Minimi quadrati pesati per la Regressione Lineare

Salto in alto... oltre le formule

Ing. Ivano Coccorullo

Perchè?

- La tabella che segue riporta il raggio medio dell'orbita R ed il periodo di rivoluzione T di alcuni pianeti del sistema solare.

Pianeta	Mercurio	Venere	Marte	Giove	Saturno	Urano	Nettuno	Plutone
R [10^6 km]	57,9	108	228	778	1430	2870	4500	5900
T [10^6 s]	7,6	19,4	59,4	374	930	2650	5200	7820

- Ricavare il periodo di rivoluzione della Terra ($R=150*10^6$ Km)
- Ricavare la legge che lega il periodo della rotazione al raggio di un pianeta (III Legge di Keplero)

Perchè?

- Il rendimento di una pressa per la produzione di manufatti polimerici dipende dalla temperatura di esercizio.
- Dal lunedì al venerdì vengono eseguite un certo numero di misurazioni, da cui si rilevano la temperatura ed il numero di pezzi prodotti in corrispondenza.
- Si vuole sapere che rendimento avrà la macchina sabato se si imposta una temperatura di 120°C?

Perchè?

- Campo economico-pratico:
  - Diametro e volume alberi, bar e cinema...
- Campo medico:
  - un campione di N città, si vuole legare la quantità di polveri fini disperse nell'aria e il tasso di incidenza delle allergie più comuni

La regressione

- Analizza i legami esistenti tra due variabili aleatorie (X, Y), partendo da un insieme di coppie di dati (x_i, y_i) rilevati, si determina, se possibile, una funzione $y=f(x)$ che rappresenti il fenomeno preso in esame. Gli scopi per cui si cerca tale funzione sono:

  - ✓ Descrivere sinteticamente la relazione fra due variabili osservate;
  - ✓ Determinare la legge di distribuzione dei dati statistici;
  - ✓ Ricavare eventuali dati intermedi mancanti;
  - ✓ Correggere valori affetti da errori accidentali o perturbati da cause secondarie.

La regressione

- La regressione è un modello che cerca di stabilire una relazione di “causalità” tra due variabili aleatorie.
- Il nesso di causalità dipende dal contesto da cui le variabili aleatorie provengono, e la “spiegazione” di questo nesso è del tutto interna alla disciplina in esame. Tuttavia la struttura statistica è la stessa indipendentemente dal contesto interpretativo.

Come funziona

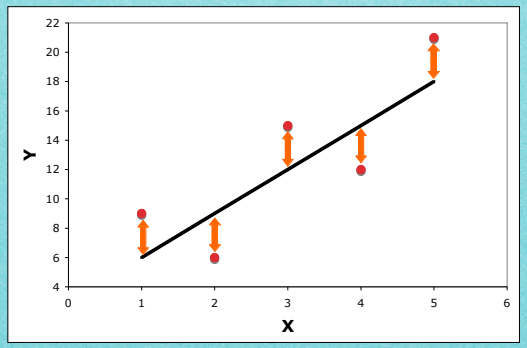
- Consideriamo due variabili aleatorie X (variabile indipendente) ed Y (variabile dipendente) sulle quali sono effettuate m rilevazioni espresse dalle coppie (x_i, y_i).
  - Le coppie dei dati rilevati si rappresentano sul piano cartesiano mediante punti (Diagramma a dispersione).
  - Si sceglie il tipo di funzione che esprime meglio la relazione
  - Si determinano i parametri della funzione scelta

Scelta della funzione...

- Non esistono regole precise per determinare priori il tipo di funzione:
- se la relazione tra X ed Y è proporzionale diretta o se gli incrementi dei valori di Y, per incrementi costanti di X, sono quasi costanti, si sceglie la retta
- se la relazione tra X ed Y è proporzionale inversa, si sceglie l'iperbole
- se le x_i sono in progressione aritmetica e le y_i sono in progressione geometrica, si sceglie la curva esponenziale

...ed i parametri? ...

