Report n.34

VARIABILI LATENTI E "SELF-SELECTION" NELLA VALUTAZIONE DEI PROCESSI FORMATIVI

Enrico GORI

Pisa, maggio 1991

1. Introduzione

La misura dei rendimenti e dell'efficacia dell'istruzione riveste un particolare interesse nell'ambito dell'economia dell'istruzione. Rinviando alla letteratura specifica (Hanushek, 1979) per una piu' completa discussione di tale concetto, si rileva che se da un lato e' possibile accettarne la definizione intuitiva di "*misura del contributo fornito al raggiungimento di un certo risultato*", dall'altro l'ambiguita' del termine "*risultato*" mette in luce le difficolta' di trovare soluzioni soddisfacenti per il problema. Ad esempio, infatti, gli economisti potranno essere interessati a risultati quantificabili monetariamente, quali il reddito, mentre i sociologi tenderanno a sottolineare maggiormente l'importanza di aspetti inerenti la sfera affettiva e dei rapporti sociali, ecc.. In generale, dunque, si possono dare tante definizioni di efficacia dell'istruzione quante sono le misure di risultato che si possono scegliere.

Queste ultime vengono usualmente distinte in misure interne ed esterne al sistema di istruzione. Per *esterni* si intendono quei risultati che lo studente raggiunge grazie al contributo dell'istruzione, ma fuori del suo ambito: tipici esempi sono il reddito ottenibile in relazione ad un certo numero di anni o tipo di istruzione, ma anche la probabilita' di trovare un lavoro soddisfacente, o, con riferimento all'istruzione media superiore, la capacita' di concludere studi universitari. Si dicono invece *interni* i risultati che lo studente raggiunge nell'ambito dell'istruzione: di questo gruppo fanno parte le votazioni conseguite in particolari esami o test, oppure le chance di conclusione degli studi fornite dal particolare istituto che lo studente frequenta. Questo secondo tipo di misure, se da un lato ha il pregio di potere essere osservato con facilita' e in campioni molto numerosi, dall'altro presenta il rischio di essere circolare e poco "oggettivo": infatti una scuola che volesse aumentare il proprio output, e quindi la sua efficacia, potrebbe farlo semplicemente riducendo il grado di selezione esercitato nei confronti degli studenti. L'uso di tali misure si giustifica pertanto solo in presenza di metodi di selezione e graduazione oggettivi, in cui siano costanti tra gli studenti gli argomenti d'esame ed i metodi di valutazione delle risposte. La difficolta' nel reperimento di misure esterne puo' comunque giustificare il ricorso a misure interne a patto che si tengano sempre presenti i limiti di cui sopra e si valutino attentamente la portata delle

eventuali incomparabilita' nelle scale di misura adottate e disomogeneita' dei risultati.

Un altro elemento su cui esiste una certa unanimita' di consensi e' che i risultati dell'istruzione, qualsiasi definizione se ne dia, vanno valutati a livello individuale, sia perche' in ultima istanza e' l'individuo che ne fruisce, sia perche' come hanno rilevato alcuni autori (Aitkin e Longford, 1986) tali risultati possono variare da individuo ad individuo: in particolare una stessa istituzione scolastica puo' essere molto efficace (nel senso di incrementare il risultato) di determinati individui, ma non di altri. Ne deriva che studi dell'efficacia dell'istruzione devono essere condotti a livello individuale. L'efficacia, inoltre, e' un concetto relativo la cui valutazione e misura scaturisce dal confronto tra istituzioni scolastiche in luoghi e/o tempi diversi. Ogni valutazione di processi formativi deve quindi basarsi su dati relativi a studenti in istituzioni temporalmente e/o spazialmente differenziate.

Cio' premesso, un modo naturale per valutare l'efficacia di un processo formativo e' quello di fare riferimento ad un modello in cui il risultato Y_n relativo all'individuo n e' "spiegato" da variabili individuali X_n e da un effetto istituzione, o tipo di istruzione, I_n, ovvero:

$$Y_n = \mathbf{x}_n \beta + \alpha I_n + U_n \qquad (A)$$

dove I_n rappresenta una variabile "dummy" che, ad esempio, vale 1 se l'individuo usufruisce dell'istruzione del tipo 1, 0 altrimenti. In questo modello il coefficiente α misura l'efficacia relativa dell'istituzione di tipo 1 rispetto alla situazione alternativa, ed e' generalizzabile al confronto tra un numero qualsiasi di istituzioni.

Ora e' un fatto ormai ampiamente riconosciuto che nelle scienze sociali l'impossibilita' di usare metodi sperimentali al fine di variare indipendentemente i trattamenti per eliminare o isolare veicoli di causalita' spuri, pone seri limiti alla possibilita' di una conoscenza oggettiva (Heckman e Robb, 1986). Cio' si verifica anche nella valutazione dei processi formativi e nella misura dell'efficacia dell'istruzione.

A parte l'estrema semplificazione del modello (A), infatti, utilizzata in pratica nei primi studi relativi all'istruzione (Hanushek, 1979), esso presuppone un'ipotesi del

tipo

$$E(U_n \mid I_n) = 0 \qquad (B)$$

tipica delle situazioni sperimentali controllate. In tali situazioni essa viene assicurata, generalmente, dall'assegnazione degli individui a caso tra i trattamenti, in modo che non sussistano correlazioni tra trattamento e fattori individuali inosservabili. La' dove tale condizione non risulti rispettata l'utilizzazione di tecniche "standard" conduce a stime distorte e inconsistenti dei parametri e in particolare dell'effetto trattamento.

E' chiaro ora che questa ipotesi non puo' essere generalmente rispettata nell'ambito del fenomeno esaminato. Nell'istruzione, infatti, lo studioso non ha mai accesso a osservazioni di tipo sperimentale. E' ovvio infatti che non e' lui ad assegnare gli individui a caso tra le istituzioni scolastiche che si intendono valutare, ma sono gli individui stessi che scelgono a quale "esperimento" partecipare. Questo non sarebbe un problema se lo studioso fosse in grado di specificare la relazione (A) in modo da rispettare la condizione (B). Il fatto e' che spesso molte delle variabili che sono rilevanti per (A) non sono note, o misurabili dal ricercatore e, purtroppo, sono sovente anche quelle che influenzano l'assegnazione dell'individuo al "trattamento": questo puo' condurre alla violazione della condizione (B). Risulta allora chiara la differenza rispetto alla situazione delle scienze sperimentali: in tale ambito all'ignoranza di variabili rilevanti per (A) si pone rimedio attraverso l'assegnazione casuale degli individui ai trattamenti, il che rende soddisfatta appunto (B).

