Concetti di modello statistico e analisi multi-variabile

P. Ravani, F. Malberti

Divisione di Nefrologia e Dialisi, Azienda Ospedaliera, Cremona

Statistical models and multivariable analysis

Most clinical research can be simplified as an investigation of an input/output relationship. The inputs are called explanatory (independent) variables or predictors and are thought to be related to the outcome, or response (independent) variable. This relationship is usually complicated by other factors related to both the input and the output (presence of confounding) and can vary according to the levels of the other variables (presence of interaction). This input/output relationship is usually described by statistical models that include a fit part and a residual component or difference between the data and the fit. The most popular models are the general linear models, which can be considered the paradigm of all models used in multi-variable analyzes. (G Ital Nefrol 2005; 22: 348-53)

KEY WORDS: Statistical model, Confounding, Interaction

PAROLE CHIAVE: Modello statistico, Confondimento, Interazione

Introduzione

Con la presente rassegna ci proponiamo di presentare alcuni concetti essenziali per la comprensione del significato e dell'utilità dell'analisi *multi-variabile* mediante la tecnica della *regressione* multipla[1]. Gli statistici definiscono regressione la variazione dei valori medi di una variabile (risposta, o variabile dipendente) – ad esempio, le misure di malattia o di effetto precedentemente incontrate, come i rischi e i rischi relativi – rispetto a un'altra (predittore, o regressore, o variabile indipendente). Va premesso che le tecniche di regressione sono applicabili non soltanto ai dati di sopravvivenza, ma a qualsiasi tipo di relazione esistente tra una malattia (fenomeno definito "esito") e i suoi determinanti (variabili esplicative del fenomeno). Ovviamente il tipo di analisi di regressione è diverso a seconda del tipo di relazione che si vuole analizzare. Ma, in generale, quando l'analisi di regressione descrive una relazione tenendo conto di più variabili indipendenti contemporaneamente, allora viene detta regressione multipla.

Concetto di funzione

Per tentare di catturare l'interesse di coloro che non hanno simpatia per la matematica, spieghiamo il significato di alcuni simboli utilizzati dagli statistici. Il simbolo che indica la variabile di risposta è la lettera y mentre il simbolo delle variabili indipendenti è la x. Ad esempio, in uno studio in cui si ipotizza che il valore della pressione arteriosa media (variabile di risposta) dipenda dal peso corporeo, dell'età e dell'abitudine di fumare, il valore di y (pressione) verrà descritto come funzione di x_1, x_2, x_3, ossia età,

[1] Esistono altre tecniche per effettuare un'analisi multi-variabile (ossia tenendo conto di più variabili contemporaneamente), come la stratificazione, ma la regressione è la tecnica più efficiente e diffusa in bio-statistica. Quando la regressione è il metodo di stima dei parametri questi vengono detti parametri di regressione.

peso, la presenza o assenza dell'abitudine di fumare[2]. Senza accorgercene, stiamo "digerendo" un concetto spesso considerato ostico: il concetto di "funzione". Abbiamo fatto un esempio: pressione (y) funzione di x_1, x_2, x_3 (età, peso e fumo). Abbiamo intuito che per y funzione di x intendiamo un legame tra la modificazione dei valori di y al variare dei valori di x. Grossolanamente possiamo immaginare una funzione come una "macchina specializzata" (un operatore matematico) che trasforma un input (le x_n) in un output (la y), ossia elabora i dati di input e restituisce i dati di output. La notazione "$y = f(x_1, x_2, x_3)$" si legge "y è funzione di x_1, x_2, x_3", nel nostro esempio la pressione (y) è funzione dell'età (x_1), del peso (x_2) e del fumo (x_3), cioè pressione = f(età, peso, fumo). Ci rimane da capire quali funzioni "f" sono utili ai nostri scopi. Lo faremo dopo avere compreso perché e quando utilizziamo la regressione multipla.