- **Metodo dei minimi quadrati**
  - (x_i, y_i) dati e $(x_i, f(x_i))$ funzione approssimante
  - ***Residuo i-esimo*** $r_i = y_i - f(x_i)$



...ed i parametri? ...

- **Metodo dei minimi quadrati**

- Best fit è la curva che rende minimo l'errore:
  - $E=\sum_i (r_i)^2=\sum_i [y_i-f(x_i)]^2$
  - la somma degli scarti quadratici tra i dati y_i misurati ed i risultati previsti $f(x_i)$ sia minima.

Regressione lineare

- Il modello più semplice per descrivere quantitativamente un tale nesso causale è quello di assumere che la variabile dipendente y sia una funzione lineare della variabile indipendente x:

$$f(x)=a+bx$$

- dove
  - a = intercetta
  - b=coefficiente angolare (pendenza)

calcolare i parametri

Retta di regressione lineare

Come calcolare il valore di a e b?

Si deve minimizzare la funzione:

$$\rho : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

$$\begin{aligned} \rho(a,b) = \sum_{i=1}^{m} r_i^2 &= \sum_{i=1}^{m} (y_i - (bx_i + a))^2 \\ &= \sum_{i=1}^{m} y_i^2 + (bx_i + a)^2 - 2(bx_i + a)y_i \end{aligned}$$

ρ è minima nei punti **a** e **b** tali che: $\begin{cases} \frac{\partial \rho}{\partial b} = 0 \\ \frac{\partial \rho}{\partial a} = 0 \end{cases}$

calcolare i parametri

Retta di regressione lineare

$$\begin{cases} \frac{\partial \rho}{\partial b} = \sum_{i=1}^{m} 2bx_i^2 - 2x_i y_i + 2x_i a = 0 \\ \frac{\partial \rho}{\partial a} = \sum_{i=1}^{m} 2a + 2bx_i - 2y_i = 0 \end{cases}$$

$$\begin{cases} b\sum_{i=1}^{m} x_i^2 + a\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} x_i y_i \\ a\sum_{i=1}^{m} 1 + b\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \end{cases}$$

ponendo:

$$S_{xx} = \sum_{i=1}^{m} x_i^2, \qquad S_x = \sum_{i=1}^{m} x_i$$

$$S_{xy} = \sum_{i=1}^{m} x_i y_i, \qquad S_y = \sum_{i=1}^{m} y_i$$

calcolare i parametri

Retta di regressione lineare

si ha:

$$\begin{cases} S_{xx}b + S_x a = S_{xy} \\ S_x b + ma = S_y \end{cases}$$

Risolvendo le equazioni:

$$\begin{cases} b = \frac{1}{d}(S_x S_y - m S_{xy}) \\ a = \frac{1}{d}(S_x S_{xy} - S_{xx} S_y) \end{cases}$$

dove:

$$d = S_x^2 - m S_{xx}$$

considerazioni sugli errori...

- Questo procedimento vale se l'errore sulla variabile indipendente è molto minore di quello sulla variabile dipendente.
- Comunque nel calcolo dell'intercetta (a) e della pendenza (b) si commette un errore che va considerato:

$$s_a = \sum \left| \frac{da}{dy_i} \right| \Delta y_i$$

$$s_b = \sum \left| \frac{db}{dy_i} \right| \Delta y_i$$

considerazioni sugli errori...

$$s_{y/x} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - (a + bX_i))^2}{n-2}} \text{(deviazione standard)}$$

$$s_a = s_{y/x}\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$s_b = \frac{s_{y/x}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

dove $\bar{X}$ è la media delle ascisse:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Bontà della regressione

Quanto la retta approssima bene i punti?
Il coefficiente di correlazione

$$R = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{(\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2)^{\frac{1}{2}}}$$

Il coefficiente di correlazione dà un'indicazione della bontà dell'adattamento dei punti alla retta: tanto piú $|R| \simeq 1$, tanto migliore è l'adattamento.

