6

L'interpolazione statistica

- Interpolazione matematica e interpolazione statistica
- Interpolazione statistica col metodo dei minimi quadrati
- Serie storiche
- Trend ed extrapolazione

Osservazioni.
In questo capitolo ci si occupa dell'interpolazione statistica, cioè del metodo statistico che permette di determinare una funzione atta ad esprimere in termini analitici il legame che intercorre fra due serie di dati. La trattazione dell'argomento si chiude con l'analisi delle serie storiche e con lo studio dell'evoluzione temporale di un fenomeno statistico.

Interpolazione

1 Si parla di interpolazione quando:

- **note alcune coppie di dati (x; y), interpretabili come punti di un piano, ci si propone di costruire una funzione, detta funzione interpolante, che sia in grado di descrivere la relazione che intercorre fra l'insieme dei valori di x e l'insieme dei valori di y.**

La ricerca della funzione interpolante può essere vista in due modi diversi.

Primo modo

Ci troviamo in presenza di una distribuzione che riteniamo lacunosa, cioè mancante di alcuni dati che non possono essere più rilevati. In questo caso la funzione interpolante deve servire a stimare i dati mancanti. Una volta fatta questa stima la distribuzione viene completata con l'inserimento dei dati trovati.

Secondo modo

Ci troviamo in presenza di una distribuzione di dati alcuni dei quali vengono ritenuti affetti da errori. In questo caso la funzione interpolante deve servire a sostituire ai dati che si ritengono affetti da errori dati che si ritengono più attendibili, cioè corretti. In sostanza, alla distribuzione di partenza viene sostituita una distribuzione approssimata ma non affetta da errori.

Nel primo caso si parla di **interpolazione matematica** o **interpolazione per punti**; nel secondo di **interpolazione statistica** o **interpolazione fra punti**. Pertanto, possiamo fare la seguente distinzione:

- **si parla di interpolazione matematica o per punti quando, dovendo inserire fra alcuni dati noti uno o più dati mancanti, si vuole che la funzione interpolante riproduca esattamente i dati noti;**
- **si parla di interpolazione statistica o fra punti quando, dovendo rettificare i dati che si ritengono affetti da errori, si vuole che la funzione interpolante fornisca una distribuzione che, rispetto a quella di cui si dispone, sia non affetta da errori (più regolare) anche se approssimata.**

Naturalmente, in entrambi i casi occorre, anzitutto, decidere il tipo di funzione che si vuole adottare: funzione lineare, funzione quadratica (o parabolica), funzione esponenziale, ecc.
Orbene, posto quanto sopra, precisiamo subito che nel corso di questo capitolo ci occupiamo della sola **interpolazione statistica**. Ciò in quanto l'interpolazione matematica costituisce oggetto di studio di una diversa parte del corso.

Interpolazione statistica: metodo dei minimi quadrati

2 Consideriamo un fenomeno statistico per il quale si dispone della seguente distribuzione di dati:

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

e supponiamo che, rappresentando tali dati in un sistema di assi cartesiani, si ottenga il diagramma a dispersione di cui alla figura 1 oppure 2.

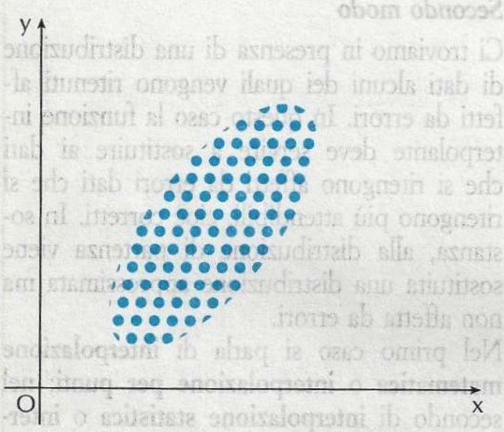

Fig. 1

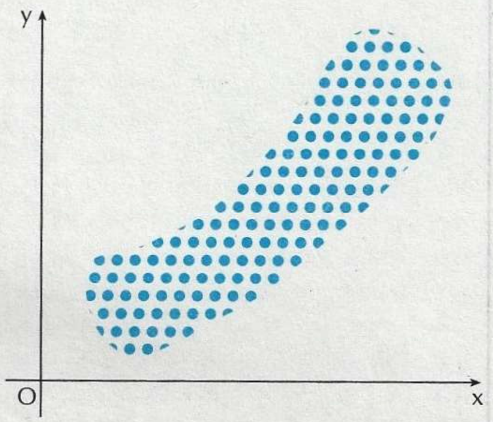

Fig. 2

Poiché, come già detto, riteniamo che i dati osservati sono affetti da errori vogliamo costruire una funzione che ci permetta di sostituire ai dati y_i osservati i dati $\overline{y}_i$ interpolati, più regolari anche se approssimati. In questo caso, il problema che si presenta è quello di scegliere la funzione interpolante in modo che l'approssimazione presenti un «buon accostamento».