Nell'istruzione, invece, l'impossibilita' di "randomizzare" rispetto alle variabili non osservabili, produce simultaneita' tra il fattore, U_n, ed il "trattamento" I_n, con conseguente distorsione nella stima dell'efficacia, e nell'interpretazione dei nessi causali tra fenomeni. Numerosi autori hanno fornito spiegazioni anche piu' sofisticate e particolari della simultaneita' nell'ambito dei fenomeni inerenti l'istruzione, sia con riferimento a misure esterne di tipo economico quali il reddito (Garen, 1984; Heckman e Robb, 1986), sia con riferimento a misure interne quali le votazioni (Murnane *et al.* 1985; Heckman e Robb, 1986; Gamoran e Mare, 1989). Scopo di questa nota e' quello di mostrare, piu' semplicemente, come la mancata considerazione di variabili quali la

motivazione dello studente, difficilmente misurabili e che di conseguenza finiscono insieme ad altri fattori inosservabili nel termine di errore, possono indurre simultaneita' del tipo descritto.

La soluzione a tali problemi consiste nello specificare un modello capace di spiegare la scelta del tipo di istruzione (trattamento) da parte dell'individuo, e nel considerarlo simultaneamente al modello (A): cosi' facendo, a condizione di specificare correttamente il modello della scelta, si pone rimedio ai problemi derivanti dall'impossibilita' di assegnare le unita' di osservazione ai trattamenti in modo "casuale".

In particolare il problema sara' affrontato con riferimento al caso in cui Y rappresenti un carattere qualitativo, ovvero il successo o insuccesso dello studente nel raggiungimento di un titolo di studio.

2. Modello simultaneo scelta-esito

Si supponga di volere mettere a confronto due tipologie di istituti medi superiori, ad esempio liceo rispetto alla scuola professionale, utilizzando quale misura (esterna) il loro apporto alla probabilita' di laurea di studenti iscritti ad una medesima facolta'. Si puo' assumere che la scelta del tipo di istituto da parte dello studente sia rappresentabile da un modello del tipo

$$I_n^* = Z_n \gamma + \epsilon_n , \qquad E(\epsilon_n \mid Z_n) = 0 \tag{1}$$

$$I_n = \begin{cases} 1 \text{ se } I_n^* > 0 \text{ (liceo)} \\ 0 \text{ se } I_n^* \leq 0 \text{ (professionale)} \end{cases} \tag{2}$$

dove Z_n sono variabili esplicative della scelta (reddito, ceto sociale, sesso ecc.), che si assumono esogene. Si suppone poi l'esistenza di una variabile latente, che rappresenta la capacita' dello studente, che determina il suo esito nella facolta':

$$Y_n^* = \mathbf{x}_n \beta + \alpha I_n + U_n, \qquad E(U_n \mid \mathbf{x}_n) = 0 \tag{3}$$

$$Y_n = \begin{cases} 1 \text{ se } Y_n^* > 0 \text{ (laurea)} \\ 0 \text{ se } Y_n^* \leq 0 \text{ (abbandono)} \end{cases} \tag{4}$$

Cio' significa ipotizzare che la capacita' individuale dipenda dall'istituto I_n, da caratteristiche individuali $\mathbf{x}_n$ osservabili, esogene, e da caratteristiche individuali inosservabili U_n. Si precisa che una tale formulazione e' alquanto restrittiva. Infatti nella (3) l'effetto "istituto" si produce solo attraverso una traslazione della capacita',

misurata da α, e non vi sono interazioni tra il tipo di istituto e le variabili osservabili $\mathbf{x}_n$ ed inosservabili U_n. Vedremo oltre come superare tali restrizioni. In questo contesto semplificato, α viene a misurare la differenza di efficacia della scuola di tipo 1 rispetto alla scuola di tipo 0: la sua stima riveste quindi un ovvio interesse. Potendo osservare Y_n^* e potendo assumere

$$E(U_n \mid I_n) = 0 \tag{5}$$

ed altre convenienti ipotesi sulla varianza, sarebbe possibile stimare α attraverso metodi standard.

Tuttavia sorgono due complicazioni, la prima e' che l'ipotesi (5) risulta irrealistica poiche' lo studente non viene assegnato a caso al tipo di istituto come in un disegno sperimentale; la seconda e' che la capacita' e' inosservabile, mentre e' possibile osservare solo la variabile dicotomica (4).

Per quanto riguarda la prima questione, infatti, si consideri che puo' essere plausibile assumere che:

$$\begin{bmatrix} \epsilon_n \\ U_n \end{bmatrix} \sim N\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \begin{bmatrix} 1 & \sigma_{\epsilon u} \\ \sigma_{\epsilon u} & \sigma_u^2 \end{bmatrix} \right\} \tag{6}$$

indipendenti tra individui, dove la covarianza tra i disturbi delle due equazioni si puo' giustificare come segue. Siano:

$$\epsilon_n = \sum a_i \, \eta_i, \quad U_n = \sum b_i \, \eta_i$$

dove η_i sono variabili casuali a media 0, con determinate varianze e covarianze, risulta allora

$$\mathrm{cov}(\epsilon_n, U_n) = \sum\sum a_i \, b_j \, \sigma_{ij}$$

che in generale potra' assumere segno positivo o negativo.

La correlazione tra i due disturbi puo' derivare quindi da fattori comuni inosservabili che possono giocare un ruolo concorde, o opposto, nelle variabili latenti della scelta e della capacita'. In particolare, nel fenomeno indagato, e' piuttosto comune il riconoscimento che le "motivazioni" dello studente hanno un ruolo importante nelle sue scelte e nei suoi risultati scolastici (Astin, 1971; Hanushek, 1979). Tuttavia, specie quando si dispone soltanto di dati amministrativi, risulta molto difficile potere misurare tale motivazione per cui in generale questa confluira' insieme ad altri fattori nei due termini di disturbo. Assumendo ad esempio che sottostante i due disturbi vi sia una unica variabile motivazionale η, con entrambe $a_i>0$ e $b_i>0$, questo indurra' correlazione positiva; con il risultato che una "motivazione" superiore alla media rendera' piu' probabile la scelta $I_n=1$ (liceo) e, allo stesso tempo, l'esito $Y_n=1$ (laurea) ([1]).

La correlazione tra i due termini di errore ϵ_n e U_n fa' si che applicando alcune regole relative ai valori attesi condizionati di variabili normali (Maddala, 1989, p.367)

$$E(U_n \mid I_n) = \sigma_{\epsilon u}\left\{ I_n \frac{\phi(-Z_n \gamma)}{1-\Phi(-Z_n \gamma)} - (1-I_n) \frac{\phi(-Z_n \gamma)}{\Phi(-Z_n \gamma)} \right\} \qquad (7)$$

per cui se $\sigma_{\epsilon u}\neq 0$ l'ipotesi (5) non e' soddisfatta, e l'applicazione di metodi standard alla stima dei parametri, qualora Y_n^* fosse osservabile, condurrebbe a risultati inconsistenti.

Situazioni di questo tipo rientrano nell'ambito dei cosidetti problemi di "*selection bias*" o "*self-selection*" (Maddala, 1983), per le quali sono state proposte varie soluzioni. Tra queste, una e' quella di ricorrere a variabili strumentali ([2]). Qualora si utilizzino invece metodi standard (minimi quadrati ad esempio) queste risulteranno distorte, in particolare, se $\rho>0$, α sara' tendenzialmente sovrastimato, viceversa risultera' sottostimato se $\rho<0$. Tale effetto e' abbastanza ovvio: supponendo che $\rho>0$, infatti, la (7) indica che i disturbi dell'equazione (3) avranno una media positiva per $I_n=1$, e negativa per $I_n=0$. Da cio' la tendenza ad una sovrastima di α. Il viceversa accadrebbe

se $\rho<0$.

