LA REGRESSIONE LINEARE NELLA RICERCA CLINICA

Fabio Provenzano, Carmine Zoccali, Giovanni Tripepi

CNR-IBIM, Unità di Ricerca di Epidemiologia Clinica e Fisiopatologia delle Malattie Renali e dell'Ipertensione Arteriosa, Reggio Calabria

KEY WORDS: Confounding, Multiple linear regression analysis, Simple linear regression analysis

PAROLE CHIAVE: Confondimento, Regressione lineare multipla, Regressione lineare semplice

Introduzione

I *trials* clinici controllati e randomizzati sono studi sperimentali disegnati per analizzare l'efficacia di uno specifico trattamento in termini quantitativi. I pazienti randomizzati ai due bracci di trattamento di un *trial* clinico risultano simili per fattori di rischio noti e non noti (1, 2) e, pertanto, alla fine del *trial*, ogni differenza osservata tra i due gruppi di pazienti può essere attribuita al solo trattamento. A differenza di quanto accade nei *trials* clinici, negli studi osservazionali gli individui esposti a un determinato fattore di rischio (per esempio, il fumo) generalmente differiscono da quelli non esposti per una serie di importanti caratteristiche (confonditori) (3) che possono alterare il rapporto tra l'esposizione oggetto dell'indagine e una determinata malattia o esito clinico. Nell'ambito della ricerca eziologica, gli epidemiologi utilizzano due principali tecniche statistiche per controllare per il confondimento: l'analisi stratificata (3) e la regressione multipla. La regressione multipla, rispetto all'analisi stratificata, ha il vantaggio di poter controllare simultaneamente per più fattori di confondimento. In questo articolo, descriveremo la regressione lineare semplice e multipla e, in un articolo successivo, la regressione logistica.

La Correlazione Lineare

La correlazione lineare analizza la forza dell'associazione tra due variabili continue e fornisce una stima di quanto la variabilità dell'una è spiegata dalla variabilità dell'altra. La regressione lineare descrive, invece, la dipendenza lineare della variabile dipendente da una o più variabili indipendenti. Consideriamo uno studio nel quale gli Autori hanno analizzato il rapporto tra i livelli circolanti di albumina e quelli di triodotironina libera (*free triiodothyronine*, fT3) in 41 pazienti in dialisi peritoneale (4). Dal momento che vi è l'evidenza sperimentale che la malnutrizione e l'infiammazione influenzano la funzione tiroidea, i ricercatori hanno considerato i livelli plasmatici di fT3 come variabile dipendente e l'albuminemia (un indicatore di malnutrizione/infiammazione) come variabile indipendente. Nell'analisi di regressione, la variabile indipendente viene sempre riportata sull'asse orizzontale (asse X), mentre la variabile dipendente sull'asse verticale (asse Y). Nella Figura 1, ogni punto rappresenta un individuo che è identificato da una coppia di valori: il valore dell'albumina sull'asse delle X e il corrispondente valore di fT3 sull'asse delle Y. Dal grafico nella Figura 1 risulta evidente come, all'aumentare dei livelli di albumina, aumentano parallelamente anche i livelli di fT3 e il rapporto tra le due variabili è descritto in maniera adeguata da una linea retta (modello lineare). In questo caso, il coefficiente di correlazione (r) tra i livelli di albumina e di fT3 è 0.52. Il quadrato del coefficiente di correlazione (0.52^2=0.27, cioè il 27%) indica che circa ¼ della variabilità nei livelli plasmatici di fT3 è spiegato dalle concomitanti variazioni dei livelli di albumina. Inoltre, sia l'albuminemia che i livelli di fT3 sono inversamente correlati all'età (Fig. 2).

