

STATISTICA E ANALISI DEI DATI

Capitolo 5 – Regressione Lineare

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

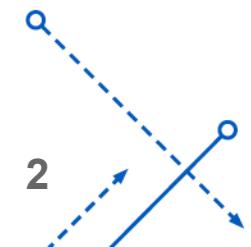
a.a. 2025-2026

Breve Recap

Il coefficiente di correlazione campionario r_{xy} misura la forza del legame di natura lineare esistente (se esiste) tra due variabili quantitative.

NOTA: Relazioni tra le variabili che assumono una **forma curvilinea** (o altra forma) non possono pertanto essere individuati con tale coefficiente (servono altri metodi).

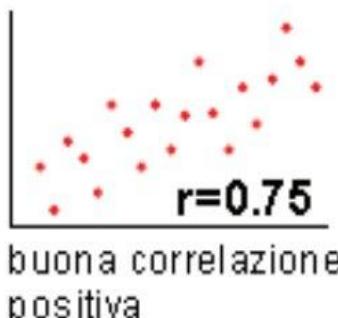
Il **segno** di r_{xy} indica sia la direzione della **retta interpolante** e sia il ritrarsi in una delle situazioni descritte nelle slide successive.



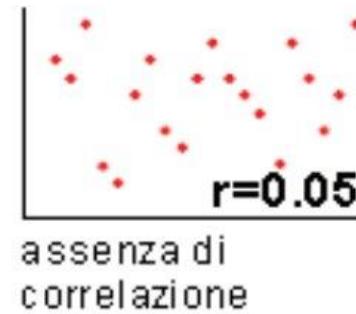
Comportamento del Coefficiente



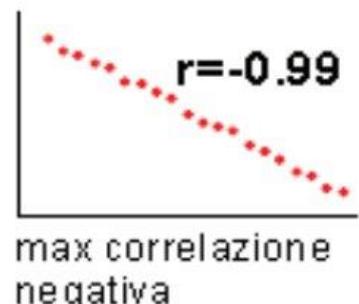
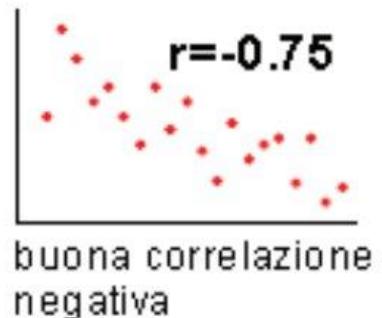
$r_{xy} = 1$
(correlazione perfetta positiva)



r_{xy} in $]0,1[$
(correlazione positiva)
I punti sono posizionati in una nuvola attorno ad una linea retta interpolante ascendente



$r_{xy} = 0$ (nessuna correlazione) i punti sono **completamente dispersi in una nuvola che non presenta** alcuna evidente direzione di natura lineare;



Interpolazione /1

L'interpolazione è un metodo matematico che permette di **stimare o calcolare valori intermedi tra due o più punti dati conosciuti.**

ESEMPIO

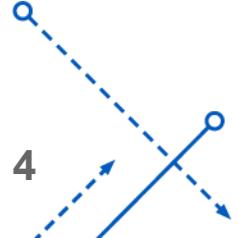
X rappresenta i voti degli studenti a LP1

Y rappresenta i voti degli studenti a LP2

Il **CdCC** tra **X** e **Y** è molto vicino a 1 (forte relazione lineare positiva)

- Lo studente s1 ha preso 23 a LP1 e 25 a Lp2
- Lo studente s3 ha preso 25 a LP1 e 27 a Lp2
- Nessuno studente ha preso 24 a LP1

Potremmo chiederci: "che voto" potrebbe raggiungere uno studente che prende 24 a LP1 a LP2 vista la correlazione tra X e Y?