Nel particolare caso considerato, $\rho>0$ indica che gli studenti piu' motivati (con ϵ_n sopra la media) tendono ad accedere con piu' probabilita' al liceo, ma al tempo stesso sono anche quelli che hanno maggiori probabilita' di laurearsi poiche' hanno capacita' inosservabili U_n sopra la media. La sovrastima di α significa che si e' indotti ad attribuire, erroneamente, tale migliore performance alla "maggiore efficacia" del liceo.

Non v'e' dubbio ora che il meccanismo descritto sia assai verosimile per quanto riguarda la realta' del fenomeno educativo nel nostro paese. Se cosi' fosse vi sarebbe una tendenza a sovrastimare l'efficacia del liceo utilizzando metodi di analisi standard (comprese le piu' semplici misure di natura descrittiva). Questo potrebbe spiegare in parte la "maggiore efficacia" di tale tipo di istituti, rispetto a quelli professionali, che si desume dai risultati di alcuni studi sia di natura descrittiva (Gori e Rampichini, 1989), che condotti con modelli (Gori e Romano, 1989) in cui non si tiene conto della eventuale simultaneita' della scelta e del risultato.

Quanto precede vale ovviamente anche qualora si disponga di una variabile indicatrice del successo o insuccesso dello studente. Tuttavia, se da un lato c'e' da attendersi un analogo risultato di inconsistenza dalla stima di (3)-(4) effettuata utilizzando un modello *probit* (o *logit*) uniequazionale, dall'altro l'applicazione della procedura basata su variabili strumentali non appare piu' ovvia. Si puo' invece ricorrere alla stima di un modello *probit bivariato*, in cui si tiene conto simultaneamente della scelta e dell'esito degli studi.

Si consideri infatti che e' possibile scrivere la seguente espressione per la probabilita' dei due eventi considerati:

$$P(I_n=1;\, Y_n=1) = P[\, \epsilon_n > -\, Z_n\, \gamma;\quad U_n > -\mathbf{x}_n\, \beta\, -\alpha\, I_n]$$

si noti ora che la precedente probabilita' resta inalterata (e cio' vale anche per quelle delle altre 3 possibili coppie di eventi) se al secondo membro si dividono ambo i termini

della seconda disuguaglianza per una costante positiva qualsiasi ed in particolare per σ_u. Cio' significa che tale parametro non e' identificabile, il che e' comprensibile in quanto non si hanno informazioni sull'unita' di misura di U_n poiche' Y_n^* e' inosservabile. Il modello e' invece identificato se si pone $\sigma_u=1$, da cui consegue che anziche' stimare $\sigma_{\epsilon u}$ si stimera' ρ, il coefficiente di correlazione tra i due disturbi, mentre β ed α saranno espressi in unita' di σ_u. Ne consegue, avendo assunto una distribuzione normale bivariata per i disturbi, che

$$\begin{aligned} &P(I_n=1;\, Y_n=1) = \\ &= 1 - \Phi(-\,Z_n\gamma) - \Phi(-\mathbf{x}_n\beta-\alpha) + \Phi_2(-\,Z_n\gamma,\, -\mathbf{x}_n\beta-\alpha;\, \rho) \end{aligned} \tag{8}$$

dove $\Phi(z)$ e $\Phi_2(z_1, z_2; \rho)$ indicano rispettivamente le funzioni di ripartizione semplice e doppia di vc normali standard, mentre le altre probabilita' sono riportate in appendice.

La stima del modello puo' essere effettuata attraverso il metodo della massima verosimiglianza, la quale risulta pari a:

$$\begin{aligned} &L(\alpha,\, \rho,\, \beta,\, \gamma;\, \{I_n,\, Y_n\}_{n=1\cdots N}) = \\ &= \prod_{\{n:\, in=1,\, yn=1\}} P(I_n=1;\, Y_n=1) \;\times \prod_{\{n:\, in=0,\, yn=1\}} P(I_n=0;\, Y_n=1) \;\times \\ &\times \prod_{\{n:\, in=1,\, yn=0\}} P(I_n=1;\, Y_n=0) \;\times \prod_{\{n:\, in=0,\, yn=1\}} P(I_n=0;\, Y_n=0) \end{aligned} \tag{9}$$

Si fa osservare che nel caso in cui $\rho\neq 0$, i precedenti 4 fattori vengono a dipendere

da tutti i parametri in gioco, il che ne impedisce la stima separata; nel caso particolare in cui $\rho=0$, la precedente verosimiglianza risulta scomponibile nel prodotto

$$L(\alpha, \rho, \beta, \gamma; \{I_n, Y_n\}_{n=1\dots N}) = L(\gamma; \{I_n\}_{n=1\dots N})\, L(\beta, \alpha; \{Y_n\}_{n=1\dots N})$$

per cui in questo caso sarebbe equivalente, oltre che piu' semplice, effettuare stime separate dei coefficienti dell'equazione della scelta e dell'esito, utilizzando due modelli *probit*.

2.1. Un esempio applicativo: le conseguenze della simultaneita'

A fini esemplificativi si riportano i risultati di simulazioni basate sul seguente modello analogo a (1)(2) (3)(4):

$$I_n^* = Z_n\, \gamma + \epsilon_n = \text{VOTO1}\ \gamma_1 + \text{REDDITO}\ \gamma_2 + \epsilon_n$$

$$I_n = \begin{cases} 1 & \text{se } I_n^* > 0 \quad \text{(liceo)} \\ 0 & \text{se } I_n^* \leq 0 \quad \text{(professionale)} \end{cases}$$

$$Y_n^* = \mathbf{x}_n\, \beta + \alpha\, I_n + U_n = \text{VOTO2}\ \beta + \alpha\ \text{LICEO} + U_n$$

$$Y_n = \begin{cases} 1 & \text{se } Y_n^* > 0 \quad \text{(laurea)} \\ 0 & \text{se } Y_n^* \leq 0 \quad \text{(abbandono)} \end{cases}$$

PARAMETRI PRESCELTI

α	β	γ_1	γ_2
0	1.5	1.0	0.5

Il modello si presta alla seguente interpretazione: la variabile latente alla scelta del tipo di maturita' dipende positivamente da VOTO1 (γ_1= 1.0), ad esempio il voto conseguito alle medie inferiori, nonche' dal REDDITO (γ_2=0.5); per quanto riguarda quella latente al successo, questa dipende positivamente da VOTO2 (β=1.5), ad esempio il voto conseguito nell'esame di maturita', mentre il tipo di maturita' non ha alcun effetto su tale variabile (α=0). Per valutare il comportamento empirico del modello al variare di ρ sono stati simulati campioni di 500 osservazioni (I_n, Y_n, Z_n, $\mathbf{x}_n$), generando prima, con numeri casuali, le coppie di variabili esplicative (Z_n, $\mathbf{x}_n$); fissato quindi ρ, sono state generate coppie di vc normali del tipo (6); dopodiche', utilizzando le regole (2) e (4) sono state generate le coppie (I_n, Y_n).

