Giornale Italiano di Nefrologia / Anno 24 n. 1, 2007 / pp. 60-65

Confondimento e interazione nella regressione multipla

P. Ravani, F. Malberti

Divisione di Nefrologia e Dialisi, Azienda Ospedaliera, Cremona

Confounding and interaction in multiple regression

In multiple regression the effect of an input (independent) variable on a continuous output (dependent or response) variable can be adjusted for the effect of confounding and modifying variables. This adjustment is useful to obtain either an unbiased estimate of the true association between an exposure and an outcome or to predict the outcome for given inputs after removing the influence of other factors. These factors are defined as confounders if they are associated with the exposure and are independent risk factors for the outcome, without being intermediates on the biological pathway between exposure and outcome. An interaction between exposure and another independent variable is present when the exposure-disease relationship varies across different values of this variable. Multivariable regression modeling removes the association between the confounder and the outcome eliminating the necessary condition for confounding. An interaction term can be also incorporated into the model to quantify any potential modifying effect. (G Ital Nefrol 2007; 24: 60-65)

KEY WORDS: Confounding, Interaction, Main effects, Effect modification, General linear model

PAROLE CHIAVE: Confondimento, Interazione, Effetti principali, Modificazioni di effetto, Modello lineare generale

Introduzione

Dopo avere introdotto il concetto di modello statistico, di regressione multipla e di modello lineare generale (1, 2), possiamo tornare al confondimento e all'interazione per verificare il significato della stima dei parametri delle variabili confondenti e/o modificatrici di effetto. Torniamo all'esempio dei dati dell'ipertensione arteriosa e applichiamo la regressione lineare. Molti concetti incontrati nelle precedenti rassegne verranno ripresi ed approfonditi. Ricordiamo che utilizzeremo un set di dati relativi a 200 soggetti con valori inventati.

Confondimento

Supponiamo di avere sottoposto a due tipi di trattamento anti-ipertensivo due gruppi di 100 pazienti e di confrontare l'effetto del farmaco A vs. l'effetto di B sui valori di pressione arteriosa media (variabile di risposta quantitativa continua), senza tener conto dell'età e di altre patologie. Nella Tabella I sono riportati i risultati di due modelli lineari: il primo in cui la variabile indipendente è l'età (regressione lineare semplice, con x_1 = età in anni), il secondo in cui il predittore è il trattamento (x_1 = 1 per il trattamento A presente, x_1 = 0 per il trattamento A assente, o trattamento B). Riportiamo per semplicità solo la parte sistematica (con la statistica R^2 che stima la variabilità spiegata dal modello) e non l'errore del modello (ma abbiamo già imparato che l'errore si riduce all'aumentare di R^2, la variabilità spiegata).

Il significato dell'intercetta (b_0) nei modelli contenenti varabili quantitative continue (come l'età) non è interpretabile: rappresenterebbe la media del valore della pressione quando l'età è zero. Andrebbe invece considerato il valore cui aggiungere il prodotto di b_1 e numero di unità di x_1 del soggetto (anni di età). Esempio: il modello dice che un soggetto di 30 anni ha, in media, una pressione media di 77.4 + 0.32*30 = 87 mmHg. Nel secondo modello invece l'intercetta è il valore di pressione nei trattati con B, mentre $b_0 + b_1$ è la media dei valori pressori dei trattati con A (nota [1]).

Dopo questa prima analisi si potrebbe concludere che i

[1] Con il termine media ci riferiamo al valore atteso, di cui la media è la migliore stima disponibile. Ricordiamo che "in media" significa tener conto solo della componente sistematica del modello e non dell'errore.

TABELLA I - MODELLI CON UN SOLO PREDITTORE PER VOLTA (ANALISI UNI-VARIABILE)

1° modello	**Pressione = b_0 + b_1*età**		**(R^2 0.321)**
	Coefficiente (b)	P (t test)	95% conf. int.
Intercetta, b_0	77.4	<0.001	73.6, 81.1
Età (in anni), b_1	0.32	<0.001	0.25, 0.38
2° modello	**Pressione = b_0 + b_1*A**		**(R^2 0.008)**
	Coefficiente (b)	P (t test)	95% conf. int.
Intercetta, b_0	94.3	<0.001	92.6, 96.1
Tratt. A vs. B, b_1	1.53	0.212	-0.88, 3.95

nostri dati confermano la nota relazione tra età e pressione ma non supportano l'esistenza di un effetto anti-ipertensivo del trattamento A verso B.

