I FONDAMENTALI

Le parole dell'epidemia spiegate dai docenti di Scienze Statistiche di Padova

Valutazione dei fattori di rischio per la salute (correlazione e regressione)

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE STATISTICHE

Prof.ssa
LAURA VENTURA

docente di
STATISTICA MEDICA
e di
MODELLI STATISTICI

I FONDAMENTALI

Le parole dell'epidemia spiegate dai docenti di Scienze Statistiche di Padova

Valutazione dei fattori di rischio per la salute (correlazione e regressione)

Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Caso di studio: Misurazioni relative a uno studio *osservazionale* su un gruppo di *n* = 300 pazienti ricoverati per Covid-19.

Variabile di interesse: Durata della degenza.

Età	Sesso	Intubazione	Febbre	Tosse	Dimissione	Durata degenza
88	m	no	si	no	casa	6
72	M	si	si	no	decesso	5
62	M	no	si	si	decesso	3
66	F	NO	si	si	decesso	11
48	M	no	si	si	altra struttura	8
48	F	si	si	si	altra struttura	9
71	F	no	si	si	decesso	2
56	M	no	si	si	decesso	6
61	F	no	si	si	decesso	19
78	m	si	si	si	altra truttura	7
56	f	no	si	si	decesso	6
45	m	no	si	no	casa	23
65	m	si	si	si	casa	10
56	m	si	si	si	altra struttura	10
54	f	no	si	si	decesso	12
78	m	no	si	si	decesso	1
56	m	si	si	si	decesso	5
88	f	no	si	si	ancora ricorevato	7
43	m	no	si	si	decesso	4
85	m	si	si	si	decesso	2
80	f	no	si	si	altra struttura	10
76	f	si	si	si	altra struttura	8
78	m	si	si	si	decesso	11
68	m	no	si	si	decesso	1
56	m	si	si	si	decesso	9
45	m	no	si	si	altra truttura	5
76	f	no	si	no	decesso	5
68	m	no	si	no	casa	2
89	m	no	si	si	casa	2
66	m	no	si	si	altra struttura	15
76	m	no	si	si	decesso	1
67	f	no	si	si	decesso	5
54	f	no	si	si	decesso	9
78	m	no	si	si	altra truttura	10
56	f	no	si	si	decesso	9
56	M	si	si	si	casa	1

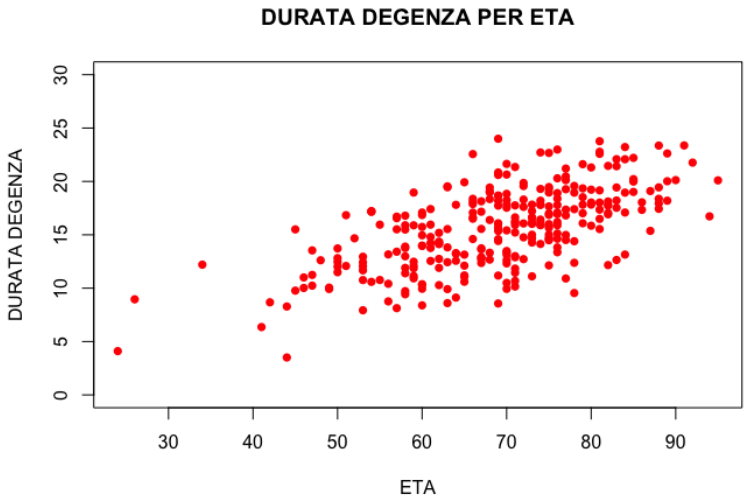
Obiettivo
Si vuole studiare l'associazione tra **età** dei pazienti e **durata** del ricovero.

