PCTO in Coding & Data Science

Modulo 3

Scuola Morgagni, Roma, 03/02/2022

CD:50/50
CODING DIVERSITY

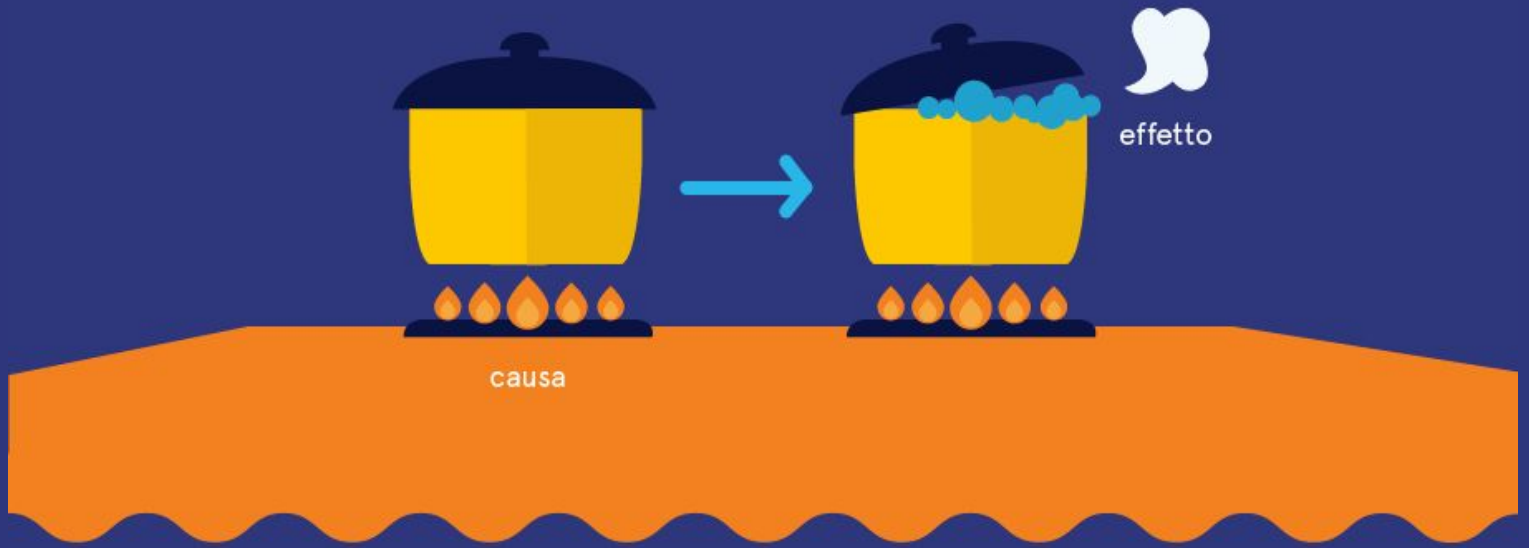
Correlazione

Che cosa vuol dire che due variabili / fenomeni sono correlati?
Vi vengono in mente degli esempi?

Correlazione vs. Causazione

CAUSAZIONE

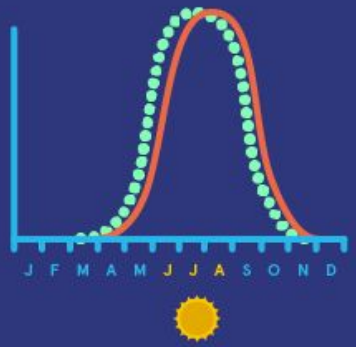
quando una cosa (una causa) provoca
il verificarsi di un'altra (un effetto)



Correlazione vs. Causazione

CORRELAZIONE

Due fenomeni «seguono lo stesso andamento»



le vendite di gelati e la percentuale di scottature solari sono correlate

questo significa che consumare gelato aumenta il rischio di scottature?

Correlazione vs. Causazione

Correlazione vs. Causazione

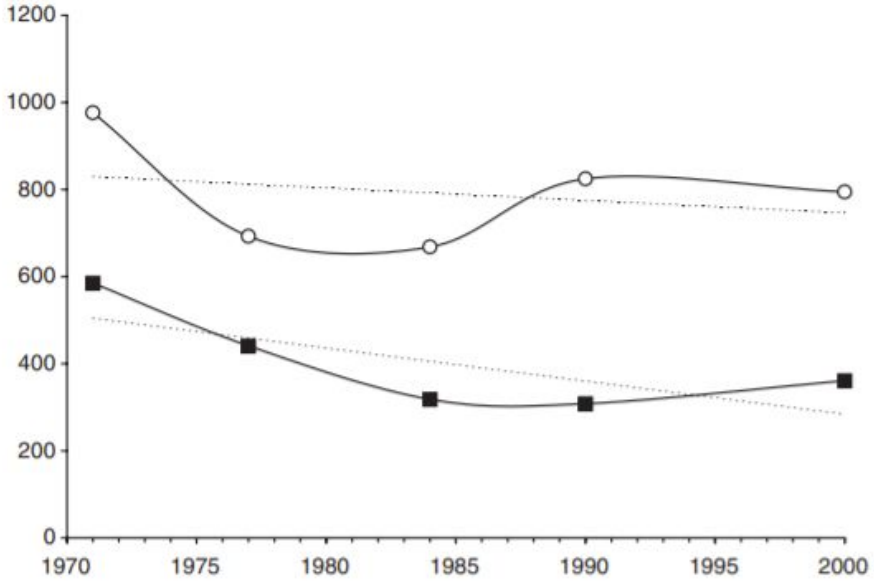


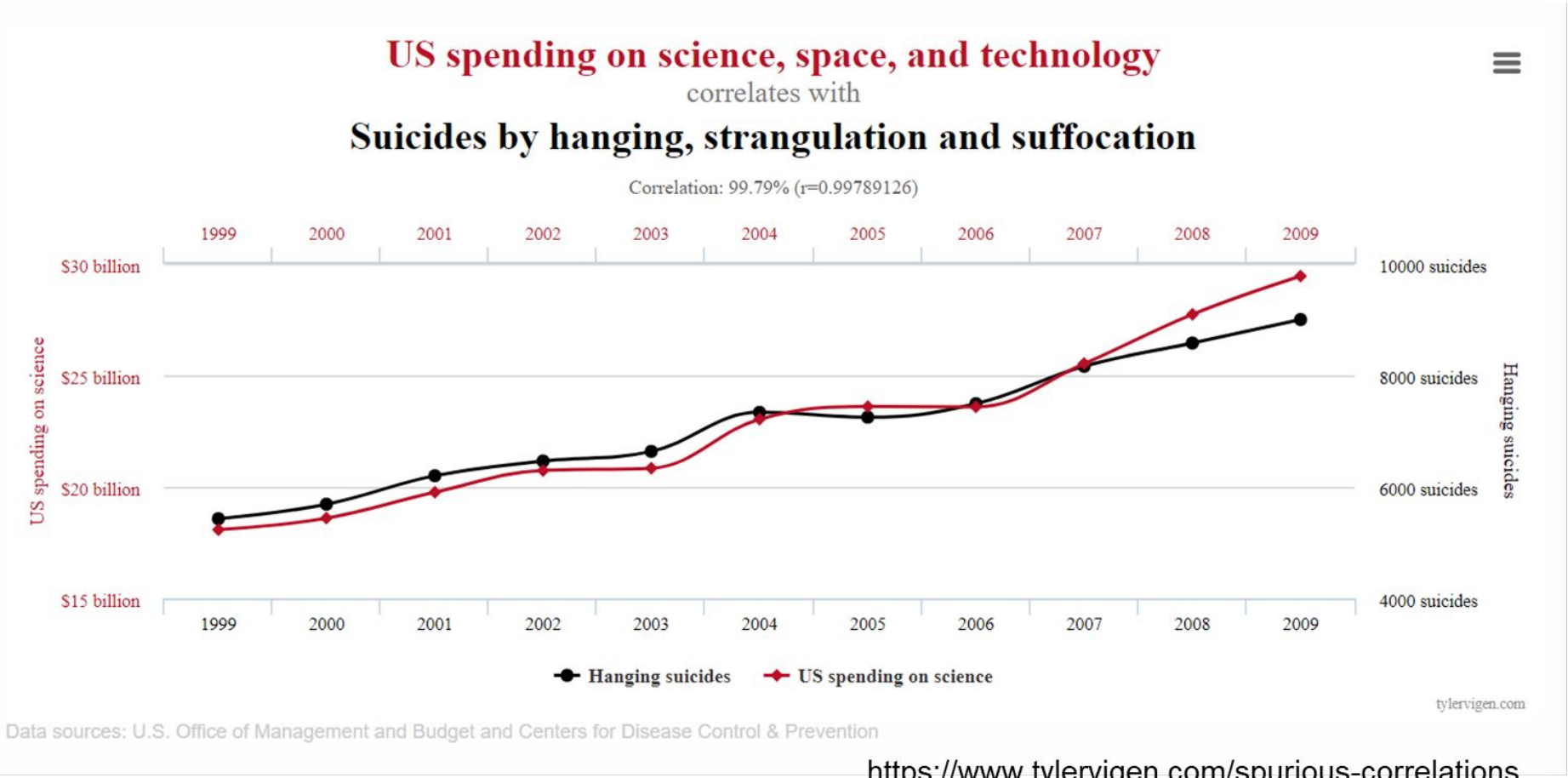
Figure 1. Storks and the birth rate in Lower Saxony, Germany (1971–2000). Open circles show yearly birthrates in hundreds in Lower Saxony. Full squares show numbers pairs of storks in Lower Saxony. Dotted lines represent linear regression trend ($y = mx + b$).