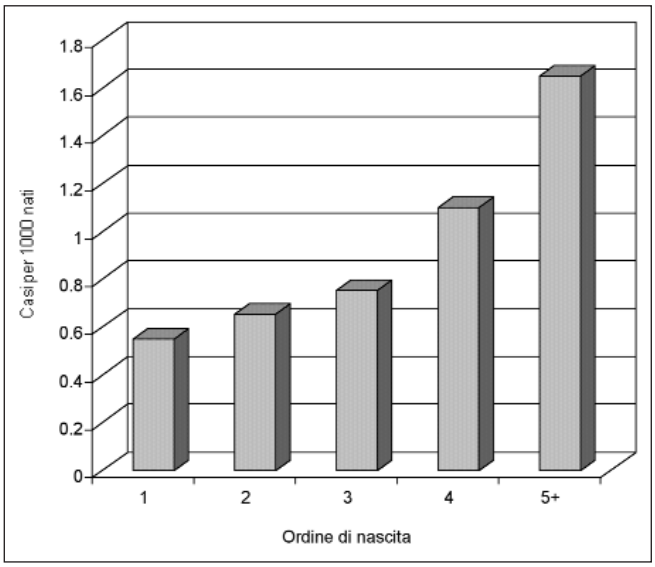

Fig. 1 - *Relazione tra ordine di nascita e prevalenza di sindrome di Down alla nascita.*

Perché la regressione multipla

Abbiamo introdotto il concetto di relazione tra una variabile y e più variabili x, una delle quali, in genere, ci interessa in modo particolare. Infatti quando intraprendiamo uno studio abbiamo un'ipotesi di lavoro. Ad esempio, supponiamo di voler testare l'ipotesi che un certo farmaco sia migliore di un altro nel ridurre i valori pressori degli ipertesi essenziali. Allora ci chiediamo se y (pressione degli ipertesi essenziali) che sappiamo essere (per semplicità) funzione di età, peso e fumo è anche funzione di una nuova x (x_4 in aggiunta a x_1, x_2, x_3) che chiamiamo *esposizione*. Infatti, non siamo così sprovveduti da non tener conto delle conoscenze precedenti, ma vogliamo aggiungerne una nuova. Pertanto consideriamo nel disegno dello studio la raccolta, per ciascun individuo, dei valori x_4 in aggiunta a quelli di x_1, x_2, x_3. Sappiamo dalle conoscenze della letteratura che la relazione potenzialmente causale tra esposizione (x_4) e risposta (y) può essere mascherata oppure modificata dalla presenza di altri fattori. Definiamo questi fattori confondenti e/o modificatori di effetto. Nell'esempio della pressione, se x_4 assume due possibili valori (uno per il farmaco A e uno per il farmaco B), allora la domanda che noi ci poniamo è: il farmaco A ha un effetto maggiore di B indipendentemente dall'età, dal peso e dall'abitudine di fumare? Con questa domanda noi chiediamo all'analisi statistica di aiutarci a capire se, controllando

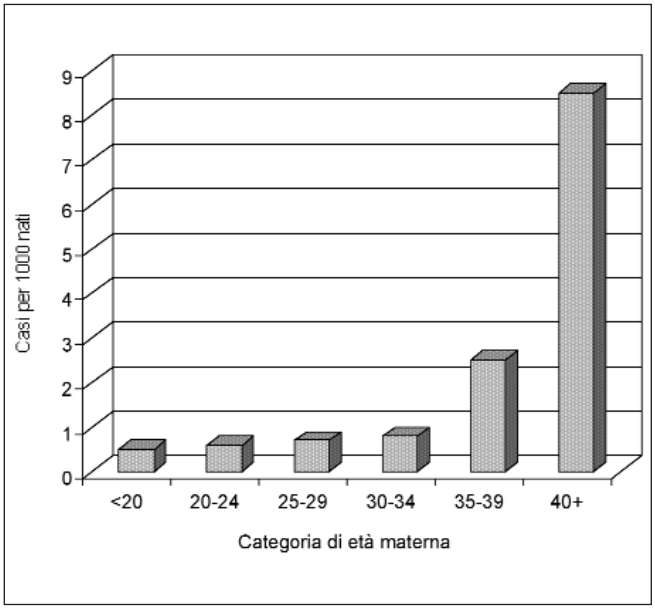

Fig. 2 - *Relazione tra età materna (categorie di età in anni) e prevalenza di sindrome di Down alla nascita.*

i potenziali confondenti (x_1, x_2, x_3) della relazione tra esposizione e malattia (tra x_4 e y), esiste ancora la relazione tra x_4 e y (controllo del confondimento). Ci poniamo inoltre la domanda: le variabili x_1, x_2, x_3 modificano l'effetto di x_4 su y? Questa domanda presuppone lo studio dell'interazione tra x_4 e x_1, x_2, x_3. Approfondiamo il concetto di confondimento e interazione separatamente.

