4. Regressione lineare parametrica (con quattro modelli di regressione)

Esistono numerose occasioni nelle quali interessa è ricostruire la relazione di funzione che lega due variabili, la variabile y (variabile dipendente, in ordinate) alla variabile indipendente (variabile x, in ascisse); se si ritiene che la relazione esistente fra le due variabili possa essere convenientemente descritta mediante una retta, l'equazione di tale retta può essere calcolata mediante la tecnica statistica nota come regressione lineare. Tale denominazione deriva dagli studi sull'ereditarietà condotti da F. Galton sul finire dell'800. Nel corso di questi studi vennero, fra l'altro, registrate le altezze dei componenti di più di 1000 gruppi familiari. Ponendo su un sistema di assi cartesiani in ascisse le altezze dei padri e in ordinate le altezze dei figli, si notò un fatto: sebbene in genere padri più alti avessero figli più alti (come del resto era atteso), padri che erano di 16 centimetri circa più alti della media dei padri, avevano figli che erano solamente 8 centimetri circa più alti della media dei figli. In altre parole sembrava che vi fosse un "tornare indietro", una regressione delle altezze dei figli rispetto a quelle dei padri : e il termine che descriveva il risultato di questa iniziale applicazione, finì con l'essere impiegato per indicare la tecnica statistica, ed è rimasto ancora oggi nell'uso.

Per adattare una retta a una serie di dati sperimentali si possono impiegare tecniche parametriche, cha fanno uso di assunzioni a priori (da verificare) sulla distribuzione dei dati, e tecniche non parametriche, che non fanno assunti a priori sulla distribuzione dei dati, ma che per questo risultano in genere meno potenti (in senso statistico).

In questo contesto farò riferimento alla più classica delle tecniche parametriche, il metodo dei minimi quadrati. Questa tecnica di approssimazione consente di minimizzare la somma dei quadrati delle differenze che residuano fra i punti sperimentali e la retta. A scopo didattico, il metodo dei minimi quadrati viene qui impiegato con quattro differenti modelli:
1) regressione x variabile indipendente (regressione lineare standard);
2) regressione y variabile indipendente;
3) componente principale standardizzata;
4) regressione lineare di Deming.

I quattro modelli corrispondono a minimizzare la somma dei quadrati delle differenze che residuano fra i punti sperimentali e la retta rispettivamente:
1) nella direzione dell’asse delle y (x variabile indipendente);
2) nella direzione dell’asse delle x (y variabile indipendente);
3) in direzione ortogonale rispetto alla retta (componente principale standardizzata);
4) in una direzione che varia tra i due limiti estremi (1) e (2), in base ad un fattore che quando è uguale a 1 porta allo stesso risulto della regressione ortogonale (regressione lineare di Deming).

A meno che non vi siano motivi particolari, che devono essere attentamente valutati, viene impiegata la regressione lineare x variabile indipendente. Il modello matematico presuppone che la x, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la y sia distribuito normalmente e con una varianza che non cambia al variare del valore della x. Tra le varie statistiche che è possibile ricavare, particolarmente interessanti sono l’errore standard del coefficiente angolare e dell’intercetta, che consentono di valutare mediante il test t di Student la significatività della differenza del coefficiente angolare da 0 (zero) e da 1 (uno), e la significatività della differenza dell'intercetta da zero (per $p < 0,05$ le differenze citate possono essere considerate significative). Il metodo è trattato su tutti i testi di statistica, fra i quali si ricordano Armitage [1] e Snedecor [2]. Vedere Westgard [3] e Pardue [4] per l'interpretazione dei risultati della regressione lineare in funzione dei diversi tipi di applicazione di questa tecnica statistica nel campo della chimica clinica.

