Marta Nai Ruscone

Modelli gerarchici: aspetti metodologici e ambiti di applicazione

Tesi di Dottorato

13 settembre 2011

Prefazione

Sempre più spesso e in vari ambiti disciplinari (come ad esempio nelle ricerche sociologiche, economiche, demografiche, epidemiologiche) si analizzano fenomeni con una struttura informativa gerarchica, in cui i dati si riferiscono a più livelli di osservazione/appartenenza: individuale, familiare, territoriale, sociale e così via. In particolare, lo studio delle relazioni tra l'individuo e il contesto che lo circonda può essere ricondotto all'analisi di fenomeni a struttura gerarchica. I modelli che si sono manifestati più idonei al trattamento di dati con struttura complessa sono i cosidetti Multilevel Model. Questa classe è caratterizzata dalle seguenti dimensioni di analisi: una dimensione micro, relativa all'individuo, e una dimensione macro, riferita al contesto in cui l'individuo vive, formalizzando l'interazione individuo/ambiente attraverso lo studio dell'effetto di opportune variabili, cosiddette macro, sulle scelte e sui comportamenti individuali. L'effetto delle variabili a livello macro su quelle a livello micro può essere definito moderante, poichè l'influenza che esso rappresenta condiziona la relazione di tipo causale tra le variabili di risposta e quelle esplicative (esse sono componenti essenziali di qualsiasi analisi di regressione). Tra i principali aspetti di questa classe di modelli è possibile individuare l'elevata complessità, in presenza di un numero elevato di livelli o di variabili esplicative. La metodologia dei modelli multilivello consente l'analisi di dati organizzati in una struttura di tipo gerarchico, ossia di dati raggruppati.
Nella teoria "classica" si introduce poi, ai fini inferenziali, l'ipotesi distributiva normale per le componenti d'errore. Tuttavia in alcuni casi tale assunzione può rilevarsi troppo restrittiva. Uno degli obiettivi di questa tesi è stato di proporre, quale alternativa alla normale, la distribuzione Skew-Normal (SN) (Azzalini, Dalla Valle, 1996), che include come caso speciale la distribuzione normale e riesce a modellare i più svariati andamenti, adattandosi in modo più appropriato alle situazioni presenti in natura. La SN permette di migliore l'approccio all'analisi potendo "manipolare", anche se non direttamente, la simmetria della distribuzione. Si metteranno in luce l'utilità di impiego della SN nell'ambito dell'analisi multilivello e si discuteranno i principali problemi

legati alla stima dei parametri.

Il lavoro di tesi è strutturato nel modo seguente.

Primo capitolo: Vengono definite le varie tipologie di struttura di tipo gerarchico, le relazioni presenti nei dati e gli strumenti classici per il loro trattamento. Questi sono elementi chiave dei temi affrontati nel primo capitolo. In particolare si fa riferimento ai concetti di livello gerarchico e di tipologia di relazione, che sintetizzano la particolare struttura dei dati. Dopo aver affrontato in maniera dettagliata le tipologie di variabili e le relazioni presenti nei dati, si effettua una disamina degli approcci classici di modellizzazione. Si cerca di inquadrare il ruolo svolto dalle metodologie multilevel cercando di evidenziarne i vantaggi, anche in riferimento ai vincoli da porre sulle componenti aleatorie.

Secondo capitolo: Nel presentare i concetti fondamentali, si fa riferimento ai modelli lineari. Si affronta quindi lo studio delle loro caratteristiche, con riferimento al numero di osservazioni, alla dimensione dei gruppi, alla misura di correlazione intraclasse. Si evidenziano inoltre gli aspetti teorici e computazionali che accomunano tutti questi modelli.

Terzo capitolo: Vengono trattati i metodi di stima più usati nelle applicazioni pratiche, tra cui, in particolare, la Quasi-Verosimiglianza Penalizzata (Goldstein e Rasbash, 1996) e la Massima Verosimiglianza con integrazione numerica (Hedeker e Gibbons, 1994).