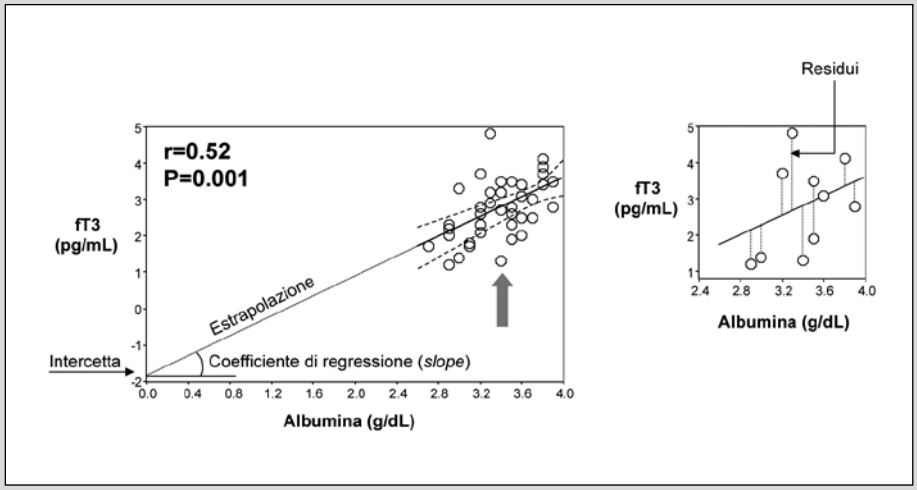

Fig. 1 - *Rapporto tra i livelli di albumina e di fT3 in 41 pazienti in dialisi peritoneale (4). Nel riquadro di destra è descritta graficamente l'analisi dei residui. Le linee tratteggiate indicano l'intervallo di confidenza al 95% della retta di regressione (vedi testo per maggiori dettagli).*

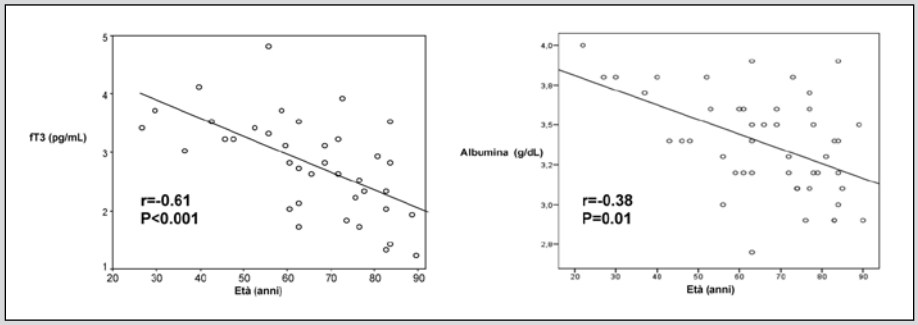

Fig. 2 - *Rapporto tra età e livelli circolanti di fT3 e di albumina in 41 pazienti in dialisi peritoneale (4).*

La regressione lineare semplice

La dipendenza lineare tra i livelli circolanti di fT3 e quelli dell'albumina può essere analizzata calcolando di quanto aumenta in media l'fT3 per ogni incremento unitario di albumina. Questa informazione può essere ottenuta tracciando la retta di regressione che descrive il rapporto fT3-albumina nei 41 pazienti dello studio. In termini generali, la retta di regressione tra due variabili è un'equazione del tipo:

$$E(y)=\beta_0+\beta_1 x$$

dove E(y) è la stima (cioè il valore predetto) della variabile dipendente Y, β_0 è l'intercetta, β_1 è il coefficiente di regressione e x è un determinato valore della variabile indipendente. L'intercetta (β_0) è il valore teorico che assume la Y quando la X è uguale a 0 (Fig. 1). Il coefficiente di regressione (β_1 o *slope*) indica di quanto aumenta in media la variabile dipendente (Y) per ogni incremento unitario della variabile indipendente (X). In termini geometrici, il coefficiente di regressione è la tangente dell'angolo che la retta di regressione forma con l'asse delle X (Fig. 1). Il metodo utilizzato per stimare l'intercetta e il coefficiente di regressione è il metodo dei minimi quadrati. Questo metodo consiste nell'identificare i parametri (β_0 e β_1) che minimizzano la distanza (residui) tra i valori osservati e quelli stimati dalla retta (vedi grafico di destra nella Fig. 1).