La regressione lineare y variabile indipendente è semplicemente l'immagine speculare della precedente. Viene utilizzata solamente a scopo esplorativo dei dati. Tanto più differisce dalla regressione x variabile indipendente, in presenza di errori nella misura delle x, tanto più si dovrà considerare di esprimere i risultati della regressione in termini di componente principale standardizzata. Infatti la regressione lineare standard non è raccomandata per l'analisi statistica di dati nei quali la variabile indipendente sia affetta da un errore di misura: in questo caso, impiegando la regressione lineare standard, si ottengono, a seconda di quale variabile sia posta in ascisse, equazioni della retta di regressione diverse. E questo fatto porta a conclusioni contraddittorie. Per i casi di questo genere (il caso tipico nel laboratorio clinico è il confronto tra due metodi per la determinazione dello stesso analita) si consiglia di impiegare, in alternativa alla regressione lineare standard, la componente principale standardizzata (per il metodo vedere Feldman [5]). In realtà esiste un altro metodo per calcolare la regressione lineare con il metodo dei minimi quadrati, ed è il metodo di Deming. Il metodo è stato portato all'attenzione del laboratorio clinico da Cornbleet e Gochman [6], ed è stato successivamente approfondito da Linnet [7] e da Martin [8].

Il foglio RegLinCalcio contiene i risultati di due ipotetici metodi per la determinazione del calcio nel siero. Sui primi tre sieri i due metodi hanno fornito identici risultati. Sui rimanenti quattro sieri hanno fornito risultati speculari. La situazione appare caratterizzata da una assoluta indecidibilità. La regressione x variabile indipendente e la regressione y variabile indipendente forniscono in effetti risultati speculari. La componente principale standardizzata più salomonicamente fa propendere per il fatto che, con le informazioni disponibili, possiamo solo concludere che i due metodi forniscono sostanzialmente gli stessi risultati. Se utilizzate Ministat, nella tabella CALCIO del file ESEMPI.MDB trovate gli stessi dati: potete visualizzare le tre rette di regressione per rendervi conto graficamente di quanto siano diverse tra loro e di quanto più "logica" sia la soluzione fornita dalla componente principale standardizzata. Dati reali, con cui sperimentare la regressione lineare, sono quelli relativi alla determinazione dell'urea (valori in mg/dL) effettuata, su una serie di sieri freschi della routine, con due strumenti diversi. I dati sono contenuti nel foglio RegLinUrea. Da notare come in questo caso la scelta di una adeguata ampiezza della dispersione dei valori di concentrazione dell'urea consente di ottenere, con la regressione x variabile indipendente e con la regressione y variabile indipendente, risultati molto simili: anche se il risultato da utilizzare dovrebbe essere quello fornito dalla componente principale standardizzata. Da notare infine che i risultati della componente principale standardizzata sono validi solamente quando i coefficienti angolari della regressione lineare x variabile indipendente e della regressione lineare y variabile indipendente sono entrambi positivi.

Vediamo ora in maggiore dettaglio questi i modelli di regressione.

4.1. Regressione lineare x variabile indipendente

Il modello matematico impiegato presuppone che la x, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la y sia distribuito normalmente e con una varianza che non cambia al variare del valore della x.

Sia allora n il numero dei punti aventi coordinate note (x_i,y_i), e siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i: il coefficiente angolare b_{yx} e l'intercetta a_{yx} dell'equazione della retta di regressione x variabile indipendente

$$y = a_{yx} + b_{yx} \cdot x$$

che meglio approssima (in termini di minimi quadrati) i dati vengono calcolati rispettivamente come

$$b_{yx} = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \Sigma (x_i - \bar{x})2$$

$$a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$$

Il coefficiente di correlazione *r* viene calcolato come

$$r = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2)}$$

E' possibile semplificare i calcoli ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma (y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

La varianza residua attorno alla regressione viene calcolata come

$$s_0^2 = (\Sigma (y_i - \bar{y})^2 - s_1^2) / (n - 2)$$

essendo

$$s_1^2 = (\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}))^2 / \Sigma (x_i - \bar{x})^2$$

Infine l'errore standard della stima s_{yx} e le deviazioni standard del coefficiente angolare (s_b) e dell'intercetta (s_a), che forniscono una misura rispettivamente della dispersione dei dati attorno alla retta calcolata, e del grado di incertezza connesso con i valori ottenuti di a_{yx} e di b_{yx}, sono calcolati come

$$s_{yx} = \sqrt{(s_0^2)}$$
$$s_b = s_{yx} \cdot \sqrt{(1 / \Sigma (x_i - \bar{x})^2)}$$
$$s_a = s_b \cdot \sqrt{(\Sigma x_i^2 / n)}$$