Un famoso esempio che viene presentato dai testi di epidemiologia per far comprendere il concetto di confondimento riguarda la relazione tra fattori di rischio materni e

[2] Gli statistici utilizzano le lettere maiuscole per indicare le variabili aleatorie o casuali (Y, X) e quelle minuscole per le loro singole realizzazioni (i valori assunti dalla variabile) nel campione (x_1, x_2, x_3, ..., x_n). Inoltre utilizzano le lettere greche e latine rispettivamente per indicare i parametri nella popolazione generale e nel campione, e il cappuccio "^" sopra le stime dei parametri. In questa rassegna utilizzeremo solo le lettere latine minuscole e non useremo altri simboli per evitare che l'eccesso di precisione confonda il lettore.

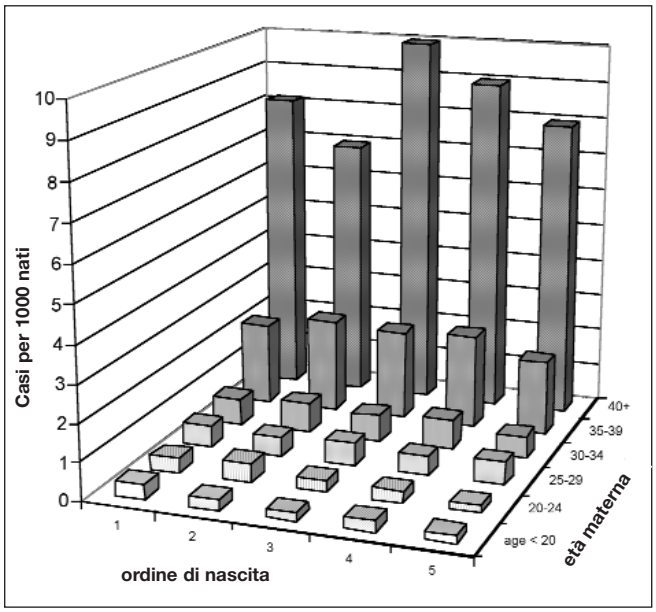

Fig. 3 - *Relazione tra ordine di nascita, età materna (in anni) e prevalenza di sindrome di Down alla nascita.*