Quarto capitolo: Per quanto concerne le componenti d'errore, si propone la "distribuzione normale asimmetrica" SN, che deriva dalla perturbazione di una distribuzione normale da parte di fattori esogeni (Azzalini, Dalla Valle, 1996). Si consideri, ad esempio, la distribuzione del peso delle persone, alcune delle quali sovrappeso, mentre la maggior parte hanno un peso normale. La classe delle normali asimmetriche multivariate include le distribuzioni normali multivariate e permette di modellare sia l'asimmetria che la curtosi. Nell'ambito dei multilevel si intrucono le distribuzioni SN sia per gli errori di primo livello, sia per gli effetti casuali.

Quinto capitolo: Si esamina, mediante simulazione, la robustezza degli stimatori, ottenuti nella assunzione di normalità, quando gli effetti casuali si distribuiscono invece come una SN. Si confrontano le varie tecniche di stima, facendo varie ipotesi sul modello. Viene inoltre osservato, sempre mediante studi simulativi, il comportamento dell'indice ICC in assenza di normalità.

Indice

Capitolo 1
La struttura di tipo gerarchico

L'approccio scientifico sperimentale è diretto alla costruzione di modelli atti a descrivere, prevedere, simulare e controllare i fenomeni reali. Per tali finalità diviene centrale la struttura logica e formale dei modelli, mediante i quali si esplicitano le relazioni funzionali tra ciò che si intende spiegare (l'effetto, la risposta, il risultato) e quello che può esserne causa (variabili esplicative o fattori). Un modello, in linea di principio, dovrebbe essere suggerito dalla teoria che studia specificatamente il fenomeno in oggetto e dalle finalità che si perseguono. Esso costituisce una rappresentazione semplificata, analogica e necessaria della realtà, derivata da deduzioni logiche e confermata dalle osservazioni sperimentali. Il modello è una semplificazione della realtà e ne esprime la complessità in modo possibilmente parsimonioso; infatti "nessuna ipotesi deve essere necessariamente complessa, quando un'altra più semplice conduce alla stessa capacità esplicativa" (principio filosofico del "Rasoio di Occam ") [1]. Il modello inoltre deve essere un'analogia della realtà: ne emula gli aspetti fondamentali, al fine di ricavare deduzioni e induzioni utili. E' un errore confondere modello e realtà in quanto il modello si specifica, si studia e si rigetta in funzione della sua utilità, mentre la realtà ha un contenuto proprio di complessità che solo in parte il modello svela e rappresenta. La costruzione di un modello si concretizza in fasi successive, che è bene distinguere: il problema reale, formulato nella fase ideativa, va successivamente "tradotto" in un modello matematico-statistico il quale, una volta formalizzatane la struttura probabilistica, viene sottoposto a opportune procedure di inferenza, utilizzando i dati osservati durante un esperimento. Tali procedure riguardano comunemente la stima e la verifica del modello statistico.

[1] Rasoio di Occam (*Ockham's razor*) è il nome con cui viene contraddistinto un principio metodologico espresso nel XIV secolo dal filosofo e frate francescano inglese William of Occam (noto in italiano come Guglielmo di Occam). Tale principio, alla base del pensiero scientifico moderno, nella sua forma più immediata suggerisce l'inutilità di formulare più assunzioni di quelle necessarie e sufficienti per spiegare un dato fenomeno: il rasoio di Ockham impone di evitare le ipotesi non strettamente necessarie

La specificazione di un modello è l'aspetto più delicato della procedura, perchè dalla sua correttezza dipendono la validità e l'efficacia di tutte le fasi successive. Essa consiste nell'esplicitare un legame tra i fenomeni di interesse: devono essere individuate le variabili in gioco e il loro ruolo. Questo aiuta a formulare più correttamente il legame funzionale, che può essere sinteticamente espresso con la notazione