Nella tabella 1 sono riportate le percentuali di laureati per tipo di maturita' ottenute in tre simulazioni con differenti valori di ρ.

Tab. 1. Percentuali di laureati in tre simulazioni con differenti valori di ρ

Matur.\ρ	−0.5	0.0	+0.5
LICEO	68.3	75.0	81.9
PROFESSIONALE	88.9	79.8	53.7

Si fa rilevare che un test χ^2 calcolato sulle tabelle di contingenza derivanti dalla classificazione delle osservazioni nelle classi (LICEO/PROFESSIONALE) e (LAUREA/ABBANDONO) porta al rifiuto dell'ipotesi di indipendenza tra esito e maturita' per $\rho=-0.5$, 0.5 mentre per $\rho=0$ l'ipotesi di indipendenza non viene rifiutata. Da un punto di vista descrittivo risulta "evidente" la maggiore "efficacia" del PROFESSIONALE per $\rho=-0.5$, e all'opposto quella del LICEO per $\rho=0.5$.

Ora, e' chiaro che tali conclusioni sono del tutto *errate*: infatti, per come sono stati simulati i dati, *entrambe i tipi di scuola forniscono lo stesso contributo alle capacita' dello studente*.

A ulteriore conferma dei rischi derivanti dalla mancata considerazione della simultaneita' tra scelta ed esito, nella tabella 2 sono stati riportati i risultati della stima dei parametri del modello ottenuti applicando stime *probit* separate alle successioni (Y_n, I_n, $\mathbf{x}_n$) e (I_n, Z_n) cioe' nell'ipotesi $\rho=0$, e quelle ottenibili massimizzando la verosimiglianza (9) ([3]).

Ebbene, anche attraverso l'uso di un modello *probit* si e' indotti erroneamente in un giudizio di maggiore (minore) efficacia del LICEO, quando $\rho\neq 0$, e solo la stima simultanea riesce a mettere in luce le vere "relazioni" tra esito e variabili esplicative.

Questo, oltre a far riflettere sull'uso di semplici strumenti di statistica descrittiva, mette in evidenza l'altrettanta inaffidabilita' di modelli mal specificati e l'importanza di piu' ampie formulazioni che possano tenere conto di un eventuale effetto simultaneo indotto dal fenomeno di "*self-selection*" e dall'impossibilita' di specificare completamente modelli del tipo (A), dove per "completa" si intende una specificazione capace di assicurare il rispetto della condizione (B).

Tab. 2. Risultati della stima del modello (1-6) per differenti valori di ρ

Valori di ρ		Parametri	α	ρ	β	γ_1	γ_2
		Valori veri	0.0	-0.5	1.5	1.0	0.5
−0.5	(1)	Stime	-0.853 !!!	n.c.	1.551	1.086	0.596
		Err.Std	*0.199*	*n.c.*	*0.230*	*0.248*	*0.084*
	(2)	Stime	-0.103	-0.515	1.471	1.051	0.589
		Err.Std	*0.128*	*0.106*	*0.223*	*0.148*	*0.074*
		Valori veri	0.0	0.0	1.5	1.0	0.5
0.0	(1)	Stime	-0.185	n.c.	1.428	1.169	0.586
		Err.Std	*0.161*	*n.c.*	*0.230*	*0.238*	*0.081*
	(2)	Stime	0.059	-0.142	1.450	0.882	0.524
		Err.Std	*0.140*	*0.127*	*0.230*	*0.146*	*0.070*
		Valori veri	0.0	0.5	1.5	1.0	0.5
+0.5	(1)	Stime	0.882 !!!	n.c.	1.638	1.085	0.469
		Err.Std	*0.148*	*n.c.*	*0.247*	*0.231*	*0.073*
	(2)	Stime	-0.024	0.606	1.622	0.907	0.477
		Err.Std	*0.123*	*0.088*	*0.183*	*0.132*	*0.062*

(1) Stime separate; (2) Stime simultanee; n.c.= non calcolato

3. Alcune estensioni

Il modello precedente si puo' estendere in vari modi. Ad esempio ammettendo un effetto differenziato del tipo di istituto per quanto riguarda le variabili individuali osservabili, il che equivale a supporre la seguente equazione, al posto della (3):

$$Y_n^* = \mathbf{x}_n \beta + \alpha I_n + \mathbf{x}_n I_n \delta + U_n$$

oppure, assumendo anche un effetto differenziato per quanto riguarda le variabili inosservabili, cioe' i termini di errore delle equazioni, il che equivale ad assumere:

$$Y_{0n}^* = \mathbf{x}_n \beta_0 + U_{0n}$$

$$Y_{1n}^* = \mathbf{x}_n \beta_1 + U_{1n}$$

Per la trattazione di questi casi nella situazione in cui Y_n^* sia osservabile si rinvia a Maddala (1983, pp. 257 e seg.); la loro estensione a dati qualitativi non pone tuttavia eccessivi problemi, a parte le restrizioni necessarie per l'identificazione dei parametri della matrice di varianze-covarianze dei disturbi, e le complicazioni nei metodi di stima.

In questa sede interessa invece valutare due altri tipi di generalizzazioni. La prima estende il modello precedente alla situazione in cui si sia interessati a confrontare due facolta' oltre che due tipi di istituto; nella seconda si considerano le conseguenze derivanti dall'introduzione di componenti di varianza per rappresentare l'effetto del particolare istituto di provenienza.

3.1. Il modello simultaneo scelta istituto - scelta facolta'- esito

Quando il problema e' confrontare due tipi di facolta' oltre che due categorie di istituti, il processo attraverso cui si formano i dati che si rendono disponibili al ricercatore, e' piu' complesso. Infatti, lo studente sceglie prima il tipo di istituto, quindi la facolta', infine si osserva l'esito. Si puo' pertanto pensare ad un sistema di tre equazioni, delle quali le prime due rappresentano le scelte e la terza la capacita'

$$I^*_{1n} = Z_{1n}\,\gamma_1 + \epsilon_{1n}\,, \qquad E(\epsilon_{1n} \mid Z_{1n}) = 0 \tag{10}$$

$$I_{1n} = \begin{cases} 1 \text{ se } I^*_{1n} > 0 \text{ (liceo)} \\ 0 \text{ se } I^*_{1n} \leq 0 \text{ (professionale)} \end{cases} \tag{11}$$

$$I^*_{2n} = Z_{2n}\,\gamma_2 + I_{1n}\,\delta + \epsilon_{2n}\,, \qquad E(\epsilon_{2n} \mid Z_{2n}) = 0 \tag{12}$$

$$I_{2n} = \begin{cases} 1 \text{ se } I^*_{2n} > 0 \text{ (facolta' 1)} \\ 0 \text{ se } I^*_{2n} \leq 0 \text{ (facolta' 0)} \end{cases} \tag{13}$$

$$Y^*_n = \mathbf{x}_n\,\beta + \alpha_1\,I_{1n} + \alpha_2\,I_{2n} + U_n, \qquad E(U_n \mid \mathbf{x}_n) = 0 \tag{14}$$

$$Y_n = \begin{cases} 1 \text{ se } Y^*_n > 0 \text{ (si laurea)} \\ 0 \text{ se } Y^*_n \leq 0 \text{ (abbandono)} \end{cases} \tag{15}$$

Qui si assume, sempre in modo semplificato, che l'istituto e la facolta' abbiano

un effetto rispettivamente pari ad α_1 ed α_2 sulla capacita' individuale, senza interazioni di sorta tra loro, con le caratteristiche individuali $\mathbf{x}_n$ e i fattori inosservabili U_n. La stima di tali parametri insieme a β riveste un ovvio interesse per la valutazione dell'efficacia degli istituti e delle facolta'.

