Concordanza e regressione

Riassunto

Quando i caratteri associati, X e Y, sono entrambi quantitativi, la distribuzione doppia può essere rappresentata in un piano cartesiano. Tale grafico prende il nome di "diagramma di dispersione" e l'insieme dei punti in esso rappresentati si chiama "nuvola di punti". Le due rette che interpolano la (cioè, passano in mezzo alla) nuvola di punti (una di Y a X, l'altra di X a Y) i cui parametri siano determinati in base al metodo dei minimi quadrati si chiamano rette di regressione. Dopo aver descritto graficamente il procedimento con cui si perviene alla determinazione dei parametri delle rette di regressione, si dimostra che i loro coefficienti angolari sono misure di concordanza.

Parole chiave. *Diagramma di dispersione, metodo dei minimi quadrati, concordanza, discordanza, indifferenza, coefficienti di regressione.*

Summary

Concordance and regression

Let X, Y be two variables observed on the same statistical units; the results can be displayed in a scatter plot. When the points are interpolated using two straight lines (one from Y to X, and the other from X to Y) according to the 'least squares method', these lines are called 'regression lines'. The procedure to calculate the parameters of the regression lines is graphically described, and it is shown that the two regression coefficients are indices of concordance.

Key words. *Scatter-plot, least squares method, concordance, discordance, indifference, regression coefficients.*

Siano X e Y due caratteri ***quantitativi*** *associati, ossia rilevati nelle stesse unità della popolazione. Il risultato della rilevazione costituisce una distribuzione doppia (v. CASCO 15, 16). Essendo quantitativi entrambi i caratteri associati, l'analisi della loro relazione è ben più sofisticata di quella che si sarebbe potuta condurre nel caso in cui almeno uno dei caratteri fosse stato qualitativo (analisi descritte nel n. 16 di CASCO).*

Nella presente nota sono forniti i concetti di base della regressione lineare riducendo al minimo l'apparato formale: un modo completamente inedito di trattare senza formule (o quasi) questo pur complesso argomento, utilissimo nello studio di fenomeni di grande interesse per la Medicina. La comprensione del testo sarà avvantaggiata dalla conoscenza delle coordinate cartesiane e dell'equazione della retta.

Lo scopo principale dell'analisi della relazione tra X e Y (entrambi quantitativi) è valutare se Y varia, e se si quanto, al variare di X ed anche di valutare se X varia, e se si quanto, al variare di Y.

Esempio. Siano X il peso e Y la statura osservati sugli stessi 5 soggetti maschi di 18 anni di età:

N. unità	1	2	3	4	5
Peso, kg	65	68	75	70	62
Statura, cm	170	178	180	172	175

Si osservi che in un foglio Excel gli stessi dati sarebbero rappresentati per colonna; qui invece li visualizziamo per riga solo per economia di spazio.

Si è riportato nella prima riga il numero d'ordine del soggetto solo per evidenziare il concetto di **caratteri associati**, cioè rilevati entrambi su ciascuna unità del collettivo. Per ogni soggetto esiste una coppia di dati (peso e statura) i cui elementi non possono essere disgiunti (cioè spostati) proprio in quanto rilevati sulla stessa unità. Ad esempio, il soggetto n. 3 pesa 75 kg ed ha statura pari a 180 cm: questi due dati sono indissolubilmente legati tra loro proprio perché rilevati entrambi sul soggetto 3. Ad ogni soggetto è associata, quindi, una coppia di dati.

Graficamente, i dati (Y = statura, X = peso) possono essere rappresentati su un sistema di coordinate cartesiane in cui sull'asse delle ascisse riportiamo (ad esempio) il peso e su quello delle ordinate la statura. In tal modo il grafico è costituito da 5 punti: ogni soggetto ha per immagine, nel grafico, un punto. Ogni punto, però, rappresenta uno o più soggetti (nel caso ci fossero più soggetti con lo stesso peso e la stessa statura).

Tale grafico prende il nome di **diagramma di dispersione** e l'insieme dei punti in esso contenuti si chiama "nuvola di punti".

In genere nelle applicazioni il numero di soggetti osservati è molto elevato. In tal caso, il diagramma di dispersione assume una forma che può essere quella riportata nella figura 1.