$$y = f(x) \tag{1.1}$$

dove x riassume l'insieme delle variabili esplicative della dipendente o delle dipendenti y. Sarebbe auspicabile, ma non è praticamente mai possibile in ambito induttivo sperimentale, ipotizzare un legame di natura deterministica, ovvero che le y dipendano solo dai fattori sperimentali sistematici individuati. Costituisce, invece, una semplificazione affermare che y è spiegata da x, in quanto nella realtà esistono interrelazioni tra le variabili che non sempre risulta agevole compendiare in modo diretto e/o variabili esplicative che il modello non ha preso in considerazione. Per questi motivi nell'approccio statistico il modello di riferimento risulta del tipo:

$$Y = f(x) + E$$

dove E è una variabile casuale di media nulla scalare o vettoriale atta a descrivere gli scostamenti tra il modello teorico,

$$y^* = f(x)$$

e la realtà osservata y. Le x sono quantità deterministiche (o aleatorie), scalari o vettoriali e la risposta Y assume, di conseguenza, la natura di variabile casuale. Indicate con:

$$y_i = f(x_i) + \varepsilon_i \quad con \;\; i = 1, ..., n \tag{1.2}$$

le osservazioni generate dalla v.c. Y, il modello introdotto ammette l'influenza sulla caratteristica Y (risposta) di fattori incontrollabili (non perfettamente prevedibili), il cui effetto si esprime in generale come contributo additivo e su cui si dovranno formulare opportune ipotesi. In alcuni contesti la specificazione della relazione funzionale $f(\cdot)$ deriva in modo immediato dalla natura del problema o dalla teoria che descrive il fenomeno. I termini di errore associati al modello sono in genere caratterizzati da ipotesi concernenti la loro indipendenza (stocastica, in media o lineare) tra di essi e rispetto alle esplicative incluse nel modello, la legge di distribuzione e l'omoschedasticità.
La stima e la verifica del modello statistico vengono successivamente eseguite utilizzando i dati raccolti attraverso un opportuno campionamento casuale. Per una analisi più efficace è bene individuare la struttura dei dati, soprattutto se questi presentano una struttura di tipo gerarchico. Si osserva comunque che i caratteri delle unità elementari sono influenzati, spesso in modo no-

tevole, dalla gerarchia: ad esempio, uno studente può avere rendimenti ben diversi a seconda della scuola in cui è inserito. È importante notare che la struttura gerarchica esercita il proprio effetto per il solo fatto di esistere, indipendentemente dalla sua genesi: infatti, anche se gli studenti non hanno scelto di frequentare una data scuola, il fatto oggettivo di condividere strutture didattiche, insegnanti e programmi scolastici rende quel gruppo di studenti diverso da quello di un'altra scuola. Talvolta il piano di campionamento si basa esplicitamente sulla gerarchia, usando metodi a più stadi; tuttavia, l'esistenza della gerarchia non è solamente legata al piano di campionamento, per cui anche i dati raccolti con il campionamento casuale semplice possono richiedere l'utilizzo di tecniche multilivello.
In alcuni casi, come precedentemente accennato, il campione che si estrae dalla popolazione potrebbe essere, ad esempio, un campione a più stadi; si pensi a tal proposito all'estrazione di un campione di studenti da utilizzare per la stima della media di una loro caratteristica, quale l'altezza in cm (Snijders, Bosker, 1999). Si può partire dall'estrazione casuale di alcuni distretti scolastici, quindi, da ognuno di essi estrarre un campione di scuole e così via. Kish (1995) evidenzia gli effetti che questo tipo di campionamento ha sulla varianza campionaria, in relazione ad altre procedure di campionamento. I modelli che vedremo in questo lavoro tengono conto correttamente di questo tipo di struttura.

