L'errore di misura negli studi retrospettivi in epidemiologia (1)

Measurement errors in Epidemiological Retrospective Studies

Monica Ferraroni
Dip. di Med. Chirurgia e Odontoiatria - Università di Milano Via A. Di Rudinì, 8 I-20142
Adriano Decarli
Istituto di Statistica Medica e Biometria - Università di Milano Via Venezian, 1 I-20133
Maura Mezzetti
Istituto di Metodi Quantitativi, Università Bocconi Via Gobbi, 5 I-20136 Milano

Abstract: The relationships between a set of covariates and a response variable Y are usually quantified by means of regressive methods. The accuracy of this quantification depends on the precision of covariate measurements. However, in many studies only a subset of explanatory variables Z is observable without error, while for some other variables X the information is not directly available. Instead of X, mis-measured surrogate variables W related to X are observable, and consequently the parameters relating Y and X cannot be directly estimated. Measurement errors (ME) related to surrogate variables may produce biased results. We propose a procedure which adjusts the estimated association for the effects of misclassification ascribable to ME. The performance of the procedure is evaluated by simulation.

Parole chiave: Measurement error, case-control study.

1. Premesse

In questo lavoro verranno trattati problemi relativi alla correzione per l'errore di misura (ME) utilizzati nell'analisi di studi di tipo caso-controllo, in cui il legame fra la variabile di risposta Y e i suoi predittori sia definibile tramite una relazione funzionale $Y = f\left(\beta_z' Z, \beta_x' X\right)$ con Z insieme dei predittori misurati senza errore e X insieme dei predittori la cui misurazione diretta senza errore non risulta possibile. In questi studi la variabile Y è un indicatore che generalmente assume valore 1 per i "casi" (soggetti malati) e 0 per i "controlli" (soggetti sani). La misura vera di X non è usualmente nota, mentre è osservata W, associata a X, chiamata "surrogato" di X. I parametri del modello non sono direttamente stimabili e la correzione per ME passa attraverso le stime indirette di tali parametri, ottenute tramite un modello in cui Y risulta funzione di (Z,W). La procedura di stima tramite la massima verosimiglianza prevede la determinazione della distribuzione congiunta di Y e W dato Z, che per $\Pr(Y|W,X,Z) = \Pr(Y|X,Z)$ è possibile scrivere come:

$$f_{Y,W|Z}(y,w|z) = \int_x \Pr(y|\beta, x, z) \times f_{W,X|Z}(w, x \mid z)\, dx \qquad (1)$$

(1) Il presente lavoro è stato finanziato dal generoso contributo dell'AIRC.

dove β è il vettore dei parametri che esprime la relazione tra la risposta Y e l'esposizione e $f_{W,X|Z}(w,x\,|\,z)$ necessita di essere definita attraverso un modello che specifichi ME[1].

2. Modelli per la definizione dell'errore di misura

I modelli per la definizione di ME vengono generalmente distinti in due classi:
'Error Models': in cui W rappresenta la vera osservazione contaminata da un errore assunto indipendente da (X,Z) e la distribuzione congiunta di W e X dato Z è definita da $f_{W,X|Z}(w,x|z)=f_{W|X,Z}(w|x,z)\times f_{X|Z}(x|z)$. La (1) diventa quindi:

$$f_{Y,W|Z}(y,w|z)=\int_x \Pr(y|\beta,x,z)\times f_{W|X,Z}(w\,|\,x,z)\times f_{X|Z}(x\,|\,z)\,dx \qquad (2)$$

Le tre componenti della (2) rappresentano rispettivamente: la relazione tra malattia e vera esposizione; il modello per ME, la distribuzione della variabile X, non osservabile. In caso di errore non differenziale, la distribuzione condizionata $f_{W|X,Z}(w|x,z)$ non dipende da Y. Questo tipo di modello è sovente utilizzato negli studi epidemiologici quando sia disponibile uno studio di validazione esterno, cioè uno studio in cui per un gruppo di soggetti "sani", non inclusi nello studio principale siano contemporaneamente disponibili la misura "vera" X e la misura "surrogato" W.
'Regression Calibration Models': in cui si specifica la distribuzione condizionale di X dato (W,Z), ad es: $X=W+Z+U_*$ con $E(U_*\,|\,W,Z)=0$. In questo caso la distribu= zione congiunta diviene $f_{W,X|Z}(w,x|z)=f_{X|W,Z}(x|w,z)\times f_{W|Z}(w|z)$, e la (1) risulta:

$$f_{Y,W|Z}(y,w|z)=\int_x \Pr(y|\beta,x,z)\times f_{X|W,Z}(x\,|\,w,z)\times f_{W|Z}(w\,|\,z)dx \qquad (3)$$