Nascite

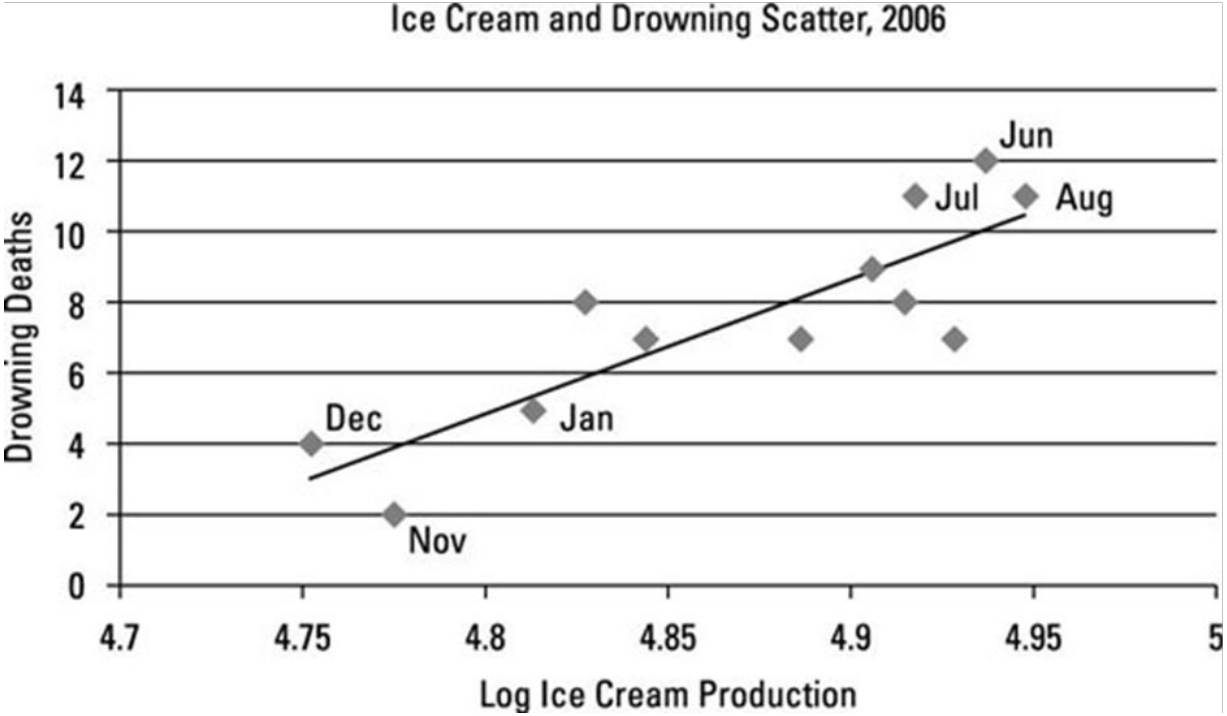
Numero di cicogne

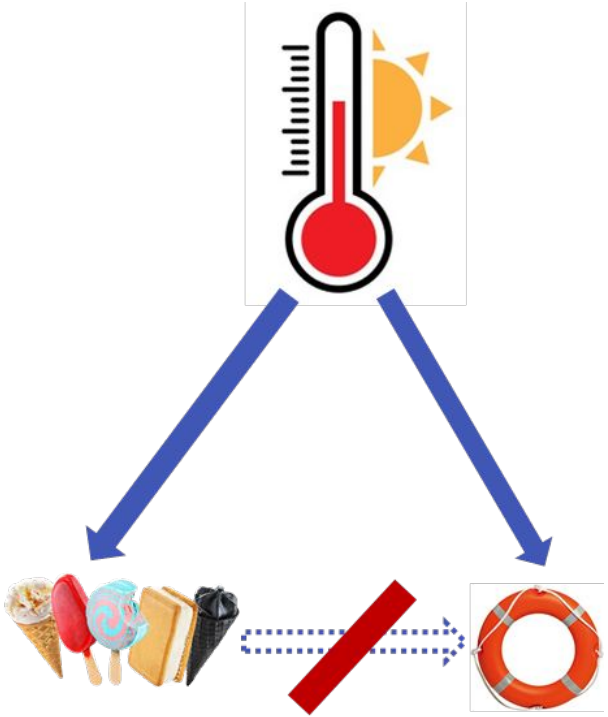
Quali fattori nascosti potrebbero coincidere con lo sviluppo dei due fenomeni?