Scelta del tipo di funzione interpolante

3 In linea generale, l'adozione di uno piuttosto che di un altro tipo di funzione può essere suggerita dall'esame del diagramma a dispersione. Per esempio, se il diagramma a dispersione si presenta come mostra la figura 1, sembra logica l'adozione di una funzione lineare: si nota,

infatti, una certa tendenza alla disposizione dei dati osservati intorno a una retta (figura 3). Se, invece, il diagramma a dispersione si presenta come mostra la figura 2, allora sembra più logica l'adozione di una funzione esponenziale: si nota, infatti, una certa tendenza alla disposizione dei dati osservati intorno a una curva esponenziale (figura 4).
Da notare che, in concreto, viene spesso adottata una funzione lineare: ciò dipende principalmente dalla sua semplicità e dalla facilità di manipolazione che ne consegue.

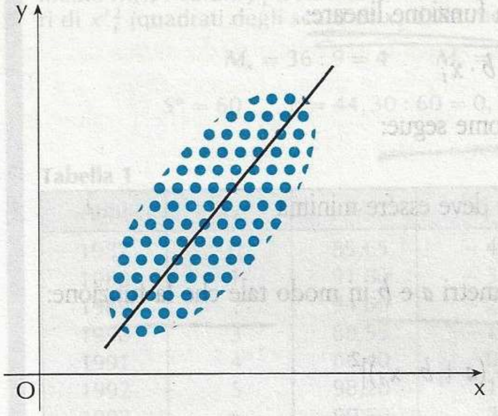

Fig. 3

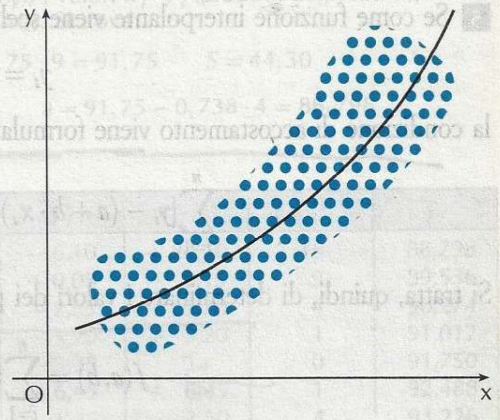

Fig. 4

Condizione per un buon accostamento

4 Ci chiediamo:

- come bisogna procedere perché, una volta scelto il tipo di funzione interpolante, possa ottenersi un buon accostamento fra la distribuzione dei valori osservati e quella dei valori teorici ottenuti interpolando?

La risposta è la seguente:

- consideriamo i valori y_i osservati e quelli teorici $\overline{y}_i$ forniti dalla funzione interpolante. Poiché gli errori commessi sostituendo a ciascun y_i il corrispondente $\overline{y}_i$ sono dati dalle differenze:

$$y_i - \overline{y}_i$$

al fine di ottenere un buon accostamento occorre minimizzare questi errori.

Naturalmente, per raggiungere l'obiettivo voluto non possiamo prendere in considerazione la somma delle differenze: essendo alcune differenze positive e altre negative, o nulle, esse potrebbero anche compensarsi. Allora, prendiamo in considerazione la somma dei quadrati delle differenze e poniamo come condizione di accostamento la seguente:

$$\sum_{i=1}^{n}(y_i - \overline{y}_i)^2 = \text{deve essere minima}$$

In questo caso parliamo di interpolazione statistica con il metodo dei minimi quadrati.
Quindi, in definitiva:

- parliamo di interpolazione statistica col metodo dei minimi quadrati quando, qualunque sia il tipo di funzione interpolante adottato, la condizione di accostamento fra valori osservati e valori teorici ottenuti interpolando viene fissata nel senso che deve essere minima la somma dei quadrati delle differenze fra valori osservati e valori teorici.

Osservazione. Precisiamo che, spesso, nelle coppie $(x; y)$ i valori di x indicano i tempi (le date) in cui vengono rilevati i valori di y. In tal caso si parla di evoluzione temporale di un fenomeno statistico. Le coppie $(x; y)$ mostrano come il fenomeno si evolve nel tempo.

Metodo dei minimi quadrati: funzione lineare

5 Se come funzione interpolante viene scelta la funzione lineare:

$$y_i = a + b \cdot x_i$$

la condizione di accostamento viene formulata come segue:

$$\sum_{i=1}^{n} [y_i - (a + b \cdot x_i)]^2 = \text{deve essere minima}$$

Si tratta, quindi, di determinare i valori dei parametri a e b in modo tale che la funzione:

$$f(a, b) = \sum_{i=1}^{n} [y_i - (a + b \cdot x_i)]^2$$

sia minima. In sostanza, si tratta di minimizzare una funzione di due variabili. Rinviando per un approfondimento al paragrafo 13 ci limitiamo qui ad affermare che i valori di a e di b che rendono minima la funzione $f(a, b)$ sono i seguenti:

$$\boxed{b = \frac{S}{S^*}} \quad (1)$$

$$\boxed{a = M_y - b \cdot M_x} \quad (2)$$

essendo:

- M_x la media dei valori x_i;
- M_y la media dei valori y_i;
- S la somma dei prodotti degli scarti di x_i da M_x e di y_i da M_y. Cioè:

$$S = \sum_{i=1}^{n} (x_i - M_x)(y_i - M_y)$$

che, ponendo:

$$x_i - M_x = x'_i \qquad y_i - M_y = y'_i$$

si scrive:

$$S = \sum_{i=1}^{n} x'_i \, y'_i$$

- S^* la somma dei quadrati degli scarti di x_i dalla media M_x. Cioè:

$$S^* = \sum_{i=1}^{n} (x_i - M_x)^2 = \sum_{i=1}^{n} x'^2_i$$

ESEMPIO

Consideriamo i dati della tabella 1: produzione di frumento (in milioni di quintali) per gli anni dal 1987 al 1995. Come si può vedere, i valori x_i sono gli anni che, per semplicità, vengono contati assumendo il 1987 come anno zero per cui il 1988 è l'anno 1, il 1989 è l'anno 2, il 1990 è l'anno 3, ..., il 1995 è l'anno 8. I valori y_i sono relativi alla quantità di frumento prodotta nei singoli anni.
Oltre ai valori x_i e y_i, sono successivamente indicati nella tabella i valori x'_i (scarti di x_i dalla media M_x), i valori y'_i (scarti di y_i dalla media M_y), i valori $x'_i \cdot y'_i$ (prodotti degli scarti), i valori di x'^2_i (quadrati degli scarti di x_i dalla media M_x). Si trova:

$$M_x = 36 : 9 = 4 \qquad M_y = 825,75 : 9 = 91,75 \qquad S = 44,30$$

$$S^* = 60 \qquad b = 44,30 : 60 = 0,738 \qquad a = 91,75 - 0,738 \cdot 4 = 88,798$$

Tabella 1

Anni	x_i	y_i	x'_i	y'_i	$x'_i \cdot y'_i$	x'^2_i	$\overline{y}_i$
1987	0	85,65	− 4	− 6,10	+ 24,10	16	88,798
1988	1	91,80	− 3	+ 0,05	− 0,15	9	89,536
1989	2	94,60	− 2	+ 2,85	− 5,70	4	90,274
1990	3	88,55	− 1	− 3,20	+ 3,20	1	91,012
1991	4	86,40	0	− 5,35	0	0	91,750
1992	5	98,20	1	+ 6,45	+ 6,45	1	92,488
1993	6	90,50	2	− 1,25	− 2,50	4	93,226
1994	7	99,35	3	+ 7,60	+ 22,80	9	93,964
1995	8	90,70	4	− 1,05	− 4,20	16	94,702
	36	**825,75**	**0**	**0**	**44,30**	**60**	**825,750**

Allora, la retta interpolante è la seguente:

$$y_i = 0,738 \cdot x_i + 88,798$$

A questo punto, dando a x_i i valori 0, 1, 2, ..., 8 si ottengono i valori teorici $\overline{y}_i$ che sono indicati nell'ultima colonna.
Rappresentando i valori teorici (sulla retta) e quelli osservati (dispersi intorno alla retta), si ottiene il diagramma della figura 5 il cui carattere è puramente indicativo.

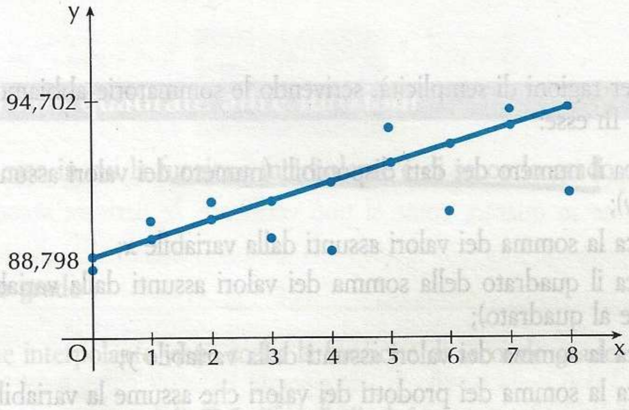

Fig. 5

Alcune puntualizzazioni

6 È il caso di fare le seguenti puntualizzazioni:

- La retta che si ottiene dando ai parametri a e b i valori calcolati in base alle (1) e (2) soddisfa la condizione di minimo posta (paragrafo 5). Ciò vuol dire che tale retta è, in base alla condizione d'accostamento adottata, la migliore funzione interpolante di tipo lineare.

Interpolando con qualsiasi altra retta l'accostamento sarebbe meno buono.

- La somma dei valori osservati e quella dei valori teorici sono uguali:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \bar{y}_i$$

Da ciò deriva che valori osservati e valori teorici hanno la stessa media aritmetica.

