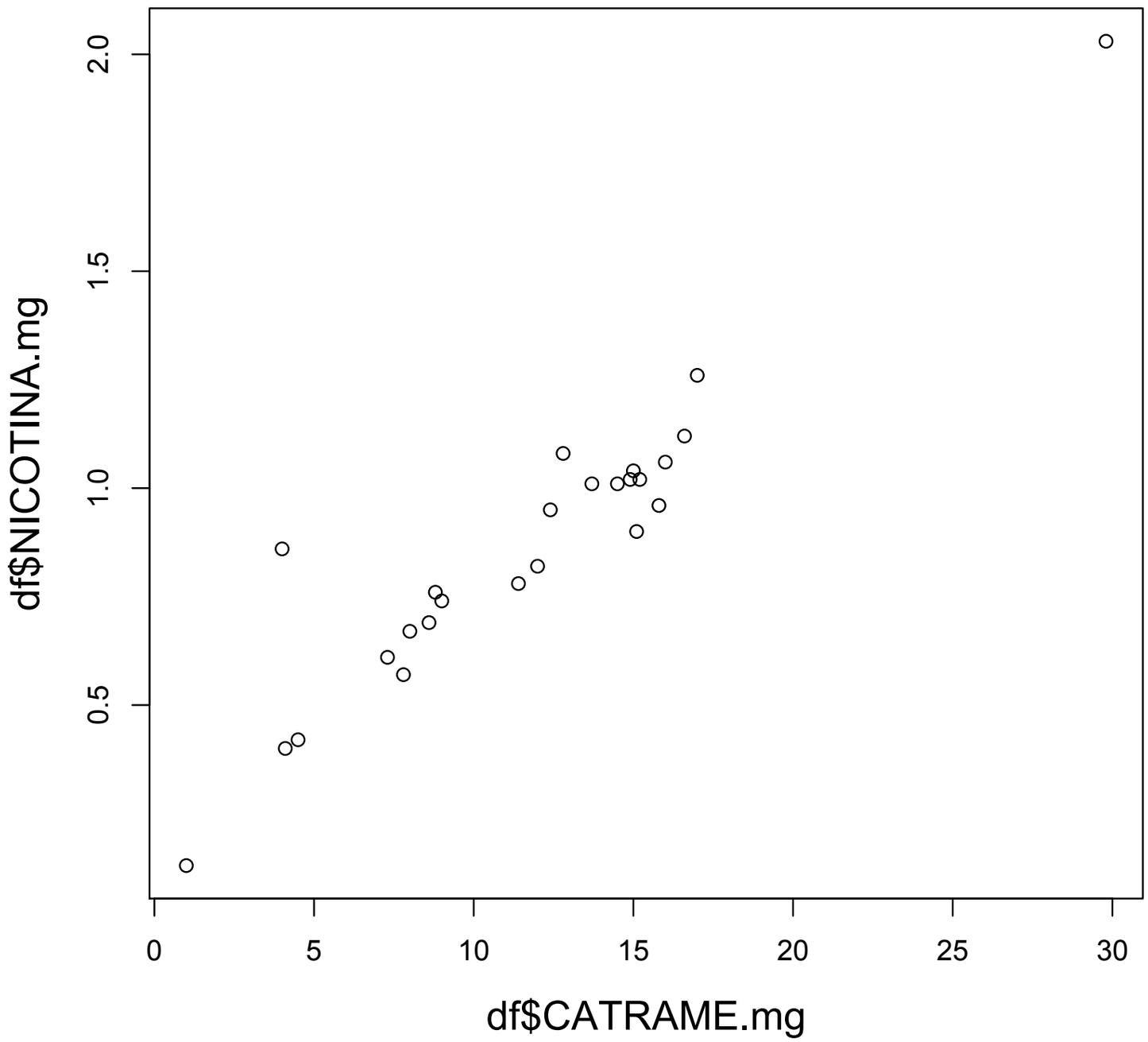

MARCA
CATRAME.mg
NICOTINA.mg
PESO.g
AC.mg
0 5 15 25
0.8 0.9 1.0 1.1
5 10 20
0.5 1.0 1.5 2.0
5 10 15 20

```
> str(df)
'data.frame':  25 obs. of  5 variables:
 $ MARCA       : Factor w/ 25 levels
"Alpine","Benson&Hedges",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ CATRAME.mg : num  4 16 29.8 8 4.1 15 8.8 12.4 16.6
14.9 ...
 $ NICOTINA.mg: num  0.86 1.06 2.03 0.67 0.4 1.04 0.76 0.95
1.12 1.02 ...
 $ PESO.g      : num  0.985 1.094 1.165 0.928 0.946 ...
 $ AC.mg       : num  13.6 16.6 23.5 10.2 5.4 15 9 12.3 16.3
15.4 ..
```

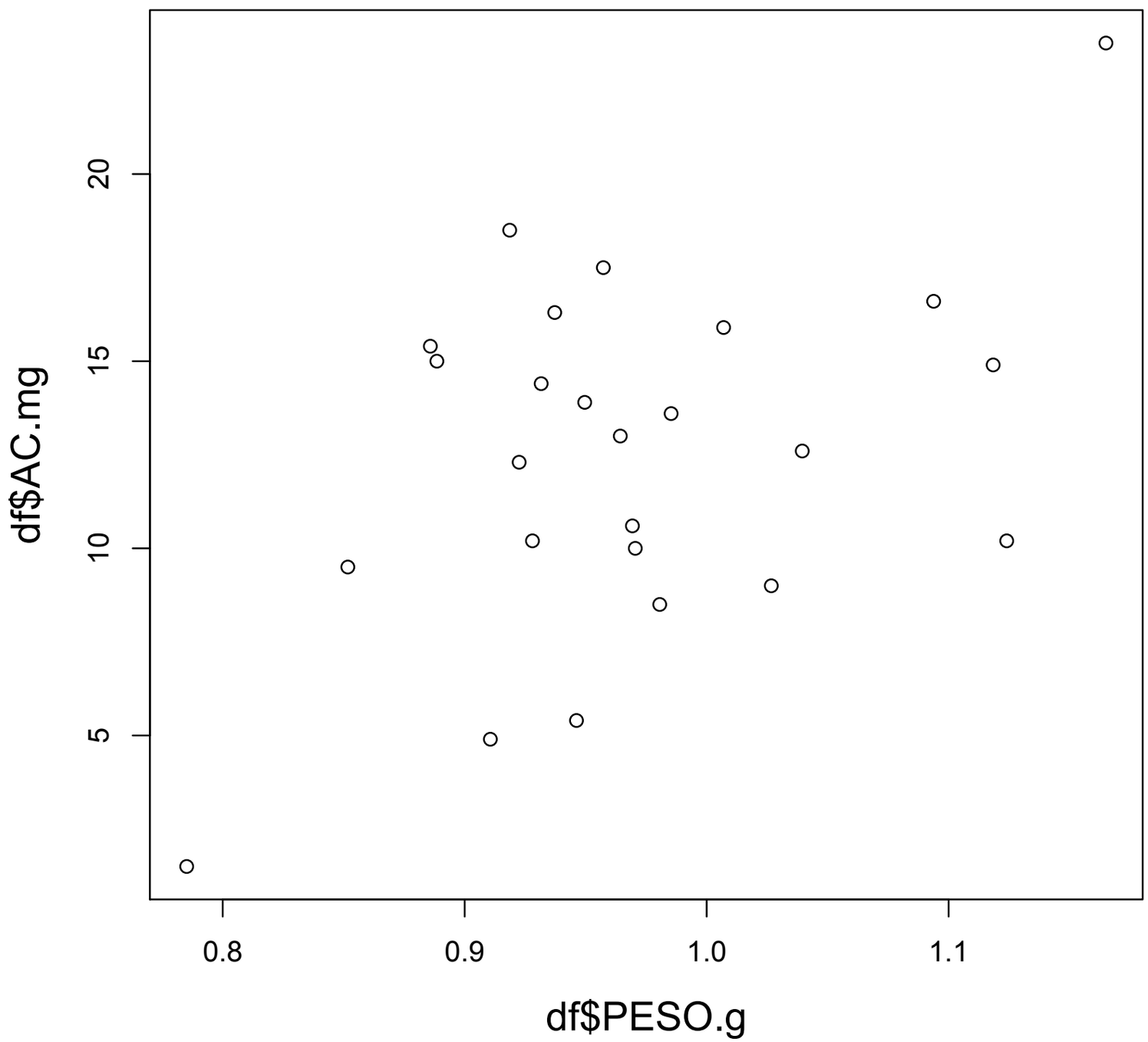
```
> plot(df$NICOTINA.mg ~ df$AC.mg, cex.lab = 1.4)
```



```
> plot(df$NICOTINA.mg ~ df$CATRAME.mg, cex.lab = 1.4)
```