STUDIO DELLA RELAZIONE TRA DUE VARIABILI QUANTITATIVE

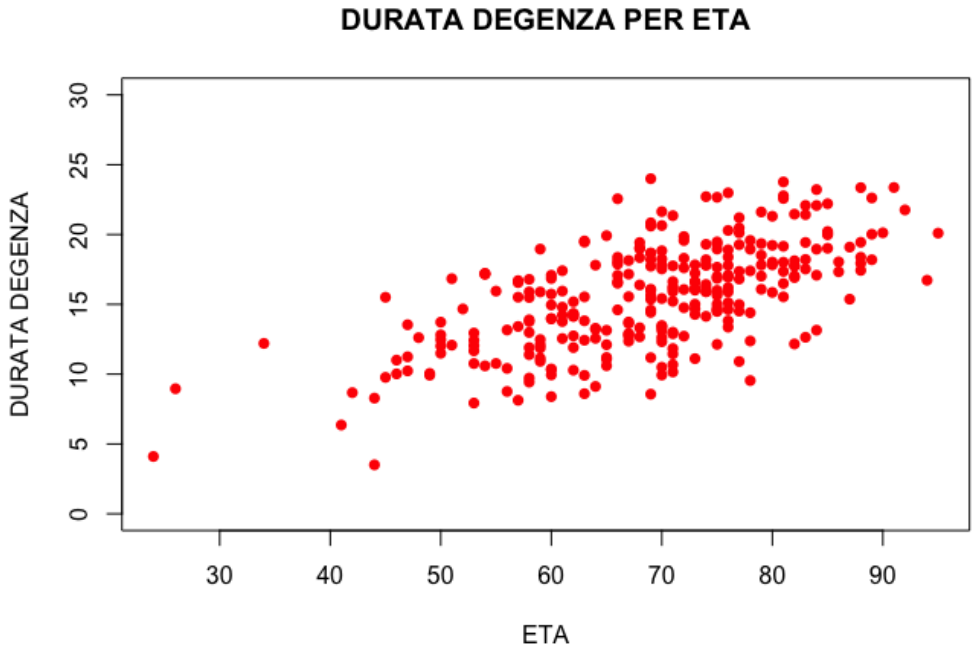
Quando le due variabili (causa e risposta) sono espresse attraverso valori numerici, la relazione che le lega può essere rappresentata graficamente in un piano cartesiano mediante un **diagramma di dispersione** (o a nuvola di punti):

- la variabile indipendente (causa) sull'asse *X*
- la variabile dipendente (risposta) sull'asse *Y.*

Nel nostro caso di studio: *X* = ETA e *Y* = DURATA.



Nel grafico sono riportate le coppie (x_i, y_i) per $i = 1, \ldots, n = 300$



Dal diagramma di dispersione si possono osservare:

- la **direzione** della relazione
- la **forma** della relazione
- la **forza** della relazione

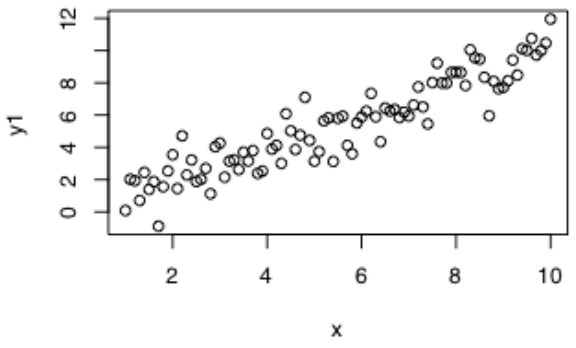
DIREZIONE

Si ha una **associazione positiva** quando al crescere dei valori della *X* crescono anche i valori della *Y.*

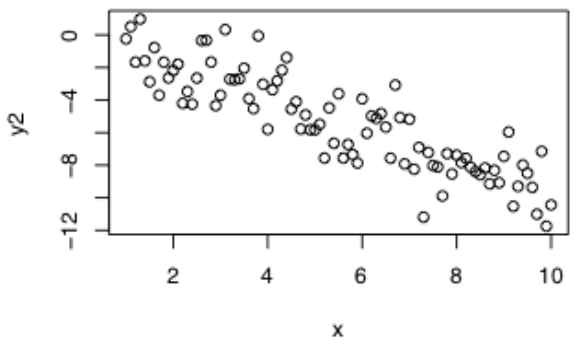
Viceversa, si ha **associazione negativa** quando al decrescere dell'una decrescono anche i valori dell'altra.