La terza componente dell'espressione (3) non dipende da X. E' quindi possibile stimare β dalla (3) massimizzando la verosimiglianza di:

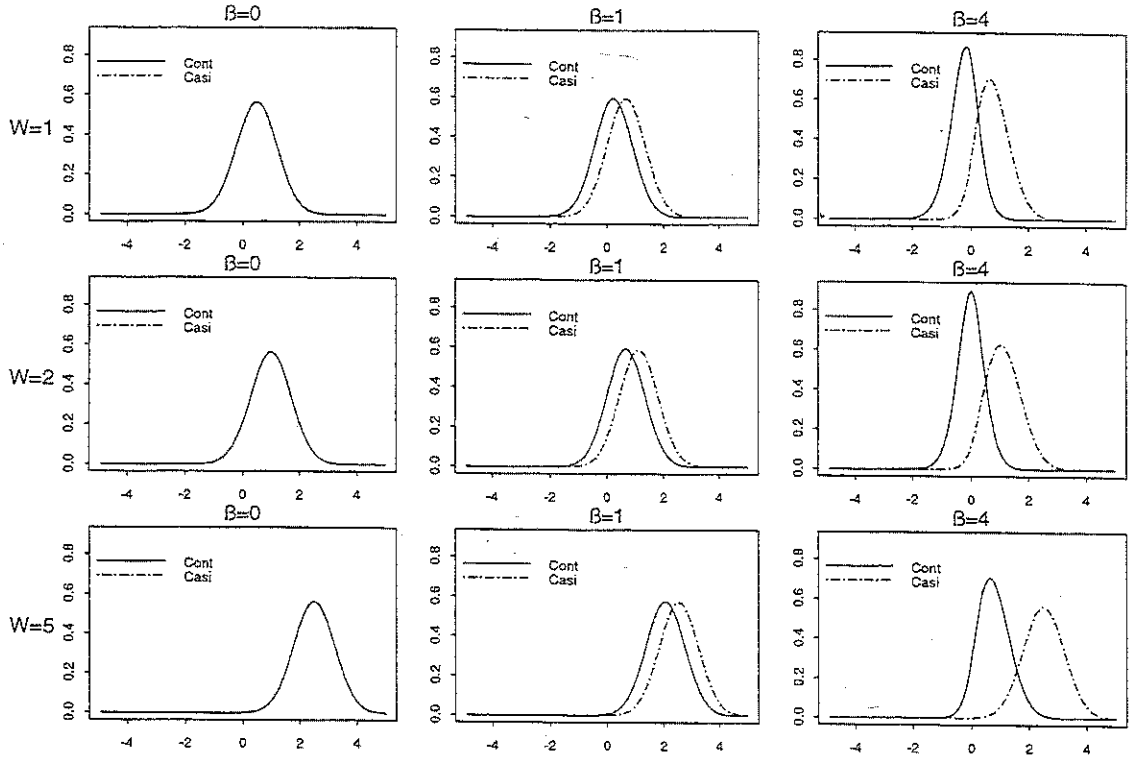
$$f_{Y|W,Z}(y\,|\,w,z)=\int_x \Pr(y|\beta,x,z)\times f_{X|W,Z}(x\,|\,w,z)\,dx \qquad (4)$$

Con questo tipo di approccio nello studio principale si stimano i parametri di interesse con gli usuali modelli sostituendo W con $E(X\,|\,W,Z)$ stimato con l'ausilio di uno studio di validazione. In questa classe di modelli la distribuzione $f_{X|W,Z}(x|w,z)$ dipende in realtà anche da Y. La distribuzione congiunta di W e X in assenza di covariate Z e con una sola variabile X, può essere scritta come $f_{W,X}(w,x)=f_{W,X}(w,x=w+u_*)=$ $=f_W(w)\times f_{u_*}(x-w)$ per cui la distribuzione condizionata è:

$$f_{X|W,Y}(x|w,y)=\frac{P(y|x)\times f_{U_*}(x-w)}{\int_x P(y|x)\times f_{U_*}(x-w)\,dx} \qquad (5)$$

Nell'ipotesi in cui $f_{U.}(x-w) \sim N(0,1)$, con $P(Y=1 \mid x)$ espressa da una funzione logistica, la rappresentazione grafica della (5) per alcuni valori del parametro β e della variabile surrogato W è riportata nella figura 1. Solo per β=0 la distribuzione condizionata è la stessa tra i due gruppi mentre al crescere di β aumenta la diversità tra le due distribuzioni. Ciò rende più difficile utilizzare uno studio di validazione esterno, nella fase di correzione per ME. Lo studio di validazione esterno permette infatti una stima di $f_{X|W,Y}(x \mid w, y)$ solo per y=0 (soggetti sani).

