Introduzione ai modelli lineari generali

P. Ravani, F. Malberti

Divisione di Nefrologia e Dialisi, Azienda Ospedaliera, Cremona

Introduction of the general linear models

General linear models can be considered the paradigm of all models used in clinical epidemiology. In these models, the independent variables combine in linear fashion to predict the values of the variable response. Since no model predicts the variable response perfectly, an error term is incorporated into the model to acknowledge what remains to be explained after getting a fit to the data. When this error term is normally distributed with constant variance, the linear models are reasonably appropriate to describe the input/output relationship of interest. (G Ital Nefrol 2005; 22: 490-3)

KEY WORDS: Linear model, Gaussian distribution, Residuals, Multivariable regression

PAROLE CHIAVE: Modelli lineari generali, Distribuzione normale, Residui, Regressione multivariabile

Introduzione

In una precedente rassegna abbiamo introdotto i concetti di modello statistico e di regressione multipla. Abbiamo anche anticipato che nella scelta di un modello statistico da utilizzare per "fittare" i nostri dati è necessario considerare sia la forma della relazione tra la y e le x che gli errori tra la variabile predetta dal modello e il suo valore osservato (componente casuale del modello). Per capire il significato di questo processo, è utile partire con un approccio pratico ai *modelli lineari generali* che rappresentano il paradigma dei modelli statistici più utilizzati nell'epidemiologia clinica.

Il modello lineare

Una relazione *lineare* tra due variabili presuppone che una sia funzione lineare dell'altra, ossia vari (aumenti o diminuisca) in modo costante (cioè in base ad un parametro fisso) al variare unitario dell'altra. La funzione lineare può essere definita (con una definizione non esaustiva) una sommatoria dei prodotti delle variabili (x_1, x_2, x_3, …, x_n) alla prima potenza (esponente = 1) per il rispettivo parametro o coefficiente (nota [1]). Lo scopo dell'analisi è quello di stimare i "parametri" (i coefficienti dell'equazione, ad esempio l'effetto del trattamento nella popolazione cui il nostro campione appartiene) e verificare che gli assunti su cui abbiamo fondato la scelta del modello siano stati rispettati. I modelli lineari sono così utili e comprensibili che, anche quando la relazione tra due variabili non è lineare, si cerca di trasformarne i valori in modo che lo diventi con lo scopo di utilizzarli. Come tutti i modelli statistici essi sono validi se non sono violati alcuni assunti su cui essi si fondano. Gli assunti su cui su cui si basa la scelta del modello lineare riguardano sia la componente sistematica che quella casuale del modello. E sono: (1) la relazione tra y e le xn deve essere lineare, ossia in media, il valore di y deve variare (aumentare o decrescere) linearmente al variare del valore dei predittori; (2) gli errori devono essere distribuiti normalmente (distribuzione gaussiana attorno alla retta di regressione, con media = 0 e varianza costante); (3) gli errori devono essere indipendenti, ossia il valore di un residuo (differenza tra y predetta e osservata) non deve in nes-

[1] Più precisamente, una funzione f(x) è lineare se dati due valori (a e b) e una costante (c) risulta che f(a) + f(b) = f(a+b) e cf(a) = f(ca). La forma geometrica della funzione lineare è una retta sugli assi cartesiani, mentre la sua forma algebrica è un'equazione di primo grado del tipo $ax + by + c = 0$. Possono esserci "n" x: $a_1x_1 + a_2x_2 + \ldots + a_nx_n + by + c = 0$. L'incognita è la y che viene stimata in base alle x. La relazione è lineare se le x entrano nell'equazione alla prima potenza. Nel testo utilizziamo il simbolo b_0 per l'intercetta della retta che l'equazione rappresenta (c) e b_1, b_2, …, bn per i parametri delle x.

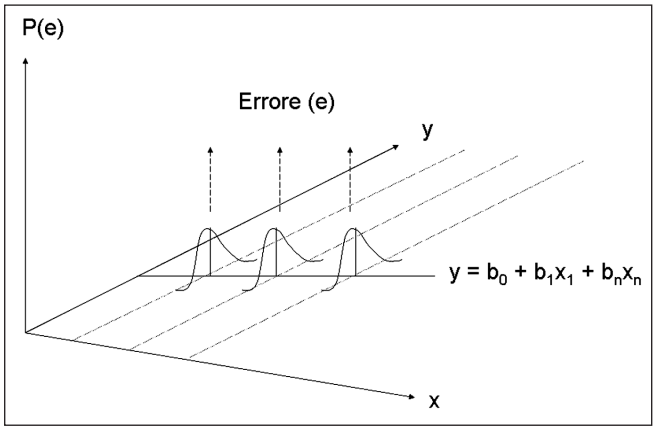
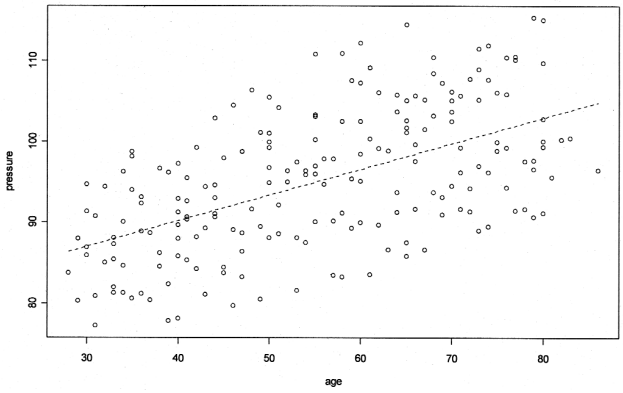