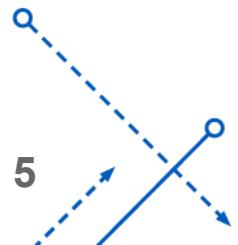


Interpolazione /2

Il processo di **interpolazione** prevede l'uso di una **formula o una funzione matematica** per calcolare il **valore approssimato** tra i punti noti.

L'obiettivo principale **dell'interpolazione** è ottenere una **stima accurata dei dati mancanti o intermedi basandosi sui dati disponibili.**

**Accettando l'idea che una correlazione forte sia mantenuta
che nei punti ignoti**



Retta interpolante e Interpolazione Lineare

Definizione "informale"

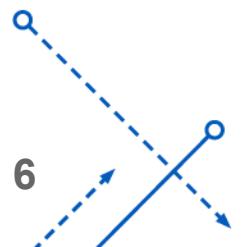
Una **retta interpolante** è un tipo specifico di **interpolazione** che coinvolge l'uso di **una retta** (una linea retta) per **stimare o approssimare valori intermedi tra due punti dati conosciuti**.

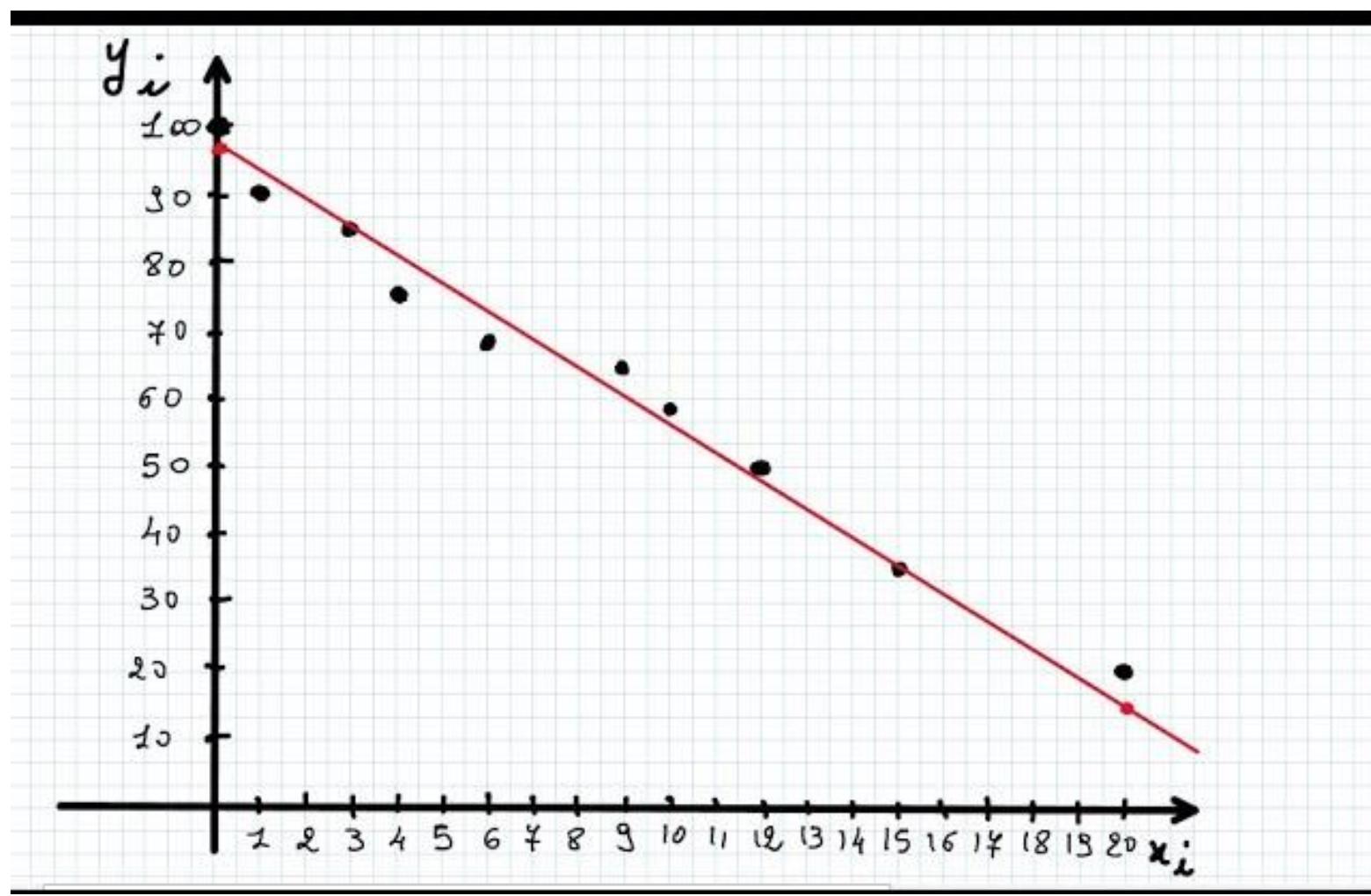
Consideriamo due punti su un grafico:

- il punto A con le coordinate (x_1, y_1)
- il punto B con le coordinate (x_2, y_2)

La **retta interpolante** collega questi due punti con una linea retta (cioè passa per entrambi i punti!).

Questa retta viene quindi utilizzata per stimare il valore di y (l'ordinata) per qualsiasi valore di x (l'ascissa) compreso tra x_1 e x_2 .





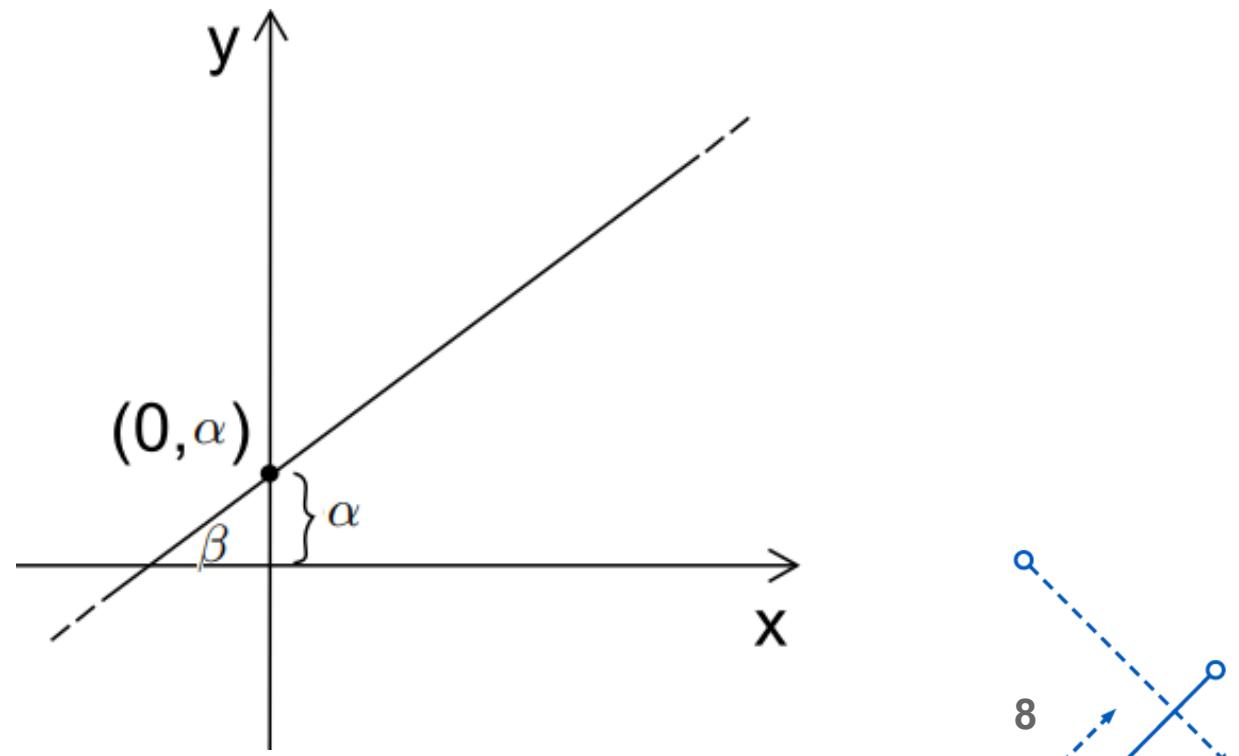
FONTE: Trenula63 (<https://www.youtube.com/watch?app=desktop&v=7M0gIYsKWm8>)