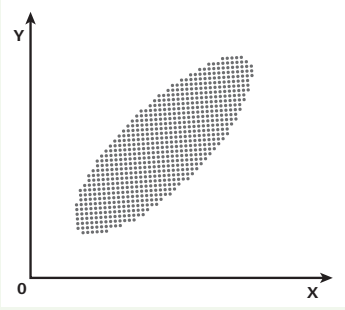

Figura 1. *Diagramma di dispersione.*

Interpolare una nuvola di punti vuol dire far passare una linea in mezzo ai punti della nuvola, in modo tale che il suo andamento esprima come varia Y al variare di X. Può anche essere interpolata una seconda linea che descriva coma varia X al variare di Y.

Osservazione. In generale, nelle applicazioni si può studiare come varia Y al variare di X, ma anche come varia X al variare di Y (nell'esempio: come varia la statura al variare del peso, ma anche come varia il peso al variare della statura). In tal caso le linee interpolatrici sono due: quella di Y a X e quella di X a Y. Talvolta però l'interesse è solo quello di valutare se e quanto varia Y (ad es. la statura) al variare di X (ad es. l'età), ma non viceversa perché avrebbe poco senso: l'età è l'antecedente logico (v. CASCO 16) e, sebbene si possa sempre tecnicamente fare, ha scarso significato misurare quanto varia l'età al variare della statura.

Come linea interpolatrice si sceglie una funzione monotòna (cioè che abbia un solo andamento), tale cioè che sia sempre crescente o sempre decrescente ovvero costante in tutto l'intervallo in cui è definita.

Se l'andamento della linea interpolatrice è crescente, cioè, se al crescere di X, Y cresce, allora si dice che tra i caratteri associati c'è **concordanza**; se è decrescente si dice che c'è **discordanza** (al crescere di X, Y decresce), se è **costante**, si dice che c'è **indifferenza** (al crescere di X, Y resta costante: la linea è parallela all'asse delle ascisse).

Riportiamo i tre casi nelle figure 2, 3 e 4.

Le linee interpolatrici di Y a X sono infinite. Tra tutte si sceglie la retta non solo per motivi di semplicità, ma soprattutto per il valore interpretativo dei suoi parametri, come sarà mostrato nel seguito.

Resta da stabilire un criterio (regola) di interpolazione perché anche le rette interpolatrici (cioè quelle che passano in mezzo ai punti della nuvola) sono infinite. Il **metodo** da seguire nell'effettuare l'interpolazione è quello **dei minimi quadrati**: tra tutte le rette del piano si sceglie quella (di Y a X) che passa il più vicino possibile ai punti dati, cioè la retta tale che la somma dei quadrati degli scarti tra le ordinate empiriche (quelle osservate nei singoli soggetti) e quelle teoriche (cioè dei punti che giacciono sopra la retta interpolatrice) sia minima.

Per studiare come varia X al variare di Y si procede analogamente scegliendo tra tutte le rette del piano che descrivono come varia X al variare di Y, quella che passa il più vicino possibile ai punti dati, cioè la retta che rende minima la somma dei quadrati degli scarti tra ascisse empiriche (osservate nei singoli soggetti) e quelle teoriche (cioè dei punti che giacciono sulla retta interpolatrice).

Nelle figure 5 e 6 sono rappresentate le rette interpolatrici: nella fig. 5 quella di Y a X e nella fig. 6 quella di X a Y; in entrambi i casi le distanze (o scarti), la cui somma dei quadrati va resa minima, sono rappresentate con linee verticali (figura 5) e orizzontali (figura 6) di minore spessore.

Riepilogando: mediante il metodo dei minimi quadrati si determinano i valori dei parametri delle equazioni delle due rette scelte per l'interpolazione.

La retta che mostra come varia Y al variare di X (nell'esempio, come varia la statura al variare del peso) ha equazione $Y = a + bX$. Quella che mostra come varia X al variare di Y (nell'esempio, come varia il peso al variare della statura) ha equazione $X = a' + b'Y$.