Fig. 1 - *Rappresentazione grafica delle condizioni di validità del modello lineare. La retta rappresenta la relazione lineare tra y e x (assunto di linearità della relazione tra y e x). Le curve gaussiane rappresentano la distribuzione dei residui attorno alla retta dei minimi quadrati (assunto di normalità dei residui). Il valore più probabile è zero (media dei residui = 0) e la varianza dei residui non cambia al variare dei predittori (che possono essere da 1 ad n), assunto di omogeneità della varianza dei residui (in pratica la forma delle curve gaussiane non variano lungo la retta).*

sun modo essere correlato con il valore di un'altra osservazione. Nella Figura 1 vengono schematizzate le condizioni di validità del modello lineare generale. Riprenderemo questi concetti in seguito.

Rappresentazione grafica dell'equazione lineare

Abbiamo precedentemente introdotto alcuni concetti epidemiologici (confondimento ed interazione) e statistici (modello probabilistico e analisi multivariata) che ora dobbiamo approfondire. Utilizziamo un set di dati assolutamente inventato con l'ausilio dei numeri casuali. Abbiamo prodotto 200 osservazioni di valori pressori in 200 soggetti (100 trattati con un farmaco anti-ipertensivo A, 100 trattati con B) con registrazione di alcune variabili indipendenti che possono influenzare la pressione arteriosa. Nella Figura 2 vediamo il grafico di dispersione dei valori pressori sull'età. Si vede una nuvola ellittica in cui è tracciabile una retta con pendenza positiva (i valori di y aumentano all'aumentare dei valori di $x_{età}$). Inoltre i residui sono uniformemente distribuiti attorno alla retta di regressione (lineare semplice) della pressione sull'età. Per comprendere la differenza tra analisi uni- o bi-variata (relazione tra 2 variabili, una y e una x) e multi-variata (relazione multipla tra una y e diverse x), proviamo ad aggiungere un'altra variabile al grafico creando una dimensione aggiuntiva (Fig. 3). Proviamo a mettere in relazione la pressione (y, sulle ordinate), l'età (x_1, sulle ascisse) e il body mass index (x_2, sull'asse zeta). Vediamo che al posto della retta di regressione abbiamo un piano di regressione in mezzo alla "nuvola" dei



Componente sistematica:

pressione descritta dal modello = $\mathbf{b_0 + b_1età}$

coefficienti stimati	SE	t value	p-value
b_0 = 77.4 mmHg	1.90	40.6	<0.001
b_1 = 0.31 anni^{-1}	0.03	9.67	<0.001

Multiple R-Squared: 0.3209, Adjusted R-squared: 0.3175
F-statistic: 93.56 on 1 and 198 DF, p-value: < 2.2e-16

Componente casuale:

pressione osservata = $b_0 + b_1$età + **residuo**

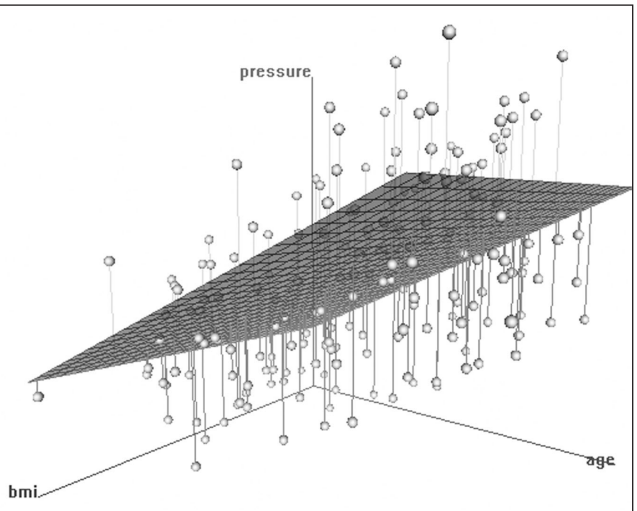
Distribuzione dei residui

Min	1Q	Median	3Q	Max
-13.40	-5.60	-0.16	5.88	16.23

Residual standard error: 7.174 on 198 degrees of freedom

Fig. 2 - *Regressione dei valori pressori (mmHg) sull'età (anni): la componente sistematica del modello riassume le informazioni relative alla variazione delle medie dei valori della variabile dipendente (y) al variare della variabile indipendente (x); la componente casuale del modello riassume le informazioni relative alla dispersione dei dati attorno alla retta di regressione.*