Le ipotesi che appare opportuno assumere sui disturbi sono ora le seguenti

$$\begin{bmatrix} \epsilon_{1n} \\ \epsilon_{2n} \\ U_n \end{bmatrix} \sim N(0, \Sigma), \qquad \Sigma = \begin{bmatrix} 1 & \rho_{12} & \rho_{1u} \\ \rho_{12} & 1 & \rho_{2u} \\ \rho_{1u} & \rho_{2u} & 1 \end{bmatrix}. \tag{16}$$

indipendenti tra individui.

L'assunzione di varianze unitarie si rende necessaria per l'identificazione dei parametri, analogamente al caso gia' considerato. Le correlazioni tra fattori inosservabili delle tre equazioni sono giustificabili oltre che da un punto di vista di maggiore generalita' del modello, anche dalla mancata considerazione del fattore motivazionale, cui si e' gia' fatto accenno, e che certamente influisce su tutte e tre le equazioni.

Se il modello che rappresenta il fenomeno e' del tipo sopra descritto, allora le variabili dummy I_{1n} e I_{2n} non possono piu' essere trattate come esogene, per cui in generale

$$E(U_n \mid I_{1n}) \neq 0, \qquad E(U_n \mid I_{2n}) \neq 0$$

e analogamente al caso gia' considerato si otterrebbero stime inconsistenti applicando un modello *probit* all'equazione dell'esito. Pertanto la stima va affrontata in modo simultaneo, per esempio attraverso il metodo di massima verosimiglianza, scrivendo le probabilita' di eventi tripli del tipo:

$$P(I_{1n}=0;\ I_{2n}=0;\ Y_n=0) =$$

$$= P[\ \epsilon_{1n} \leq - Z_{1n}\ \gamma_1;\ \ \epsilon_{2n} \leq - Z_{2n}\ \gamma_2;\ \ U_n \leq -\mathbf{x}_n\ \beta\ ] =$$

$$= \Phi_3(-\ Z_{1n}\ \gamma_1, -\ Z_{2n}\ \gamma_2, -\mathbf{x}_n\ \beta;\ \ \rho_{12}, \rho_{1u}, \rho_{2u})$$

dove $\Phi_3(z_1, z_2, z_3;\ \rho_{12}, \rho_{1u}, \rho_{2u})$ indica la funzione di ripartizione di una vc normale standard tripla, con correlazioni specificate in precedenza. Le espressioni per le altre probabilita' sono riportate in appendice. La verosimiglianza si costruisce in modo analogo all'espressione (9), ma in questo caso sara' uguale al prodotto di 8 termini, corrispondenti alle prodottorie delle probabilita' delle 2^3 possibili triplette (I_{1n}, I_{2n}, Y_n). La stima puo' essere effettuata utilizzando le routine del programma GAUSS per il calcolo delle funzioni di ripartizione normali standard triple e doppie che compaiono nella verosimiglianza, e la routine MAXLIK per la sua massimizzazione. Quali valori iniziali dei parametri si puo' porre $\rho_{12} = \rho_{1u} = \rho_{2u} = 0$, mentre per gli altri si possono utilizzare le stime ottenute adattando modelli *probit* separati alle tre equazioni.

3.2. Self-selection e modelli a componenti di varianza

In un recente lavoro Aitkin e Longford (1986) hanno proposto l'uso di modelli a componenti di varianza per la misura dell'efficacia di istituzioni scolastiche. In tale studio si utilizza tuttavia un output quantitativo (votazioni) anziche' qualitativo. Inoltre viene presupposta l'assenza di simultaneita' tra risultato scolastico e scelta dell'istituto da parte dello studente. Se una tale ipotesi puo' essere sostenuta nel particolare caso considerato dai due autori (risultati di studenti in scuole medie superiori), essa appare quantomai irrealistica quando si voglia valutare l'efficacia di

istituti di istruzione secondaria utilizzando come "risultato" la performance dello studente nella carriera universitaria successiva al diploma. La seguente generalizzazione ha lo scopo di estendere il modello di Aitkin e Longford a dati qualitativi e di tenere conto al tempo stesso della eventuale simultaneita' tra scelta e risultato.

Per introdurre questa generalizzazione si ipotizzera' una situazione semplificata in cui gli individui vengono assegnati casualmente agli istituti medi superiori, per cui si pone solo un problema di simultaneita' tra la scelta della facolta' e l'esito ([4]). Supporremo inoltre di porre a confronto due sole facolta', come nel caso precedente. Lo studente sceglie dunque la facolta' secondo equazioni del tipo (1)(2), mentre il suo sucesso o insuccesso si determina attraverso un modello analogo a (3)(4). E' opportuno sottolineare che in questa formulazione la scelta riguarda la facolta' e non piu' il tipo di istituto, mentre, nell'equazione della capacita', α misura l'effetto facolta'.

Si osserva tuttavia che nell'equazione della capacita' non compare piu' l'effetto istituto, per cui il modello e' certamente carente. Per recuperare tale aspetto ed andare oltre la dicotomizzazione tra liceo e scuola professionale, e' opportuno modificare le equazioni della scelta e della capacita' come segue, dove i indica l'istituto di provenienza:

$$I^*_{in} = Z_{in}\,\gamma + \epsilon_{in}\,, \tag{17}$$

$$Y^*_{in} = \mathbf{x}_{in}\,\beta + \alpha\, I_{in} + \eta_i + U_{in}, \tag{18}$$

dove η_i indica l'effetto dell'istituto. Si assumeranno sui disturbi le seguenti ipotesi:

$$\begin{bmatrix} \epsilon_{in} \\ U_{in} \\ \eta_i \end{bmatrix} \sim N(0,\,\Sigma), \qquad \Sigma = \begin{bmatrix} 1 & \sigma_{\epsilon u} & 0 \\ \sigma_{\epsilon u} & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_\eta^2 \end{bmatrix}. \tag{19}$$

da cui

$$V(Y^*_{in}) = \sigma^2_u + \sigma^2_\eta.$$

Assumendo l'indipendenza tra disturbi relativi a studenti di scuole differenti, si ha

