Lezione 10: Interpolazione lineare

Corso di Statistica
Facoltà di Economia
Università della Basilicata

Prof. Massimo Aria

- aria@unina.it

Il concetto di interpolazione

In matematica, e in particolare in analisi numerica, per interpolazione (**interpolazione per punti**) si intende un metodo per individuare nuovi punti del piano cartesiano a partire da un insieme finito di punti conosciuti, *nell'ipotesi che tutti i punti si possano riferire ad una funzione f(x)* di una data famiglia di funzioni di una variabile reale.

In altre parole, si cerca una funzione che sia in grado di interpolare (cioè "di passare") per tutti i punti disponibili in un piano cartesiano.

In statistica, il termine interpolazione assume un diverso significato.

Si parla di **interpolazione statistica** (o *interpolazione attraverso i punti*) quando si intende rappresentare in maniera sintetica una relazione funzionale tra due (o più) variabili statistiche attraverso una funzione f(x).

Questa funzione non ha l'obiettivo di passare per tutti i punti, ma di rappresentare "al meglio", anche se in via sintetica, la relazione esistente tra i due caratteri X e Y.

Obiettivo dell'interpolazione

Gli scopi per cui si cerca una funzione di interpolazione statistica 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.

Ruolo delle variabili

Si parla di interpolazione statistica quando è possibile riconoscere un *legame di causalità* tra le due variabili statistiche considerate.

Più chiaramente, nell'interpolazione si definiscono:

- **Variabile esplicativa** (indicata solitamente con X)

La variabile che, nel nesso logico di casualità, può essere considerata l'elemento "antecedente" della relazione.

- **Variabile dipendente** (indicata solitamente con Y)

La variabile che gioca invece il ruolo di "conseguente", cioè di variabile le cui variazioni dipendono dalla variabile esplicativa.

Esempi:

- Le precipitazioni in un bacino idrogeologico (X) e il livello del fiume che lo attraversa (Y);
- La velocità di percorrenza di un veicolo (X) e il consumo medio per percorrere un determinato tragitto (Y)
- Il reddito di una famiglia (X) e il livello del consumo della stessa (Y)
- ecc.

Interpolazione lineare

Tra le funzioni f(x) che normalmente vengono utilizzate nell'interpolazione statistica, quella lineare gioca un ruolo di primo piano.

Infatti per la semplicità di determinazione e di interpretazione, la funzione lineare rappresenta il legame funzionale a cui normalmente si fa riferimento nella descrizione di una variabile doppia.

Si parla quindi di **interpolazione statistica lineare** per indicare la scelta della funzione lineare come elemento interpolante.

Nel linguaggio comune lo stesso concetto di interpolazione statistica, se non accompagnato da altro aggettivo, sottintende una interpolazione di tipo lineare.

Funzione lineare

Con l'interpolazione lineare si intende quindi descrivere, in maniera sintetica, la relazione esistente tra due caratteri statistici attraverso una retta:

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

- dove *a* rappresenta l'intercetta, cioè il valore assunto dalla funzione quando x=0. Geometricamente è il punto in cui la retta interseca l'asse delle ordinate;
- dove *b* è invece il coefficiente angolare, che esprime la pendenza della retta in termini di variazione della funzione dovuta ad una variazione unitaria della X.

Criterio dei minimi quadrati

Dato un collettivo di N unità statistiche su cui sono state osservate le variabili X e Y,

L'interpolazione consiste nella determinazione della retta che "meglio interpola in senso statistico" la nube dei punti osservati.

Il criterio adottato per determinare la migliore retta prende il nome di **criterio dei minimi quadrati**.

Esso perviene alla *identificazione della coppia di parametri (a,b) la cui retta passa il più vicino possibile ai punti osservati.*

La vicinanza è valutata come differenza tra i valori osservati della Y e i valori teorici Y^* determinati con la funzione f(x).

$$y_l^* = f(x_l) \longrightarrow y_l^* = a + b \cdot x_l$$

$$scarto = \left(y_l - y_l^*\right) \text{ con } l = 1,2,\ldots,N$$

valore osservato — valore teorico

Definizione di scarto

Determinazione dei parametri della retta

I parametri vengono quindi determinati attraverso la minimizzazione della somma degli scarti al quadrato.