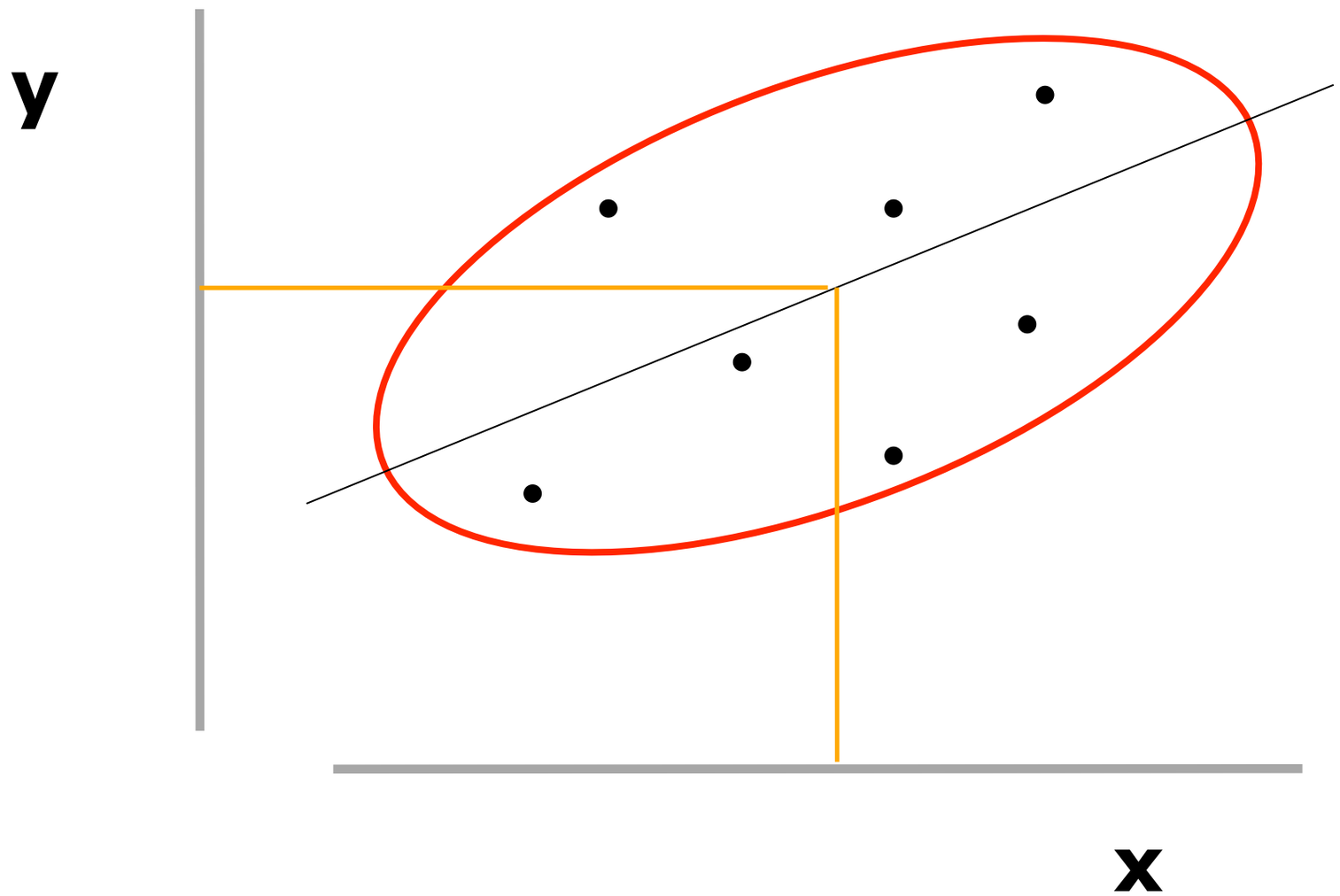


```
> plot(df$AC.mg ~ df$PESO.g, cex.lab = 1.4)
```



il punto delle medie e la retta delle DS

baricentro e asse di simmetria



punto delle medie

è il punto che ha coordinate (media di x, media di y)

rappresenta il “baricentro” della distribuzione bivariata

in una distribuzione bivariata lineare, attorno al pdm c’è la maggiore densità di punti

retta delle DS

è la retta i cui punti stanno a un ugual numero di deviazioni standard dalle rispettive medie

rappresenta l'asse di simmetria della distribuzione bivariata

l'associazione è tanto più forte quanto più i dati sono vicini alla retta delle DS

esempio

M(x) = 10, DS(x) = 2

M(y) = 20, DS(y) = 3

il punto (12, 23) sta sulla retta?

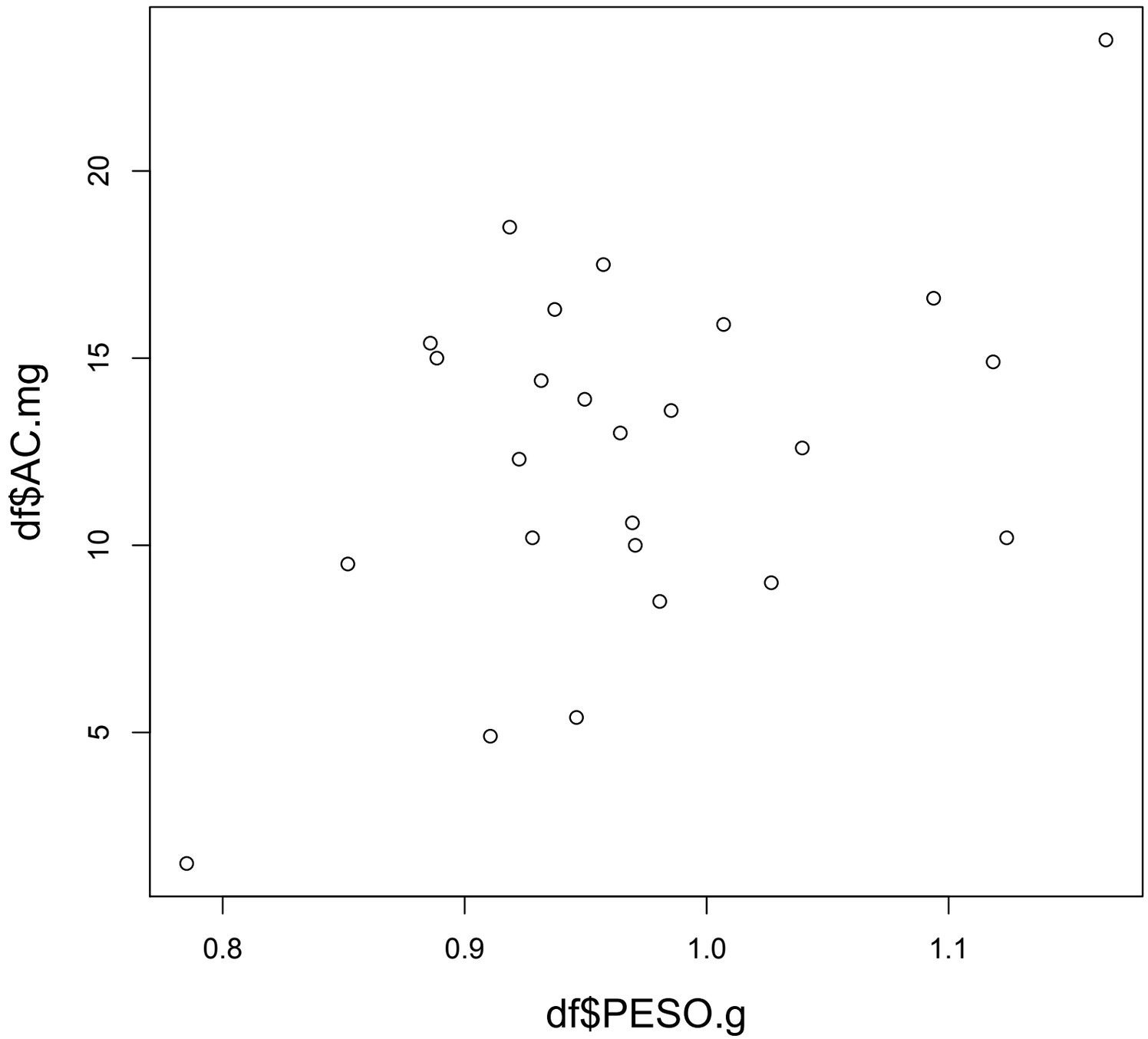
il punto (6, 14) sta sulla retta?