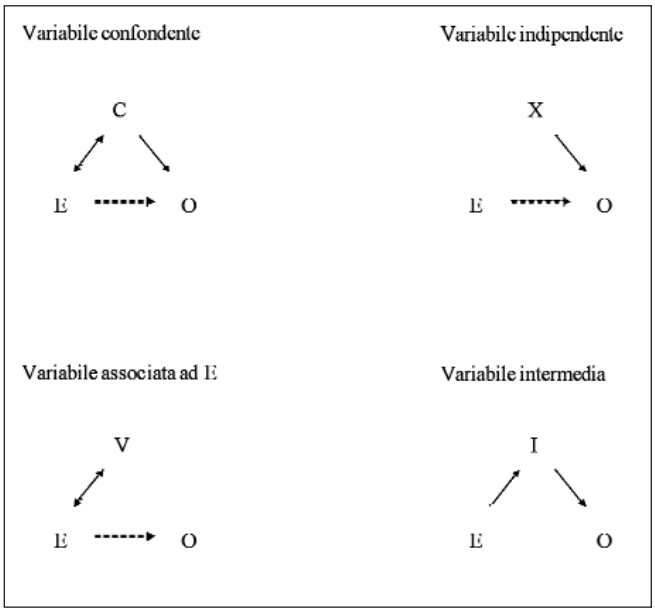

Fig. 4 - "Epidemiologic triangle." *Una variabile con effetto confondente (C) è associata sia alla risposta (O, outcome) che all'esposizione (E). Ciò implica un'associazione multipla che produce un effetto misto e una distorsione (bias) dell'associazione osservata tra E ed O. Nell'esempio della sindrome di Down, l'associazione tra ordine di nascita (E) e malattia (O) è il risultato di un "effetto misto" determinato dall'associazione E-C, ossia tra ordine di nascita ed età materna. Questo effetto produce un'associazione E-O (ordine di nascita e malattia) che in realtà dipende da C (età). Poiché in presenza dell'età l'effetto dell'ordine di nascita scompare, allora l'età è il vero fattore di rischio e non l'ordine di nascita. In molti casi E e C si dimostrano entrambi associati ad O, ma con effetti diversi nei modelli multipli (effetti aggiustati) rispetto alle associazioni uni-variate. Gli altri tre triangoli sottolineano che non c'è confondimento se un ulteriore predittore è associato solo alla malattia (X-O); oppure all'esposizione (V-E). Inoltre, non è un confondente una variabile che spiega completamente l'effetto di E in quanto è una variabile intermedia nel processo con cui E determina O (E-I-O).*

sindrome di Down nel neonato (1). La sindrome di Down, misurata in termini di prevalenza (misura di malattia espressa in numero di casi per 1000 nati), è la variabile di risposta (y) mentre i fattori di rischio considerati, ordine di nascita ed età materna sono le variabili indipendenti o predittori (x_1, x_2). Nella Figura 1 vediamo che esiste una relazione (non lineare in questo caso, ma non importa) tra ordine di nascita e prevalenza di malattia. Inoltre il rischio (prevalenza) di malattia aumenta con l'aumentare dell'età (Fig. 2). Entrambi i fattori predicono il valore di y (prevalenza di malattia) oppure uno confonde l'effetto dell'altro? La Figura 3 chiarisce il problema: la prevalenza di malattia aumenta con l'aumentare dell'età, mentre nell'ambito di ciascun ordine di nascita il rischio di sindrome di Down è altamente variabile (quindi valori maggiori di ordine di nascita non predicono prevalenza maggiore di malattia). L'ordine di nascita non ha un effetto indipendente sul rischio di malattia, ma confonde la relazione tra malattia e il vero fattore di rischio (età materna). A volte il confondente ha un effetto indipendente, ma possiede anche un'azione confondente perché comunque la "forza" dell'associazione del fattore di esposizione è diversa quando se ne tiene conto. Nella Figura 4 è schematizzato il possibile ruolo di una terza variabile nella relazione tra esposizione e malattia (triangolo epidemiologico).

Un esempio di modificazione di effetto (interazione) è offerto dalla Figura 5, che si riferisce a uno studio della relazione tra spazio morto alveolare (y) ed altezza (x_1) in bambini asmatici e non asmatici (x_2) (2). Si nota che all'aumentare dell'altezza lo spazio morto aumenta meno negli asmatici rispetto ai non asmatici. La relazione tra altezza e spazio morto alveolare è modificata in presenza di asma. In termini statistici esiste un'interazione tra altezza e asma (diversa pendenza delle due rette di regressione di y su x). Ritorneremo sul confondimento e sull'interazione con l'esempio della pressione arteriosa dopo avere introdotto il concetto di modello statistico.

Test di verifica

1) Per y funzione di x si intende:
a. Che la variabile y è una variabile indipendente
b. Che i valori di y si modificano al variare di x
c. Che la relazione è lineare
d. Che i valori di x si modificano al variare di y
e. Che y è pari alla somma dei valori di x.