https://web.stanford.edu/class/hrp259/2007/regression/storke.pdf

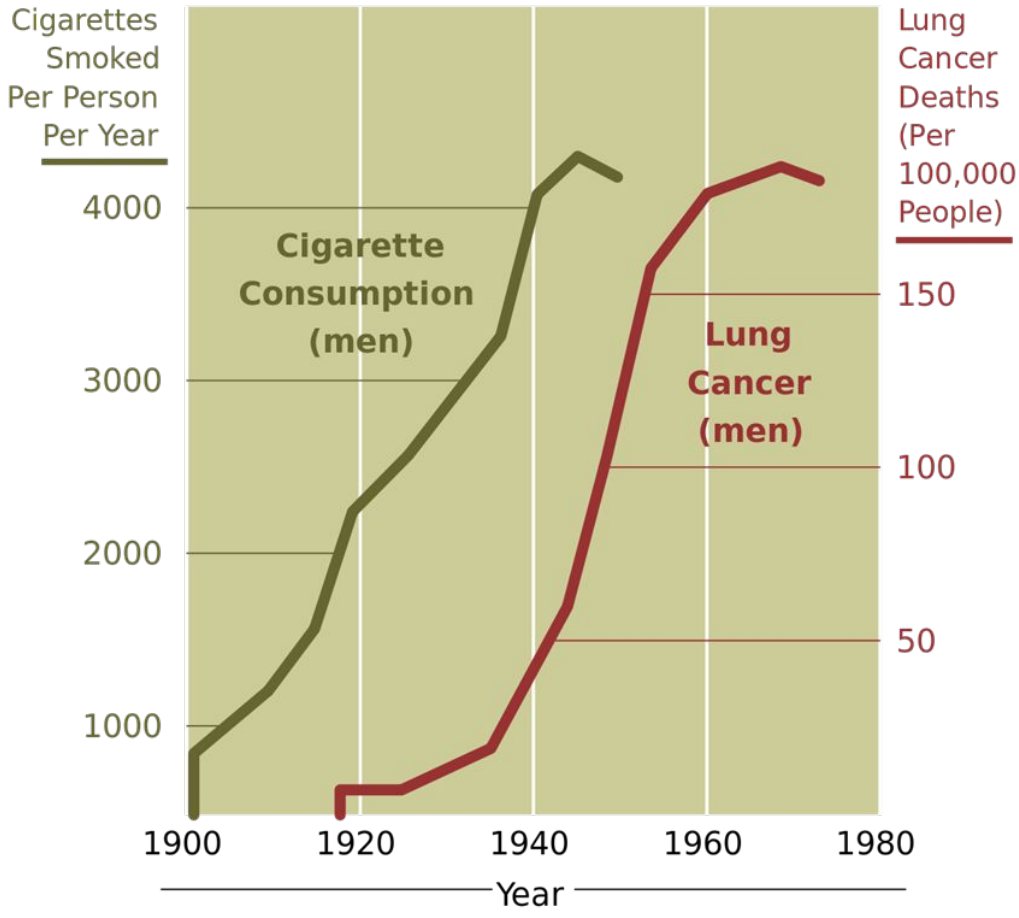


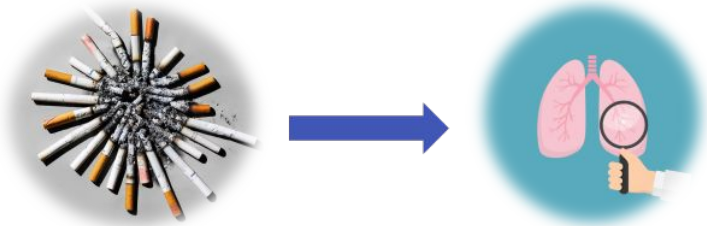
https://www.tylervigen.com/spurious-correlations


Ice Cream and Drowning Scatter, 2006
Drowning Deaths
14
12
10
8
6
4
2
0
4.7
4.75
4.8
4.85
4.9
4.95
5
Log Ice Cream Production
Jun
Jul
Aug
Dec
Jan
Nov



20-Year Lag Time Between Smoking and Lung Cancer





Correlazione vs. Causazione

La causazione è un problema difficile, che richiede esperimenti in ambienti controllati e analisi statistiche avanzate per determinare se una certa correlazione tra gli eventi è puramente frutto del caso oppure c’è effettivamente un nesso causale.

Covarianza

$$\operatorname{cov}(A, B) = \frac{(x_1 - \mu_A)(y_1 - \mu_B) + \ldots + (x_s - \mu_A)(y_t - \mu_B).}{n - 1}$$

Settimana	Lung. omero	Lung. femore
12	9	8
16	20	20
20	30	31
24	40	42
28	48	52
32	55	61
36	61	68
40	66	74

L'idea di fondo è:

1) date due variabili A e B sulla stessa popolazione [per esempio, peso e altezza misurati sulle stesse persone]
2) calcolo, per ogni campione delle due popolazioni, la sua distanza dalla media ($x-\mu_A$) e ($y-\mu_B$) [quanto è più grande o più piccolo rispetto alla media?)
3) moltiplico queste grandezze [relative allo stesso campione]
4) se hanno segno concorde (entrambe positive o negative), il contributo alla somma è positivo, altrimenti negativo.
5) Divido per "n" perché voglio fare la media.
6) Se Cov > 0 => le variabili crescono e decrescono insieme
7) Se Cov < 0 => alla crescita di una corrisponde la decrescita dell'altra
8) Se Cov = 0 => non c'è una regola, le variabili seguono andamenti diversi

Covarianza

$$\operatorname{cov}(A, B) = \frac{(x_1 - \mu_A)(y_1 - \mu_B) + \ldots + (x_s - \mu_A)(y_t - \mu_B).}{n - 1}$$

Settimana	Lung. omero	Lung. femore
12	9	8
16	20	20
20	30	31
24	40	42
28	48	52
32	55	61
36	61	68
40	66	74

$$\frac{1}{7}[(9 - 41.12)(8 - 44.5) + (20 - 41.12)(20 - 44.5) +$$

$$+(30 - 41.12)(31 - 44.5) + (40 - 41.12)(42 - 44.5) +$$

$$+(48 - 41.12)(52 - 44.5) + (55 - 41.12)(61 - 44.5) +$$

$$+(61 - 41.12)(68 - 44.5) +$$

$$(66 - 41.12)(74 - 44.5)] = 474.92$$

Covarianza

$$\text{cov}(A, B) = \frac{(x_1 - \mu_A)(y_1 - \mu_B) + \ldots + (x_s - \mu_A)(y_t - \mu_B).}{n - 1}$$

Settimana	Lung. omero	Lung. femore
12	9	8
16	20	20
20	30	31
24	40	42
28	48	52
32	55	61
36	61	68
40	66	74

$$\frac{1}{7}[(9 - 41.12)(8 - 44.5) + (20 - 41.12)(20 - 44.5) +$$

$$+(30 - 41.12)(31 - 44.5) + (40 - 41.12)(42 - 44.5) +$$

$$+(48 - 41.12)(52 - 44.5) + (55 - 41.12)(61 - 44.5) +$$

$$+(61 - 41.12)(68 - 44.5) +$$

$$(66 - 41.12)(74 - 44.5)] = 474.92$$

E' tanto o è poco?

Come per il CV, ci piacerebbe una misura più facile da interpretare (che non dipenda dalla scala).

Correlazione

$$\rho_{AB} = \frac{\operatorname{cov}(A, B)}{\sigma_A \cdot \sigma_B}$$

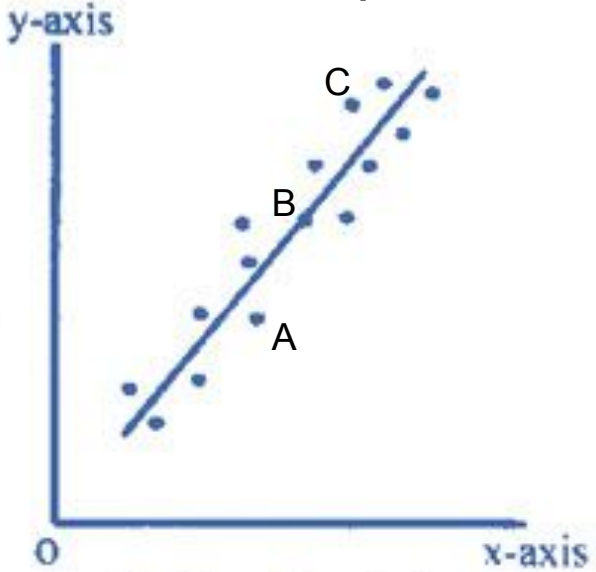
Dividiamo la covarianza per le due deviazioni standard delle variabili.
Questo valore va da -1 a 1:
1 => correlazione lineare diretta (le variabili hanno lo stesso andamento)
-1 => correlazione lineare inversa (le variabili hanno andamento inverso)
0 => nessuna correlazione

Settimana	Lung. omero	Lung. femore
12	9	8
16	20	20
20	30	31
24	40	42
28	48	52
32	55	61
36	61	68
40	66	74

In questo caso, il risultato è attorno allo 0,99, quindi le due variabili sono molto correlate linearmente!
=> Persone che hanno l'omero lungo avranno spesso anche il femore lungo, persone che hanno l'omero corto avranno spesso anche il femore corto.