- Il coefficiente angolare b della retta interpolante indica la variazione media annua del fenomeno: di anno in anno il valore iniziale $y_0 = a$ aumenta in misura costante uguale a b. Riprendendo l'esempio precedente e osservando i valori teorici si riscontra che il valore iniziale 88,798 aumenta ogni anno in misura costante pari a 0,738 (quintali).

Metodo semplificato

7 L'applicazione del metodo dei minimi quadrati può essere semplificato esprimendo i parametri a e b, definiti rispettivamente dalla (1) e dalla (2), in modo diverso e ciò come segue:

$$a = \frac{\sum y_i \cdot \sum x_i^2 - \sum x_i \cdot \sum (x_i \cdot y_i)}{n \cdot \sum x_i^2 - (\sum x_i)^2} \tag{3}$$

$$b = \frac{n \cdot \sum (x_i \cdot y_i) - \sum x_i \cdot \sum y_i}{n \cdot \sum x_i^2 - (\sum x_i)^2} \tag{4}$$

Come si vede, per ragioni di semplicità, scrivendo le sommatorie abbiamo omesso l'indicazione degli estremi. In esse:

n	indica il numero dei dati disponibili (numero dei valori assunti dalle due variabili x e y);
$\sum x_i$	indica la somma dei valori assunti dalla variabile x;
$(\sum x_i)^2$	indica il quadrato della somma dei valori assunti dalla variabile x (valore precedente al quadrato);
$\sum y_i$	indica la somma dei valori assunti dalla variabile y;
$\sum (x_i \cdot y_i)$	indica la somma dei prodotti dei valori che assume la variabile x per i corrispondenti valori che assume la variabile y;
$\sum x_i^2$	indica la somma dei quadrati dei valori che assume la variabile x.

La semplificazione del calcolo sta nel fatto che, usando le (3) e (4), si evita di calcolare gli scarti e il prodotto degli scarti.
Tralasciando di dimostrare come la (1) e la (2) vengono trasformate rispettivamente nelle (3) e (4), consideriamo l'esempio che segue.

ESEMPIO

Riprendiamo l'esempio della tabella 1 disponendo i calcoli che interessano come mostra la tabella 2.

Tabella 2

Anni	x_i	y_i	x_i^2	$x_i \cdot y_i$
1987	0	85,65	0	0
1988	1	91,80	1	91,80
1989	2	94,60	4	189,20
1990	3	88,55	9	265,65
1991	4	86,40	16	345,60
1992	5	98,20	25	491,00
1993	6	90,50	36	543,00
1994	7	99,35	49	695,45
1995	8	90,70	64	725,60
	36	**825,75**	**204**	**3.347,30**

Si ottiene:

$$n = 9 \qquad \sum x_i = 36 \qquad (\sum x_i)^2 = 36^2 = 1.296$$

$$\sum y_i = 825,75 \qquad \sum (x_i \cdot y_i) = 3.347,30 \qquad \sum x_i^2 = 204$$

Quindi, applicando la (3) e la (4), si ritrova:

$$a = \frac{825,75 \cdot 204 - 36 \cdot 3.347,30}{9 \cdot 204 - 1.296} = 88,798$$

$$b = \frac{9 \cdot 3.347,30 - 36 \cdot 825,75}{9 \cdot 204 - 1.296} = 0,738$$

e l'equazione della retta interpolante è ancora:

$$y_i = 0,738 \cdot x_i + 88,798$$

Metodo dei minimi quadrati: altre funzioni

8 Consideriamo il caso in cui la funzione interpolante è di secondo grado oppure esponenziale.

Funzione di secondo grado

9 Se come funzione interpolante viene scelta la funzione di secondo grado:

$$y_i = a \cdot x_i^2 + b \cdot x_i + c$$

occorre determinare i parametri a, b e c. Dalla condizione di accostamento si ottiene una funzione delle tre variabili a, b, c. Il fatto è che, non soltanto minimizzare tale funzione di tre variabili è molto complicato, ma che anche la determinazione dei valori numerici di a, b e c risulta piuttosto laboriosa. Stando a ciò, ci limitiamo a considerare la più semplice ipotesi in cui:

$$y_i = a \cdot x_i^2$$

In questo caso, la condizione di accostamento viene formulata come segue:

$$\sum_{i=1}^{n}(y_i - a \cdot x_i^2)^2 = \text{deve essere minima}$$

e si tratta di determinare l'unico parametro a in modo tale che la funzione:

$$f(a) = \sum_{i=1}^{n}(y_i - a \cdot x_i^2)^2$$

sia minima. Questa volta ci troviamo in presenza di una funzione della sola variabile a e, quindi, sappiamo risolvere il problema: basta trovare la derivata rispetto ad a e porla uguale a zero. Applicando la regola di derivazione di una funzione di funzione si ha:

$$\frac{d}{da}\sum_{i=1}^{n}(y_i - a \cdot x_i^2)^2 = 2 \cdot \sum_{i=1}^{n}(y_i - a \cdot x_i^2) \cdot \frac{d}{da}(y_i - a \cdot x_i^2) =$$

$$= 2 \cdot \sum_{i=1}^{n}(y_i - a \cdot x_i^2) \cdot (-x_i^2) = -2 \cdot \sum_{i=1}^{n}(y_i - a \cdot x_i^2) \cdot x_i^2$$

e, quindi, ponendo uguale a zero:

$$-2 \cdot \sum_{i=1}^{n}(y_i - a \cdot x_i^2) \cdot x_i^2 = 0$$

Dividendo ambo i membri per -2 e svolgendo i calcoli si ottiene:

$$\sum_{i=1}^{n}(y_i - a \cdot x_i^2) \cdot x_i^2 = 0 \quad \text{cioè} \quad \sum_{i=1}^{n} x_i^2 \cdot y_i = a \cdot \sum_{i=1}^{n} x_i^4$$

Infine, si ricava:

$$\boxed{a = \frac{\sum x_i^2 \cdot y_i}{\sum x_i^4}}$$

avendo omesso per semplicità gli indici della sommatoria. Il valore di a così trovato rende minima la funzione in quanto, come si può verificare, la derivata seconda della funzione è positiva.

ESEMPIO

Consideriamo i dati della tabella 3. Disponendo i calcoli come mostrano la terza, quarta e quinta colonna si trova:

$$\sum x_i^2 \cdot y_i = 2.371,4 \qquad \sum x_i^4 = 979$$

per cui:

$$a = 2.371,4 : 979 = 2,42$$

Allora, la funzione di secondo grado interpolante è:

$$y_i = 2,42 \cdot x_i^2$$

e dando a x_i i valori 1, 2, 3, 4, 5 si ottengono i valori teorici $\overline{y}_i$ indicati nell'ultima colonna.

Tabella 3

x_i	y_i	x_i^2	$x_i^2 \cdot y_i$	x_i^4	$\bar{y}_i$
1	2,3	1	2,3	1	2,42
2	9,5	4	38,0	16	9,68
3	21,2	9	190,8	81	21,78
4	38,3	16	612,8	256	38,72
5	61,1	25	1.527,5	625	60,50
			2.371,4	**979**	

Funzione esponenziale

10 Se come funzione interpolante viene scelta una funzione esponenziale:

$$y_i = a \cdot b^{x_i} \quad \text{con} \quad a > 0 \quad b > 0 \quad e \quad b \neq 1 \tag{1}$$

la condizione di accostamento viene formulata come segue:

$$\sum_{i=1}^{n} (y_i - a \cdot b^{x_i})^2 = \text{deve essere minima}$$

Anche ora, come nel caso di funzione lineare, occorre determinare i due parametri a e b. In più, le cose vengono complicate dal fatto che la manipolazione della funzione esponenziale comporta qualche difficoltà. Stando a ciò, per determinare i due parametri conviene ricorrere a un artificio: trasformiamo la (1) prendendo il logaritmo (decimale) di ambo i membri. In questo modo otteniamo:

$$Log\ y_i = Log\ a + x_i \cdot Log\ b \tag{2}$$

Come si vede, la relazione esponenziale viene trasformata in una relazione lineare. Allora, posto:

$$Log\ a = n \qquad Log\ b = m \qquad Log\ y_i = z_i$$

scriviamo la (2) come segue:

$$z_i = n + m \cdot x_i \tag{3}$$

A questo punto:

- applichiamo il metodo dei minimi quadrati, nel caso di funzione lineare, usando la distribuzione di x_i e dei corrispondenti $z_i = Log\ y_i$;
- a questo scopo usiamo la (3) e la (4) del paragrafo 7. Esse diventano:

$$n = \frac{\sum Log\ y_i \cdot \sum x_i^2 - \sum x_i \cdot \sum (x_i \cdot Log\ y_i)}{n \cdot \sum x_i^2 - (\sum x_i)^2}$$

$$m = \frac{n \cdot \sum (x_i \cdot Log\ y_i) - \sum x_i \cdot \sum Log\ y_i}{n \cdot \sum x_i^2 - (\sum x_i)^2}$$

Trovati i valori di n ed m li sostituiamo nella (3);

- ciò che interessa è, però, la funzione esponenziale. Per scrivere l'espressione di quest'ultima, da:

$$Log\ b = m \quad \text{e} \quad Log\ a = n$$

ricaviamo:

$$b = 10^m \quad \text{e} \quad a = 10^n$$

che sostituiamo nella (1).

ESEMPIO

Consideriamo i dati della tabella 4 disponendo i calcoli della terza, quarta e quinta colonna.