2) Per analisi multi-variabile si intende:
a. Studio della relazione esistente tra una variabile dipendente e più variabili indipendenti
b. Studio di una relazione al netto di eventuali confondenti

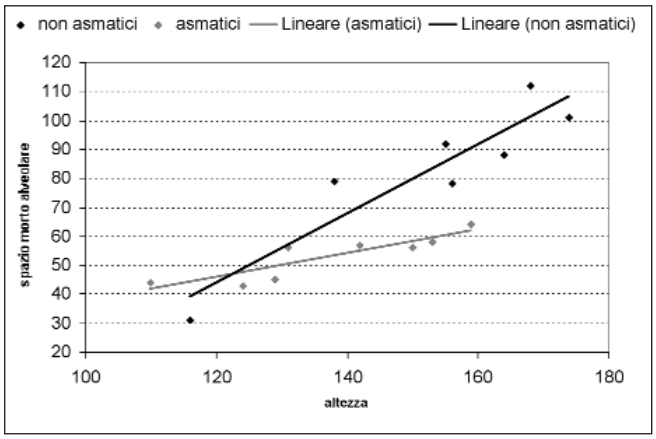

Fig. 5 - *Effetto dell'altezza (cm) sullo spazio morto alveolare polmonare (ml) negli asmatici e nei non asmatici.*

c. Studio di una relazione tenendo conto di modificazioni di effetto (interazioni)
d. Studio della relazione tra due variabili tenendo conto dell'effetto di altre variabili
e. Tutte le precedenti.

3) Lo scopo della regressione multipla è:
a. Stabilire la causa di una malattia
b. Formulare modelli statistici
c. Quantificare l'effetto indipendente delle variabili x coinvolte in una relazione multipla con la variabile dipendente y
d. Escludere effetti causali
e. Quantificare effetti casuali.

La risposta corretta alle domande sarà disponibile sul sito internet www.sin-italy.org/gin e in questo numero del giornale cartaceo dopo il Notiziario SIN

Concetto di modello

Un modello è una rappresentazione della struttura essenziale di un oggetto o un processo reale. Per esempio, la terra è approssimata a una sfera nei calcoli astronomici e geografici anche se è schiacciata ai poli. Tuttavia le imprecisioni generate dai calcoli effettuati sotto l'assunto di sfericità sono sicuramente accettabili e vantaggiose rispetto a precisi calcoli basati su misure reali (vere). L'esempio può essere esteso anche a processi o fenomeni, come il moto di un corpo terrestre o la relazione tra diverse caratteristiche di individui. Per esempio, intuiamo che all'aumentare dell'altezza di un soggetto aumenta anche il suo peso corporeo. Questa intuizione è generata dall'idea di un modello (una funzione) che abbiamo in testa: l'esistenza di una relazione tra peso e altezza (il peso funzione dell'altezza). Pensiamo infatti che il peso aumenti di una quota fissa all'aumentare dell'altezza, anche se questa quantità non permette di stabilire precisamente il peso dopo aver misurato l'altezza. Ciò accade perché ogni modello è *imperfetto* in quanto, essendo una rappresentazione della realtà, non può includere ogni suo aspetto e presuppone delle assunzioni (che devono essere accettabili) sulla struttura essenziale e le correlazioni fra gli oggetti e gli eventi del mondo reale. Ad esempio, la sfericità del pianeta non sarebbe accettabile se la terra fosse un cono. Tuttavia la rappresentazione di un processo per mezzo di un modello, pur essendo semplificata (e imperfetta) rispetto alla realtà, aiuta ad individuare il funzionamento intimo del processo stesso. Con l'uso di modelli possiamo evitare di tener conto di tutte le caratteristiche non rilevanti del fenomeno e concentrarci sui suoi aspetti fondamentali. Ovviamente il modello è tanto migliore quanto meglio simula la realtà, ma è anche tanto più utile quanto più è semplice. Un buon modello è un compromesso accettabile tra il numero di informazioni relative alle conoscenze *a priori* che esso include e gli assunti su cui si fonda.

Modelli statistici

I modelli matematici utilizzati in epidemiologia sono i modelli statistici. Si tratta di modelli simbolici costituiti da equazioni (funzioni) in cui compaiono i parametri coinvolti nella genesi e nell'evoluzione di un fenomeno biologico (una malattia per esempio). Nell'esempio precedente la quantità da moltiplicare per la variazione unitaria di altezza per ottenere la variazione unitaria di peso, è il parametro che ci interessa (la caratteristica che lega il peso all'altezza).