Se però consideriamo la distribuzione dell'età in base al trattamento vediamo che i trattati con A sono più anziani: età media nel gruppo A 64 anni, in B 47 anni (differenza media 17 anni, da 13.5 a 20.5 anni, P < 0.001). L'età è associata sia alla variabile di risposta che all'esposizione e, pertanto, potrebbe confonderne l'effetto. I *confondenti* sono infatti variabili contemporaneamente associate all'esposizione e all'*outcome* senza essere un passaggio intermedio nel meccanismo con cui l'esposizione determina la risposta. Devono, inoltre, poter essere tenute sotto controllo nel disegno sperimentale. Supponiamo, per esempio, che il gruppo trattato con il farmaco anti-ipertensivo A abbia, in media, valori pressori inferiori al gruppo B, ma nel gruppo B l'età sia minore (come è accaduto nel nostro campione). In tal caso dovremmo escludere che l'età confonda la relazione tra tipo di trattamento e valori pressori in quanto l'età è un potenziale confondente: è associata ad esposizione ed *outcome*, non è un passaggio intermedio nel meccanismo con cui il trattamento agisce sulla pressione ed è controllabile in uno studio sperimentale. Da ciò deriva la necessità di aggiustare l'effetto del trattamento (esposizione) per l'età, ossia depurare l'effetto dell'esposizione da quello del confondente.

Nella Tabella II riportiamo il modello contenente le due variabili (esposizione e confondente).

Questo modello (nota [2]) spiega meglio la variabilità della pressione media osservata nel campione (R^2 è maggiore, quindi il modello si adatta "meglio" ai dati: rimane una minor quota di variabilità di *y* da spiegare dopo che il modello è stato fittato). Inoltre mostra come l'effetto del trattamento sia statisticamente significativo (e clinicamente rilevante): il gruppo trattato con A ha, in media, valori pressori inferiori di oltre 5 mmHg. L'effetto dell'età è maggiore quando si considera insieme all'effetto del trattamento. Pertanto il confondimento comportava, nel nostro caso, una sottostima sia dell'effetto di A che dell'età. Ad esempio, un soggetto di 30 anni, avrà una pressione di 74.6 + 0.42*30 = 87.2 mmHg se non trattato con A e di 87.2 – 5.6 = 81.6 mmHg se trattato con A. Nella Figura 1 sono rappresentate le rette di regressione della pressione sull'età in base al tipo di trattamento.

Test di verifica

1) Un confondente è:
a. Una variabile che cambia valore in modo imprevedibile
b. Una variabile che va esclusa da un modello statistico
c. Una variabile che è associata sia alla risposta che all'esposizione
d. Una variabile che causa la malattia
e. La causa dell'esposizione.

2) Il confondimento:
a. È un fenomeno che non si riduce con l'aumentare della dimensione del campione
b. È di frequente riscontro negli studi clinici
c. È meglio "trattato" nei disegni sperimentali
d. Può costituire un *bias* nel disegno degli studi
e. Tutte le precedenti.

3) I parametri del modello lineare generale stimano:
a. Quanto varia la y al variare unitario della specifica x
b. L'effetto delle interazioni
c. L'effetto dei confondenti
d. L'intercetta
e. L'errore.

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

[2] Questo tipo di modello è chiamato ANCOVA (*analysis of covariance*) ed è utilizzato molto spesso nella ricerca medica.