il punto (11, 29) sta sulla retta?

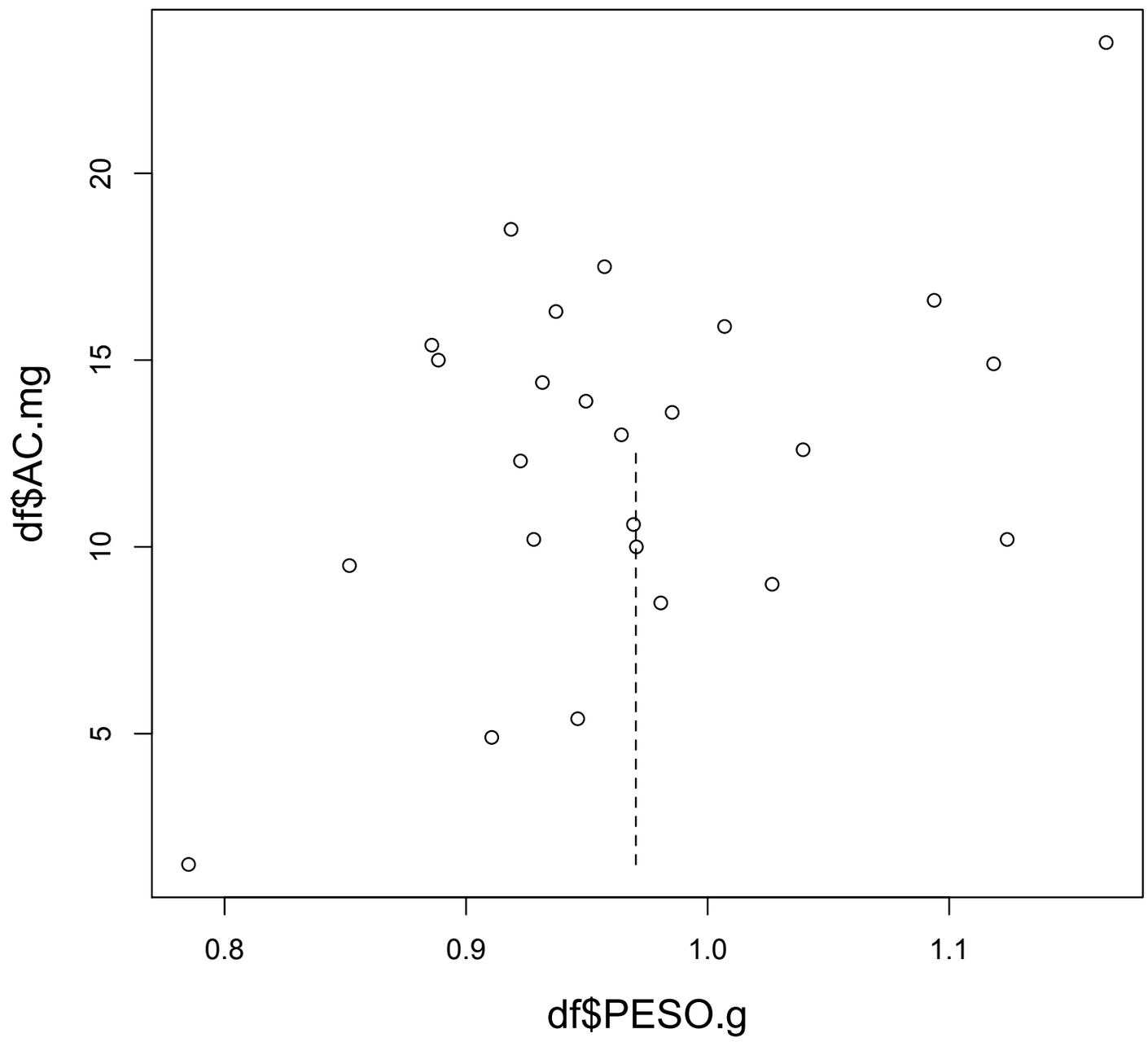
il punto (4, 26) sta sulla retta?

il punto (10, 20) sta sulla retta?

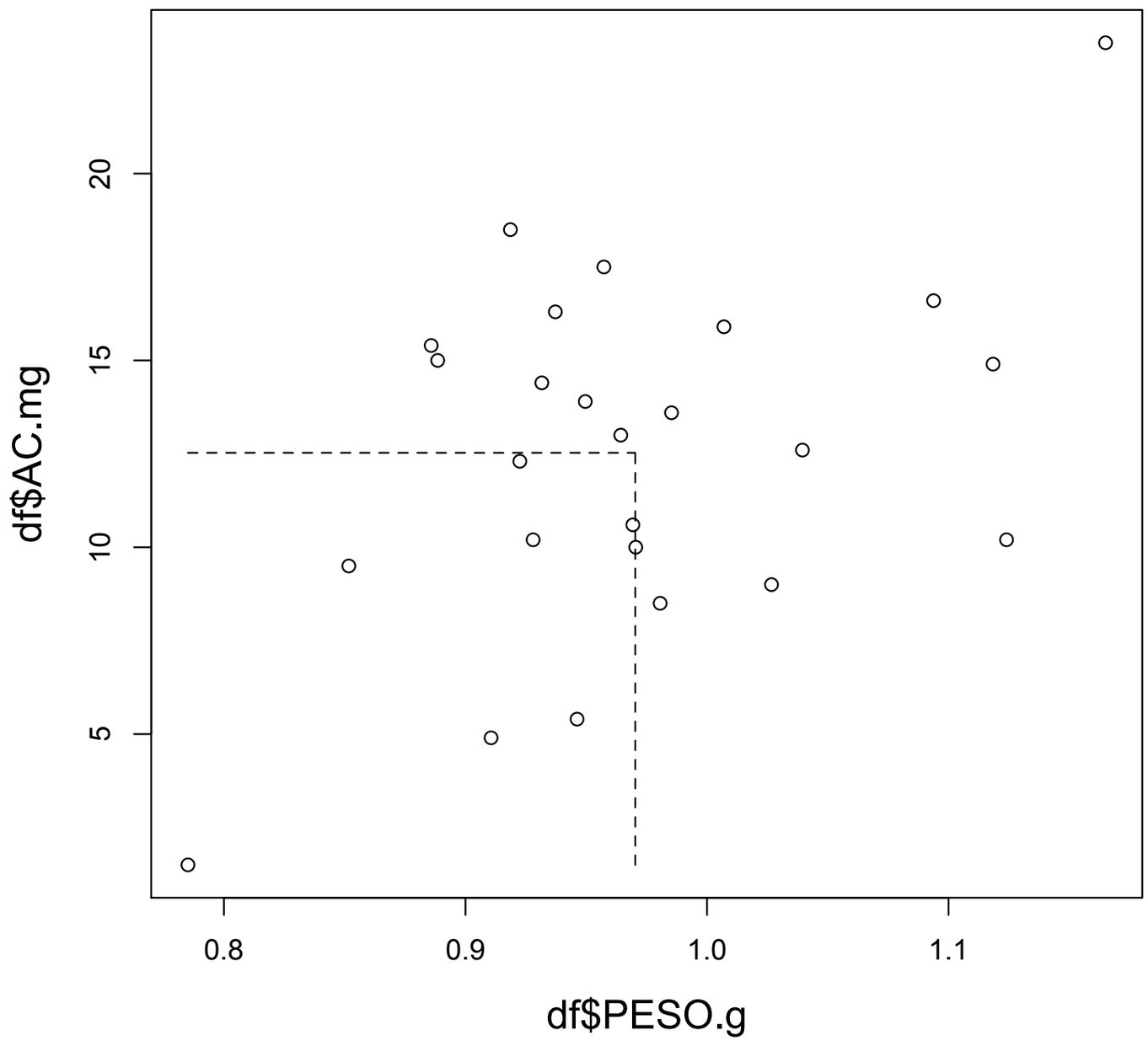
```
> plot(df$AC.mg ~ df$PESO.g, cex.lab = 1.4)
```