Nella ricerca bio-medica vengono spesso utilizzati i modelli di regressione multi-variabile (meno correttamente multi-variata) perché permettono di studiare contemporaneamente la relazione tra più variabili e sono uno strumento utilissimo per la comprensione del confondimento e dell'interazione tra variabili prognostiche. Cercheremo pertanto di comprendere il significato di questi fenomeni epidemiologici mediante l'utilizzo di modelli statistici. Esistono moltissimi modelli statistici, anche se soltanto alcuni sono ampiamente utilizzati in campo clinico. Per gli scopi delle rassegne che il GIN dedica alla bio-statistica, è utile partire con un approccio pratico ai *modelli lineari generali* perché rappresentano il paradigma dei modelli statistici più utilizzati nell'epidemiologia clinica.

Ma come si sceglie il tipo di modello adatto ai nostri scopi? Per esempio come decidiamo di utilizzare il modello lineare? Lo sviluppo scientifico si fonda sulla ripetizione infinita di un processo a tre stadi successivi: l'osservazione di un fenomeno, la formulazione di un modello per descrivere (spiegare) il fenomeno e l'utilizzo del modello per predire osservazioni successive. I modelli matematici utilizzati allo scopo sono di due tipi: i modelli deterministici e i modelli probabilistici (o stocastici), i primi familiari a chi si occupa di fisica e chimica e i secondi utilizzati da

bio-statistici ed epidemiologi. I modelli deterministici permettono di descrivere e prevedere il valore della variabile dipendente con esattezza (o con piccole imprecisioni dovute ad errori di calibrazione, misura ecc). Infatti il fenomeno deterministico si ripete sempre allo stesso modo se ripetuto nelle stesse condizioni[3]. I fenomeni di interesse biologico sono di tipo probabilistico, in quanto non permettono di predire osservazioni future con assoluta certezza, né di ottenere stime definitive dei parametri. Per questi fenomeni i *modelli probabilistici* sono appropriati in quanto contengono indici di incertezza delle stime dei parametri (le caratteristiche della popolazione stimate nel campione, nel nostro caso i coefficienti delle equazioni) e, quindi, intervalli di probabilità delle predizioni. Il modello statistico (probabilistico) produce una stima dei parametri e un indice di errore di questa stima in base al quale è possibile costruire intervalli di confidenza e test statistici. Ogni modello statistico è infatti costituito da una componente sistematica (da non confondere con l'errore sistematico o bias) e una casuale. La parte *sistematica* è la componente del modello che spiega (o predice) la variabile di risposta, mentre la parte *casuale* è la componente della variabile di risposta non spiegata dal modello. Nell'esempio della pressione arteriosa se, per esempio, dovessimo ottenere, in base al nostro campione, una stima della pressione (y) = 100 + 0.5*anni di età (x_1), allora un soggetto di 35 anni avrà una pressione stimata dal modello di 100 + 0.5*35 = 117.5 mmHg. Questo valore difficilmente sarà il valore realmente registrato (osservato). Un soggetto di 35 anni potrebbe avere una pressione di 125 oppure di 110 mmHg. Significa allora che il modello è sbagliato? No, significa che il modello è utile per riassumere informazioni, ma non è esatto. È un modello probabilistico, ossia un modello in cui esiste anche un termine di errore. L'errore è pari alla differenza tra la y stimata dalla componente sistematica del modello e quella osservata nei dati. La differenza si chiama *residuo* ed esiste nel campione così come nella popolazione da cui il campione proviene. Il termine residuo suggerisce l'idea di una variabilità (di y) che rimane da spiegare dopo che il modello è stato fittato (adattato) ai dati. Il tipo di relazione tra la variabile dipendente (y) e le indipendenti (x_n) è soltanto uno degli elementi in base ai quali si opera la scelta del modello statistico più opportuno per l'analisi dei dati. Gli statistici chiamano questa scelta (e la verifica della sua validità) *specificazione del modello*. Essa dipende fondamentalmente dal disegno dello studio, dalla variabile di risposta che interessa al ricercatore (dalla distribuzione dei valori della variabile di risposta nella popolazione) e dalla "forma" della sua relazione con le variabili indipendenti considerate. Per esempio, per il confronto dell'effetto di diversi trattamenti anti-ipertensivi (variabile indipendente categorica) sui livelli di pressione arteriosa (variabile dipendente quantitativa), possiamo verificare l'esistenza di una relazione lineare tra esposizione (categoria di farmaci) e outcome (valori pressori) e *assumere un modello*, in questo caso lineare, per l'analisi. Il punto di partenza è l'ipotesi che un modello matematico possa descrivere la relazione in esame cercando di utilizzare il modello più semplice e intuitivo possibile.