Regressione lineare semplice /1

Il **modello di regressione lineare** semplice è esprimibile attraverso l'**equazione di una retta** che riesce ad **interpolare la nuvola** di punti dello scatterplot **meglio di tutte e altre possibili rette**.

$$Y = \alpha + \beta X$$

Intercetta Coefficiente angolare



Regressione lineare semplice /2

Il **modello lineare** viene di solito utilizzato per spiegare, descrivere, o **anche prevedere un andamento futuro** sulla base della **relazione** che si instaura tra una variabile **Y**, chiamata **variabile dipendente**, e una o più altre variabili che assumono il significato di **variabili indipendenti X_1, X_2, \dots, X_p** .

DA TENERE A MENTE SEMPRE

Se $p = 1$, l'analisi prende il nome di **regressione semplice**

Se $p = 2, 3 \dots$ si parla di **regressione multipla**.

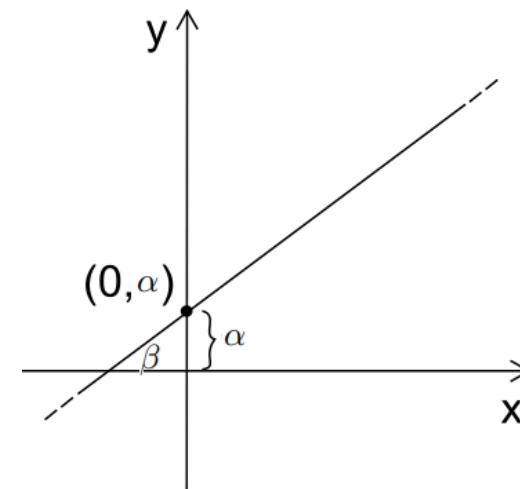
Per poter utilizzare un modello di regressione è **fondamentale individuare**

- variabili indipendenti
- variabile dipendente (se esiste)

Regressione Lineare Semplice /3

Per **applicare la regressione lineare semplice** è dunque necessario:

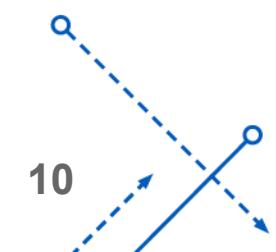
- Identificare la variabile dipendente e le indipendenti;
- Capire come calcolare α e β in modo tale che la **retta interpolante risultante** sia la retta che interpola nel miglior modo possibile la nuvola di punti noti.



Semplificato: la retta interpolante deve passare alla minor distanza possibile tra tutti i punti della nuvola.