NOTA: evidentemente, la lunghezza dell'omero non causa la lunghezza del femore né viceversa!

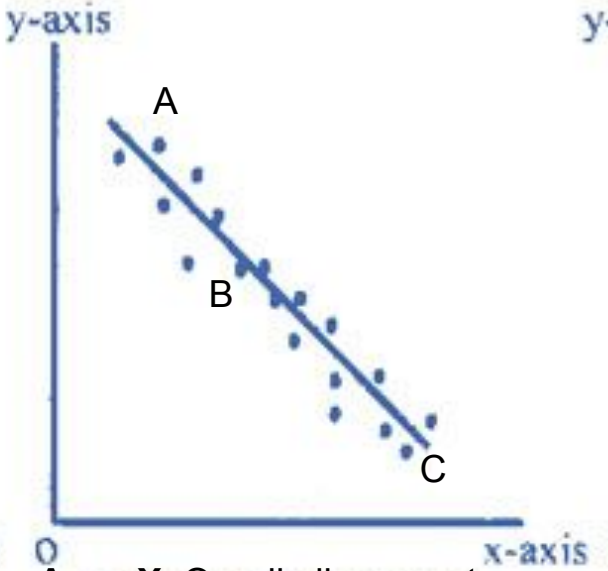
Correlazione positiva



Asse X: Numero di partite giocate
Asse Y: Goal fatti

Giocatore	**Asse X**	**Asse Y**
A	5	4
B	7	7
C	10	12

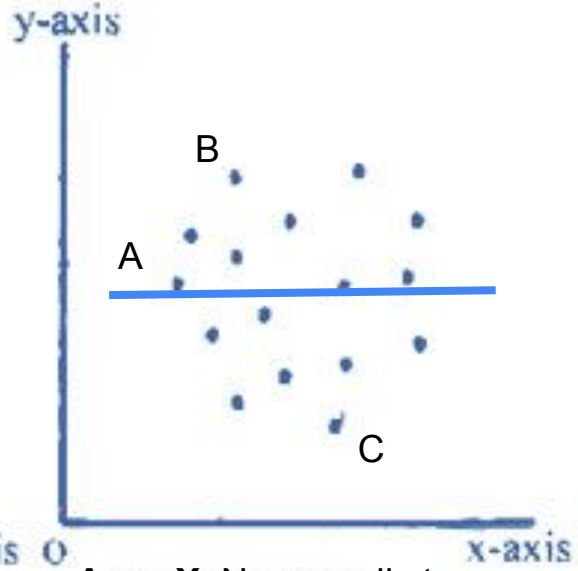
Correlazione negativa



Asse X: Ore di allenamento
Asse Y: Errori fatti nella canzone

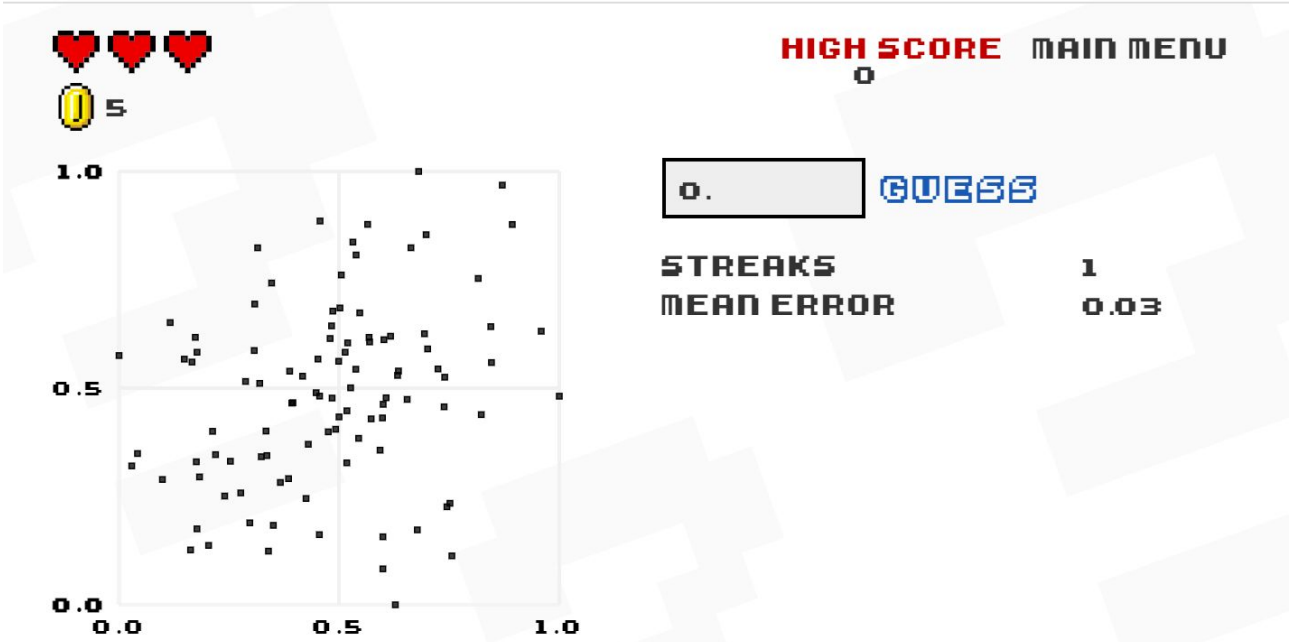
Allievo	**Asse X**	**Asse Y**
A	2	16
B	10	9
C	15	1

Correlazione nulla



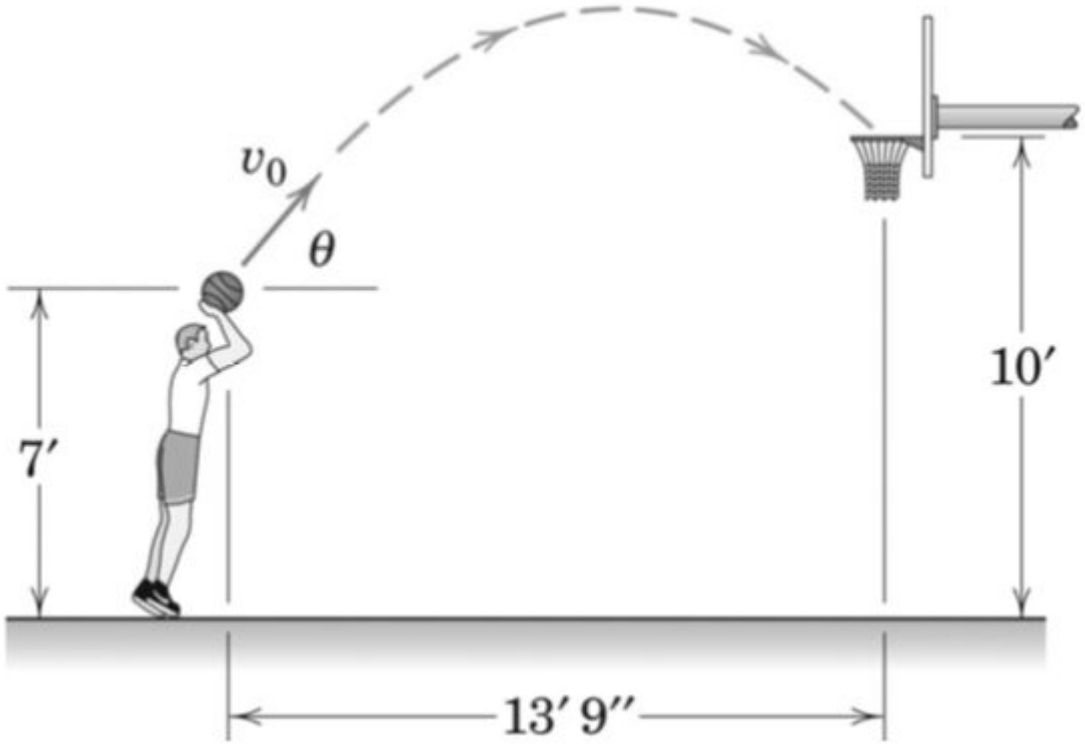
Asse X: Numero di stanze
Asse Y: Numero di quadri appesi