1.1 Struttura dei dati

Ricordiamo che una delle finalità più comuni di un modello è la specificazione delle relazioni di tipo causa-effetto, allo scopo di interpretare, prevedere, simulare, controllare i fenomeni reali. Per questo, è importante enucleare, da una molteplicità di informazioni ottenute su numerose unità statistiche, gli aspetti essenziali presenti nei dati.
Una disamina accurata delle informazioni disponibili agevola l'applicazione dei metodi statistici più idonei per le analisi. Per questo è importante premettere ad ogni elaborazione una descrizione accurata del fenomeno in esame, del contesto in cui viene studiato e dei limiti che inevitabilmente condizionano l'ottenimento dei dati reali.
La scelta della struttura da usare per l'analisi dipende dagli obiettivi della ricerca (Tacq, 1986) e dalla oggettiva natura delle interrelazioni tra le variabili in gioco. Vedremo dapprima i modi di rappresentare in forma grafica e tabulare le varie possibili strutture dei dati.
In base ai dati di cui si dispone e agli obiettivi dell'analisi vengono individuati i livelli di osservazione. La struttura dei dati può essere semplice o complessa, e ciò condiziona anche la struttura della componente aleatoria accidentale. I dati a struttura semplice sono quelli per i quali non si rilevano particolari tipi di dipendenze o l'esistenza di particolari raggruppamenti delle osservazioni. I

dati a struttura complessa sono quelli per i quali le unità statistiche si trovano suddivise in sottoinsiemi (sia in maniera naturale, che a seguito delle ipotesi formulate per l'analisi o delle modalità di osservazione dei dati), all'interno dei quali possono essere specificate ipotesi del tutto generali sulle componenti di errore.
Tali raggruppamenti si possono presentare a uno o più livelli o stadi.
Una prima distinzione delle strutture complesse è tra le cosiddette *nested* e quelle non-*nested*.

1.1.1 Struttura Nested

Una struttura *nested* è quella in cui la gerarchia comporta l'esistenza di sottoinsiemi nidificati che contengono sotto-gruppi definiti a livelli inferiori. Ad esempio, facendo riferimento agli studenti della scuola primaria (scuola elementare) essi possono essere raggruppati in classi, istituti e distretti scolastici. Una struttura di questo tipo corrisponde a una serie di sottoinsiemi innestati (propri) da cui il nome *nested* (vedi Figura 1.1). In termini matematici è una partizione in gruppi di un insieme di unità.

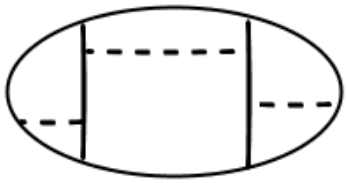

Figura 1.1: Rappresentazione di struttura di tipo *nested*.

Una caratteristica dei dati con struttura *nested* è che gli individui che fanno parte del medesimo gruppo sono più somiglianti fra loro rispetto a quelli appartenenti a gruppi diversi: per esempio, gli studenti con attitudini e motivazioni affini si trovano ad essere riuniti nelle stesse scuole a seguito di processi di selezione oppure, anche nel caso in cui il raggruppamento venga fatto senza tenere in considerazione le caratteristiche degli individui, gli alunni della stessa scuola condividono la stessa realtà e subiscono le medesime influenze; le persone che vivono nella stessa area geografica o amministrativa sono soggette alle stesse politiche locali e manifestano uno stile di vita e un comportamento più simile rispetto a persone residenti in contesti differenti.
Si osservi che le strutture *nested* sono in genere indicate anche con la denominazione alternativa di "gerarchiche" (si pensi, ad esempio, ad un organigramma aziendale, oppure alla gerarchia militare). I dati hanno struttura di tipo gerarchico se le entità appartengono a gruppi che a loro volta possono essere

contenuti in altri gruppi di ampiezza/livello superiore. In Tabella 1.1 è rappresentato il data frame di dati con struttura gerarchica a quattro livelli, che possono essere rappresentati con un diagramma (Figura 1.2) che evidenzia le relazioni e le classificazioni delle unità nei vari livelli: al livello-1 si trovano gli studenti, al livello-2 le classi, al livello-3 le scuole, al livello-4 i distretti. Spesso, ma non necessariamente, la struttura gerarchica è in tutto o in parte rispecchiata dal piano di campionamento (il campionamento a più stadi, ad esempio, riflette in genere la struttura gerarchica che caratterizza i dati).