$$Y = \alpha + \beta X$$

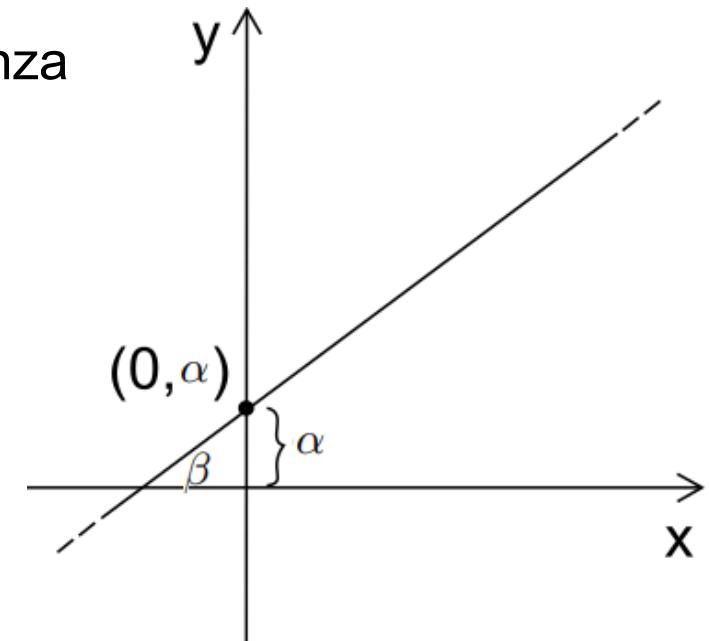
Nota: nel "miglior modo possibile" dovrebbe già farci pensare a **qualcosa da minimizzare... cosa?**



Come determiniamo α e β ?

Il coefficiente angolare β esprime **quantitativamente** la pendenza (inclinazione) della retta:

- un coefficiente angolare positivo ($\beta > 0$) indica una retta di regressione crescente;
- un coefficiente angolare negativo ($\beta < 0$) indica una retta decrescente;
- un coefficiente angolare nullo ($\beta = 0$) indica una retta orizzontale.



L'intercetta α corrisponde all'ordinata del punto di intersezione della retta interpolante (di regressione) con l'asse delle ordinate.

Statistica e analisi dei dati

Come determiniamo α e β ? /2

L'identificazione di questa retta viene ottenuta applicando il **metodo dei minimi quadrati**.

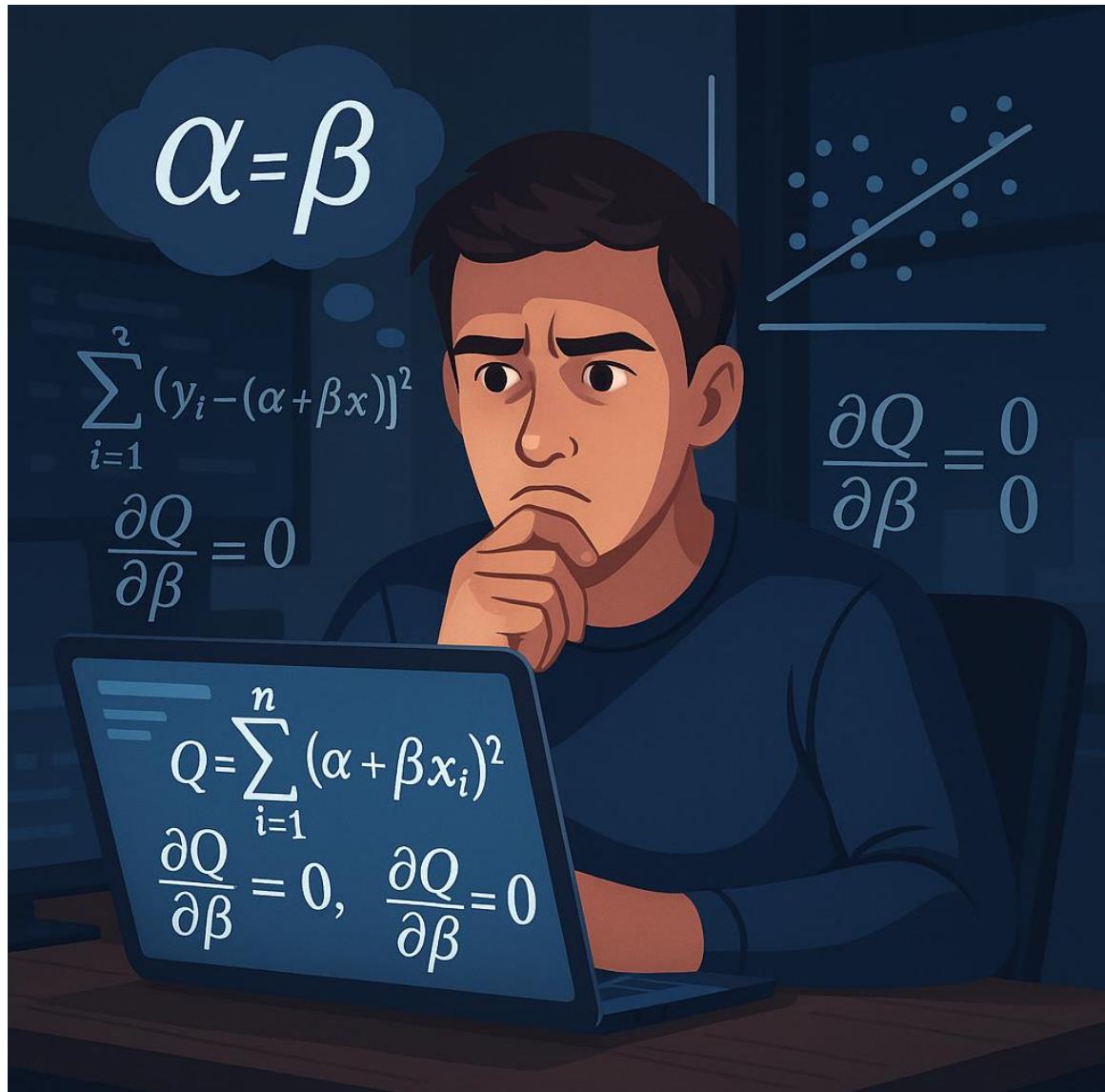
In particolare, i **coefficienti di regressione** sono i valori α e β per i quali la somma **Q** dei quadrati degli errori è minima.