Tabella 4

x_i	y_i	$Log\ y_i$	x_i^2	$x_i \cdot Log\ y_i$	$\overline{y}_i$
1	45,4	1,65706	1	1,65706	45,67468
2	47,2	1,67394	4	3,34788	47,84879
3	51,5	1,71181	9	5,13543	50,12640
4	53,8	1,73078	16	6,92312	52,51241
5	52,6	1,72099	25	8,60495	55,01200
6	58,4	1,76641	36	10,59846	57,63058
21		**10,26099**	**91**	**36,26690**	

Si ha:

$$n = 6 \qquad \sum x_i = 21 \qquad \left(\sum x_i\right)^2 = 21^2 = 441$$

$$\sum Log\ y_i = 10,26099 \qquad \sum (x_i \cdot Log\ y_i) = 36,26690 \qquad \sum x_i^2 = 91$$

e, quindi:

$$n = \frac{10,26099 \cdot 91 - 21 \cdot 36,26690}{6 \cdot 91 - 441} = 1,63948$$

$$m = \frac{6 \cdot 36,26690 - 21 \cdot 10,26099}{6 \cdot 91 - 441} = 0,02020$$

Allora, la retta interpolante, cioè la (3), si scrive come segue:

$$z_i = 1,63948 + 0,02020 \cdot x_i$$

D'altra parte, da:

$$Log\ a = 1,63948 \quad si\ ha \quad a = 10^{1,63948} = 43,59935$$

$$Log\ b = 0,02020 \quad \text{si ha} \quad b = 10^{0,02020} = 1,04760$$

e, quindi, la funzione esponenziale interpolante è:

$$y_i = 43,59935 \cdot 1,04760^{x_i}$$

Da questa, dando a x_i i valori 1, 2, ..., 6 si trovano i valori teorici $\overline{y}_i$ che sono indicati nell'ultima colonna.

Stima del grado di accostamento

11 Per stimare il grado di accostamento fra valori osservati e valori teorici si usa, di solito, un indice detto errore standard. Precisamente:

• **l'errore standard è la media quadratica delle differenze fra valori osservati e valori teorici.**

Cioè:

$$E_s = \sqrt{\frac{\sum(y_i - \bar{y}_i)^2}{n}}$$

Da notare che:

• se i valori teorici coincidono esattamente con quelli osservati tutte le differenze $y_y - \bar{y}_i$ (1 = 1, 2, ..., n) sono nulle. Quindi, si ha:

$$\sum(y_i - \bar{y}_i)^2 = 0 \quad \text{e} \quad E_s = 0$$

• in qualsiasi altro caso si ha:

$$\sum(y_i - \bar{y}_i)^2 > 0 \quad \text{e} \quad E_s > 0$$

In definitiva, possiamo dire che:

• **migliore è l'accostamento fra valori osservati e teorici più piccolo è il valore (positivo) dell'errore standard. In particolare l'errore standard vale zero quando l'accostamento è perfetto.**

ESEMPI

1 Riprendiamo la distribuzione della tabella (1) e consideriamo le differenze fra valori osservati e valori teorici. Tali differenze, nonché i rispettivi quadrati, sono indicati nella tabella 5.

Tabella 5

x_i	y_i	$\bar{y}_i$	$y_i - \bar{y}_i$	$(y_i - \bar{y}_i)^2$
0	85,65	88,798	− 3,148	9,9099
1	91,80	89,536	+ 2,264	5,1257
2	94,60	90,274	+ 4,326	18,7143
3	88,55	91,012	− 2,462	6,0614
4	86,40	91,750	− 5,350	28,6225
5	98,20	92,488	+ 5,712	32,6269
6	90,50	93,226	− 2,726	7,4311
7	99,35	93,964	+ 5,386	29,0090
8	90,70	94,702	− 4,002	16,0160
				153,5168

Si ottiene:

$$E_s = \sqrt{153,5168 : 9} = 4,1301$$

Come si può vedere, siamo lontani da un perfetto accostamento. Anzi, possiamo dire che l'accostamento lascia a desiderare.

2 Riprendiamo la distribuzione della tabella 3 e consideriamo anche ora le differenze al quadrato fra valori osservati e valori teorici: tabella 6.

Tabella 6

x_i	y_i	$\overline{y}_i$	$y_i - \overline{y}_i$	$(y_i - \overline{y}_i)^2$
1	2,3	2,42	− 0,12	0,0144
2	9,5	9,68	− 0,18	0,0324
3	21,2	21,78	− 0,58	0,3364
4	38,3	38,72	− 0,42	0,1764
5	61,1	60,50	+ 0,60	0,3600
				0,9196

Si trova:

$$E_s = \sqrt{0,9196 : 5} = 0,4289$$

che si può ritenere indicativo di un buon accostamento.