Casa	**Asse X**	**Asse Y**
A	2	4
B	3	6
C	6	1

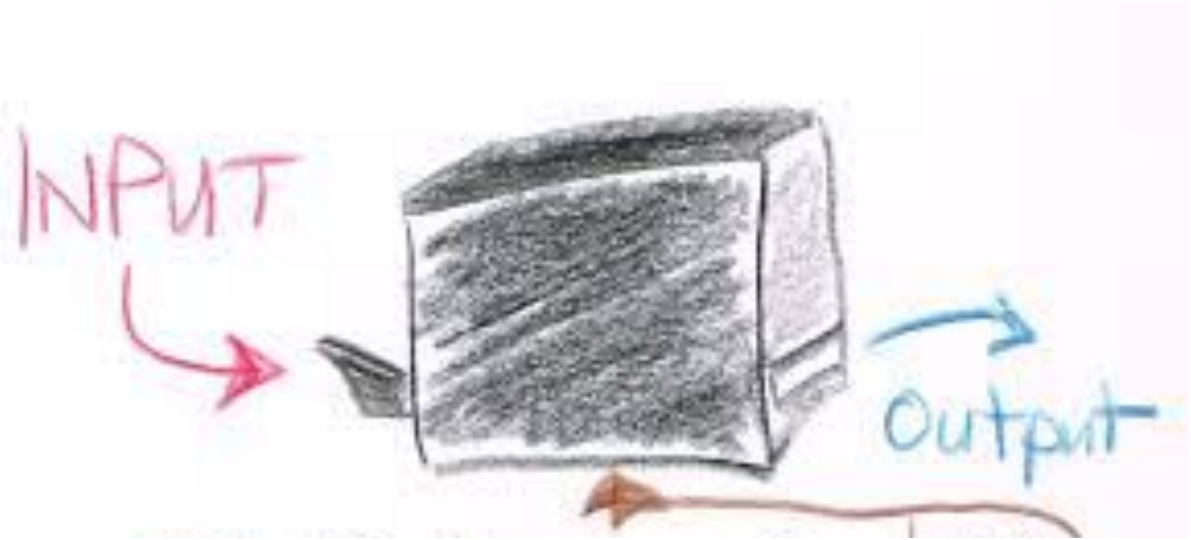


http://guessthecorrelation.com/

Modellare il mondo

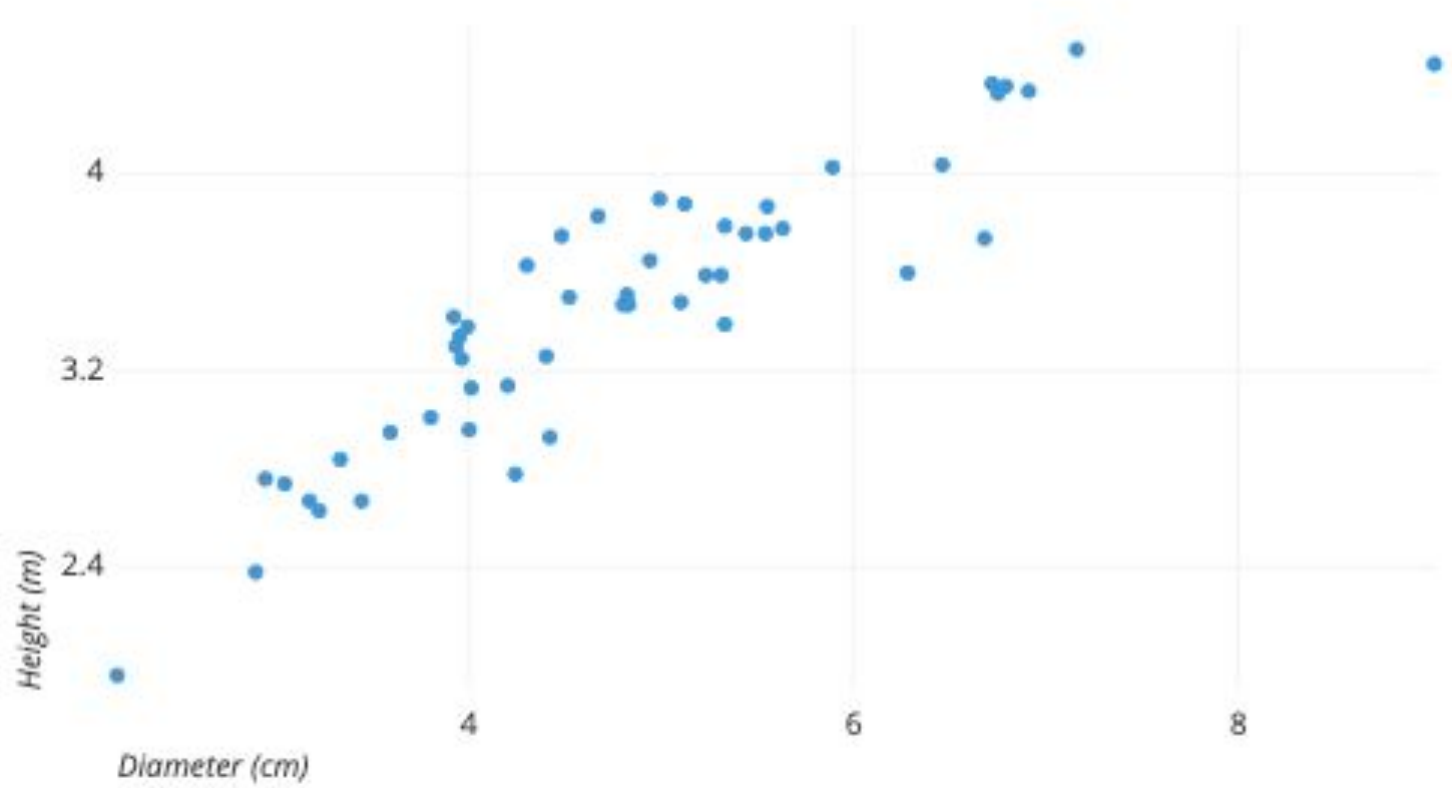


Creare modelli

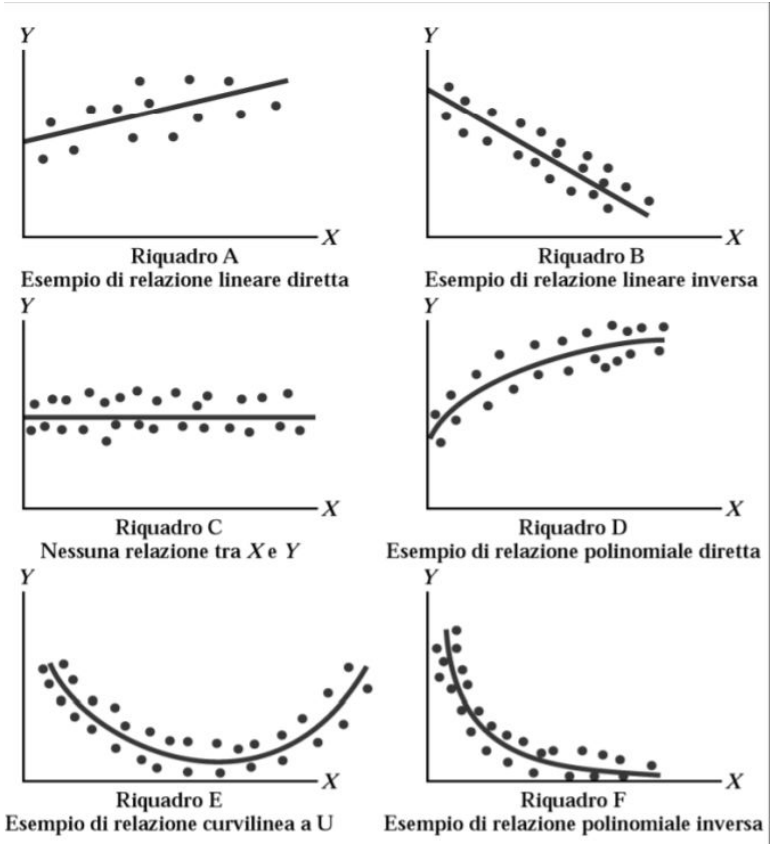


Qualcosa succede qui dentro

Creare modelli

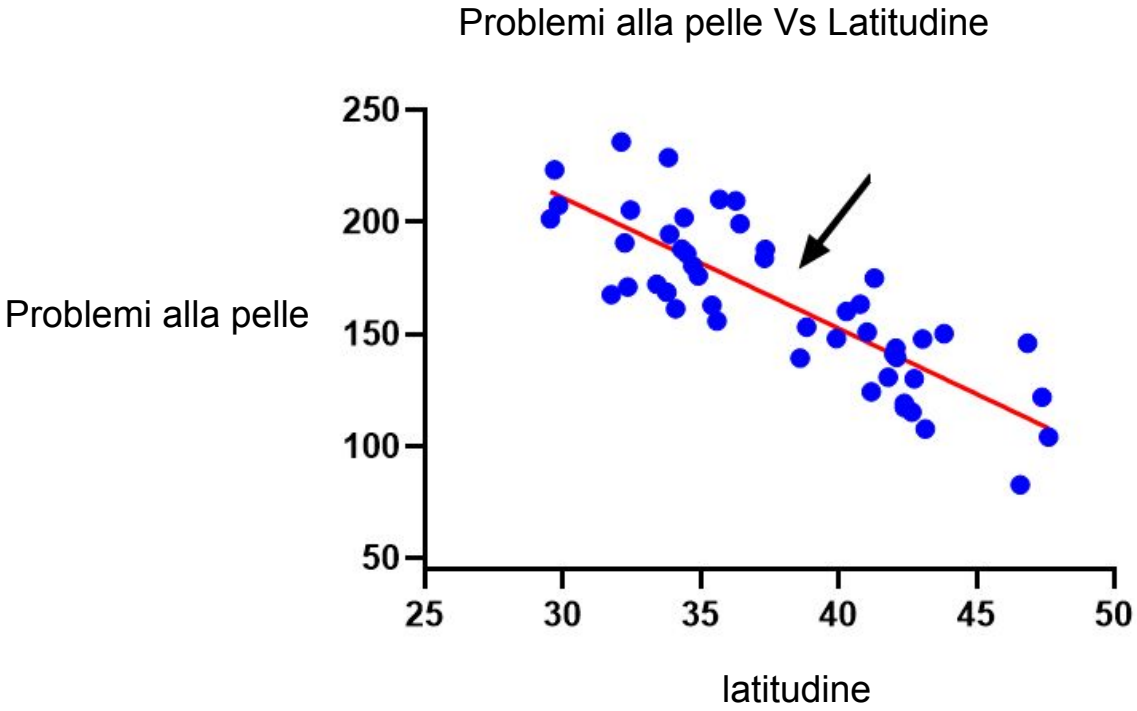


Creare modelli

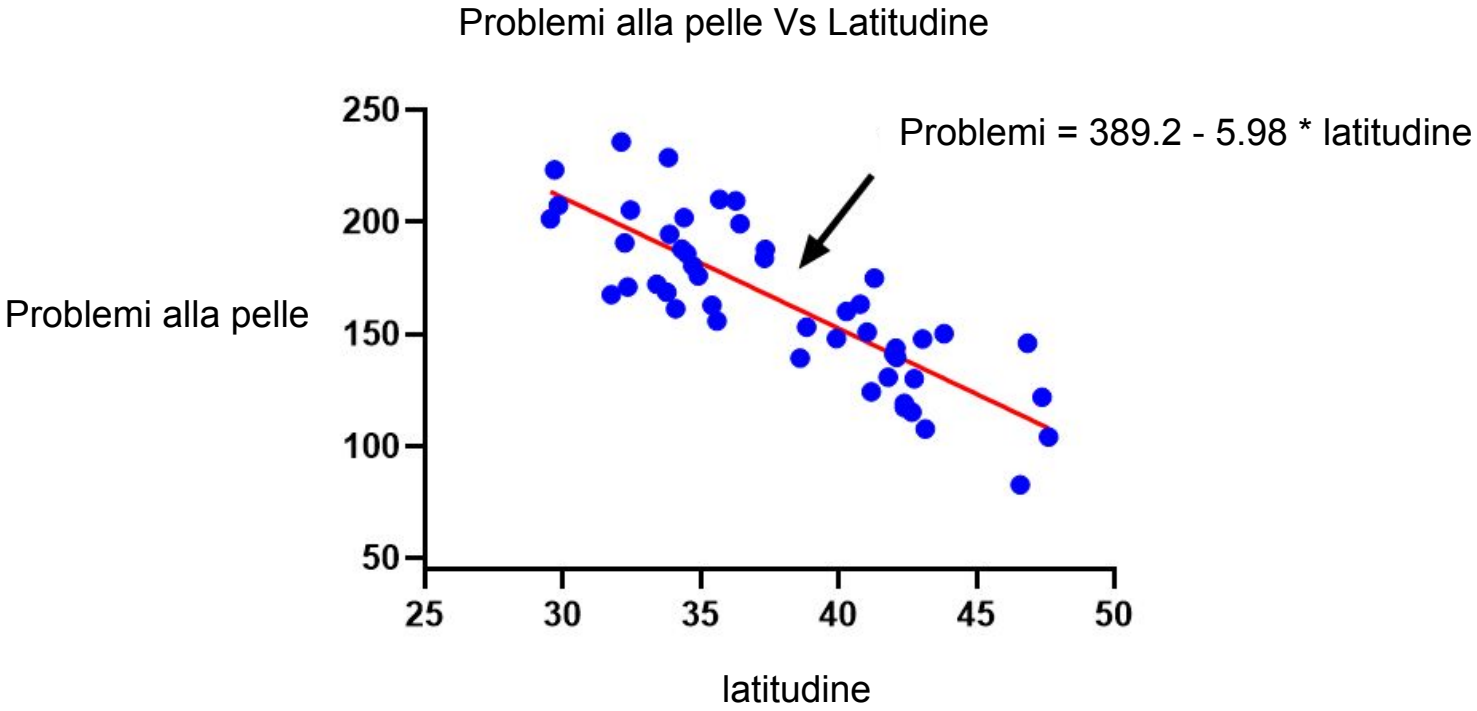


Dato un insieme di osservazioni, vorremmo provare a trovare un buon modello che rappresenti la nostra popolazione.
Per modello intendiamo una regola (per esempio un'equazione) che ben rappresenti i dati che abbiamo osservato.
Ammettendo un errore contenuto, si possono ottenere dei modelli molto generali che ben descrivono le nostre osservazioni, anche quelle che non abbiamo visto.