$$Q = \sum_{i=1}^n \left[y_i - (\alpha + \beta x_i) \right]^2$$

n è il numero di osservazioni del nostro campione **C**

(y_1, y_2, \dots, y_n) sono i valori osservati della variabile Y

(x_1, x_2, \dots, x_n)
sono i valori osservati della v
ariabile X nel campione C



$$Q = \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2$$

Q è una **funzione di due parametri** (quindi il tasso di variazione di Q può dipendere da come ci muoviamo nello spazio (α, β));

Poiché Q dipende da due variabili usiamo **le derivate parziali**.

Statistica e analisi dei dati

Un altro santo uomo (Taylor – approssimazione di primo ordine) ci ha assicurato che:

$$Q(\alpha + \Delta\alpha, \beta + \Delta\beta) \approx Q(\alpha, \beta) + \frac{\partial Q}{\partial \alpha} \Delta\alpha + \frac{\partial Q}{\partial \beta} \Delta\beta.$$

Praticamente, ci ha dimostrato che, quando ci spostiamo di poco da (α , β) in un intorno:

il nuovo valore è “quasi” il vecchio valore più due piccoli aggiustamenti lineari, uno dovuto allo spostamento in α e uno allo spostamento in β , pesati dalle rispettive pendenze (derivate parziali).

ATTENZIONE: quel "circa" indica che anche qui introduciamo un errore ma è trascurabile.

Statistica e analisi dei dati

Che succede ora?