Classificazione				*Risposta*	*Variabili Esplicative*		
Studente	*Classe*	*Scuola*	*Distretto*	*Voto esame*	*Sesso*	*Insegnamento*	*Tipo scuola*
1	1	1	1	75	M	Formale	Statale
2	1	1	1	71	M	Formale	Statale
3	1	1	1	91	F	Formale	Statale
4	2	1	1	68	F	Informale	Statale
5	2	1	1	37	M	Informale	Statale
6	1	2	1	67	M	Formale	Privata
7	1	2	1	82	F	Formale	Privata
8	1	2	1	85	F	Formale	Privata
9	1	3	1	54	M	Informale	Statale

Tabella 1.1: Data Frame di un modello gerarchico a quattro livelli

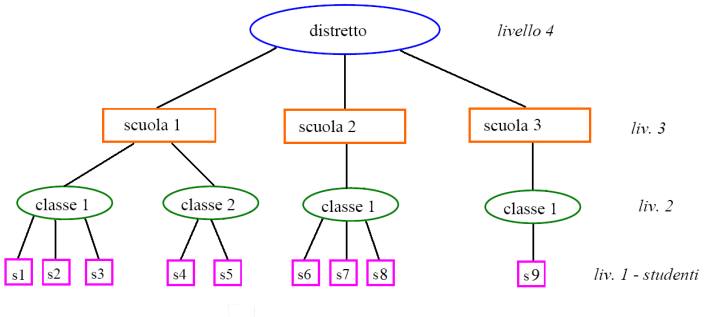

Figura 1.2: Rappresentazione di una struttura gerarchica a quattro livelli

Brown et al. (2001) proposero, per rappresentare dati con struttura complessa, uno ”schema di rappresentazione grafica” (vedi Figura 1.3) che facilita la descrizione della loro struttura: i box rappresentano i livelli ai quali le unità sono classificate, mentre la relazione gerarchica esistente viene evidenziata da una freccia.

Gli studenti della scuola elementare (livello-1) di una città, sono *nested* nelle classi (livello-2) in cui studiano, a loro volta *nested* nelle scuole di ap-

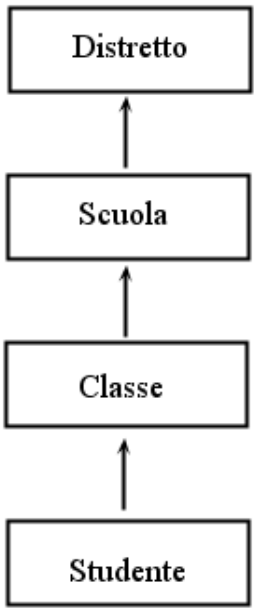

Figura 1.3: Schema di modello gerarchico a quattro livelli (Brown et al., 2001)

partenenza (livello-3), *nested* nel distretto di riferimento (livello-4). Le osservazioni individuali non risultano generalmente indipendenti: gli studenti di una stessa classe tendono, infatti, ad avere un livello di formazione simile, a causa dei processi di selezione (per esempio, alcune scuole attraggono individui appartenenti ad una medesima classe sociale) o a causa della comune storia che condividono vivendo nella medesima realtà scolastica.

1.1.1.1 *Misure ripetute*

Esistono, inoltre, strutture gerarchiche in cui sono presenti le cosiddette misure ripetute [2], quando la stessa variabile è misurata in più di una occasione per ogni soggetto (vedi Tabelle 1.2, 1.3, 1.4 e Figura 1.4); si pensi, ad esempio alle indagini longitudinali/panel in cui l'obiettivo è rivolto a misurare l'evoluzione nel tempo delle caratteristiche di interesse mediante l'espediente di ricontattare le unità per analizzarne i cambiamenti. Nell'analisi di dati longitudinali gli individui possono essere pensati come unità di secondo livello e le osservazioni ripetute come unità di primo livello. Se poi gli individui fanno parte di gruppi, questi rappresentano le unità di terzo livello. In Tabella 1.2

[2] L'esempio classico è quello di un pre e post trattamento medico, in cui si desidera misurare l'effetto del trattamento nel tempo.

vengono riportati i voti di un test somministrato ad alcuni soggetti prima e dopo aver seguito uno specifico corso. In Tabella 1.3 sono invece indicati per ogni riga: i soggetti facenti parte del campione, i risultati ottenuti ad un test (somministrato più volte) e l'età del soggetto quando il test è stato somministrato. Nella tabella 1.4 vengono invece indicate le rilevazioni effettuate per ciascun soggetto in diversi momenti: ad es. il peso del primo soggetto è stato rilevato a 5, 6 e 7 anni.