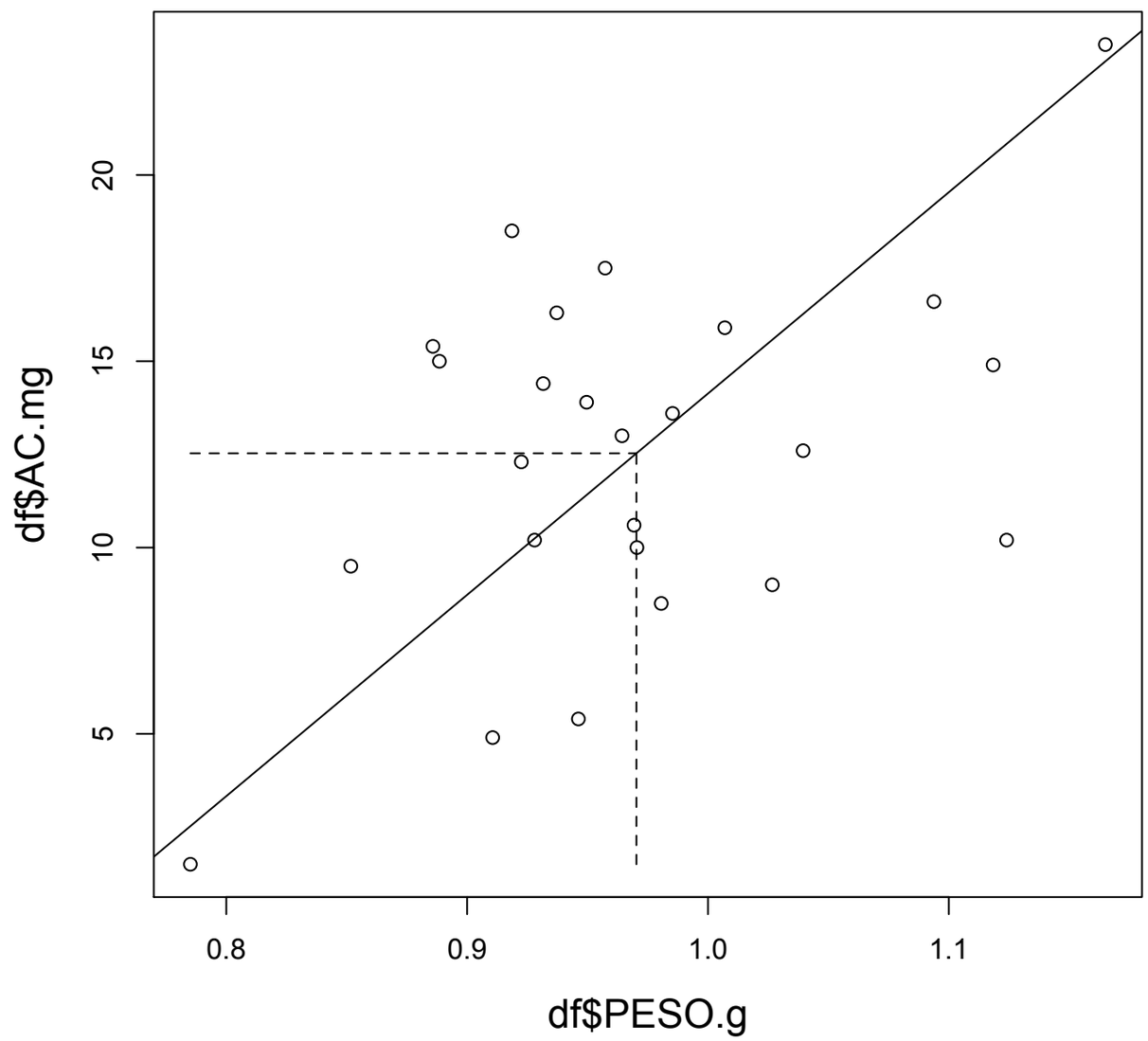
```
> xm <- mean(df$PESO.g)
> ym <- mean(df$AC.mg)
> segments(xm, min(df$AC.mg), xm, ym, lty
= "dashed")
```



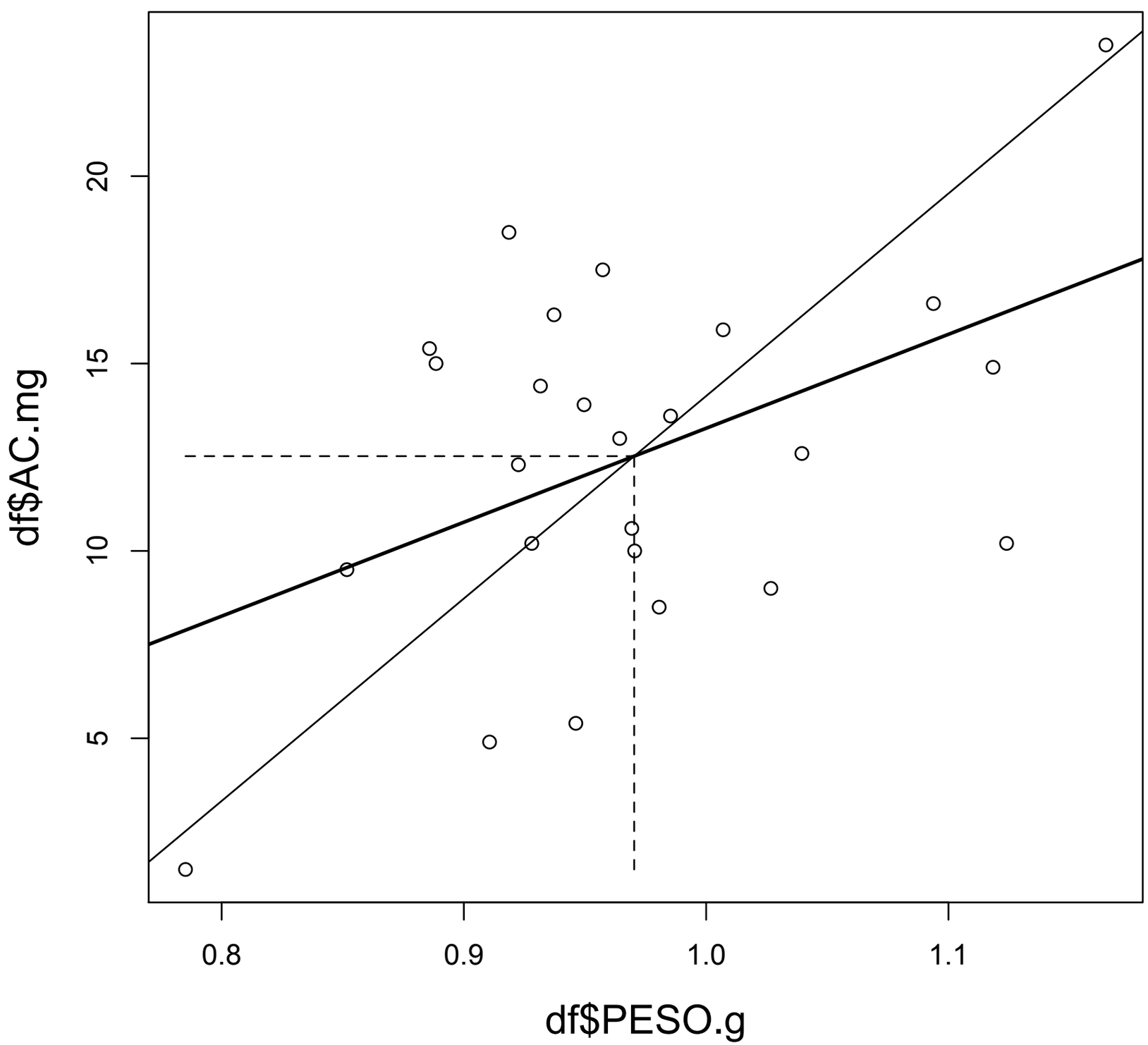
```
> segments(min(df$PESO.g), ym, xm, ym, lty
= "dashed")
```



```
> x1 <- xm - 3 * sd(df$PESO.g)
> x2 <- xm + 3 * sd(df$PESO.g)
> y1 <- ym - 3 * sd(df$AC.mg)
> y2 <- ym + 3 * sd(df$AC.mg)
> segments(x1, y1, x2, y2)
```