dati. Se aggiungessimo un'ulteriore variabile x_3, avremmo un'iperpiano in uno spazio tetradimensionale. Aggiungendo una quarta variabile x_4 l'iperpiano di regressione si troverebbe in uno spazio a 5 dimensioni e così via. La regressione lineare costruisce i valori stimati della variabile dipendente (y) sulla retta (o piano, o iperpiano) alla distanza minima da tutti i punti osservati. Il metodo di stima dei parametri (coefficienti) di questa retta (o piano, o iperpiano) è detto *metodo dei minimi quadrati* perché il valore delle stime è quello che minimizza i residui. Il metodo funziona (è valido) se sono validi gli assunti che abbiamo elencato in precedenza (e riassunti in Fig. 1). Nella Figura 3 si vede che il piano di regressione, quello dei valori stimati di y, si trova alla distanza minima da tutti i punti (le osservazioni). Ossia alla distanza che minimizza i residui. Sotto i grafici delle Figure 2 e 3 è riportato l'output dell'analisi di regressione con il pacchetto statistico R, gratuitamente scaricabile dal sito CRAN (1), utilizzato per analizzare i dati dei nostri 200 ipertesi. Negli out-puts sono



Componente sistematica:
pressione descritta dal modello = $\mathbf{b_0 + b_1età + b_2BMI}$

coefficienti stimati	SE	t value	p-value
b_0 = 72.6 mmHg	4.69	15.48	<0.001
b_1 = 0.32 anni^{-1}	0.03	9.64	<0.001
b_2 = 0.19 mmHg	0.17	1.109	0.269

Multiple R-Squared: 0.3251, Adjusted R-squared: 0.3183
F-statistic: 47.45 on 2 and 197 DF, p-value: < 2.2e-16

Componente casuale:
pressione osservata = $b_0 + b_1$età + **residuo**

Distribuzione dei residui

Min	1Q	Median	3Q	Max
-13.16	-5.34	-0.3152	5.87	16.14

Residual standard error: 7.169 on 197 degrees of freedom

Fig. 3 - *Piano di regressione dei valori di pressione arteriosa (valori predetti) stimati dal modello con body mass index (BMI) ed età come predittori. I puntini sopra e sotto il piano sono le osservazioni (valori osservati). Le distanze dal piano sono i residui, ciò che resta da stimare dopo aver fittato il modello.*

riportate informazioni relative alle componenti sistematica e casuale del modello statistico. Nella parte finale delle rassegne dedicate ai modelli lineari approfondiremo la parte sistematica (significato dei coefficienti). La parte casuale offre informazioni sulla varianza dei residui (errore *standard* residuo, o radice quadrata della varianza). Questa quantità tende a ridursi quanto meglio il modello si adatta ai dati, quanto più la stima del modello è vicina al valore dei dati osservati. A questo punto bisogna capire il significato epidemiologico dei coefficienti della retta costruita dal modello lineare generale.

Test di verifica

1) La relazione esistente tra y e le x è lineare se:
a. y è funzione lineare delle x
b. y è il risultato dei prodotti delle x tra loro
c. Le variabili x entrano al quadrato nell'equazione
d. Non è possibile fittare i dati
e. Le variabili x sono tra loro in relazione lineare.

2) La scelta del un modello lineare si basa:
a. Sia sul tipo di relazione nella componente sistematica che sulla distribuzione dei residui
b. Sul tipo di analisi consentito dal *software* statistico adottato
c. Solo sulla distribuzione dei valori della variabile di risposta
d. Sulla distribuzione dei valori dei predittori
e. Solo sulla distribuzione dei residui.

3) Gli errori sono:
a. La prova che il modello non è esatto
b. Una ulteriore dimostrazione che la statistica non serve ai nostri scopi
c. La differenza tra il valore della y predetta dal modello e il suo valore osservato
d. Assolutamente utili per poter usare i modelli lineari
e. Raramente distribuiti gaussianamente.