Classificazioni		Risposta	Variabili esplicative		
Studenti	Scuole	Voto test finale	Voto test iniziale	Sesso studenti	Tipo scuola
1	1	75	56	M	Statale
2	1	71	45	M	Statale
3	1	91	72	F	Statale
1	2	68	49	F	Privata
2	2	37	36	M	Privata
3	2	67	56	M	Privata
1	3	82	76	F	Statale

Tabella 1.2: Esempio di Data Frame misure ripetute

Persona	Voto-Occ1	Voto-Occ2	Voto-Occ3	Età-Occ1	Età-Occ2	Età-Occ3	Sesso
1	75	85	95	5	6	7	F
2	82	91	*	7	8	*	M
3	88	93	96	5	6	7	F

Tabella 1.3: Esempio di Data Frame misure ripetute

Classificazione		Risposta	Variabili esplicative	
$Rilevazione_i$	$Soggetto_j$	$Peso_{ij}$	Et_{ij}	$Sesso_j$
1	1	75	5	F
2	1	85	6	F
3	1	95	7	F
1	2	82	7	M
2	2	91	8	M
1	3	88	5	F
2	3	93	6	F
3	3	96	7	F

Tabella 1.4: Esempio di Data Frame misure ripetute

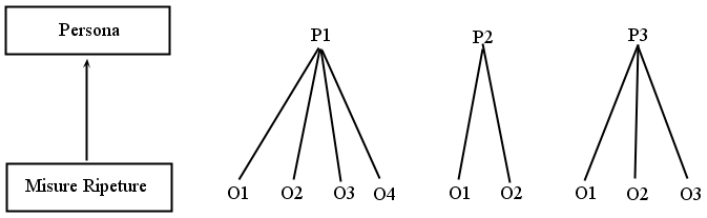

Figura 1.4: Rappresentazone di dati con strutture di misure ripetute.

1.1.2 Struttura Non-Nested

I dati hanno struttura *non nested* quando la condizione di contenimento dei livelli più bassi non è soddisfatta. Nella struttura non *nested* non è cioè definibile una partizione.
Un esempio potrebbe derivare dai dati sullo studio dei redditi di un insieme di persone fisiche caratterizzate dal tipo di occupazione, il luogo di residenza e il luogo di lavoro. Questo è un caso *non-nested* in quanto la classificazione delle unità statistiche in base alle diverse variabili sopra considerate non produce la stessa suddivisione. Altri esempi sono forniti dai dati con struttura cosidetta *cross-classified* e quelli con struttura *multiple-membership*.

1.1.2.1 Struttura *Cross-classified*

I dati hanno struttura cosiddetta *cross-classified* quando ogni unità è classificata in base a due o più criteri tra loro non ordinati gerarchicamente. Ad esempio, gli studenti distinti per area/quartiere di residenza oppure per scuola di appartenenza (vedi Tabella 1.5 e Figura 1.5): gli studenti sono *cross-classified* con le scuole e l'area di provenienza. Anche per i dati con struttura *cross-classified* è possibile la rappresentazione di Brown (vedi Figura 1.6), ma con questa rappresentazione non si intende che tutte le unità siano *cross-classified*, ma che almeno una unità lo sia (purtroppo non è ben evidente quale sia *cross-classified*).
Si considerino, ad esempio, i bambini che frequentano la scuola elementare. Sia il quartiere/area che la scuola frequentata possono avere effetti sui risultati educativi degli stessi. Una scuola può essere frequentata da bambini che risiedono in quartieri diversi, ma i bambini che abitano nello stesso quartiere/area possono anche frequentare scuole diverse (vedi Figura 1.7). Si nota quindi che i bambini sono *nested* rispetto al quartiere o rispetto alla scuola; tuttavia, se si considerano congiuntamente sia il quartiere di residenza che la scuola si ha una struttura *cross-classified*, quindi non- *nested*. Questo evidenzia che la struttura dati dipende anche dagli obiettivi e dal tipo di analisi che si intende svolgere.