retta DS ≠ retta di regressione

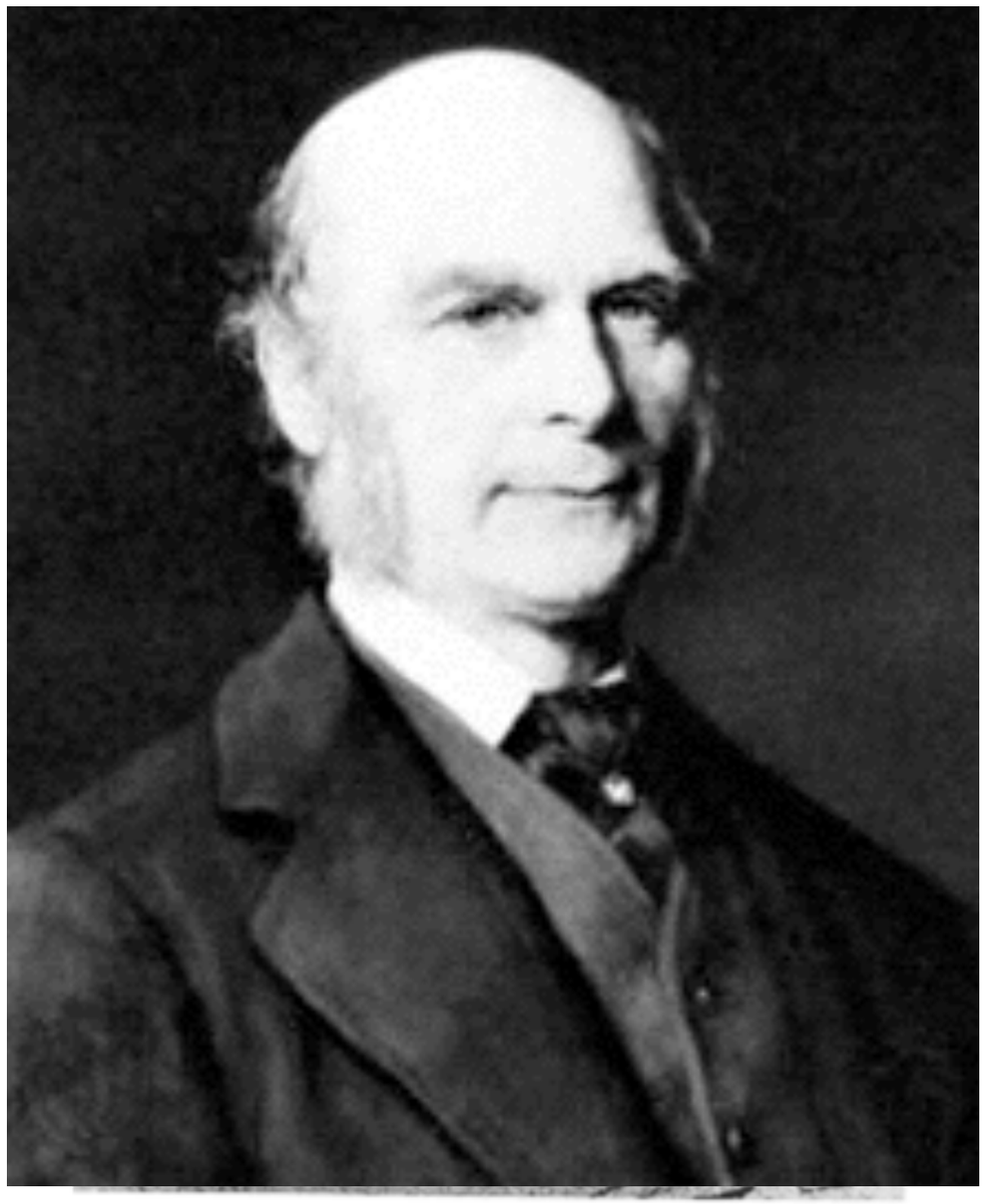


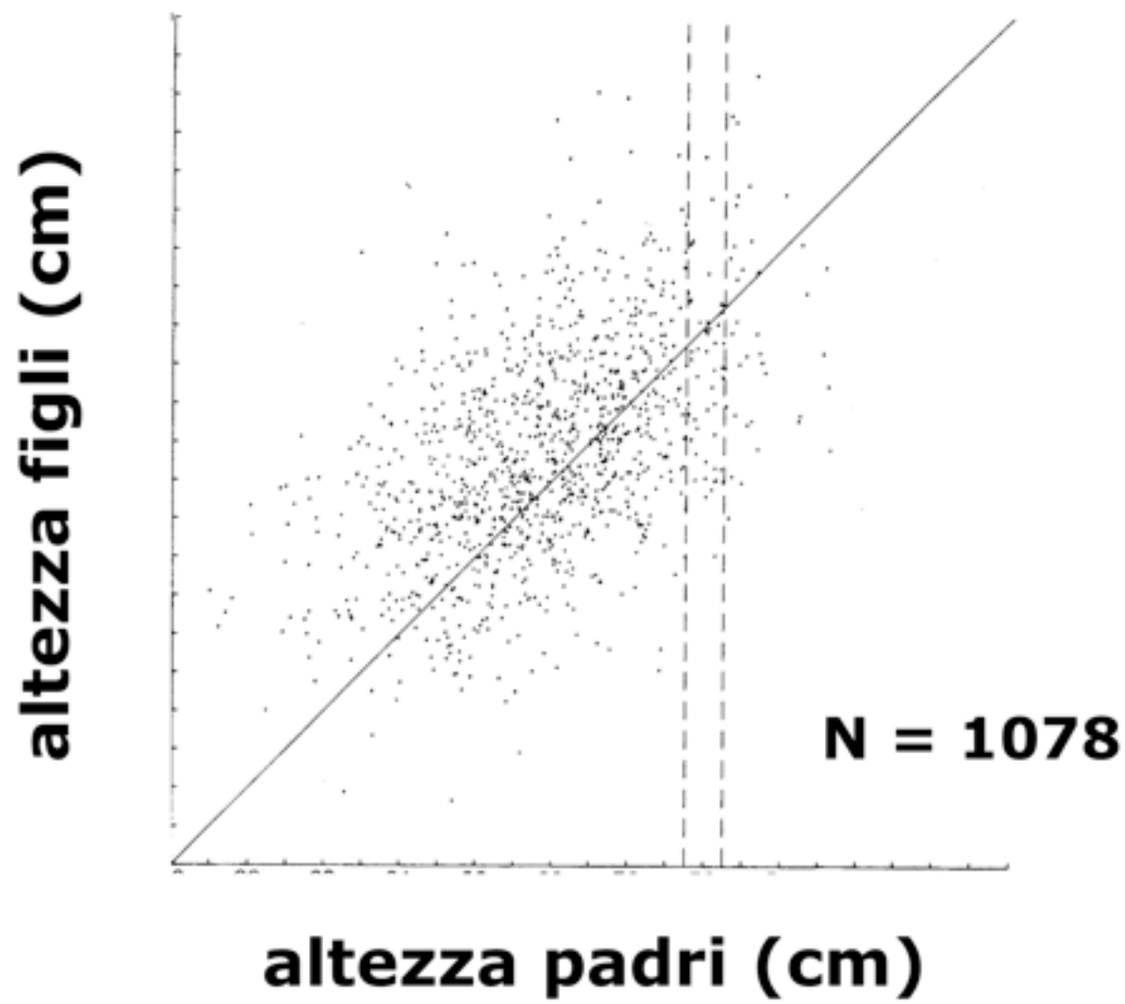
DS

regressione

misurare l’associazione (lineare) fra due variabili numeriche

Francis Galton
(1822-1911)