$$Cov(Y^*_{in}, Y^*_{i'n'}) = 0, \qquad \text{se } i \neq i'$$

mentre per studenti provenienti dallo stesso istituto, in virtu' della comune componente η_i, risulta

$$Cov(Y^*_{in}, Y^*_{in'}) = \sigma^2_\eta, \qquad \text{se } n \neq n'$$

per cui mentre individui di istituti diversi sono incorrelati, studenti provenienti dallo stesso istituto sono tra loro correlati positivamente. Inoltre il coefficiente di correlazione tra individui dello stesso istituto, noto anche come coefficiente di correlazione intraclasse, e' pari a

$$Corr(Y^*_{in}, Y^*_{in'}) = \frac{\sigma^2_\eta}{\sigma^2_u + \sigma^2_\eta}$$

ed e' interpretabile come la *parte di variabilita' della capacita' dovuta agli istituti medi superiori*. Il ricorso a componenti di varianza e' necessario nella misura in cui gli istituti sono numerosi e/o il numero di studenti proveniente da ciascun istituto e' limitato: si ottiene cosi' un notevole risparmio di parametri e un guadagno di efficienza nella stima. In particolare e' sufficiente introdurre l'ulteriore parametro σ^2_η.

Anche in questo caso la covarianza tra i fattori individuali inosservabili, ϵ_{in} e U_{in}, nelle equazioni della scelta e della capacita' puo' essere spiegata da fattori comuni,

quali la motivazione, che non e' possibile specificare completamente. L'assenza di correlazione tra gli altri disturbi e' assunta per semplicita', ma potrebbe essere rimossa, a prezzo di qualche ulteriore complicazione formale.

Poiche' gli studenti non sono assegnati casualmente alle due facolta', ma attraverso un'equazione analoga a (2), risulta valida un'espressione simile alla (7), indotta dalla *"self-selection"* dello studente: questo rende inconsistente la stima dell'equazione del risultato separatamente da quella della scelta, ed occorre nuovamente ricorrere ad una stima simultanea che tuttavia risulta complicata dalla presenza delle componenti di errore.

In primo luogo occorre osservare che

$$P(I_{in}=1;\, Y_{in}=1) = P[\, \epsilon_{in} > -\, Z_n\, \gamma;\quad \eta_i + \, U_{in} \, > -\mathbf{x}_n\, \beta\, -\alpha\,] =$$

$$= P[\, \epsilon_{in} > -\, Z_n\, \gamma;\quad \frac{\sigma_\eta}{\sigma_u}\, \frac{\eta_i}{\sigma_\eta} + \, \frac{U_{in}}{\sigma_u} \, > -\mathbf{x}_n\, \frac{\beta}{\sigma_u} - \frac{\alpha}{\sigma_u}\,]$$

per cui σ_u e σ_η non sono identificabili. Una possibile restrizione che consente di eliminare tale indeterminazione e' quella suggerita nell'ultima equazione che equivale ad assumere

$$\sigma_u = 1, \qquad \frac{\sigma_\eta}{\sigma_u} = \theta$$

e la seguente trasformazione sul modello originario

$$I^*_{in} = Z_{in}\, \gamma + \epsilon_{in}\, , \tag{20}$$

$$Y^*_{in} = \mathbf{x}_{in}\, \beta + \alpha\, I_{in} \, + \, \theta\, \eta_i + \, U_{in}, \tag{21}$$

$$\begin{bmatrix} \epsilon_{in} \\ U_{in} \\ \eta_i \end{bmatrix} \sim N(0, \Sigma), \qquad \Sigma = \begin{bmatrix} 1 & \rho_{\epsilon u} & 0 \\ \rho_{\epsilon u} & 1 & 0 \\ 0 & 0 & \theta^2 \end{bmatrix} \tag{22}$$

il che equivale ad esprimere i parametri originari α, β, σ_η e $\sigma_{\epsilon u}$ in unita' di σ_u.

Si fa rilevare che i parametri del modello permettono di sottoporre a test le seguenti ipotesi

H_0: $\rho_{\epsilon u} = 0 \Leftrightarrow$ assenza di *"self-selection"*,

H_0: $\theta = 0 \Leftrightarrow$ assenza di effetto istituto

H_0: $\alpha = 0$ assenza di effetto facolta'

si noti inoltre che il coefficiente di correlazione intraclasse diventa

$$\text{Corr}(Y^*_{in}, Y^*_{in'}) = \frac{\sigma^2_\eta}{\sigma^2_u + \sigma^2_\eta} = \frac{\theta^2}{1+\theta^2}$$

che, anche in questa riparametrizzazione, continua a misurare la parte di variabilita' della capacita' individuale spiegata dagli istituti.

Per il modello (20)(21)(22), si ottiene allora:

$$P(I_{in}=1;\ Y_{in}=1) = P[\ \epsilon_{in} > -\ Z_n\ \gamma;\quad \theta\ \eta_i + \ U_{in}\ > -\mathbf{x}_n\ \beta\ -\alpha\ ] =$$

$$= \int P[\ \epsilon_{in} > -\ Z_n\ \gamma;\quad \theta\ \eta_i + \ U_{in}\ > -\mathbf{x}_n\ \beta\ -\alpha\ |\ \eta_i\ ]\ \ \phi(\eta_i)\ d\eta_i =$$

$$= \int P[\ \epsilon_{in} > -\ Z_n\ \gamma;\quad U_{in}\ > -\mathbf{x}_n\ \beta\ -\alpha\ -\ \theta\ \eta_i\ |\ \eta_i\ ]\ \ \phi(\eta_i)\ d\eta_i$$

Si noti ora che nella probabilita' condizionata, all'interno dell'integrale, le variabili casuali ϵ ed u sono indipendenti da η per ipotesi, per cui in vista del risultato (8) si ottiene:

$$P(I_{in}=1;\ Y_{in}=1) =$$

$$= E_\eta\Big\{\ 1\ -\ \Phi(-\ Z_{in}\gamma)\ -\ \Phi(-\mathbf{x}_{in}\beta-\alpha-\theta\ \eta)\ \ +\ \Phi_2(-\ Z_{in}\gamma,\ -\mathbf{x}_{in}\beta-\alpha\ -\ \theta\eta;\ \rho_{\epsilon u})\Big\}$$

dove il valore atteso e' calcolato rispetto alla variabile $\eta \sim N(0, 1)$.