I valori (a, b), soluzione dei minimi quadrati, rappresentano i parametri della migliore retta secondo il criterio adottato.

$$\min_{a,b} S = \sum_{l=1}^{N} \left[y_l - y_l^*\right]^2$$

$$\min_{a,b} S = \sum_{l=1}^{N} \left[y_l - (a + bx_l)\right]^2$$

Risolvendo il sistema

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{Cod(XY)}{Dev(X)}$$

$$a = \mu_y - b\mu_x$$

Determinazione dei parametri della retta

Scomposizione della devianza

Identificata la retta di interpolazione, è possibile scomporre la variabilità della Y in due componenti:

- **Dev(S)** *Devianza Spiegata*

E' la somma delle differenze al quadrato tra i valori teorici della retta e la media di Y.

Essa esprime l'ammontare di variabilità della Y spiegata dalle variazioni della variabile esplicativa X.

- **Dev(E)** *Devianza Residua*

È la somma degli scarti al quadrato tra i valori osservati e teorici della Y.

Essa esprime l'ammontare di variabilità residua della Y non spiegata dalle variazioni della X.

Questa è la variabilità che non dipende dal legame lineare tra i due caratteri ma da fattori diversi (errori di misurazione, altre variabili che influenzano la Y, ecc.)

$$Dev(Y) = \sum_{i=1}^{N} (y_i - \mu_y)^2 =$$

$$= \sum_{i=1}^{N} (y_i - y_i^*)^2 + \sum_{i=1}^{N} (y_i^* - \mu_y)^2 + 2\left[\left(\sum y_i - \sum y_i^*\right)\ \left(\sum y_i^* - N\mu_y\right)\right]$$

Il doppio prodotto si dimostra essere nullo!!! $=0$

$$Dev(Y) = \sum_{i=1}^{N} (y_i - \mu_y)^2 = \sum_{i=1}^{N} (y_i^* - \mu_y)^2 + \sum_{i=1}^{N} (y_i - y_i^*)^2$$

$$Dev(Y) = Dev(S) + Dev(E)$$

Devianza Totale Dev(Y)	=	Devianza Spiegata Dev(S)	+	Devianza Residua Dev(E)

Scomposizione della devianza

Bontà di adattamento

Attraverso la scomposizione della devianza è possibile derivare un indice per valutare la bontà della sintesi ottenuta con l'interpolazione lineare.

L'**indice di bontà di adattamento R^2** (o *indice di determinazione lineare*) è ottenuto rapportando la devianza spiegata alla devianza totale.

Elevati valori della Dev(S), e quindi di R^2, indicano un buon adattamento in quanto larga parte della variabilità di Y è spiegata (linearmente) dalle variazioni della X.

Al contrario elevati valori della Dev(E), e quindi un R^2 prossimo a zero, indicheranno invece una scarsa bontà di adattamento della retta alla vera relazione esistente tra i caratteri X e Y.

L'indice R^2 è un numero puro che varia tra 0 e 1.

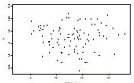
Permette di misurare la bontà di adattamento:

$$R^2 = \frac{Dev(S)}{Dev(Y)} = 1 - \frac{Dev(E)}{Dev(Y)} \qquad 0 \le R^2 \le 1$$

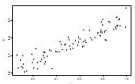
- Indica quanta parte della devianza di Y è spiegata dalla retta di interpolazione
- Dalla scomposizione di $Dev(Y)$ si ricava che:

Casi:

R^2 prossimo a 0 scarso adattamento



R^2 prossimo a 1 adattamento quasi perfetto



Indice di determinazione lineare

Indice di determinazione lineare e correlazione

Tra l'indice di determinazione lineare R^2 e il coefficiente di correlazione ρ esiste un interessante relazione:

$$\mathbf{R^2 = \rho^2}$$

"*l'indice di determinazione lineare è pari al quadrato del coefficiente di correlazione lineare*"

Questa relazione consente di misurare la bontà di adattamento senza dover scomporre la devianza ma unicamente calcolando il coefficiente di correlazione lineare.

Interpolazione e ruolo delle variabili

Il coefficiente **ρ** è un indice simmetrico, quindi il suo valore è costante a prescindere dal ruolo giocato da X e Y.

Viceversa, per ogni distribuzione doppia (X, Y) esistono due rette di interpolazione:

- la retta che spiega le variazioni di Y rispetto a X
- la retta che spiega le variazioni di X rispetto a Y

Ciò significa che invertendo il ruolo delle variabili cambieranno i valori dei parametri *a* e *b,* ma ovviamente rimarrà invariato il segno del coefficiente angolare (che è determinato dalla covarianza, cioè dalla correlazione tra X e Y)

Alcune formule abbreviate

Nella determinazione dei parametri della retta e del coefficiente di correlazione lineare è possibile utilizzare le formule abbreviate per la varianza e la covarainza.