Figura 1: *$f_{X|W,Y}(x \mid w, y)$ al variare del valore del parametro β per fissi valori di W.*



3. Un modello iterativo

Per stimare in modo corretto gli effetti di interesse (β_x) in generale sono stati proposti metodi appartenenti alla classe dei '*Regression Calibration Models*' che ipotizzano $f_{X|W}(x \mid w)$ uguale tra i due gruppi. La procedura proposta non ipotizza $f_{X|W,Y}(x \mid w, y) = f_{X|W}(x \mid w)$ e consente di giungere ad una stima di β in modo iterativo quando sia disponibile uno studio di validazione esterno in cui W e X siano congiuntamente rilevati. La procedura è costituita dai seguenti passi:

1. Stima di β utilizzando i dati relativi allo studio principale, attraverso la regressione logistica, come avviene usualmente in questi studi.

2. Stima per ogni soggetto di $E(X|W,Y) = \int_x x\, f_{X|W,Y}(x|w,y)\,dx$, utilizzando le infor=

mazioni relative allo studio di validazione, con $f_{X|W,Y}(x|w,y)$ espressa dalla (5) in cui $P(y \mid x)$ è sostituito dalle stime ottenute in 1.

3. Stima di β_X come in 1. utilizzando i valori $E(X|W,Y)$ ottenuti in 2.:

$$P(Y|E(X|W,Y)) = \frac{\left[\exp\left(\alpha + \beta_x^{(i)} E(X|W,Y)\right)\right]^y}{1+\exp\left(\alpha + \beta_x^{(i)} E(X|W,Y)\right)} \qquad [y=0,1]$$

4. Ripetizione dei punti 2. e 3 finché $\left|\beta_X^{(i)} - \beta_X^{(i-1)}\right| \sim 0$

4. Validazione dei modelli di correzione per l'errore di misura.

Il metodo proposto è stato saggiato analizzando una serie di 500 studi tipo caso-controllo generati da una popolazione in cui erano noti la distribuzione di X, la relazione funzionale tra X e Y e la struttura dell'errore che lega W a X. I risultati ottenuti sono stati confrontati con quelli ottenibili con altri metodi presenti in letteratura.

In tabella 1 si riportano il valor vero di β, quello stimato senza correzione per ME $(\hat{\beta}_W)$, quello ottenuto con la correzione proposta da Rosner[2] $(\hat{\beta}_1)$, con la correzione proposta da Reeves[3] $(\hat{\beta}_2)$ e quello ottenuto con il metodo iterativo esposto nella sezione 3 $(\hat{\beta}_3)$. Il modello proposto compensa solo in parte la sottostima dovuta a ME, ma sembra fornire risultati migliori del modello proposto da Rosner. Il modello di Reeves fornisce una stima più prossima al valor vero, benché sovrastimata e con valore più elevato della deviazione standard. Il metodo proposto sarà applicato a dati relativi ad uno studio caso-controllo su abitudini alimentari e tumore alla mammella.

Tabella 1: *Confronto tra valor vero* (β), *valore stimato senza correzione* $(\hat{\beta}_W)$, *valore stimato con il metodo di Rosner* $(\hat{\beta}_1)$, *il metodo di Reeves* $(\hat{\beta}_2)$ *e con il metodo proposto* $(\hat{\beta}_3)$. *Risultati ottenuti su 500 simulazioni (deviazioni standard in parentesi).*

β	$\hat{\beta}_W$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
0.01790	0.0111	0.0132	0.0184	0.0152
	(0.0018)	(0.0021)	(0.0024)	(0.0019)

Riferimenti bibliografici

[1]Carroll R.J., Gail M.H., Lubin J.H. (1993) Case-control sudies with errors in covariates, *Journal of American Statistical Association,* 88, 185-199.

[2]Reeves G.K., Cox D.R., Darby S.C. and Whitley E. (1999) Some aspects of measurement error in explanatory variables for continuous and binary regression models, *Statistics in Medicine*, 17, 2157-2177.

[3]Rosner B.A. (1996) Measurement error models for ordinal exposure variables measured with error. *Statistics in Medicine*, 15, 293-303.

SIS 2000
@ds.unifi.it

Società Italiana di Statistica

XL RIUNIONE SCIENTIFICA

Firenze, 26-28 aprile 2000

Sessioni plenarie e specializzate: sintesi
Sessioni comunicazioni spontanee
Sessioni satellite