- dove $\bar{x} = \sum_{i=1}^{m} x_i$ $\bar{y} = \sum_{i=1}^{m} y_i$

Kart Pearson, matematico inglese 1857-1936

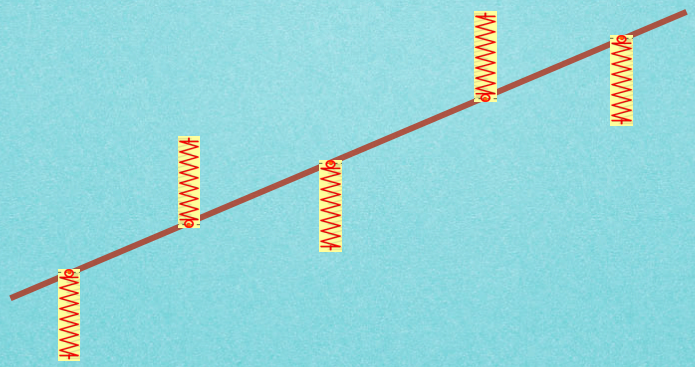
Laboratorio di informatica

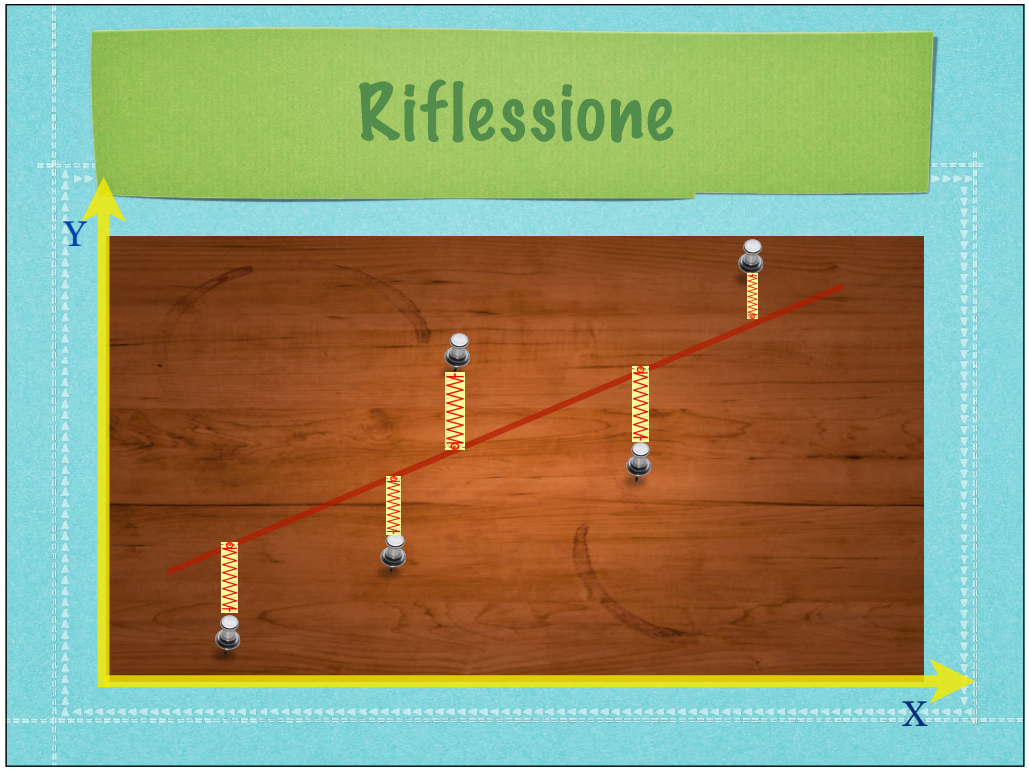
II fase


Riflessione
Y
X


Riflessione
Y
X

Riflessione




Riflessione
Y
X

Riflessione guidata

- Come si disporrà il listello?
- Il listello di legno si disporrà esattamente secondo la regressione lineare ottenuta applicando il metodo dei minimi quadrati
- Perché?

Riflessione guidata

- Suggerimento:
  - Cosa rappresentano gli allungamenti della molla?
  - Quanto vale l'energia di una molla e del sistema?
  - Qual'è la configurazione energetica a cui tende un sistema?
  - Quindi.....