Regressione lineare



Regressione lineare



Regressione lineare

Date delle osservazioni (X,Y), con X variabile indipendente e Y variabile dipendente, cerchiamo la miglior regola nella forma:

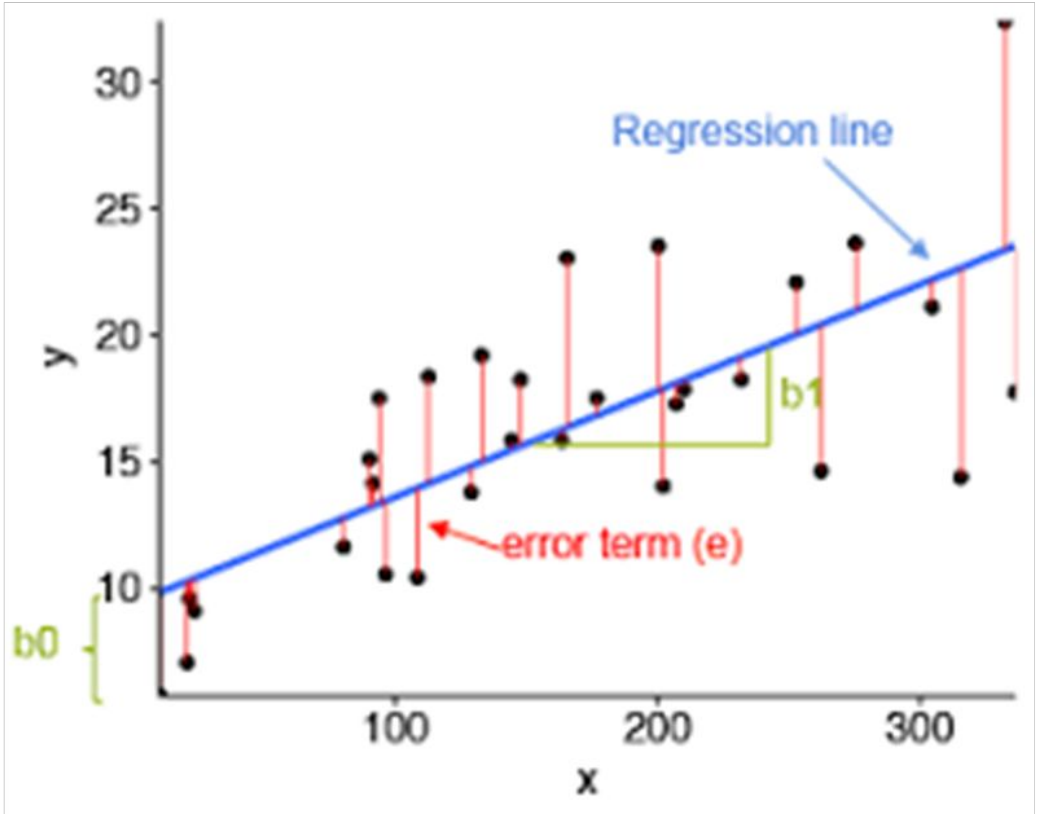
$$Y = a * X + b$$

Ovvero:

cerchiamo “a” e “b” tali per cui si minimizzi l’errore di predizione.

Questa equazione costituisce **un modello:** abbiamo scritto una legge matematica che in qualche modo modella un fenomeno.

Regressione lineare



Data una variabile indipendente x e una variabile dipendente y, cerchiamo la retta che meglio approssima la funzione y = f(x)
con
y* = f(x) = ax+b (il nostro modello, una retta)

Stimeremo quindi dei valori y* della nostra variabile dipendente e potremo calcolare l'errore come

Errore = |y*-y| (quello che abbiamo stimato con il nostro modelle meno il valore reale, misurato)

=> Tanto più è piccolo l'errore, tanto meglio la nostra retta modella in nostri dati

Capire i fattori che influenzano una variabile

ESEMPIO

Vogliamo considerare quali fattori influenzano il **costo di una casa.**
Per cui, decidiamo di fare una regressione lineare con le seguenti variabili:

variabile dipendente: costo_della_casa

variabili indipendenti: metri_quadri,
età_immobile

La nostra formula sarà:

costo_casa = a + b1*metri_quadri + b2*età_immobile

Calcolando i coefficienti dai nostra dati abbiamo ottenuto:

a = 50000
b1 = 3000
b2 = -2000

Per cui, il nostro modello è:

costo_casa = 50000 + 3000*metri_quadri - 2000*età_immobile

A parole: Il costo “di base” di una casa è 50’000€, più 3’000€ al metro quadro, ma ogni anno di vita della casa la deprezza di 2’000€.

Capire i fattori che influenzano una variabile

il nostro modello

costo_casa = 50000 + 3000*metri_quadri - 2000*età_immobile

- Che cosa succede se aumentiamo i metri quadri?
- E se aumentiamo l’età dell’immobile?

Predire la variabile indipendente date le variabili dipendenti

il nostro modello

costo_casa = 50000 + 3000*metri_quadri - 2000*età_immobile

1) Voglio vendere una casa di 60 metri quadri, costruita 10 anni fa. Qual è il prezzo “giusto”, secondo il mio modello?

   costo_casa = 50000 + 10*60 - 20*10 = 50400 euro

2) Sono intenzionato a comprare una casa: mi viene proposta una casa vecchia di 24 anni a 300’000€. Quanti metri quadri deve avere per risultare una soluzione conveniente per me?

   300000 = 50000 + 3000 * metri_quadri - 2000 * 22

   300000 = 50000 + 3000 * metri_quadri - 44000

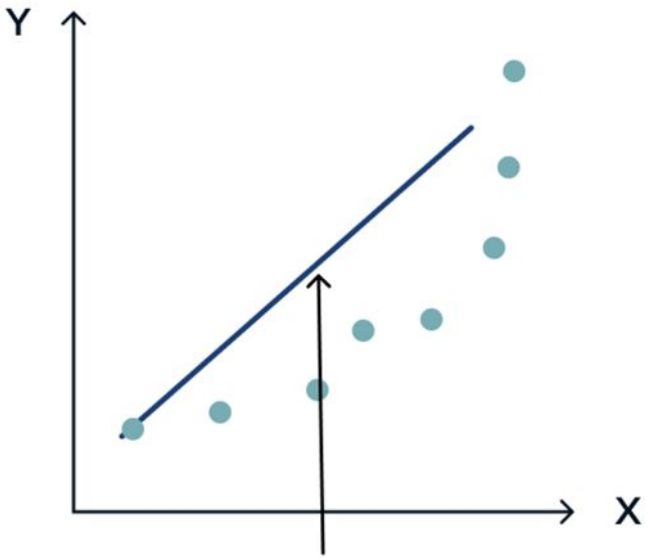
   300000 = 6000 + 3000 * metri_quadri

   300000 - 6000 = 3000 * metri_quadri

   294000 / 3000 = metri_quadri

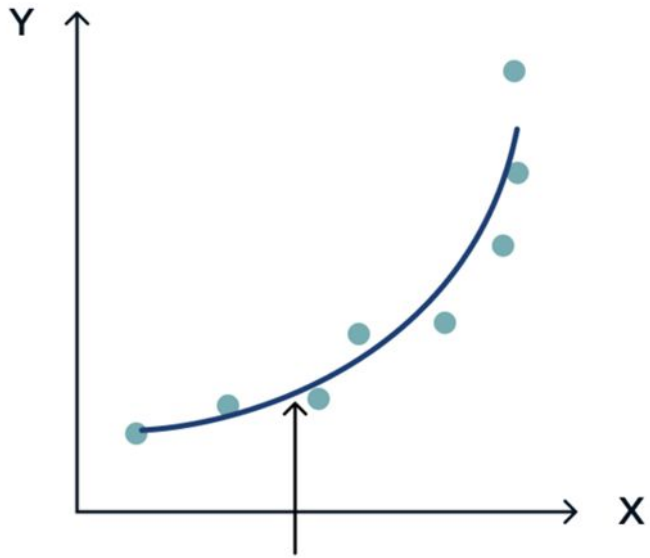
   L’affare è equo se: metri_quadri = 98
   Quindi da 99 metri_quadri in su ci guadagno!