Come varia un fenomeno statistico: serie storiche, trend ed extrapolazione

12 La tabella 1, ma anche la 3 e la 4, mostrano come varia un fenomeno statistico in tempi x_i successivi. In particolare, la tabella 1 mostra come varia la produzione annuale di frumento negli anni dal 1987 al 1995. Ebbene, le tabelle considerate sono esempi di serie storiche. In generale:

- **si dice serie storica una raccolta di rilevazioni eseguite in tempi successivi.**

In presenza di una serie storica è molto importante potere cogliere quelle regolarità, se esistono, che ne determinano il **comportamento tendenziale** o **trend**. A ciò mira, in sostanza, l'interpolazione statistica. Una volta individuato il trend è poi possibile formulare previsioni sul futuro andamento del fenomeno. Infatti, la conoscenza di una funzione interpolante soddisfacente fornisce la possibilità di stimare, in modo meccanico, i valori del trend al di là dell'intervallo temporale osservato. Si parla in tal senso di extrapolazione della serie storica.
Occorre precisare subito che la possibilità di previsione fornita dall'extrapolazione di una serie storica è soggetta a condizioni logiche ben precise. Occorre quantomeno che:

- la funzione interpolante scelta per rappresentare il fenomeno nel periodo passato di osservazioni sia «valida»: l'errore standard deve assumere valori prossimi a zero;
- le condizioni che hanno determinato l'evoluzione del fenomeno osservato nel periodo passato permangono anche nel futuro.

In particolare, questa seconda condizione può manifestarsi poco realistica nel caso di fenomeni economici e ciò tanto più quanto più lungo è l'orizzonte temporale sul quale viene estesa la proiezione.
Stando a quanto detto possiamo trarre la seguente:

Conclusione. **Più che fornire una previsione, i valori futuri stimati con l'extrapolazione di una serie storica forniscono una proiezione sul futuro. Tali valori debbono essere visti non come valutazioni di ciò che accadrà ma come valutazioni di ciò che dovrebbe accadere nel caso in cui dovessero sussistere tutti quegli elementi che hanno determinato l'evoluzione passata del fenomeno. Da ciò deriva l'opportunità di aggiornare periodicamente la funzione che esprime il trend rideterminando i parametri che la caratterizzano in modo da potere riformulare proiezioni basate su valori storici sempre recenti.**

Chiariamo quanto detto considerando l'esempio che segue.

ESEMPIO

La serie storica della tabella 7 riguarda l'andamento della mortalità infantile (morti per 1.000 nati vivi) in Italia negli anni dal 1971 al 1981. Con riferimento a tale serie storica:

Tabella 7

Anni	Indici mortalità
1971	28,5
1972	27,0
1973	26,2
1974	22,9
1975	21,2
1976	19,5
1977	18,1
1978	17,1
1979	15,7
1980	14,6
1981	14,1

- costruiamo la funzione che esprime il trend lineare;
- costruiamo la funzione che esprime il trend esponenziale;
- confrontiamo i valori teorici che si ottengono nei due casi misurando il grado di accostamento;
- extrapoliamo la serie storica per gli anni dal 1982 al 1986 confrontando le proiezioni con i valori effettivi rilevati successivamente al 1981 (per gli anni dal 1982 al 1986);
- infine, procediamo all'aggiornamento delle funzioni di trend utilizzando i dati effettivi relativi agli anni dal 1977 al 1986.

Funzione di trend lineare - Disponiamo i calcoli come mostra la tabella 8. Si ha:

$$n = 11 \qquad \sum x_i = 55 \qquad \left(\sum x_i\right)^2 = 55^2 = 3.025$$

$$\sum y_i = 224,9 \qquad \sum (x_i \cdot y_i) = 956,7 \qquad \sum x_i^2 = 385$$

$$a = \frac{224,9 \cdot 385 - 55 \cdot 956,7}{11 \cdot 385 - 3.025} = 28,07273$$

$$b = \frac{11 \cdot 956,7 - 55 \cdot 224,9}{11 \cdot 385 - 3.025} = -1,52545$$

$$y_i = -1,52545 \cdot x_i + 28,07273$$

Tabella 8 - Ipotesi di trend lineare

Anni	x_i	y_i	x_i^2	$x_i \cdot y_y$	$\overline{y}_i$	$(y_i - \overline{y}_i)^2$
1971	0	28,5	0	0	28,07273	0,18256
1972	1	27,0	1	27,0	26,54728	0,20496
1973	2	26,2	4	52,4	25,02183	1,38808
1974	3	22,9	9	68,7	23,49638	0,35567
1975	4	21,2	16	84,8	21,97093	0,59433
1976	5	19,5	25	97,5	20,44548	0,89393
1977	6	18,1	36	108,6	18,92003	0,67245
1978	7	17,1	49	119,7	17,39458	0,08678
1979	8	15,7	64	125,6	15,86913	0,02860
1980	9	14,6	81	131,4	14,34368	0,06570
1981	10	14,1	100	141,0	12,81823	1,64293
	55	**224,9**	**385**	**956,7**		**6,11599**

