LA INTERPOLAZIONE

Appartamenti venduti nel 2006 da un'agenzia immobiliare di Treviso.

superficie (mq)	prezzo (k€)
120	160
80	95
64	98
120	170
100	150
115	165
98	145
70	105
60	110
40	60
70	100
80	100
42	75
60	100
75	110

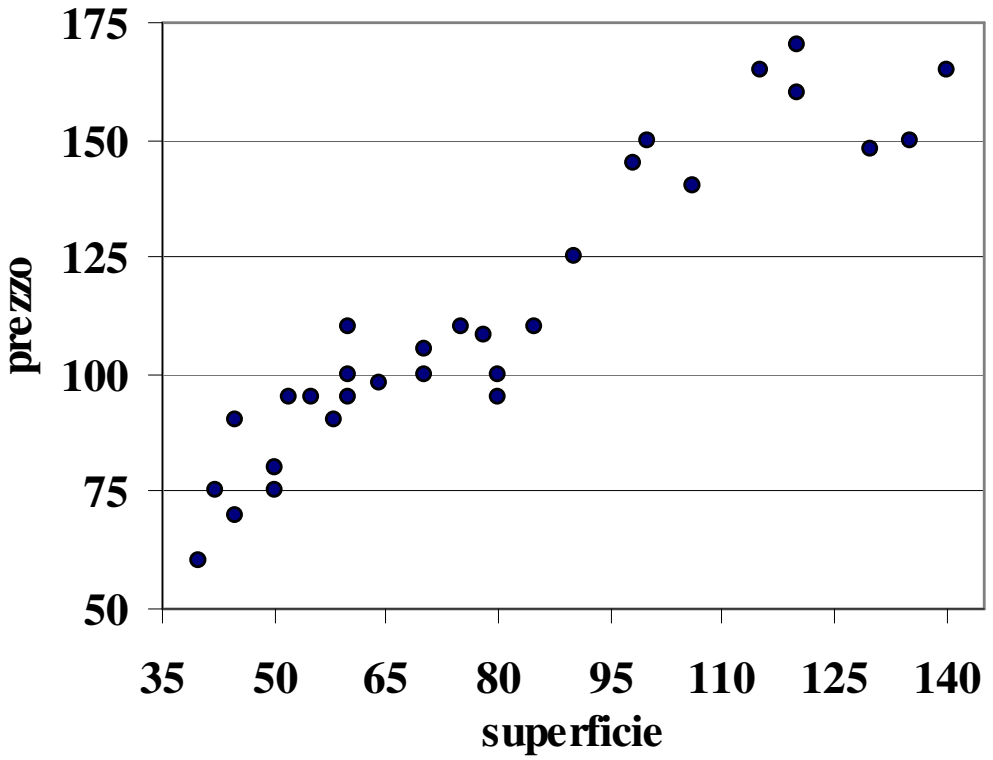
continua…

…segue

45	70
85	110
106	140
50	80
90	125
140	165
45	90
50	75
60	95
55	95
130	148
135	150
58	90
78	108
52	95

Obiettivo: studiare la relazione tra i due caratteri.

Diagramma a dispersione (scatter plot).



L'andamento del prezzo di vendita è tendenzialmente crescente al crescere della superficie.

In altre parole: al crescere della superficie di un appartamento il prezzo tende a crescere.

Tale relazione non è esatta, ma tendenziale, visto che ad es. un appartamento da 60mq è stato venduto a prezzo più alto di uno da 80mq.

Se la relazione fosse esatta, il prezzo di un appartamento sarebbe univocamente determinato dalla sua superficie (a prescindere ad es. dalla posizione dell'appartamento o dal suo stato di usura).

Obiettivo: descrivere il legame esistente tra due caratteri quantitativi, uno dei quali (indicato di norma ma non sempre con Y) si ritiene dipenda dall'altro (indicato di norma ma non sempre con X).

Di fronte a due caratteri, la scelta di quale assumere come dipendente e quale come indipendente dipende da considerazioni teoriche oppure dall'esperienza passata.

L'osservazione congiunta di X e Y su n unità produce una sequenza di n coppie di valori (x_j,y_j) $j=1,\ldots,n$.

La rappresentazione cartesiana delle coppie (x_j,y_j) produce una nuvola di punti la cui conformazione è spesso di aiuto nella interpretazione dell'andamento di Y in corrispondenza di X.

L'andamento congiunto che si riscontra generalmente tra due fenomeni reali non è regolare come quello di una funzione matematica $y=f(x)$, ma presenta delle irregolarità dovute a fattori accidentali oppure all'influenza di altri caratteri non considerati.

Al fine di accrescere la comprensione dei fenomeni empirici è in genere utile fare riferimento ad un modello, inteso come astrazione (rappresentazione semplificata) della realtà.

Interpolazione: operazione tramite la quale è possibile descrivere il legame tra due caratteri quantitativi utilizzando una funzione analitica.

La funzione analitica è lo strumento astratto che ci permette di passare dall'irregolarità dei punti empirici alla regolarità di una funzione matematica $y=f(x)$.

LA REGRESSIONE LINEARE

$$y=f(x;a,b)=a+bx.$$

Consideriamo il metodo di interpolazione detto dei minimi quadrati.