TABELLA II - MODELLO MULTIVARIABILE CON DUE PREDITTORI INSIEME (L'ESPOSIZIONE DI INTERESSE E UN POTENZIALE CONFONDENTE)

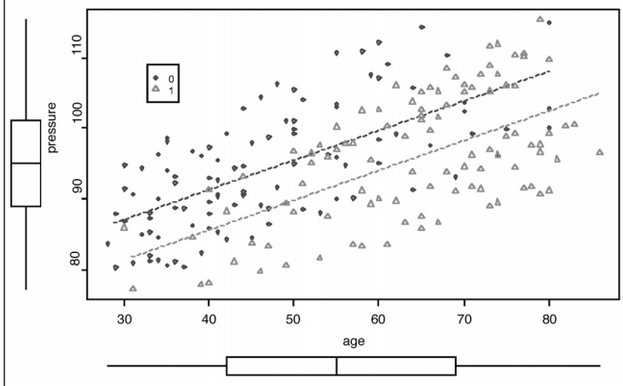
3° modello	**Pressione = b_0 + b_1*A + b_2*età**		**(R^2 0.393)**
	Coefficiente (b)	P (t test)	95% conf. int.
Intercetta, b_0	74.6	<0.001	70.8, 78.3
Tratt. A vs. B, b_1	-5.6	<0.001	-7.88, -3.32
Età in anni, b_2	0.42	<0.001	0.34, 0.49

Interazione

Un'altra possibilità da considerare (sia in presenza che in assenza di un contemporaneo effetto confondente) è l'esistenza dell'*interazione* tra x_1 e x_2. Ossia la presenza di una modificazione dell'effetto di x_1 determinato da x_2. La variabile x_3 (generabile con il prodotto x_1*x_2) può essere introdotta nel modello per testare l'effetto dell'interazione (modificazione di effetto). Pertanto, se i valori di x_2 sono disponibili, un modello del tipo $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$ sarà più conveniente ed interessante ai nostri scopi. Il quesito cui risponde il test dell'interazione è il seguente: verificato che il trattamento A riduce i valori pressori in modo superiore a B indipendentemente dall'età, l'effetto di A è modificato dall'età? Ossia, esiste una modificazione dell'effetto di A dovuta all'età (ovvero, che è la stessa cosa, una modificazione dell'effetto dell'età dovuta al trattamento)? A questa domanda rispondiamo creando la variabile x_3 (interazione) calcolata come il prodotto di x_1 e x_2 (trattamento ed età). La variabile assume valori pari a 0 per i trattati con B e pari all'età nei trattati con A. Il modello verifica se un ulteriore effetto, oltre ad età e trattamento, è presente nei trattati con A rispetto ai trattati con B. Il modello è riportato in Tabella III.

TABELLA III - MODELLO CON I DUE PREDITTORI (EFFETTI PRINCIPALI) E IL LORO TERMINE DI INTERAZIONE (MODIFICATORE DI EFFETTO): TEST DELL'INTERAZIONE

4° modello	**Pressione = b_0 + b_1*A + b_2*età + b_3*interaz.**		**(R^2 0.393)**
	Coefficiente (b)	P (t test)	95% conf. int.
Intercetta, b_0	74.6	<0.001	69.6, 79.6
Tratt. A vs. B, b_1	-5.82	0.185	-14.4, 2.8
Età in anni, b_2	0.41	<0.001	0.31, 0.52
Interaz. (A*età), b_3	0.003	0.958	-0.14, 0.15



Componente sistematica:
pressione descritta dal modello = **b_0 + $b_1$età + $b_2$trattamento**

coefficienti stimati	SE	t value	p-value
b_0 = 74.6 mmHg	1.89	39.4	<0.001
b_1 = 0.42 mmHg per anno	0.03	11.19	<0.001
b_2 = -5.6 mmHg	1.15	-4.85	<0.001

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

Distribuzione dei residui

Min	1Q	Median	3Q	Max
-11.56	-5.63	-1.06	5.6	13.1

Errore standard residuo: 6.796 (197 gradi di libertà)
R2 multiplo: 0.3935, R2 aggiustato: 0.3873
Statistica - F: 63.91 (2 e 197 GDL, p < 2.2e-16)

Fig. 1 - *Regressione lineare dei valori pressori medi sull'età stratificata per trattamento A (assente = 0, cioè trattamento B; presente = 1, cioè trattamento A). Da notare che nelle 2 regressioni lineari semplici (una per strato) la pendenza è la stessa.*

Il modello dice che l'effetto dell'interazione non è statisticamente significativo (e si può escludere dal modello). Pertanto l'età non modifica l'effetto del trattamento. Gli effetti dei termini di interazione sono detti effetti *principali* e vanno sempre mantenuti nel modello insieme al loro prodotto per poter interpretare l'effetto dell'interazione (nota [3]).