L'equazione matematica che descrive la retta di regressione del rapporto

fT3-albumina nei 41 pazienti in dialisi peritoneale è la seguente:

fT3=-1.84+1.36 x albumina (g/dL)

Un coefficiente di regressione di 1.36 indica che, a un aumento di 1 g/dL di albumina, corrisponde un incremento medio di 1.36 pg/mL di fT3 [ciò implica che, a un aumento di 2 g/dL di albumina, corrisponde un aumento medio di 2.72 pg/mL di fT3 (cioè 1.36x2)]. Un coefficiente di regressione positivo indica l'esistenza di un rapporto diretto tra fattore di rischio (variabile indipendente) e variabile di risultato (variabile dipendente), mentre un coefficiente di regressione negativo indica un rapporto inverso. Il valore dell'intercetta (-1.84 pg/mL) corrisponde al valore teorico di fT3 quando l'albumina è uguale a 0 (Fig. 1). Risulta evidente che un valore negativo di fT3 (-1.84 pg/mL) e un livello di albumina pari a zero sono dei valori puramente teorici. L'intercetta è utile perché può essere utilizzata, insieme al coefficiente di regressione, per predire il valore di fT3 per un dato paziente del quale conosciamo solo i livelli di albumina.

Per esempio, il valore stimato di fT3 per un paziente in dialisi peritoneale con un'albuminemia di 3.4 g/dL (vedi il punto indicato dalla freccia nella Fig. 1) può essere facilmente calcolato con l'equazione:

fT3=-1.84+1.36x3.4=2.78 pg/mL

Pertanto, utilizzando la retta di regressione costruita in 41 pazienti in dialisi peritoneale, è possibile predire un valore di fT3 pari a 2.78 pg/mL per un paziente di cui è nota l'albuminemia (3.4 g/dL). Per tale paziente, il residuo (-1.48 pg/mL) è calcolato come differenza tra il valore osservato di fT3 (1.30 pg/mL) e quello stimato dalla retta di regressione (2.78 pg/mL). Ripetendo questo calcolo per tutti i valori osservati e predetti di fT3 nei 41 pazienti in dialisi peritoneale, si ottiene la distribuzione dei residui. L'analisi dei residui è fondamentale per la verifica degli assunti sottostanti all'uso della regressione lineare: 1) ad ogni valore della variabile indipendente (asse X) deve corrispondere un *set* di valori normalmente distribuiti della variabile dipendente (asse delle Y), 2) la deviazione *standard* di questo *set* di valori deve essere identica per ogni valore della variabile indipendente e 3) il rapporto tra le due variabili considerate deve essere di tipo lineare. Se si verificano tutte queste ipotesi, i residui avranno una distribuzione normale.

Nel nostro esempio, i residui hanno una distribuzione normale, il che implica che gli assunti sottostanti all'uso della regressione lineare sono soddisfatti.

Intervallo di confidenza al 95% della retta di regressione

La retta di regressione del rapporto fT3-albumina tracciata nei 41 pazienti in dialisi peritoneale è una stima della *vera* retta di regressione tra le due variabili, cioè della retta di regressione che descrive il rapporto fT3-albumina nell'insieme teorico di tutti i pazienti in dialisi peritoneale (universo campionario). Perciò, è necessario stimare il grado di incertezza della retta di regressione calcolando l'intervallo di confidenza al 95% (vedi linee tratteggiate nella Fig. 1). Il concetto di *intervallo di confidenza* può essere spiegato in maniera semplice: se analizziamo 100 campioni di pazienti in dialisi peritoneale, ciascuno dei quali ha la stessa dimensione del campione sotto esame (n=41), e tracciamo per ogni campione la retta di regressione

del rapporto fT3-albumina otteniamo una serie di 100 (lievemente differenti) rette di regressione. L'intervallo di confidenza al 95% è l'intervallo che include il 95% delle rette di regressione dei 100 campioni analizzati. Nel nostro esempio, l'intervallo di confidenza al 95% è abbastanza stretto (Fig. 1), il che implica che il modello lineare descrive in maniera adeguata il rapporto fT3-albumina.

La regressione lineare multipla

La regressione lineare multipla permette di stimare l'effetto di una determinata variabile indipendente (per esempio, x_1) su livelli di una specifica variabile dipendente (Y), controllando per l'effetto confondente di altri fattori, o covariate (per esempio, x_2, x_3…, x_n). In termini generali, la regressione lineare multipla ha un'equazione del tipo:

$$E(y)=\beta_0+\beta_1x\beta_1+\beta_2x\beta_2+\beta_3x\beta_3+\ldots+\beta_nx_n$$

dove E(y) è la stima o valore predetto di Y, β_0 è l'intercetta (cioè il valore di Y quando x_1, x_2, x_3 sono uguali a 0) e β_1, β_2, β_3 e β_n sono i coefficienti di regressione di x_1, x_2, x_3 e x_n.