Da un punto di vista formale, pertanto, l'introduzione delle componenti di varianza ha l'effetto di rendere necessaria la sostituzione delle probabilita' congiunte scelta-esito (8) con il loro valore atteso rispetto a tali componenti. La verosimiglianza delle osservazioni sara' inoltre differente dalla (9) in quanto studenti dello stesso istituto non sono indipendenti tra loro. La verosimiglianza tuttavia fattorizza nel prodotto delle verosimiglianze a livello di istituto:

$$L(\alpha,\ \rho,\ \beta,\ \gamma,\ \theta;\ \{I_n,\ Y_n\}_{n=1\cdots N}) = \prod L_i(\{I_{in},\ Y_{in}\}_{n=1\cdots N_i})$$

mentre

$$L_i(\{I_{in}, Y_{in}\}_{n=1\cdots N_i}) = E_\eta\Big\{\Lambda_i(\eta)\Big\}$$

dove $\Lambda_i(\eta)$ e' un'espressione analoga ad (9)

$$\Lambda_i(\eta) = \prod P(I_{in}=1;\, Y_{in}=1) \times \prod P(I_{in}=0;\, Y_{in}=1) \times$$

$$\times \prod P(I_{in}=1;\, Y_{in}=0) \times P(I_{in}=0;\, Y_{in}=0)$$

Per la massimizzazione di tale espressione e' necessario ricorrere al calcolo numerico dei vari integrali che vi compaiono, ma il problema a parte la presenza della distribuzione normale doppia, dovuta alla necessita' della stima simultanea dell'equazione scelta-esito, e' analogo a quello considerato da Anderson e Aitkin (1985) per cui e' possibile applicare una soluzione simile.

4. Conclusioni

Gli esempi ed i modelli presentati mettono in luce l'importanza del ricorso a sistemi di equazioni simultanee nell'analisi dell'efficacia dell'istruzione. L'innegabile evidenza che gli studenti non vengono assegnati a caso agli istituti e alle facolta', insieme all'impossibilita' di misurare tutti i fattori che ne determinano scelte e risultati, possono indurre infatti simultaneita' tra fattori inosservabili e variabili esplicative del risultato scolastico. In particolare simultaneita' tra errori e "dummy" relative all'effetto istituto. La conseguenza e' che l'applicazione di metodi standard alla stima dei parametri, in particolare quelli dell'efficacia, conduce a risultati distorti e inconsistenti.

Per evitare tali inconvenienti nel lavoro si propongono vari modelli per il caso in cui il risultato sia di tipo qualitativo. Il caso quantitativo e' infatti gia' trattato in letteratura. Il lavoro non ha pero' la pretesa di essere esaustivo delle possibili situazioni per quanto riguarda esigenze conoscitive e dati disponibili. Al contrario, l'elemento che emerge e' che ognuna di queste richiede un modello ad hoc che tenga conto delle specificita' del problema. Gli esempi e la metodologia proposta dovrebbero tuttavia avere chiarito l'impostazione da seguire quando si desideri affrontare la questione della misura dell'efficacia dell'istruzione al riparo dai rischi derivanti dalla *self-selection*, sempre presenti.

Un elemento da sottolineare e' il pesante uso della distribuzione normale, semplice e multipla, nella formalizzazione e stima dei modelli relativi alle variabili latenti. Tale ipotesi conduce tipicamente all'impiego di modelli *probit* multivariati per variabili qualitative. Cio' puo' essere senz'altro un limite. Limite attenuato tuttavia, da un lato, dalla semplicita' di trattazione di modelli con variabili dipendenti qualitative, e dall'altro dalla considerazione che forse e' "preferibile" un modello simultaneo, malspecificato nella distribuzione degli errori, ad uno non simultaneo che condurrebbe sicuramente a distorsioni. Resta da dire che, al dila' della simulazione presentata, i modelli proposti dovranno essere "testati" in situazioni reali. A tali applicazioni saranno devoluti prossimi lavori.

Appendice

Espressioni per le probabilita' delle coppie (I_n, Y_n) nella verosimiglianza (9).

Indicando rispettivamente con $\Phi(z)$ e $\Phi_2(z_1,z_2;\rho)$ le funzioni di ripartizione di una normale standard e di una normale doppia con media 0, varianza 1 e coefficiente di correlazione pari a ρ, risulta:

$$P(I_n=1;\ Y_n=1) = P[\ \epsilon_n > -\ Z_n\ \gamma;\ U_n > -\mathbf{x}_n\ \beta - \alpha\ I_n] =$$

$$= 1 - \Phi(-\ Z_n\gamma) - \Phi(-\mathbf{x}_n\beta-\alpha) + \Phi_2(-\ Z_n\ \gamma,\ -\mathbf{x}_n\ \beta\ -\ \alpha;\ \rho)$$

$$P(I_n=1;\ Y_n=0) = P[\ \epsilon_n > -\ Z_n\gamma\ ;\ U_n \leq -\mathbf{x}_n\ \beta\ -\ \alpha\ I_n\ ] =$$

$$= \Phi(-\mathbf{x}_n\ \beta\ -\ \alpha) - \Phi_2(-Z_n\ \gamma,\ -\mathbf{x}_n\ \beta - \alpha;\ \rho)$$

$$P(I_n=0;\ Y_n=1) = P[\ \epsilon_n \leq -\ Z_n\ \gamma\ ;\ U_n > -\mathbf{x}_n\ \beta\ ] =$$

$$= \Phi(-Z_n\ \gamma) - \Phi_2(-\ Z_n\ \gamma,\ -\mathbf{x}_n\ \beta\ ;\ \rho)$$

$$P(I_n=0;\ Y_n=0) = P[\ \epsilon_n \leq -\ Z_n\ \gamma\ ;\ U_n \leq -\mathbf{x}_n\ \beta] =$$

$$= \Phi_2(-\ Z_n\ \gamma,\ -\ \mathbf{x}_n\ \beta\ ;\ \rho).$$

Espressioni per le probabilita' delle triplette (I_{1n}, I_{2n}, Y_n) per la verosimiglianza del modello (10–16). Ricordando che per una variabile casuale tripla (X, Y, Z) con funzione di ripartizione congiunta F(x, y, z) risulta $\forall\{\ \mathbf{x}_1 < \mathbf{x}_2;\ y_1 < y_2;\ z_1 < z_2\}$:

$$P(\ \mathbf{x}_1 < X \leq \mathbf{x}_2;\ y_1 < Y \leq y_2;\ z_1 < Z \leq z_2) =$$

$$= F(\mathbf{x}_2, y_2, z_2) - F(\mathbf{x}_1, y_2, z_2) - F(\mathbf{x}_2, y_1, z_2) - F(\mathbf{x}_2, y_2, z_1) +$$

$$+ F(\mathbf{x}_1, y_1, z_2) + F(\mathbf{x}_1, y_2, z_1) + F(\mathbf{x}_2, y_1, z_1) - F(\mathbf{x}_1, y_1, z_1)$$

e indicando con $\Phi(z)$, $\Phi_2(z_1,z_2;\rho)$ ed $\Phi_3(z_1, z_2, z_3;\ \rho_{12},\ \rho_{13},\ \rho_{23})$ le funzioni di ripartizione di

variabili casuali normali standard semplici, doppie e triple con media 0, varianza 1 e coefficiente di correlazione opportuni, risulta:

$$P(I_{1n}=1,\ I_{2n}=1,\ Y_n=1) =$$

$$=1 - \Phi(-Z_{1n}\ \gamma_1) - \Phi(-Z_{2n}\ \gamma_2-\delta) - \Phi(-\mathbf{x}_n\beta\ -\alpha_1-\alpha_2) +$$

$$+\Phi_2(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2-\delta;\ \rho_{12}) +\Phi_2(-Z_{1n}\ \gamma_1,\ -\mathbf{x}_n\ \beta\ -\alpha_1-\alpha_2;\ \rho_{1u}) +$$