Correlazione

- Per misurare l'associazione lineare fra due variabili numeriche si utilizza il coefficiente di correlazione (r)
- r misura quanto i dati sono vicini alla retta delle DS

Definizione

- **Il coefficiente di correlazione (r) è la media dei prodotti z(x)z(y)**

- **Quindi per il calcolo:**
  - **standardizzare ogni caso x, y**
  - **moltiplicare ogni z(x), z(y)**
  - **calcolare la media dei prodotti**

Esempio

x	y
1	5
3	9
4	7
5	1
7	13

Mx = 4, DS = 2

My = 7, DS = 4

z(x) e z(y)

x	y	(x - 4)/2	(y-7)/4
1	5	-1.5	-0.5
3	9	-0.5	0.5
4	7	0	0
5	1	0.5	-1.5
7	13	1.5	1.5

Prodotti

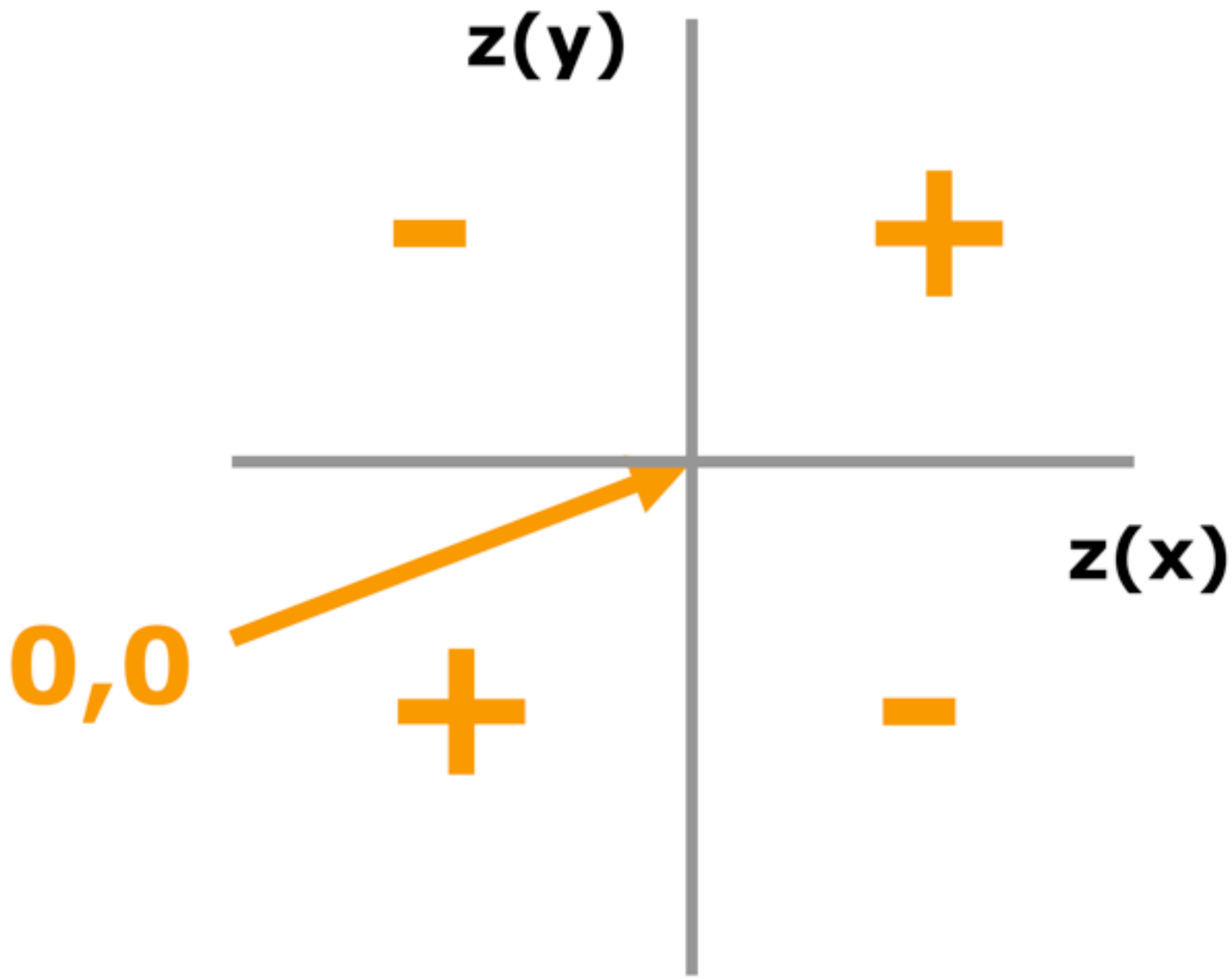
z(x)	z(y)	Z(x)Z(y)
-1.5	-0.5	0.75
-0.5	0.5	-0.25
0	0	0
0.5	-1.5	-0.75
1.5	1.5	2.25

media = 0.4

$-1 < r < 1$

- il segno misura la direzione dell'associazione
- il valore assoluto, la forza dell'associazione
- r non ha unità di misura
  - perché?