I valori teorici sono indicati nella penultima colonna mentre nell'ultima colonna sono indicati i quadrati delle differenze fra valori teorici e valori effettivi. L'errore standard che fornisce il grado di accostamento è:

$$E_s = \sqrt{6,11599 : 11} = 0,74565$$

Funzione di trend esponenziale - Disponiamo i calcoli come mostra la tabella 9. Si ha:

$$n = 11 \qquad \sum x_i = 55 \qquad \sum (x_i)^2 = 55^2 = 3.025$$

$$\sum Log\ y_i = 14,28086 \qquad \sum (x_i \cdot Log\ y_i) = 67,81862 \qquad \sum x_i^2 = 385$$

$$Log\ a = \frac{14,28086 \cdot 385 - 55 \cdot 67,81862}{11 \cdot 385 - 3.025} = 1,46125$$

$$Log\ b = \frac{11 \cdot 67,81862 - 55 \cdot 14,28086}{11 \cdot 385 - 3.025} = -0,03260$$

$$a = 10^{1,46125} = 28,92344 \qquad b = 10^{-0,03260} = 0,92768$$

$$y_i = 28,92344 \cdot 0,92768^{x_i}$$

Tabella 9 - Ipotesi di trend esponenziale

Anni	x_i	y_i	$Log\ y_i$	x_i^2	$x_i \cdot Log\ y_i$	$\overline{y}_i$	$(y_i - \overline{y}_i)^2$
1971	0	28,5	1,45484	0	0	28,92344	0,17930
1972	1	27,0	1,43136	1	1,43136	26,83170	0,02832
1973	2	26,2	1,41830	4	2,83660	24,89123	1,71288
1974	3	22,9	1,35984	9	4,07952	23,09109	0,03652
1975	4	21,2	1,32634	16	5,30536	21,42115	0,04891
1976	5	19,5	1,29003	25	6,45015	19,87197	0,13836
1977	6	18,1	1,25768	36	7,54608	18,43483	0,11211
1978	7	17,1	1,23300	49	8,63100	17,10162	0,00000
1979	8	15,7	1,19590	64	9,56720	15,86483	0,02717
1980	9	14,6	1,16435	81	10,47915	14,71749	0,01380
1981	10	14,1	1,14922	100	11,49220	13,65312	0,19970
	55		14,28086	385	67,81862		2,49707

Anche ora i valori teorici sono indicati nella penultima colonna e i quadrati delle differenze fra valori teorici e valori effettivi sono indicati nell'ultima. L'errore standard che fornisce il grado di accostamento è:

$$E_s = \sqrt{2,49707 : 11} = 0,47645$$

Confronto fra i due tipi di trend - Nel caso lineare l'errore standard è 0,74565 mentre nel caso esponenziale è 0,47645. Quindi, essendo esso più vicino a zero nel caso esponenziale, vuol dire che l'ipotesi di trend esponenziale meglio si adatta al fenomeno studiato che non quella di trend lineare. Il grado di accostamento può essere ritenuto abbastanza buono.

Extrapolazione della serie storica e confronto fra proiezioni e valori reali - Extrapolando la serie storica, relativamente agli anni dal 1982 al 1986, otteniamo le proiezioni indicate nella colonna 3 (ipotesi di trend lineare) e nella colonna 4 (ipotesi di trend esponenziale) della tabella 10. Nell'ultima colonna sono, inoltre, indicati i valori effettivamente riscontrati per gli stessi anni (mediante le rilevazioni fatte dopo il 1981).

Tabella 10 - Extrapolazione

Anni	x_i	valori extrapolati		Valori reali
		ipotesi lineare	ipotesi esponenziale	
1982	11	11,29278	12,66573	12,90000
1983	12	9,76733	11,74974	12,30000
1984	13	8,24188	10,90000	11,30000
1985	14	6,71643	10,11171	10,50000
1986	15	5,19098	9,38043	10,10000

Ovviamente, i valori della colonna 3 sono ottenuti da:

$$y_i = -1,52545 \cdot x_i + 28,07273 \qquad \text{per} \qquad x_i = 11,12,13,14,15$$

e quelli della colonna 4 sono ottenuti da:

$$y_i = 28,92344 \cdot 0,92768^{x_i} \qquad \text{per} \qquad x_i = 11,12,13,14,15$$

Come si può vedere, la proiezione esponenziale fatta nel 1981 è più vicina alla realtà che viene riscontrata (valori reali dell'ultima colonna).

Aggiornamento della funzione di trend - Procediamo all'aggiornamento della funzione di trend utilizzando i dati (osservati) negli ultimi 10 anni, cioè dal 1977 al 1986, ponendo il 1977 uguale a zero: tabella 11.
Limitandoci alla sola ipotesi esponenziale e tralasciando l'esposizione dei calcoli, troviamo:

$$y_i = 18,07903 \cdot 0,93601^{x_i}$$

con la quale si possono fare nuove proiezioni per gli anni 1987, 1988, ...

Tabella 11

Anni	indici mortalità
1977	18,1
1978	17,1
1979	15,7
1980	14,6
1981	14,1
1982	12,9
1983	12,3
1984	11,3
1985	10,5
1986	10,1