La varianza è anche pari alla media quadratica (media delle x_I al quadrato) *meno la media al quadrato*.

La covarianza è anche pari alla media dei prodotti meno il prodotto delle medie.

Formule abbreviate per Varianza e Covarianza

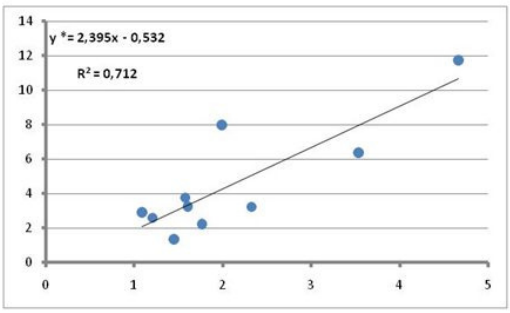
$$Var(X) = \sum_{I=1}^{N} (x_I - \mu_x)^2 = \frac{1}{N}\sum_{I=1}^{N} x_I^2 - \mu_x^2$$

$$Var(Y) = \sum_{I=1}^{N} (y_I - \mu_y)^2 = \frac{1}{N}\sum_{I=1}^{N} y_I^2 - \mu_y^2$$

$$Cov(X,Y) = \sum_{I=1}^{N} (x_I - \mu_x)\cdot(y_I - \mu_y) = \frac{1}{N}\sum_{I=1}^{N} x_I \cdot y_I - \mu_x \cdot \mu_y$$

Un esempio applicativo

Azienda	Fatturato (X)	Addetti (Y)	X^2	Y^2	XY
1	1,21	2,60	1,46	6,76	3,15
2	1,09	2,92	1,19	8,53	3,18
3	2,33	3,23	5,43	10,43	7,53
4	1,99	8,00	3,96	64,00	15,92
5	3,54	6,40	12,53	40,96	22,66
6	1,45	1,35	2,10	1,82	1,96
7	4,67	11,76	21,81	138,30	54,92
8	1,77	2,25	3,13	5,06	3,98
9	1,61	3,26	2,59	10,63	5,25
10	1,58	3,78	2,50	14,29	5,97
Totali	21,24	45,55	56,71	300,78	124,51



Media(X)=2,124; Media(Y)=4,555;

$Var(X) = \frac{1}{10} \cdot 56{,}71 - (2{,}1214)^2 = 1{,}1592$; $Var(Y) = \frac{1}{10} \cdot 300{,}78 - (4{,}555)^2 = 9{,}3298$

$Cov(X,Y) = \frac{1}{10} \cdot 124{,}51 - (2{,}124 \cdot 4{,}555) = 2{,}77627$

$b = \frac{Cov(X,Y)}{Var(X)} = \frac{2{,}77627}{1{,}1592} = 2{,}3950$ $a = \mu_y - b \cdot \mu_x = 4{,}555 - (2{,}3950 \cdot 2{,}124) = -0{,}5320$

$\rho = \frac{2{,}77627}{\sqrt{1{,}1592 \cdot 9{,}3298}} = 0{,}8442$ $R^2 = \rho^2 = 0{,}8442^2 = 0{,}712$

Esempio di interpolazione della distribuzione doppia "Fatturato - Addetti"

Interpretazione dei risultati

Nell'esempio precedente, i parametri della retta possono così essere interpretati:

La relazione lineare tra X e Y può essere espressa dalla funzione di sintesi **Y*=-0,5320+2,3950X**

b=2,3950 significa che ad un incremento unitario del fatturato (X) il numero di addetti di un'azienda cresce in media di 2,3950 unità.

a=-0, 5320 significa che per un fatturato nullo, le aziende hanno in media un numero di addetti pari a -0,5320. Ovviamente in questo caso, dove la variabile addetti è non può assumere valori negativi, l'interpretazione dell'intercetta perde di significato.

R^2= 0,712 evidenzia come il 71,2% della variabilità totale del numero di addetti (Y) sia spiegato dalle variazioni della variabile fatturato (X).

Si può concludere che vi è un buon adattamento della retta di interpolazione ai dati.

Nella prossima lezione

Nella prossima lezione si affronteranno i seguenti argomenti:

- rapporti statistici
- numeri indice