Se i punti del diagramma sono dispersi casualmente nel piano, non c'è associazione tra le le due variabili.

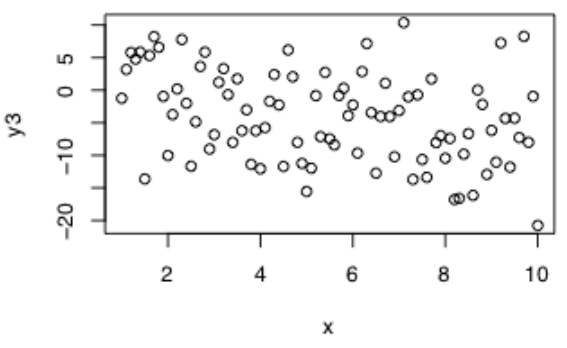
associazione positiva



associazione negativa



nessuna associazione

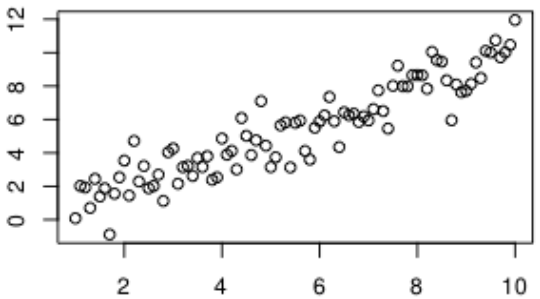


FORMA

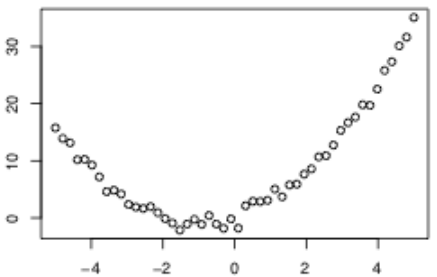
La forma viene desunta dalla disposizione dei punti nel diagramma.

Si parla di **relazione lineare** quando i punti si dispongono approssimativamente in linea retta.

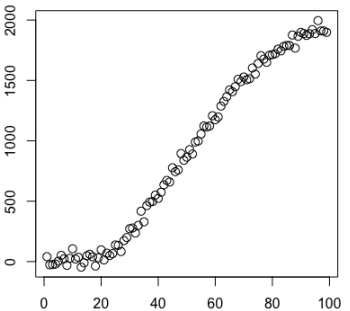
relazione lineare



relazione quadratica



relazione non-lineare

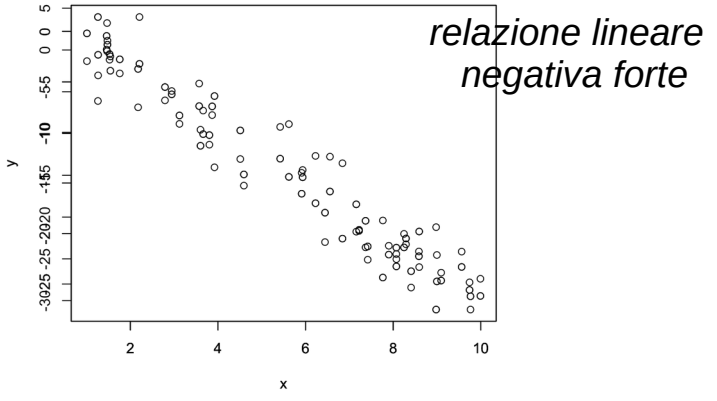


FORZA

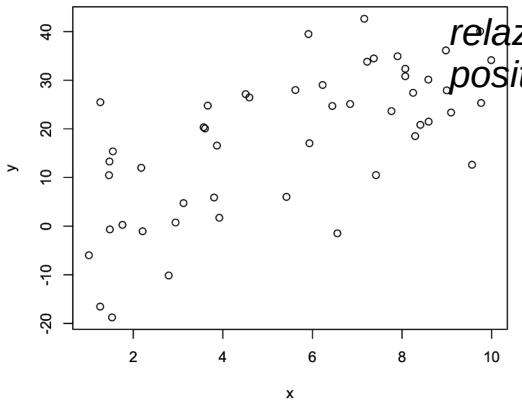
La forza si può desumere dalla dispersione dei punti nel diagramma.