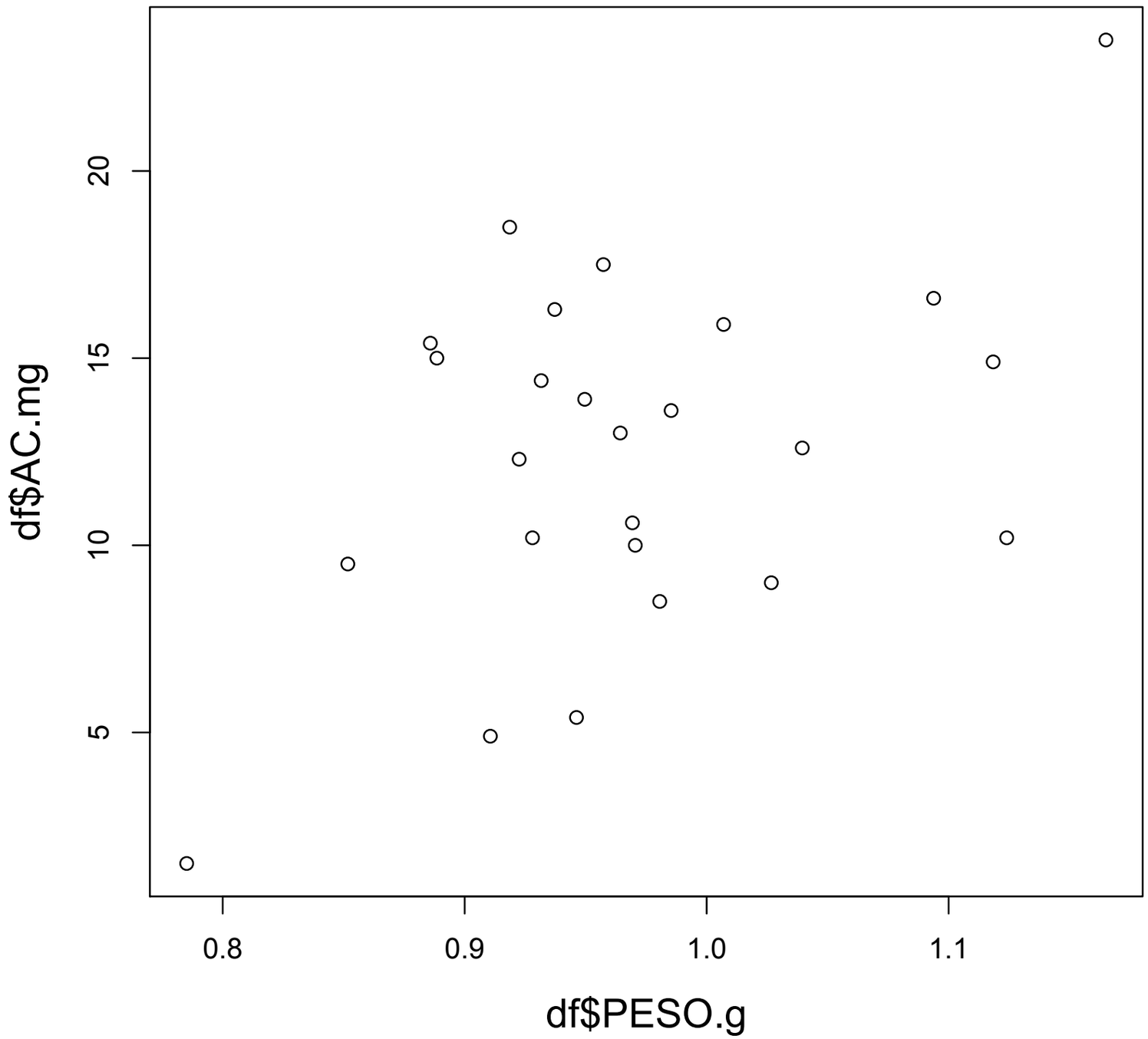
Il segno di r



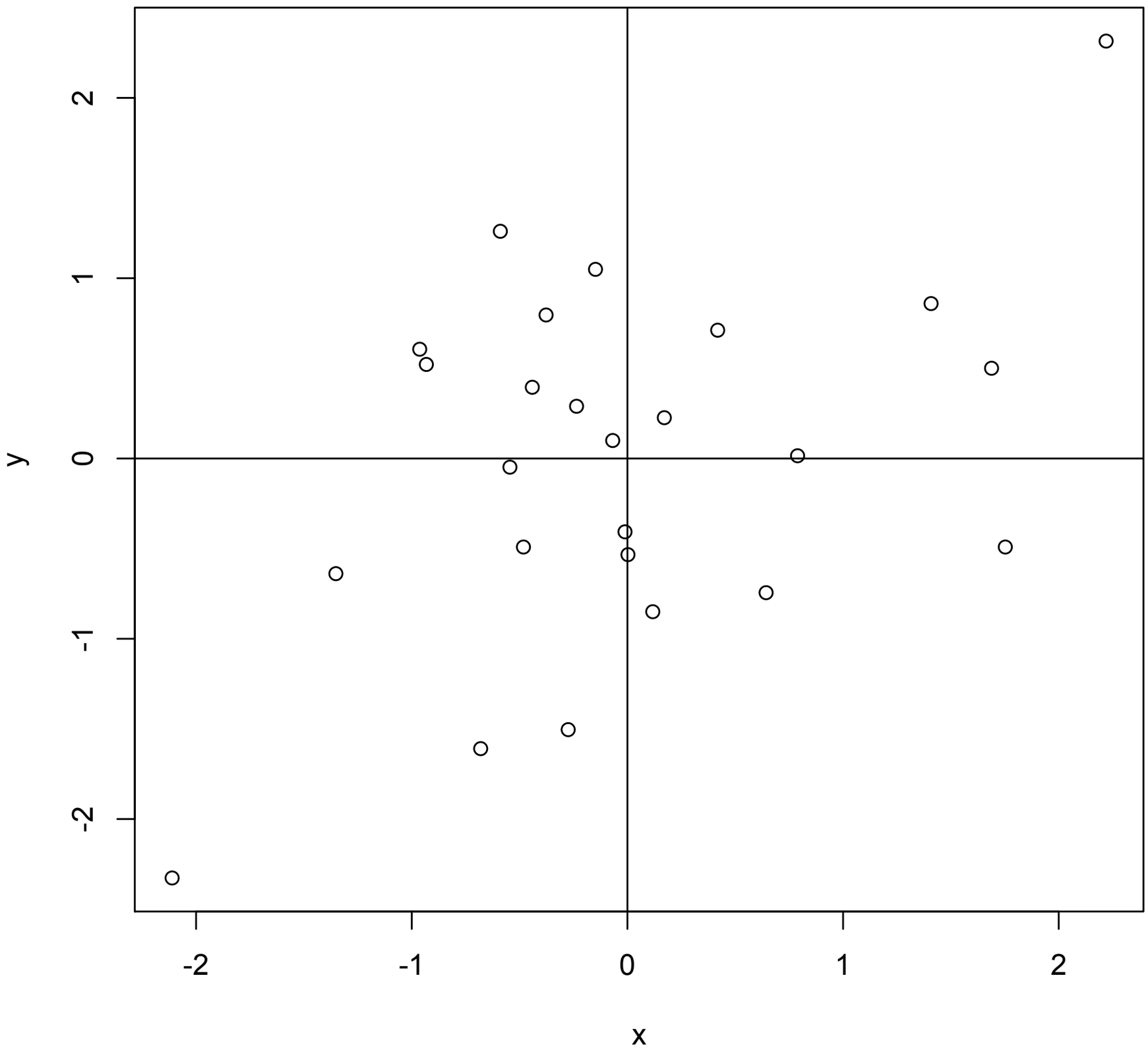
Il valore assoluto di r

- in valore assoluto, $0 < |r| < 1$
- questo perché
  - se i prodotti negativi e positivi si equivalgono, la media -> 0
  - questo accade quando i dati stanno su tutti i quadranti
  - quanti più i dati si concentrano solo sui quadranti + o -, tanto più la media -> 1

```
> plot(df$AC.mg ~ df$PESO.g, cex.lab = 1.4)
```