Si consideri che la retta di regressione campionaria

$$y = a_{yx} + b_{yx} \cdot x$$

rappresenta la migliore stima possibile della retta di regressione della popolazione

$$y = \alpha + \beta \cdot x$$

Si consideri che il test t di Student per una media teorica nella forma già vista

$$t = (\bar{x} - \mu) / \sqrt{(s^2 / n)}$$

può essere riscritto tenendo conto delle seguenti identità

$$\bar{x} = a_{yx}$$

$$\mu = \alpha$$
$$\sqrt{(s^2 / n)} = s_a$$

assumendo quindi la forma

$$t = (a_{yx} - \alpha) / s_a$$

Questo consente di sottoporre a test la differenza dell'intercetta *a* rispetto a un valore atteso (per esempio rispetto a 0, cioè all'intercetta di una retta passante per l'origine). Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza dell'intercetta rispetto al valore atteso.

Si consideri che il test t di Student per una media teorica può anche essere riscritto tenendo conto delle seguenti identità

$$\bar{x} = b_{yx}$$
$$\mu = \beta$$
$$\sqrt{(s^2 / n)} = s_b$$

assumendo quindi la forma

$$t = (b_{yx} - \beta) / s_b$$

Questo consente di sottoporre a test la differenza del coefficiente angolare *b* rispetto a un valore atteso (per esempio rispetto a 0, cioè al coefficiente angolare di una retta orizzontale, oppure rispetto a 1, cioè al coefficiente angolare di una retta a 45 gradi). Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza del coefficiente angolare rispetto al valore atteso.

4.2. Regressione lineare y variabile indipendente

Il modello matematico impiegato presuppone che la *y*, cioè la variabile indipendente, sia misurata senza errore, e che l'errore con cui si misura la *x* sia distribuito normalmente e con una varianza che non cambia al variare del valore della *y*. Si noti che in questo caso inizialmente la *y* (variabile indipendente) viene posta in ascisse e la *x* (variabile dipendente) viene posta in ordinate.

Sia allora *n* il numero dei punti aventi coordinate note (x_i,y_i), e siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i: il coefficiente angolare b_{xy} e l'intercetta a_{xy} dell'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

che meglio approssima (in termini di minimi quadrati) i dati vengono calcolati rispettivamente come

$$b_{xy} = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \Sigma (y_i - \bar{y})^2$$

$$a_{xy} = \bar{x} - b_{xy} \cdot \bar{y}$$

Il coefficiente di correlazione *r* viene calcolato come

$$r = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2)}$$

(si noti che, come atteso, esso risulta identico a quello calcolato mediante la regressione *x* variabile indipendente).

E' possibile semplificare i calcoli ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma (y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

Per riportare i dati sullo stesso sistema di coordinate cartesiane utilizzato per la regressione *x* variabile indipendente, si esplicita l'equazione della retta di regressione *y* variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

rispetto alla *y*, ottenendo

$$x - a_{xy} = b_{xy} \cdot y$$

e quindi, dividendo entrambi i membri per b_{xy}

$$y = - a_{xy} / b_{xy} + 1 / b_{xy} \cdot x$$

Quindi l'intercetta *a* e il coefficiente angolare *b* che consentono di rappresentare la regressione *y* variabile indipendente sullo stesso sistema di coordinate cartesiane della regressione *x* variabile indipendente saranno rispettivamente uguali a

$$a = - a_{xy} / b_{xy}$$
$$b = 1 / b_{xy}$$

4.3. Componente principale standardizzata

Il modello matematico impiegato presuppone tanto la *x* quanto la *y* siano affette da un errore di misura equivalente.

Sia allora *n* il numero dei punti aventi coordinate note (x_i,y_i), siano $\bar{x}$ la media dei valori delle x_i e $\bar{y}$ la media dei valori delle y_i, sia b_{yx} il coefficiente angolare dell'equazione della retta di regressione *x* variabile indipendente, e sia b_{xy} il coefficiente angolare dell'equazione della retta di regressione *y* variabile indipendente.