Simple linear model



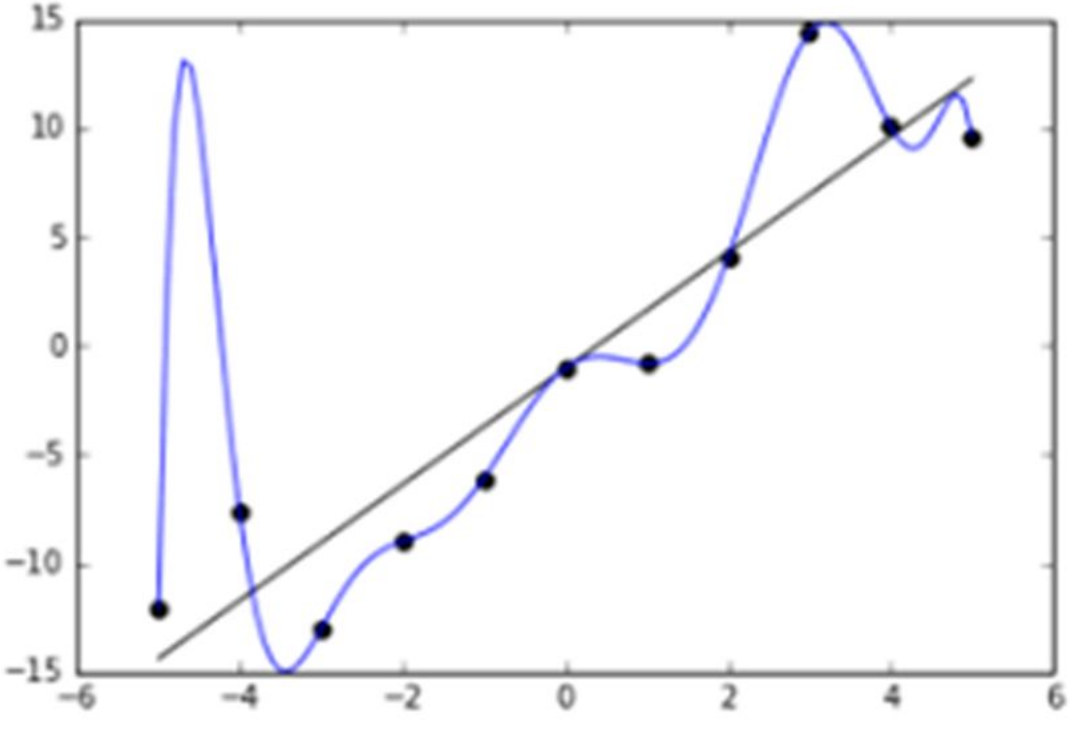
$$y=b_0+b_1x$$

Polynomial model

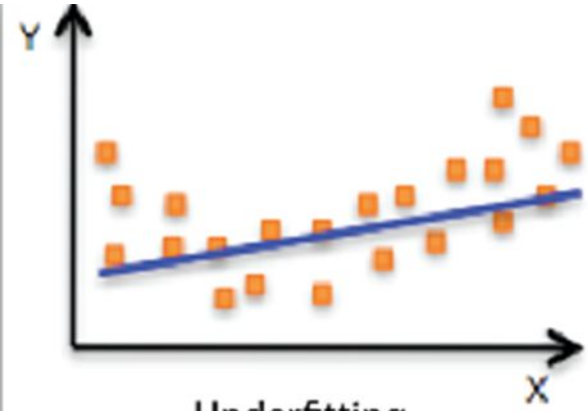


$$y=b_0+b_1x+b_2x_1^2$$

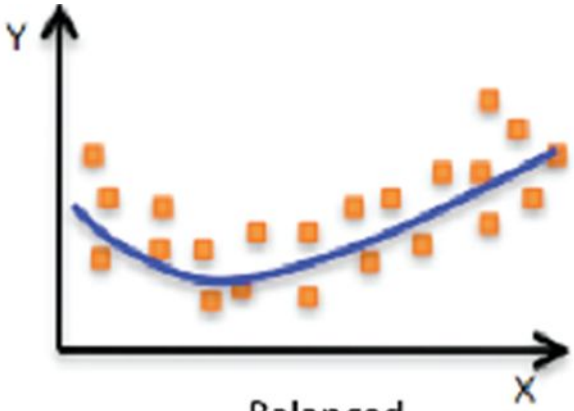
Non sempre la relazione che lega x e y è una retta…



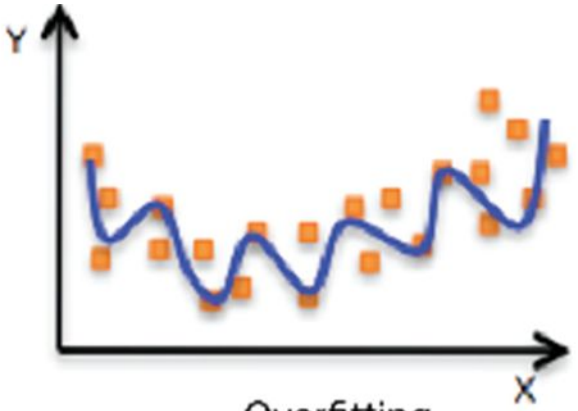
Ma attenzione a non cercare relazioni troppo complesse!



Underfitting



Balanced



Overfitting

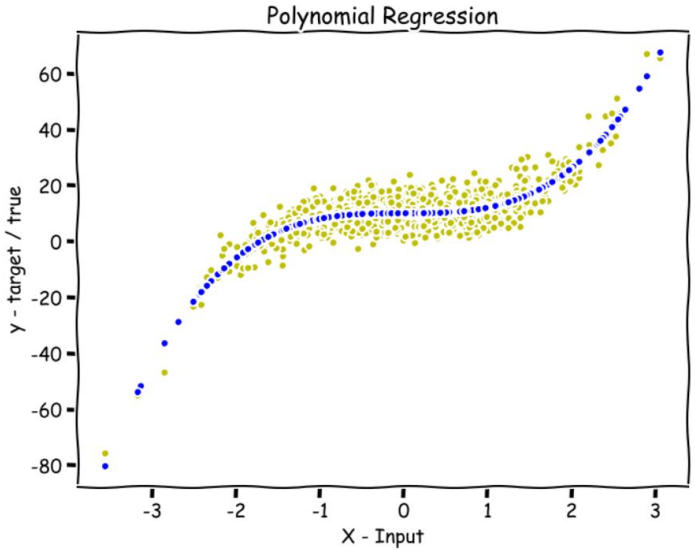
Relazione «troppo poco complessa»: non è espressiva abbastanza per rappresentare i dati, il modello commette un errore alto

Relazione «troppo complessa»: il modello commette pochissimo (o nessun) errore sui nostri dati, ma va bene sono per i nostri dati specifici, perde di generalità sui dati della stessa popolazione

Overfitting vs underfitting

Ovviamente è possibile usare funzioni più espressive come i polinomi:

```
mymodel = numpy.poly1d(numpy.polyfit(x, y, 3))
```



Ma modelli troppo potenti rischiano di non essere molto rappresentativi. In questo caso si parla di “Overfitting”: abbiamo cioè fatto troppo affidamento sui nostri dati, e ora il modello non è più generale