L'analisi multi-variata ci permette quindi di conoscere la differenza dei valori della variabile y in base a x_1 aggiustata per l'effetto di x_2 e per l'eventuale presenza di una interazione tra x_1 e x_2 (modificazione dell'effetto di x_1 in presenza di x_2). Il concetto epidemiologico di effetto indipendente dell'esposizione (indipendente da confondimento e interazione) è espresso da un semplice passaggio matematico: $y - (b_2x_2 + b_3x_3) = b_0 + b_1x_1 + e$. In questo modo ritorniamo alla regressione lineare semplice e possiamo conoscere quanto varia la variabile di risposta (stimare il parametro che ci interessa insieme ai suoi intervalli di confidenza) al variare di un solo predittore al netto dell'effetto degli altri. Ossia, la regressione multipla ci permette di stimare il valore della variabile di risposta "depurato dall'effetto del confondimento e dell'interazione".

L'interazione è possibile non solo tra una variabile categorica e una continua, ma anche tra variabili categoriche e tra variabili continue. Vediamo il significato dei coefficienti dei termini di interazione dell'equazione lineare in questi casi.

Supponiamo che nel nostro *trial* metà dei trattati con A e metà dei trattati con B siano diabetici. Vogliamo considerare l'effetto del diabete sulla pressione e, a questo punto, l'interazione tra età e diabete e tra diabete e trattamento. Inoltre consideriamo anche il *body mass index* (BMI) come potenziale confondente e modificatore di effetto.

In base ai nostri dati e al modello assunto le stime dei parametri (b) sono le seguenti (5° modello, finale, con R^2 0.899 Tabella IV):

pressione = b_0 + b_1*A + b_2*età + b_3*DM + b_4*BMI + b_5*D_età + b_6*D_A + b_7*BMI_età

Si vede come la variabilità spiegata dal modello è aumentata notevolmente. Inoltre si conferma l'effetto di A: indipendentemente da età, diabete, BMI e dalle interazioni si associa a riduzione importante della pressione media (di circa 6 mmHg). I diabetici, i soggetti con maggior BMI e gli anziani tendono ad avere valori pressori maggiori. Il significato dei coefficienti dei 3 termini di interazione risultati significativi è il seguente. La retta di regressione dei valori pressori sull'età ha una pendenza aumentata del 30% circa (0.2/0.6) nei diabetici rispetto ai non diabetici. Ossia mentre nei non diabetici la pressione è pari a 0.64 moltiplicato per il numero di anni di età, nei diabetici la pressione è 0.64+0.19 per ogni anno di età. Esiste anche una modificazione di effetto del diabete sulla pressione determinato dal trattamento (e viceversa): la pressione nei diabetici trattati con A è 2.22 mmH in meno rispetto ai diabetici non trattati con A (2.8-2.22 vs. 2.8). Il valore del coefficien-

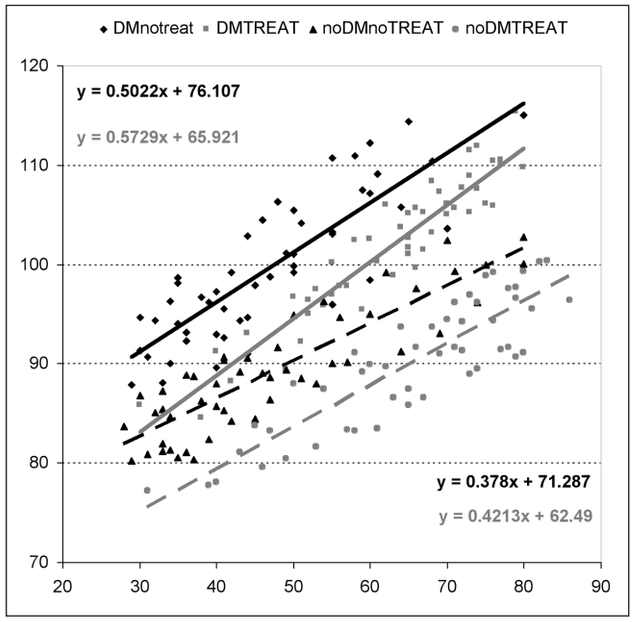


Fig. 2 - *Regressione lineare dei valori pressori medi sull'età stratificata per diabete e tipo di trattamento. Da notare la diversa pendenza delle rette dei diabetici rispetto ai non diabetici.*