Se i punti sono molto dispersi tra loro, la relazione tra le due variabili è debole.

Se invece i punti sono poco dispersi, allora la relazione è forte.



relazione lineare negativa forte



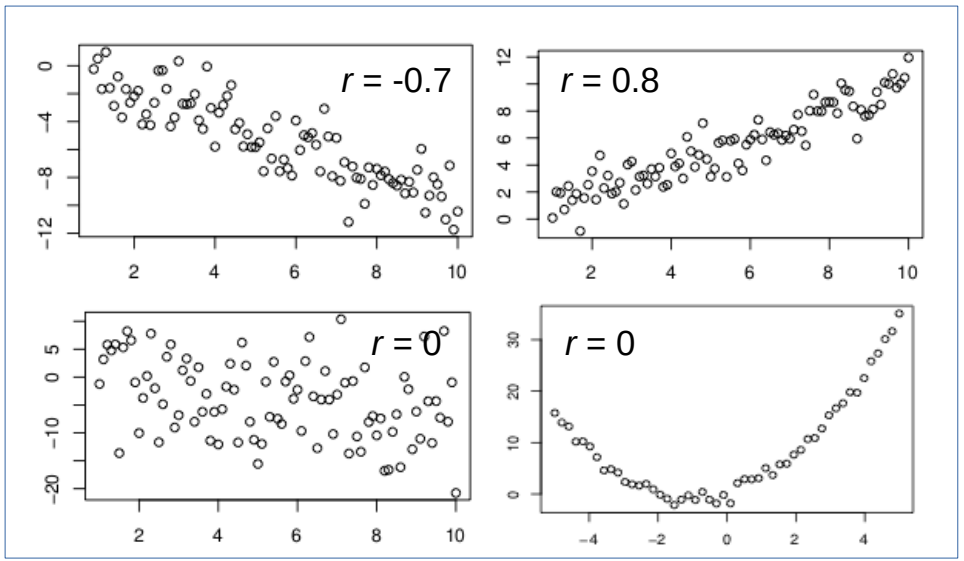
relazione lineare positiva debole

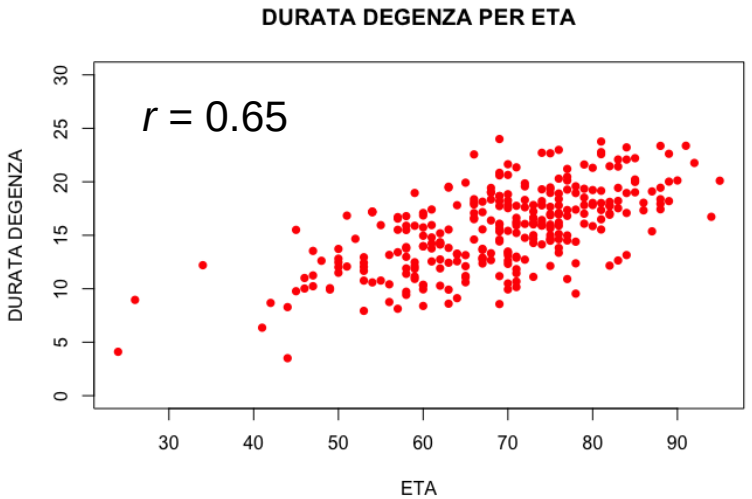
CORRELAZIONE

Per avere una valutazione analitica del grado di **associazione lineare** tra due variabili quantitative, esiste un indice che misura la direzione e la forza di una relazione lineare: l'**indice di correlazione *r*,** che assume valori nell'intervallo [-1,1]:

- se $r = \pm 1$: correlazione positiva/negativa perfetta (tutti i punti su una retta: crescente o descescente, rispettivamente)
- se $r > 0$: correlazione positiva
- se $r < 0$: correlazione negativa
- se $r = 0$: assenza di relazione lineare