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

Significato epidemiologico delle stime dei parametri del modello lineare

Dopo avere intuito che il modello lineare può essere un buon modello per i nostri dati, vediamo cosa rappresentano le stime dei parametri che abbiamo visto produrre nelle Figure 2 e 3. Nell'esempio del trattamento farmacologico degli ipertesi essenziali assumiamo che i valori di pressione arteriosa si modifichino in modo lineare in base al trattamento farmacologico secondo una relazione descrivibile graficamente da una retta che intercetta l'asse delle ordinate nel punto b_0 ed ha pendenza b_1 sull'asse delle x_1 (se il modello è lineare semplice) o nello spazio multi-dimensionale (se il modello è multi-variato, ossia se più variabili indipendenti x si combinano in modo lineare). La *forma matematica* di questo modello multi-dimensionale (nella popolazione generale e nel nostro campione) è $y = b_0 + b_1x_1 + b_2x_2 + b_nx_n + e$, dove n indica il numero delle variabili indipendenti del modello. Nel modello, e è l'errore, ossia la differenza tra il valore della variabile di risposta predetta dal modello che assumiamo e quello osservato (i nostri dati). I coefficienti b (stime dei parametri) sono le differenze medie dei valori della variabile di risposta y al variare

unitario dei predittori *x*. Se, fondamentalmente in base ad *e* (variabile con media = 0 e distribuzione normale attorno alla retta di regressione), oltre che al disegno dello studio, i nostri dati si adattano bene al modello, allora potremo ragionevolmente fidarci delle conclusioni suggerite dai risultati. Abbiamo imparato che la componente $b_0 + b_1x_1 + b_2x_2 + b_nx_n$ è chiamata *parte sistematica* del modello ("predittore lineare"), mentre *e* è la componente casuale (non spiegata). Per ragioni di spazio, concentreremo la nostra attenzione sul significato dei parametri stimati dal modello lineare generale mentre trascureremo la statistica relativa alla componente casuale (diagnostica dei residui e analisi di sensibilità) dando per scontato che i dati non violino gli assunti del modello e non ci occuperemo del metodo dei minimi quadrati, con cui vengono stimati i coefficienti, e degli errori dei modelli lineari generali.

Abbiamo anticipato che i confronti tra gruppi possono considerare la relazione tra esposizione e *outcome* soltanto (analisi uni-variata) oppure tener conto anche di altri fattori (analisi multi-variabile) al fine di evidenziare l'effetto del predittore sulla variabile di risposta *al netto* dell'effetto dei cosiddetti *confondenti*. Ad esempio, se la pressione arteriosa media è la nostra y e la variabile di gruppo (tipo di trattamento) è x_1, allora il modello $y = b_0 + b_1x_1 + e$, è una semplice analisi della varianza ad una via, detta *One Way ANOVA* (oppure un *t* test se i gruppi sono soltanto 2), o una regressione lineare semplice se la variabile x_1 è continua (esempio, la dose di un farmaco). Questo modello è semplice: il valore della pressione è uguale all'intercetta se x_1 è = 0 e a $b_0 + b_1$ se $x_1 = 1$, mentre è = $b_0 + b_1*x_1$ se x_1 assume valori diversi da 0/1. Purtroppo, però, questo modello non ci permette di conoscere l'effetto di altre variabili (x_2) che potrebbero *confondere* o *modificare* l'effetto della variabile di interesse (x_1). Vedremo successivamente come la regressione multipla ci permette di stimare i parametri della variabile di esposizione principale aggiustati per l'effetto del confondimento e dell'interazione.

Test di verifica

1) Nel modello lineare generale:
a. Esiste solo una componente sistematica
b. La componente casuale è minima
c. È contemplata sia una componente sistematica che casuale
d. La componente sistematica è il residuo
e. La differenza tra le x stimano gli errori.

2) I coefficienti delle x del modello lineare generale:
a. Sono i valori per cui bisogna dividere le x per ottenere la y
b. Rimangono costanti applicando il modello a diversi campioni
c. Sono il risultato del caso
d. Si distribuiscono spesso in modo non-gaussiano
e. Sono la differenza tra i valori medi di y stimato al variare unitario dei valori di x.

3) L'intercetta di un'equazione lineare:
a. Non è la stima di un parametro
b. Non esiste
c. È il valore stimato di y quando le x sono pari a zero
d. È molto utile
e. Non serve.

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 rappresentano il paradigma dei modelli più utilizzati in epidemiologia clinica. In questi modelli le variabili indipendenti o predittori si combinano in modo lineare per predire la variabile di risposta. Dal momento che nessun modello è perfetto esso include un termine di errore che rappresenta ciò che rimane da spiegare dopo che il modello stesso è stato fittato (adattato) ai dati. Quando questo termine di errore è distribuito Normalmente e presenta una variabilità costante lungo la retta di regressione i modelli lineari sono ragionevolmente appropriati per descrivere la relazione tra le variabili esplicative e la variabile di risposta.

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. http://cran.r-project.org/