Il coefficiente angolare b_{cps} dell'equazione della retta di regressione calcolata come componente principale standardizzata è allora uguale a

$$b_{cps} = \sqrt{(b_{yx} \cdot b_{xy})}$$

cioè alla media geometrica tra il coefficiente angolare b_{yx} della regressione *x* variabile indipendente e il coefficiente angolare b_{xy} della regressione *y* variabile indipendente, cioè

$$b_{xy} = \sqrt{(\Sigma (y_i - \bar{y})^2 / \Sigma (x_i - \bar{x})^2)}$$

mentre l'intercetta a_{cps} dell'equazione della retta di regressione calcolata come componente principale standardizzata è uguale a

$$a_{cps} = \bar{y} - b_{cps} \cdot \bar{x}$$

Infine il coefficiente di correlazione *r* viene calcolato come

$$r = \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{(\Sigma (x_i - \bar{x})^2 \cdot \Sigma (y_i - \bar{y})^2)}$$

(si noti che, come atteso, esso risulta identico sia a quello calcolato mediante la regressione *x* variabile indipendente sia a quello calcolato mediante la regressione *y* variabile indipendente).

E' possibile sempre semplificare i calcoli ricordando che

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$
$$\Sigma (y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$
$$\Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \Sigma x_i \cdot y_i - (\Sigma x_i) \cdot (\Sigma y_i) / n$$

4.3. Regressione lineare di Deming

Il modello matematico proposto da Deming è sottile, e parte dal fatto che in una regressione lineare con il metodo dei minimi quadrati la regressione x variabile indipendente e la regressione y variabile indipendente sono due casi limite: il primo assume che la x sia misurata senza errore, il secondo assume che la y sia misurata senza errore. Come detto esistono dei casi nei quali nessuno dei due assunti è completamente valido. Tipicamente se sia la x sia la y sono misurate, entrambe saranno affette da un errore di misura. L'idea è di adottare un fattore di correzione che misura l'entità dell'errore dal quale sono affette le due variabili. Il fattore è indicato con λ e viene definito come

$$\lambda = \Sigma (x_i - \bar{x})^2 / \Sigma (y_i - \bar{y})^2$$

ovvero come rapporto tra la varianza della x e la varianza della y. In pratica il rapporto λ può essere calcolato impiegando la varianza della x e la varianza della y ottenute da misure replicate di un campione con una concentrazione prossima alla media dei valori osservati, misure effettuate con i due metodi x e y in esame.

Quando $\lambda = 1$ la regressione di Deming fornisce un risultato esattamente a metà strada tra quello della regressione x variabile indipendente e la regressione y variabile indipendente: ed è lo stesso risultato della componente principale standardizzata. Quando $\lambda \approx 500$ il risultato è sostanzialmente identico a quella della regressione y variabile indipendente, mentre quando $\lambda \approx 0{,}002$ il risultato è sostanzialmente identico a quello della regressione x variabile indipendente.

Il coefficiente angolare b_{xy} della regressione lineare secondo Deming è quindi calcolato come

$$b_{yx} = U + \sqrt{[U^2 + (1/\lambda)]}$$

essendo

$$U = [\Sigma (y_i - \bar{y})^2 - (1/\lambda) \cdot (\Sigma (x_i - \bar{x})^2)] / [2 \cdot (\Sigma (x_i - \bar{x}) \cdot \Sigma (y_i - \bar{y})]$$

Infine l'intercetta a_{yx} viene calcolata a partire dal coefficiente angolare b_{xy} nel modo usuale, ovvero come

$$a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$$

Bibliografia

1. Armitage P. Statistica medica. Milano: Feltrinelli Editore, 1979:151-167 e 264-266.

2. Snedecor GW, Cochran WG. Statistical methods. Seventh Edition. Ames: The Iowa State University Press, 1980:149-174.

3. Westgard JO, Hunt MR. Use and interpretation of common statistical tests in method-comparison studies. Clin Chem 1973;19:49-57.

4. Davis RD, Thompson JE, Pardue HL. Characteristics of statistical parameters used to interpret least-squares results. Clin Chem 1978;24:611-20.

5. Feldman U, Schneider B, Klinkers H. A multivariate approach for the biometric comparison on analytical methods in clinical chemistry. J Clin Chem Clin Biochem 1981;19:131-7.

6. Corbleet PJ, Gochman N. Incorrect least-squares regression coefficients in method-comparison analysis. Clin Chem 1979;25:432-8.

7. Linnet K. Evaluation of regression procedures for method comparison studies. Clin Chem 1993;39:424-32.

8. Martin RF. General Deming regression for estimating systematic bias and its confidence interval in method-comparison studies. Clin Chem 2000;46:100-4.