Nell'esempio precedente, abbiamo descritto il rapporto tra fT3 e albumina in 41 pazienti in dialisi peritoneale e abbiamo trovato che le due variabili erano strettamente associate. L'obiettivo che gli Autori del lavoro si sono posti (4) è stato anche quello di analizzare il rapporto fT3-albumina correggendo per l'effetto confondente dell'età, una variabile che risulta linearmente correlata sia ai livelli di fT3 ($r=-0.61$, $P<0.001$) che a quelli dell'albumina ($r=-0.38$, $P=0.001$) (Fig. 2). L'età può essere considerata un potenziale fattore di confondimento, in quanto soddisfa tutti i criteri della definizione di "confonditore" (3). Infatti, l'età influenza sia i livelli di fT3 (la variabile dipendente) sia quelli dell'albumina (la variabile indipendente) (Fig. 2): non è un effetto dell'esposizione (l'età non è una conseguenza dei livelli di albumina) e non vi sono evidenze che dimostrino che l'età si trovi nella catena patogenetica tra l'esposizione (albumina) e la variabile di risultato (fT3). Introducendo l'età nel modello lineare multiplo, la retta di regressione ha la seguente equazione:

fT3=1.41+0.87 x albumina (g/dL)-0.024 x età (anni)

Un coefficiente di regressione di 0.87 indica che, per ogni aumento di 1 g/dL di albumina, si ha un corrispondente aumento di 0.87 pg/mL di fT3. Il coefficiente di regressione aggiustato per l'effetto confondente dell'età (0.87) differisce in misura importante da quello non corretto (1.36), il che indica che l'età è un confonditore di cui tenere conto nell'analisi del rapporto tra fT3 e albumina nei pazienti in dialisi peritoneale.

In un modello di regressione multipla, il numero delle variabili da testare deve essere in rapporto con il numero dei pazienti inclusi nello studio. Una regola pratica è quella di includere una covariata ogni 10 osservazioni. Pertanto, in un campione di 41 pazienti, è possibile inserire nel modello al massimo 4 covariate.

Conclusione

La regressione lineare è un importante strumento per testare ipotesi negli studi epidemiologici. L'uso della regressione lineare dipende criticamente dalla verifica di specifici assunti: 1) ad ogni valore della variabile indipendente (asse X) deve corrispondere un *set* di valori normalmente distribuiti della variabile dipendente (asse delle Y), 2) la deviazione *standard* di questo *set* di valori deve essere identica per ogni valore della variabile indipendente e 3) il rapporto tra le due variabili considerate deve essere di tipo lineare. Se si verificano tutte queste ipotesi, i residui avranno una distribuzione normale.

Indirizzo degli Autori:
Dr. Giovanni Tripepi, Statistician, MSc (Epidemiology)
CNR-IBIM, Istituto di Biomedicina
Epidemiologia Clinica e Fisiopatologia
delle Malattie Renali e dell'Ipertensione Arteriosa
Via Vallone Petrara 55/57
89124 Reggio Calabria
e-mail: gtripepi@ibim.cnr.it

Dichiarazione di conflitto di interessi

Gli Autori dichiarano di non avere conflitto di interessi.

Bibliografia

1. Provenzano F, Tripepi G, Zoccali C. [Clinical *trials* (Part I)]. G Ital Nefrol 2010; 27 (4): 396-8.
2. Provenzano F, Tripepi G, Zoccali C. [Clinical *trials* (Part II)]. G Ital Nefrol 2010; 27 (5): 536-9.
3. Provenzano F, Versace MC, Tripepi R, Zoccali C, Tripepi G. [Confounding in epidemiology.]. G Ital Nefrol 2010; 27 (6): 664-7.
4. Enia G, Panuccio V, Cutrupi S, et al. Subclinical hypothyroidism is linked to micro-inflammation and predicts death in continuous ambulatory peritoneal dialysis. Nephrol Dial Transplant 2007; 22 (2): 538-44.