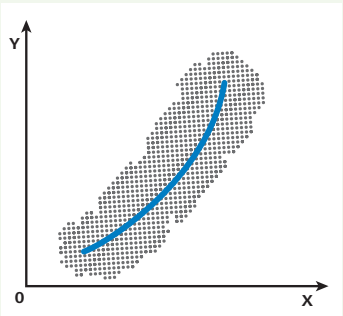

Figura 2. *Concordanza.*

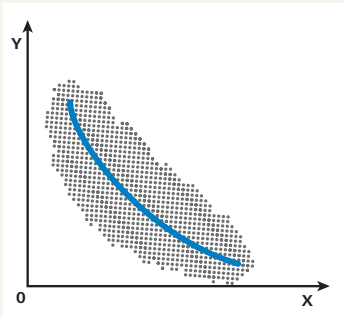

Figura 3. *Discordanza.*

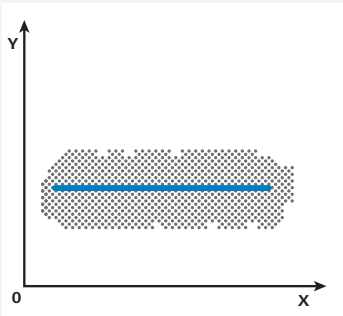

Figura 4. *Indifferenza.*

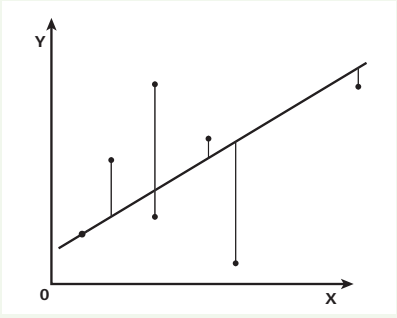

Figura 5. *Retta di equazione $y = a + bx$.*

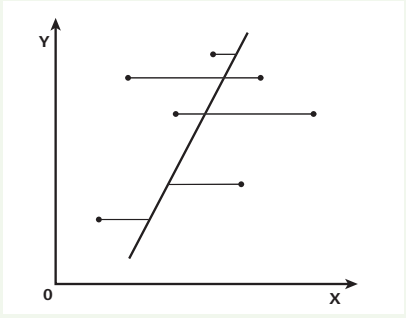

Figura 6. *Retta di equazione $x = a' + b'y$.*

In entrambi i casi i parametri sono determinati in base al metodo dei minimi quadrati; in particolare, per la retta di Y a X (cioè Y = a + bX), a e b sono determinati rendendo minimi i quadrati degli scarti (o distanze) tra ordinate empiriche (quelle osservate nei vari soggetti) e ordinate teoriche (quelle dei punti che giacciono sulla retta interpolatrice); per la retta di X a Y, a' e b' sono determinati rendendo minima la somma dei quadrati degli scarti tra ascisse empiriche (osservate nei soggetti) e ascisse teoriche (quelle dei punti che si trovano sulla retta interpolatrice), come rappresentato nelle figg. 5 e 6.

Le rette di equazione

Y = a + bX

X = a' + b'Y

con i parametri a, b, a' e b' determinati in base al metodo dei minimi quadrati, prendono il nome di **rette di regressione** e, tra tutte le rette del piano, sono quelle che passano il più vicino possibile ai punti empirici.

Riepilogando: esistono due rette di regressione, una di Y a X (cioè che mostra come varia Y al variare di X) e una di X a Y (che mostra come varia X al variare di Y). Tali rette hanno equazioni i cui parametri sono sempre calcolabili. Tuttavia, quando X è l'antecedente logico (cioè la "causa") di Y ha senso solo considerare la retta di Y a X (cioè Y = a + bX, v. ***Osservazione***); i parametri dell'altra sono sempre tecnicamente calcolabili, ma, in tal caso, la retta di equazione X = a' + b'Y (nell'***Osservazione***, come varia l'età al variare della statura) non ha interesse.

Per tener fede al titolo della rubrica, evitiamo di appesantire l'esposizione con dettagli matematici, tanto più che i parametri delle equazioni delle rette di regressione sono calcolati anche da Excel. Occupiamoci invece dell'interpretazione dei valori dei parametri.

Per studiare come varia Y al variare di X, si consideri l'equazione Y = a + bX che descrive come varia (lungo la retta di regressione) Y al variare di X.