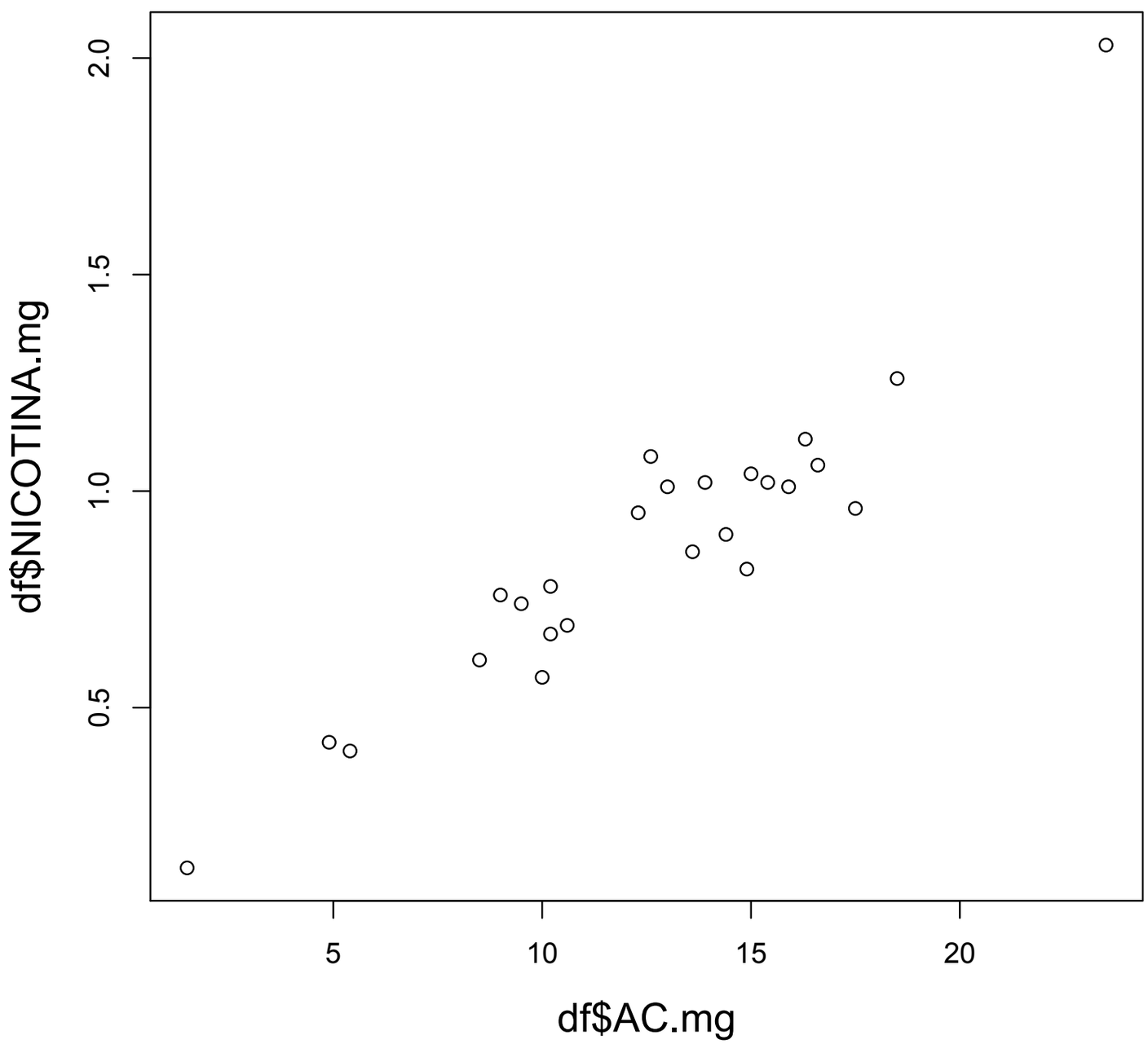
```
> x <- scale(df$PESO.g)
> y <- scale(df$AC.mg)
> plot(x, y)
> segments (0, -3, 0, 3)
> segments (-3, 0, 3, 0)
```


y
2
1
0
-1
-2
-2
-1
0
1
2
x

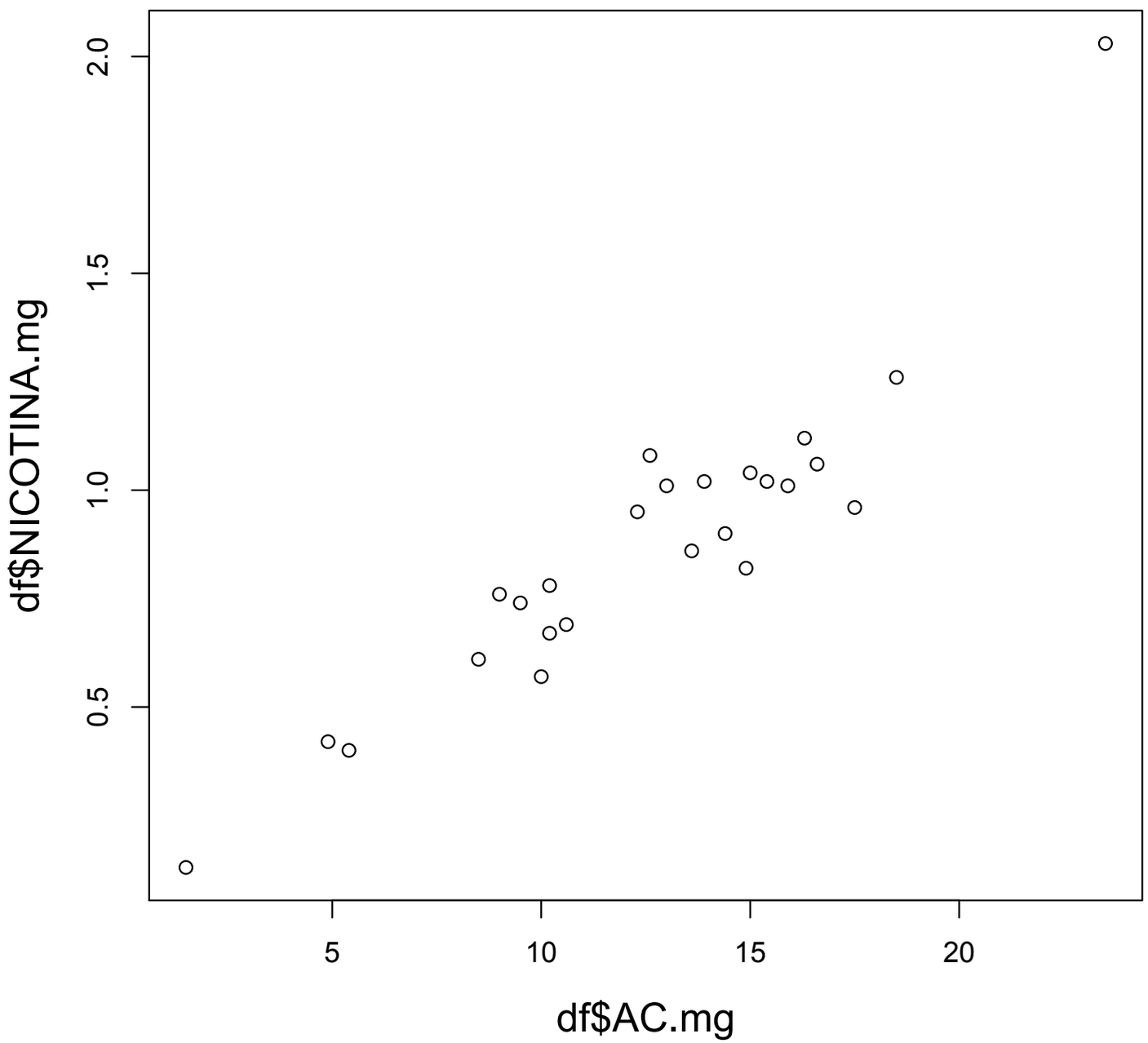
```
> r <- mean(x * y)
> r
[1] 0.4454008

> cor(df$PESO.g, df$AC.mg)
[1] 0.4639592

> n <- length(x)
> r <- sum(x * y)/ (n - 1)
> r
[1] 0.4639592
```

```
> cor(df$PESO.g, df$AC.mg)
[1] 0.4639592

> cor(df$AC.mg, df$PESO.g)
[1] 0.4639592

> cor(df$AC.mg, df$PESO.g, method =
"pearson")
[1] 0.4639592

> cor(df$AC.mg, df$PESO.g, method =
"spearman")
[1] 0.2169648
```

```
> plot(df$NICOTINA.mg ~ df$AC.mg, cex.lab = 1.4)
```



```
> cor(df$NICOTINA.mg, df$AC.mg, method = "pearson")
[1] 0.9259473
```



altra definizione

correlazione come covarianza standardizzata

$$r = \frac{\mathrm{cov}(x, y)}{\mathrm{sd}(x)\ \mathrm{sd}(y)} = \frac{\sum[(x - Mx)(y - My)]/n}{\mathrm{sd}(x)\ \mathrm{sd}(y)}$$

```
> x <- rnorm(30)
> y <- x

> cov(x,y)
[1] 1.339416

> sd(x) * sd(y)
[1] 1.339416

> var(x)
[1] 1.339416

> cov(x,y)/sd(x) * sd(y)
[1] 1
```

```
> x <- rnorm(30)
> y <- rep(0, 30)
> cov(x, y)
[1] 0

> sd(x)
[1] 1.157332
> sd(y)
[1] 0
> sd(x) * sd(y)
[1] 0
> cov(x, y)/ sd(x) * sd(y)
[1] 0
```