Come misura di scostamento tra il valore reale y_j e quello teorico $y_j^*=a+bx_j$, si usa $|y_j - y_j^*|$ che si può vedere come l'errore che si commette nell'interpolare y_j con y_j^*.

L'errore complessivo viene definito attraverso la somma degli scostamenti al quadrato tra valori empirici e teorici

$$\sum_{j=1}^{n}\left(y_j - y_j^*\right)^2 = \sum_{j=1}^{n}\left(y_j - a - bx_j\right)^2$$

e i parametri *a* e *b* vengono determinati in modo da rendere minima questa quantità.

La retta così ottenuta è la retta del piano *XY* che meglio si adatta ai punti (x_j,y_j) secondo il criterio dei minimi quadrati, ovvero la retta per cui è minima la somma delle distanze al quadrato tra valori empirici e teorici.

E' possibile dimostrare che i parametri della retta ai minimi quadrati valgono

$$b = \frac{\sum_{j=1}^{n} y_j x_j - n\overline{x}\,\overline{y}}{\sum_{j=1}^{n} x_j^2 - n\overline{x}^2} \qquad \textbf{e} \qquad a = \overline{y} - b\overline{x}\,\textbf{.}$$

Il parametro *a* è interpretabile graficamente come l'ordinata del punto di intercetta della retta con l'asse delle ordinate e indica il valore teorico del carattere dipendente *Y* in corrispondenza di un valore nullo del carattere indipendente *X*.

Il parametro *b* ha invece a che fare con la inclinazione della retta, essendone il coefficiente angolare. Esso indica la variazione teorica del carattere dipendente in corrispondenza di una variazione unitaria del carattere indipendente.

***b*>0 retta crescente (concordanza)**

***b*<0 retta decrescente (discordanza)**

***b*=0 retta parallela all'asse delle ascisse (indifferenza)**

Esempio dei 30=*n* appartamenti.

x	y	xy	x^2
120	160	19200	14400
80	95	7600	6400
…	…	…	…
78	108	8424	6084
52	95	4940	2704
2373	3379	292896	213627

$\sum_j y_j x_j = 292896;$ $\quad \bar{x} = 79{,}1;$ $\quad \bar{y} = 112{,}63;$

$\sum_j x_j^2 = 213627.$

$$b = \frac{292896 - 30 \cdot 79{,}1 \cdot 112{,}63}{231627 - 30 \cdot 79{,}1^2} = 0{,}988;$$

$$a = 112{,}63 - 0{,}988 \cdot 79{,}1 = 34{,}466.$$

Equazione della retta ai minimi quadrati

$$y=a+bx=34{,}466+0{,}988x.$$

E' una retta crescente, ad indicare che al crescere della superficie tende a crescere il prezzo.

a=34,466 indica come il prezzo di un teorico appartamento di 0 mq sia di 34466€ (è una sorta di costo fisso).

b=0,988 indica che, in base al modello, all'aumentare di 1mq di superficie il prezzo aumenta di 988€.

Bontà di adattamento: ANALISI DEI RESIDUI

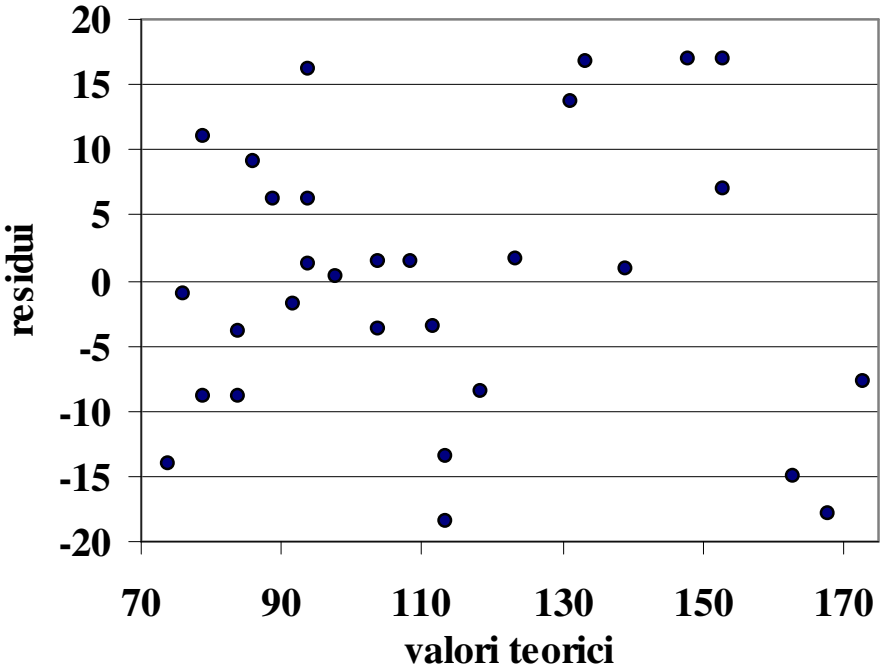
Residuo corrispondente all'unità *j*: $e_j = y_j - y_j^*$.

$$\sum_{j=1}^{n} e_j = \sum_{j=1}^{n} \left(y_j - y_j^* \right) = 0.$$

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} \left(y_j - y_j^* \right)^2 = \min.$$

Un andamento casuale dei residui indica che il modello lineare funziona bene.

Esempio dei 30 appartamenti



Un indice normalizzato di adattamento è l'indice di determinazione lineare R^2.