TABELLA IV - MODELLO FINALE CON I DUE PREDITTORI E I LORO TERMINI DI INTERAZIONE (DI SECONDO ORDINE)

	Coefficiente (b)	**P (t test)**	**95% conf. int.**
Intercetta, b_0	53.8	<0.001	43.2, 64.4
Tratt. A vs. B, b_1	-6.2	<0.001	-7.5, -4.9
Età in anni, b_2	0.64	<0.001	0.45, 0.83
Diabete (si vs. no), b_3	2.8	0.013	0.62, 5.1
BMI, b_4	0.71	0.002	0.25, 1.16
Int. (diab*età), b_5	0.19	<0.001	0.14, 0.24
Int. (diab*A), b_6	-2.22	0.024	-4.15, -0.29
Int. (BMI*età), b_7	-0.010	0.011	-0.018, -0.002

[3] L'assenza di significatività dell'interazione dimostra che il modello è *additivo*, ossia l'effetto dovuto alla presenza di due covariate è uguale alla somma degli effetti di ciascuna. Quando è presente un'interazione (test sul parametro significativo) allora l'effetto dovuto alla presenza di due covariate è inferiore o superiore alla somma degli effetti di ciascuna di esse.

te dell'interazione tra trattamento e diabete va sommato a b_0 (come quello del trattamento e del diabete), mentre il coefficiente dell'interazione tra diabete ed età va sommato a quello dell'età (modifica la pendenza della retta) in presenza di diabete (è zero nei non diabetici). Il coefficiente delle due variabili continue BMI ed età va sommato a quello di età e BMI (sia nei diabetici che nei non diabetici). Andrebbe sommato solo in caso di diabete se fosse risultata significativa l'interazione BMI*età*diabete (mantenendo comunque nel modello anche il termine a due fattori BMI*età) (nota [4]).

Pertanto le possibili rette tracciabili in base al modello finale sono:

Diabetici trattati con A:
pressione = $(b_0 + b_1 + b_3 + b_6) + (b_2 + b_5)$*età + b_4*BMI + b_7*(età*BMI)

Diabetici trattati con B:
pressione = $(b_0 + b_3) + (b_2 + b_5)$*età + b_4*BMI + b_7*(età*BMI)

Non diabetici trattati con A:
pressione = $(b_0 + b_1) + b_2$*età + b_4*BMI + b_7*(età*BMI)

Non diabetici trattati con B:
pressione = $b_0 + b_2$*età + b_4*BMI + b_7*(età*BMI)

Da notare che il coefficiente dell'interazione nei modelli lineari rappresenta "la differenza delle differenze": la differenza che esiste tra diabetici esposti ad A vs. B dopo aver considerato l'effetto principale del diabete; la differenza di pendenza dopo aver considerato la pendenza dell'età. Ecco perché gli effetti principali sono indispensabili per interpretare il significato dell'interazione.

Nella Figura 2 possiamo vedere l'analisi multi-variata della relazione tra pressione ed età. Nei grafici è stata adottata la stratificazione invece della regressione multipla (per rappresentare le relazioni su due dimensioni). Il risultato conferma la diversa pendenza delle rette in presenza e assenza dello stato diabetico e la differenza di intercetta dovute al diabete e al trattamento. Nella stratificazione, tuttavia, si perde potenza in quanto la regressione è condotta su 4 gruppi di 50 pazienti. Con l'aumentare delle variabili indipendenti aumentano gli strati e si riduce la potenza dei tests. I risultati di una regressione multipla, invece, ci consentono di costruire uno spazio immaginario multi-densionale (un iperpiano) con tante dimensioni quante sono le variabili del modello sviluppato tenendo conto di tutte le osservazioni del campione. L'iperpiano di regressione giace in questo spazio multi-dimensionale.

Test di verifica

1) Per interazione si intende:
a. L'interazione tra la y e le x del modello
b. La modificazione dell'effetto di una x sulla y in base al valore di un'altra x
c. La violazione degli assunti del modello lineare generale
d. La variazione della cinetica di un farmaco
e. L'effetto causale di una variabile indipendente.