Classificazione o livelli			Risposta	Variabili esplicative		
$Studente_i$	$Scuola_j$	$Area_k$	$Voto\ esame_{i(jk)}$	$Sesso_{i(jk)}$	$Area\ IMD_k$	$Tipo\ di\ scuola_j$
1	1	1	75	M	24	Statale
2	1	2	71	F	46	Statale
3	1	1	91	F	24	Statale
4	2	2	68	M	46	Privata
5	2	1	37	M	24	Privata
6	3	2	67	F	46	Privata
7	3	2	82	F	46	Statale
8	3	3	85	M	11	Statale
9	4	3	54	M	11	Privata
10	4	2	91	M	46	Privata
11	4	3	43	F	11	Privata
12	4	3	66	M	11	Privata

Tabella 1.5: Data Frame di una struttura cross-classified a due livelli (scuole/area di provenienza).

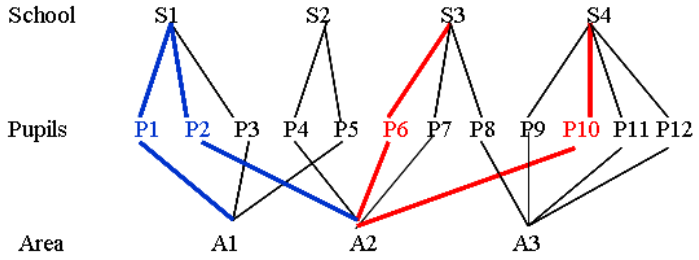

Figura 1.5: Rappresentazione di una struttura cross-classified a due livelli

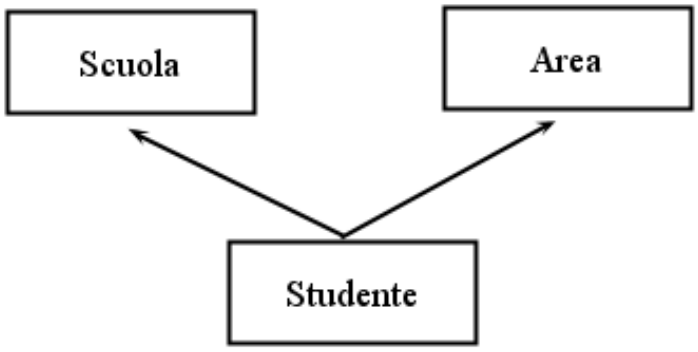

Figura 1.6: Rappresentazione di una struttura *cross-classified* a due livelli.

1.1.2.2 Struttura *Multiple-membership*

I dati possono avere struttura *multiple-membership* quando alcune unità di livello più basso appartengono a più unità del livello superiore. Un esempio classico (vedi Tabella 1.6) è rappresentato dagli studenti che cambiano scuola durante il periodo di osservazione o cambiano residenza. Ad esempio, se si considera lo studente $P8$ si nota che durante il periodo di rilevazione risiede sempre nell'area-3 ma ha frequentato sia la scuola-3 che la scuola-4. Se invece si considera lo studente $P7$ si può osservare che durante il periodo di rilevazione frequenta sempre la stessa scuola ma risiede nell'area-2 e successivamente nell'area-3. Anche in questo caso Brown propone una rappresentazione sintetica (vedi Figura 1.8), per la quale si presenta lo stesso problema evidenziato in precedenza per i dati *cross-classified*: non viene esplicitato quali e quanti dati sono *multiple-membership*. Anche qui la doppia linea (che indica relazione *multiple membership*) suggerisce che per almeno una unità si ha una relazione di tipo multiple membership. Ad esempio, con riferimento alla Tabella 1.6, lo studente $P1$ che risiede nell'$Area1$ frequenta inizialmente la $Scuola1$ e successivamente la $Scuola2$; lo studente $P7$ frequenta per tutto il periodo della rilevazione la $Scuola3$ ma prima risiede nell'$Area2$ e in un secondo momento nell'$Area3$.