Se facciamo crescere di una unità X, Y crescerà (o diminuirà se b è negativo) di una quantità H:

Y + H = a + b(X + 1).

Sottraendo membro a membro da questa equazione quella di partenza:

Y = a + bX,

si ha (Y + H) − Y = (a + bX + b) − (a + bX) → **H = b**.

Il coefficiente angolare della retta Y = a + bX, cioè **b, misura quanto cresce in media** (cioè lungo la retta di regressione) **Y al crescere unitario di X**. Come tale, quindi, **"b" è un indice (o misura) di concordanza**.

Tornando ai dati riportati nell'esempio, l'equazione della retta di regressione della statura rispetto al peso è Y = 144,4 + 0,45 X. Vediamo le informazioni che fornisce.

Anzitutto il segno di b è positivo, il che indica che c'è concordanza, cioè che al crescere del peso, anche la statura cresce: la retta è inclinata positivamente rispetto all'asse X.

Il valore del parametro "a" è irrilevante perché è l'ordinata all'origine (cioè la statura di un soggetto con peso nullo). Invece è assai importante il valore di "b" perché indica che la statura cresce in media di 0,45 cm per ogni accrescimento unitario di peso (cioè al crescere di 1 Kg di peso).

Analogamente, la retta di regressione del peso rispetto alla statura è X = - 45,75 + 0,65 Y:

il segno "+" di "b'" mostra che c'è concordanza; il valore di "a'" è irrilevante (sarebbe il peso di un soggetto con statura pari a 0); b' = 0,65 misura quanto cresce il peso, in media, per ogni incremento unitario di statura (per ogni cm in più di statura, il peso cresce, in media, di 0,65 Kg: in quel gruppo di soggetti, ogni aumento di un centimetro di statura comporta, in media, un incremento ponderale di 0,65 chilogrammi).

Si osservi che quando c'è concordanza nella relazione di Y a X, c'è necessariamente concordanza anche nella relazione di X a Y. Questa affermazione si può provare matematicamente, ma trova anche una giustificazione in base al ragionamento che se al crescere della statura il peso cresce (cioè i soggetti più pesanti hanno anche, in media, una statura superiore) per forza i soggetti con peso superiore hanno, in media, una statura più elevata: b e b' hanno sempre lo stesso segno.

Se è b = 0, l'equazione della retta di regressione di Y a X diventa: Y = a, che è l'equazione di una retta parallela all'asse delle ascisse: al crescere di X, Y (valutato sulla retta di regressione), resta costante. In tal caso si dice che tra i caratteri associati c'è indifferenza: al crescere di un carattere, l'altro (sulla retta di regressione) resta costante. Inoltre, se è b = 0, è anche b' = 0.

In conclusione, b e b' sono indici di concordanza perché misurano, rispettivamente, quanto cresce in media Y al crescere unitario di X e quanto cresce in media X al crescere unitario di Y.

Le applicazioni in medicina sono numerosissime e vanno, ad esempio, dallo studio della relazione tra due parametri emotobiochimici (ad es. glicemia e colesterolemia in un gruppo omogeneo di pazienti), all'analisi di quanto la qualità di vita (misurata con un test psicometrico) possa influire sulla sopravvivenza, all'esame di quanto la PFS possa dipendere dall'età del paziente.

Da ultimo, occorre segnalare la possibilità di eseguire previsioni mediante una retta di regressione. Ad esempio, sia Y la riduzione di glicemia dopo due ore dall'assunzione di un farmaco e X la dose del farmaco stesso (nel *range* 200 – 800 mg, stabilito perché è noto che il farmaco sotto 200 mg è inefficace e sopra 800 mg può dar luogo ad effetti collaterali gravi); se l'equazione della retta di regressione di Y a X è Y = 5 + 0,1X, si può dare una stima dell'effetto di una certa dose del farmaco. Ad esempio, 500 mg produrrà una riduzione dei valori di glicemia pari a Y(500) = 5 + 0,1 x 500 = 55.

Nel prossimo numero di CASCO, esporremo un altro indice di concordanza e verrà fornito un metodo per valutare la qualità dell'interpolazione eseguita mediante le rette di regressione.

Enzo Ballatori