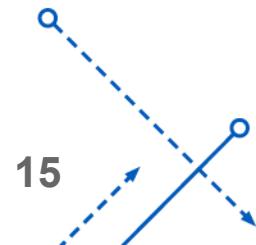
Se "imponiamo" che le due derivate parziali siano zero, i **termini lineari spariscono** e le variazioni piccole non cambiano Q (*al primo ordine*)

$$Q(\alpha + \Delta\alpha, \beta + \Delta\beta) \approx Q(\alpha, \beta) + \cancel{\frac{\partial Q}{\partial \alpha} \Delta\alpha} + \cancel{\frac{\partial Q}{\partial \beta} \Delta\beta}.$$

$$Q = \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2 \quad \rightarrow$$

$$\frac{\partial Q}{\partial \alpha} = -2 \sum_{i=1}^n [y_i - (\alpha + \beta x_i)] = 0,$$

$$\frac{\partial Q}{\partial \beta} = -2 \sum_{i=1}^n x_i [y_i - (\alpha + \beta x_i)] = 0,$$



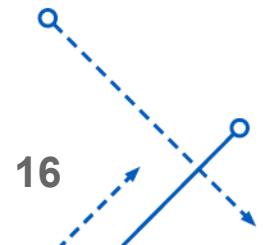
Statistica e analisi dei dati

Qualcuno tempo fa (*un santo*) ha definito e studiato **il gradiente**, che è il vettore delle derivate parziali con delle proprietà fantastiche. Se consideriamo un punto \mathbf{x} e una funzione f :

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_k} \right).$$

Il vettore gradiente calcolato in \mathbf{x} :

- Indica **la direzione di massima crescita** di f in quel punto.
- La sua norma è la **pendenza massima**;
- Ha altre proprietà che non ci interessano.



Statistica e analisi dei dati

Nel nostro caso abbiamo due variabili. Possiamo definire i termini seguenti:

$$\boldsymbol{\theta} = (\alpha, \beta) \quad \Delta\boldsymbol{\theta} = (\Delta\alpha, \Delta\beta).$$

Raggruppiamo i coefficienti dei due incrementi in un vettore:

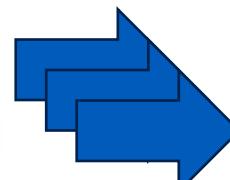
$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_k} \right) \rightarrow \nabla Q(\boldsymbol{\theta}) = \left(\frac{\partial Q}{\partial \alpha}, \frac{\partial Q}{\partial \beta} \right).$$

$$Q(\alpha + \Delta\alpha, \beta + \Delta\beta) \approx Q(\alpha, \beta) + \cancel{\frac{\partial Q}{\partial \alpha} \Delta\alpha} + \cancel{\frac{\partial Q}{\partial \beta} \Delta\beta}.$$

Annnullare il gradiente determina che Q sia minima

$$Q = \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2 \quad \rightarrow \quad \nabla Q(\theta) = \left(\frac{\partial Q}{\partial \alpha}, \frac{\partial Q}{\partial \beta} \right).$$

$$\begin{aligned} \frac{\partial Q}{\partial \alpha} &= -2 \sum_{i=1}^n [y_i - (\alpha + \beta x_i)] = 0, \\ \frac{\partial Q}{\partial \beta} &= -2 \sum_{i=1}^n x_i [y_i - (\alpha + \beta x_i)] = 0, \end{aligned}$$



$$\beta = \frac{s_y}{s_x} r_{xy},$$

$$\alpha = \bar{y} - \beta \bar{x}.$$

Osservazione importante

$$\beta = \frac{s_y}{s_x} r_{xy},$$

Le **medie campionarie**, le **deviazioni standard campionarie** e il **coefficiente di correlazione** permettono di stimare i parametri α e β della retta di regressione.

$$\alpha = \bar{y} - \beta \bar{x}.$$

Medie, deviazioni standard e correlazione **sintetizzano tutta l'informazione necessaria** per stimare α e β nella regressione lineare a una variabile.

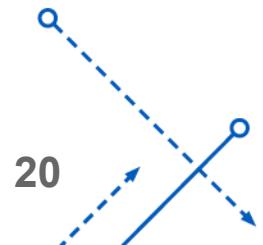
Quanto siamo bravi a "interpolare"?