2) L'interazione è un fenomeno che:
a. È meglio trascurare
b. Non serve ai nostri scopi
c. Non ha impatto clinico
d. Non ha mai un chiaro significato clinico
e. È sempre necessario considerare nella costruzione di un modello statistico.

3) L'interazione e il confondimento:
a. Come i modelli stessi, sono fenomeni da ricercare e studiare su basi cliniche
b. Possono aiutare a spiegare meglio il significato della relazione tra diversi fattori
c. Aiutano, se presenti, ad adattare meglio il modello matematico ai dati
d. Sono fenomeni quantificabili attraverso la regressione multipla
e. Tutte le precedenti.

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

Generalizzazione del modello lineare

I concetti discussi a proposito dei modelli lineari sono generalizzabili per modelli matematici di "forma" diversa (non lineare). Molti studi epidemiologici generano dati in cui la variabile di risposta non è quantitativa ma è un *outcome* binario, ossia una variabile con 2 valori possibili (comparsa o meno di un evento come morte, infarto, guari-

[4] L'interpretazione dei coefficienti di due variabili continue e del loro termine di interazione è complicata dalla difficoltà di una loro rappresentazione grafica. Dovremmo immaginare un certo numero di soggetti della stessa età ma diverso BMI e un gruppo dello stesso BMI ma di diversa età. Allora, secondo il modello, noi ci aspettiamo che la pressione aumenti di b_4 mmHg per ogni unità di BMI tra soggetti della stessa età e di b_2 mmHg per ogni anno di età tra soggetti dello stesso BMI. Ogni incremento va corretto per il valore del coefficiente dell'interazione (b_7) che rappresenta la differenza di pressione per unità di BMI quando considero la differenza per unità di età (e viceversa per unità di età quando considero la differenza per unità di BMI). I coefficienti sono stimati dal modello anche se nessuno dei soggetti ha esattamente lo stesso BMI o la stessa età.

gione) oppure è una conta (numero di ospedalizzazioni, decessi, ecc.). In questi casi le regressioni più utilizzate sono la regressione logistica, la regressione di Poisson (modelli lineari generalizzati) e quella di Cox (modello semiparametrico). I concetti di modello probabilistico, di esposizione, confondente e interazione sono validi anche in questi casi. Ciò che cambia è la variabile di risposta (e quindi la forma del modello da utilizzare) e il significato dei coefficienti delle variabili indipendenti (le *x*) che compaiono nell'equazione. I concetti discussi nelle rassegne dedicate ai modelli lineari e i principi validi per i modelli lineari generali sono validi anche in questi modelli probabilistici.

Riassunto

Nella regressione multipla l'effetto di una variabile indipendente su una variabile di risposta (dipendente) continua può essere aggiustato per l'effetto di confondenti e modificatori di effetto. Questo è molto utile per ottenere stime non distorte della vera associazione tra esposizione e malattia o per predire l'esito conoscendo i valori del predittore al netto dell'influenza di altri fattori. Questi fattori sono definiti confondenti se sono associate all'esposizione e all'*outcome* senza essere un passaggio intermedio nel meccanismo patogenetico con cui l'esposizione influenza l'*outcome*. Un'interazione tra esposizione e altri fattori è presente se l'effetto dell'esposizione cambia a seconda del valore assunto da questi fattori. L'analisi di regressione multipla permette di rimuovere l'associazione tra confondente e *outcome* eliminando la condizione necessaria per il confondimento. Un termine di interazione può inoltre essere incluso nel modello per quantificare eventuali modificazioni di effetto.

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@ospedale.cremona.it

Bibliografia

1. Ravani P, Malberti F. Introduction to the general linear models. G Ital Nefrol 2005; 22: 490-3.
2. Ravani P, Malberti F. Statistical models and multivariable analysis. G Ital Nefrol 2005; 22: 348-53.

Testi utili per approfondimenti

Campbell M. Statistics at square two. BMJ books, 2001. Described at: http://www.bmjbookshop.com/shop/product_display.asp?productid=0727913948&productname=Statistics+at+Square+Two

Glantz S, Slinker BK. Appled regression and Analysis of Variance, McGraw-Hill, Inc. 2001, second edition. Described at. http://www.vetmed.wsu.edu/AppliedRegression/