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad \text{con} \quad \bar{x} = \frac{1}{n}\sum x_i \quad \text{e} \quad \bar{y} = \frac{1}{n}\sum y_i$$



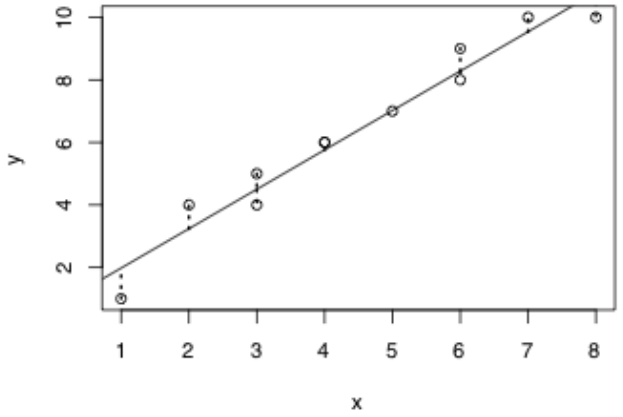


LA RETTA DI REGRESSIONE

- La regressione lineare si usa quando le variabili in studio hanno fra loro una relazione lineare, e quindi i punti del diagramma a dispersione tendono a disporsi secondo una linea retta.

Qual è la retta $Y = a + bX$ "giusta"?

- La retta ottimale è quella più “vicina” a tutti i punti del diagramma.
- Poichè le distanze positive e negative (punti sopra e punti sotto la retta *ideale)* si compensano, la loro somma, e quindi la media delle distanze, è nulla.
- Si considerano allora le distanze al quadrato. La retta “giusta” è allora quella che rende minima la somma delle distanze al quadrato.
- Per questo, tale metodo si chiama ***metodo dei minimi quadrati***.

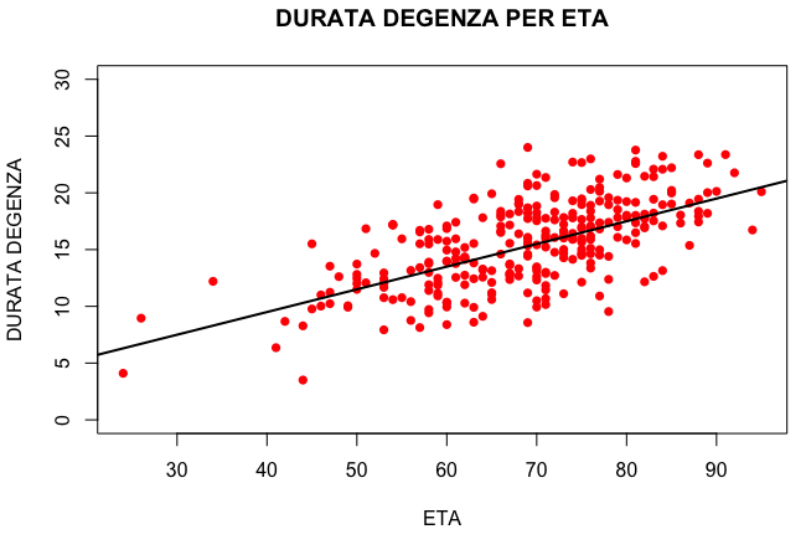


distanza = $\sum(y_i - a - bx_i)^2$

minima

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$



Nel caso di studio che stiamo considerando, si trova: **DURATA = 1.5 + 0.2 ETA**

Si deduce che:

- esiste una relazione lineare tra ETA' e DURATA della degenza
- l'associazione è positiva (a un aumento dell'età corrisponde un aumento della durata)
- l'associazione è *moderatamente* forte (r = 0.65)
- la pendenza della retta (b = 0.2) ci dice con quale velocità la degenza cresce con l'età

Per un paziente di 60 anni, la degenza è in media di 13.5 giorni (Y = 1.5 + 0.2 x 60 = 13.5).