Test di verifica

1) Assumere un modello significa:
a. Utilizzare un tipo di relazione matematica a cui adattare i dati della ricerca
b. Descrivere dei dati
c. Fare previsioni
d. Che le variabili x entrano al quadrato nell'equazione
e. Stimare parametri per stabilire relazioni causali.

2) Un modello probabilistico si distingue da un modello deterministico:
a. Perché le funzioni dei modelli statistici sono solo lineari
b. Perché è composto da 2 parti: una sistematica e una casuale (errore)
c. Perché il valore della variabile y stimata è sempre uguale a quella osservata
d. Perché è totalmente casuale
e. Perché è sempre causale.

3) La scelta di un modello lineare si basa:
a. Solo sulla forma della relazione tra la variabile di risposta e i predittori
b. Sulla tecnica di regressione lineare multipla
c. Sulla distribuzione dei predittori
d. Non solo sul tipo di relazione della componente sistematica
e. Il modello lineare si può usare sempre perché è il più semplice.

La risposta corretta alle domande sarà disponibile sul sito internet www.sin-italy.org/gin e in questo numero del giornale cartaceo dopo il Notiziario SIN

Riassunto

I modelli lineari generali sono il paradigma di tutti i modelli utilizzati per le analisi di regressione multi-variabile. Nell'epidemiologia clinica è infatti spesso necessario studiare la relazione tra esposizione (un fattore di rischio o

[3] Un esempio di *modello deterministico* è offerto dalla legge di Ohm applicata alla relazione lineare tra flusso di corrente (I) e potenziale elettrico (V) attraverso un conduttore a conduttanza (reciproco della resistenza) nota (y). In questa relazione la variabile dipendente I è funzione della variabile indipendente V secondo il modello I = y*V (con intercetta = 0), dove la quantità "y" è il "parametro della funzione", ossia una quantità da "fissare" per poter applicare una regola generale ad un caso specifico.

un trattamento) ed esiti (malattia o evento di qualsiasi genere) al netto (tenendo conto) dell'effetto di altri fattori potenzialmente associati all'esposizione e all'esito (confondenti) o modificatori dell'effetto dell'esposizione (interazione). Nei modelli statistici la relazione tra esposizione (variabile indipendente o predittore) ed esito (variabile di risposta o dipendente) è una funzione in cui possono intervenire diverse altre variabili indipendenti. La scelta del tipo di modello statistico dipende fondamentalmente dalla forma della relazione esistente tra esposizione e risposta e dalla distribuzione degli errori o residui, la componente non spiegata del modello. Gli errori sono infatti ciò che rimane da spiegare dopo che il modello è stato adattato ai dati, mentre la componente del modello che spiega la variabile di risposta in base ai valori delle variabili indipendenti viene definita componente sistematica del modello.

Indirizzo degli Autori:
Dr. Pietro Ravani
Divisione di Nefrologia e Dialisi
Azienda Istituti Ospitalieri di Cremona
Largo Priori, 1
26100 Cremona
e-mail: p.ravani@libero.it

Bibliografia

1. Stark CR, Mantel N. Effects of maternal age and birth order on the risk of mongolism and leukemia. J Natl Cancer Inst 1966; 37 (5): 687-98.
2. Swinscow TDV. Statistics at square one, ninth Edition, Revised by M J Campbell, University of Southampton, Copyright BMJ Publishing Group 1997 at http://bmj.bmjjournals.com/statsbk/ (last access date 28-9-2004)