$$+\Phi_2(-Z_{2n}\ \gamma_2\ -\delta,\ -\mathbf{x}_n\ \beta\ -\alpha_1-\alpha_2;\ \rho_{2u}) +$$

$$-\Phi_3(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2-\delta,\ -\mathbf{x}_n\ \beta\ -\alpha_1-\alpha_2;\ \rho_{12},\ \rho_{1u},\ \rho_{2u})$$

$$P(I_{1n}=1,\ I_{2n}=1,\ Y_n=0) =$$

$$= \Phi(-\mathbf{x}_n\beta\ -\alpha_1-\alpha_2) - \Phi_2(-Z_{1n}\ \gamma_1,\ -\mathbf{x}_n\ \beta\ -\alpha_1-\alpha_2;\ \rho_{1u}) -$$

$$- \Phi_2(-Z_{2n}\ \gamma_2\ -\delta,\ -\mathbf{x}_n\ \beta\ -\alpha_1-\alpha_2;\ \rho_{2u}) +$$

$$+ \Phi_3(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2-\delta,\ -\mathbf{x}_n\ \beta\ -\alpha_1-\alpha_2;\ \rho_{12},\ \rho_{1u},\ \rho_{2u})$$

$$P(I_{1n}=1,\ I_{2n}=0,\ Y_n=1) =$$

$$= \Phi(-Z_{2n}\ \gamma_2-\delta) - \Phi_2(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2-\delta;\ \rho_{12}) -$$

$$-\Phi_2(-Z_{2n}\ \gamma_2\ -\delta,\ -\mathbf{x}_n\ \beta\ -\alpha_1;\ \rho_{2u}) +$$

$$+ \Phi_3(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2-\delta,\ -\mathbf{x}_n\ \beta\ -\alpha_1;\ \rho_{12},\ \rho_{1u},\ \rho_{2u})$$

$$P(I_{1n}=1,\ I_{2n}=0,\ Y_n=0) =$$

$$= \Phi_2(-Z_{2n}\ \gamma_2\ -\delta,\ -\mathbf{x}_n\ \beta\ -\alpha_1;\ \rho_{2u}) -$$

$$- \Phi_3(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2-\delta,\ -\mathbf{x}_n\ \beta\ -\alpha_1;\ \rho_{12},\ \rho_{1u},\ \rho_{2u})$$

$$P(I_{1n}=0,\ I_{2n}=1,\ Y_n=1) =$$

$$= \Phi(-Z_{1n}\ \gamma_1) - \Phi_2(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2;\ \rho_{12}) - \Phi_2(-Z_{1n}\ \gamma_1,\ -\mathbf{x}_n\ \beta\ -\alpha_2;\ \rho_{1u}) +$$

$$+ \Phi_3(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2,\ -\mathbf{x}_n\ \beta\ -\alpha_2;\ \rho_{12},\ \rho_{1u},\ \rho_{2u})$$

$$P(I_{1n}=0,\ I_{2n}=1,\ Y_n=0) =$$

$$= \Phi_2(-Z_{1n}\ \gamma_1,\ -\mathbf{x}_n\ \beta\ -\alpha_2;\ \rho_{1u}) -\ \Phi_3(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2,\ -\mathbf{x}_n\ \beta\ -\alpha_2;\ \rho_{12},\ \rho_{1u},\ \rho_{2u})$$

$$P(I_{1n}=0,\ I_{2n}=0,\ Y_n=1) =$$

$$= \Phi_2(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2;\ \rho_{12}) - \Phi_3(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2,\ -\mathbf{x}_n\ \beta\ ;\ \rho_{12},\ \rho_{1u},\ \rho_{2u})$$

$$P(I_{1n}=0,\ I_{2n}=0,\ Y_n=0) =$$

$$= \Phi_3(-Z_{1n}\ \gamma_1,\ -Z_{2n}\ \gamma_2,\ -\mathbf{x}_n\ \beta\ ;\ \rho_{12},\ \rho_{1u},\ \rho_{2u})$$

Note

(1) Si noti che questo tipo di meccanismo puo' effettivamente rappresentare la situazione in cui $I_n=1$ rappresenta un istituto liceale, il quale, come noto, viene scelto da coloro che hanno intenzione (motivazione) di continuare gli studi universitari: il segno atteso per ρ sarebbe pertanto positivo.

(2) In sostanza dopo avere stimato, consistentemente, i parametri γ della (7), la si aggiunge all'equazione (3) e si tratta $\sigma_{\epsilon u}$ come un'ulteriore parametro da stimare: cio' consente di "correggere" il termine di errore, riportandone a zero la media condizionata, e permette di ottenere stime consistenti di β, α e $\sigma_{\epsilon u}$.

(3) La stima puo' essere effettuata utilizzando le routines del programma GAUSS per il calcolo delle funzioni di ripartizione normali standard semplici e doppie che compaiono nella verosimiglianza, e la routine MAXLIK per la sua massimizzazione. Come valori iniziali dei parametri si puo' porre $\rho = 0$, mentre gli altri sono ottenuti adattando modelli *probit* separati alle due equazioni della scelta e del risultato.

(4) Tale ipotesi restrittiva puo' essere tuttavia rimossa, anche se a prezzo di qualche complicazione formale oltre che nei metodi di stima. Il modello proposto, quindi, rappresenta un caso particolare di piu' ampie versioni, che comunque non e' il caso di considerare in questa sede.

Bibliografia

Aitkin M., Longford N. (1986). Statistical modelling issues in school effectivness studies. *JRSS A* 149, 1-43.

Anderson D.A., Aitkin M. (1985). Variance components models with binary response: interviewer variability. *JRSS B* 47, 203-210.

Astin A. W. (1971). *Predicting Academic Performance in College.* New York, Free Press.

Gamoran A., Mare R.D. (1989) Secondary school tracking and educational inequality: compensation, reinforcement, or neutrality?. *American Journal of Sociology 94,* p. 1146.

Garen J. (1984), The returns to schooling: a selectivity bias approach with a continuous choice variable. *Econometrica* 52, p. 1199-1218.

Gori E., Romano M.F. (1990). I risultati dell'istruzione universitaria: il ruolo dei fattori individuali e di struttura. *Economia e Diritto del Terziario, n. 2,* Genova.

Gori E., Rampichini C. (1989). Par. 3 in "Gli studenti dell'Ateneo fiorentino". *Notiziario dell'Universita' degli Studi di Firenze,* n. 10.

Heckman J.J., Robb R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In Wainer H. (ed.) (1986) *Drawing Inferences from Self-Selected Samples,* Springer-Verlag, New York.

Hanushek E.A. (1979). Conceptual and empirical issues in the estimation of educational production function. *Journal of Human Resources* 14, 351-388.

Maddala G.S. (1983). *Limited-dependent and qualitative variables in econometrics.* Cambridge Univ. Press, New York.

Murnane R.J. et al. (1985) Comparing public and private schools: the puzzling role of selectivity bias. *Journ. of Business & Economic Statistics,* 3, p.23.