Una volta calcolati i valori dei coefficienti α e β e disegnata la retta di regressione che interpola la nuvola dei punti nel corrispondente scatterplot, è possibile osservare quanto questa retta **si adatta ai punti** che individuano le osservazioni.

Tradotto: è possibile vedere graficamente di quanto sbagliamo ogni predizione (anche graficamente)

Denotiamo con $\hat{y}_i = \alpha + \beta x_i \quad (i = 1, 2, \dots, n)$ i valori stimati ottenuti mediante la retta di regressione.

sono posizionati sulla
retta di regressione



Residui /1

In generale, esisteranno degli scostamenti (residui) tra le ordinate dei punti y_i (valori osservati) e i corrispondenti valori stimati \hat{y}_i .

La media campionaria dei valori stimati $(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$ è uguale alla media campionaria \bar{y} delle osservazioni (y_1, y_2, \dots, y_n) .

$$\frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n (\alpha + \beta x_i) = \alpha + \beta \bar{x} = (\bar{y} - \beta \bar{x}) + \beta \bar{x} = \bar{y}.$$

I residui sono così definiti

$$E_i = y_i - \hat{y}_i = y_i - (\alpha + \beta x_i) \quad (i = 1, 2, \dots, n)$$

Residui /2

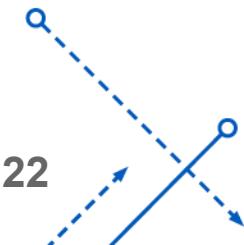
In generale, esisteranno degli scostamenti (residui) tra le ordinate dei punti y_i (valori osservati) e i corrispondenti valori stimati \hat{y}_i .

I residui mostrano di quanto si discostano i valori osservati y_i dai valori stimati \hat{y}_i con la retta di regressione.

La media campionaria dei residui E, negli regressori lineari con intercetta è sempre nulla, ossia in media gli scostamenti positivi e negativi si compensano. Infatti, risulta:

$$E_i = y_i - \hat{y}_i = y_i - (\alpha + \beta x_i) \quad (i = 1, 2, \dots, n)$$

$$\bar{E} = \frac{1}{n} \sum_{i=1}^n E_i = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) = \bar{y} - \frac{1}{n} \sum_{i=1}^n \hat{y}_i = 0.$$



Residui /2

La varianza campionaria dei residui `

$$s_E^2 = \frac{1}{n-1} \sum_{i=1}^n (E_i - \bar{E})^2 = \frac{1}{n-1} \sum_{i=1}^n E_i^2,$$

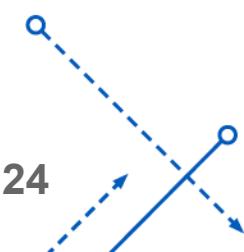
Coefficiente di determinazione

Poiché si è interessati a vedere quanto la retta si adatta ai dati, l'accento può essere posto sul quadrato del coefficiente di correlazione e su quanto esso si avvicini ad uno

E chiaro che un coeff. di Cc. tende a 1 tutti i punti tenderanno ad allinearsi lungo la retta di regressione, mentre se è prossimo a 0 **esprime una completa incapacità della retta di rappresentare la distribuzione dei dati considerati.**

Il coefficiente di determinazione per la regressione lineare semplice è il rapporto tra la varianza dei valori stimati tramite la retta di regressione e la varianza dei valori osservati.

$$D^2 = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$



Coefficiente di determinazione

Questo coefficiente ci dice **quanta parte della variabilità di Y è spiegata dalla relazione lineare con X.**

$D^2 = 0$: la retta non spiega nulla; **conviene usare la media di y per fare predizioni.**

$D^2 = 1$ adattamento perfetto ai dati.

$D^2 = 0,64$: il 64% della variabilità osservata in Y è spiegata dalla retta; il 36% resta nei residui.

È una **misura di forza dell'associazione lineare**, non di causalità.