	Area 1	Area 2	Area 3
Scuola 1	P1,P3	P1,P2	
Scuola 2	P5	P4	
Scuola 3		P6,P7	P7,P8
Scuola 4		P10	P8,P9,P11,P12

Tabella 1.6: Data Frame di una struttura multiple membership

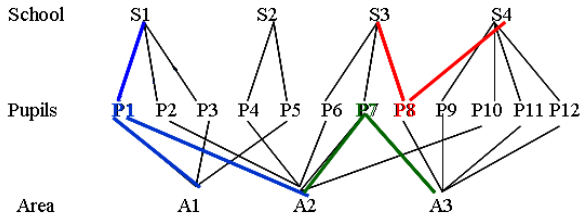

Figura 1.7: Rappresentazione grafica di struttura multiple membership

La struttura multilivello incorpora tutte quelle viste in precedenza (Pkewis, 1997, su Multilevel Modelling Newsletter, vol. 9 No. 1) (Goldstein, 2003, "*Multilevel Statistical Models* ").

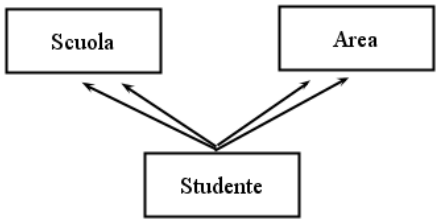

Figura 1.8: Rappresentazioni grafica dei dati con struttura multiple membership secondo Brown.

Si osserva da ultimo che, per i fenomeni analizzati in letteratura, la struttura più complessa non va in genere oltre i 3 livelli; tuttavia, le strutture viste possono essere combinate in un modello per ottenere modelli complessi con struttura combinata (vedi un esempio in Figura 1.9).

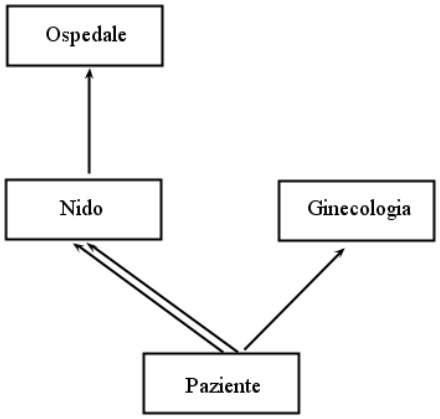

Figura 1.9: Rappresentazione grafica secondo Brown di dati con struttura combinata

1.2 Ragioni di utilizzo della struttura gerarchica

La metodologia multilevel fornisce un insieme di strumenti adatti ad analizzare simultaneamente variabili classificate a livelli differenti di gerarchia, con riferimento a modelli statistici che specificano le varie possibili forme di dipendenza. Le osservazioni all'interno di un gruppo sono infatti fra loro

più simili rispetto a quelle di altri gruppi. I modelli multilivello considerano i vari livelli di osservazione: quello relativo all'individuo e quello cosiddetto contestuale, che può derivare sia da aggregazioni di individui che da caratteristiche proprie dell'area cui l'individuo appartiene. Storicamente, le analisi di dati gerarchicamente organizzati sono state inizialmente realizzate mediante le tecniche standard, come l'analisi della varianza o la regressione multipla, spostando tutte le variabili su un solo livello di interesse. Ciò avveniva mediante due distinte procedure: aggregazione e disaggregazione. L'aggregazione è lo spostamento di variabili originariamente osservate su un livello basso della gerarchia verso un livello superiore. Al contrario, la disaggregazione è lo spostamento di variabili verso un livello più basso della gerarchia.
Ad esempio, con la regressione aggregata (pooled regression) si ignora la eventuale struttura gerarchica dei dati. Si ipotizza che le differenze tra i gruppi siano spiegate solo dalle esplicative X (covariate), ignorando i possibili effetti della struttura gerarchica nei dati. In tal modo, con la regressione su dati aggregati, tutta la variabilità viene attribuita alle differenze tra le medie dei gruppi; all'interno di ciascun gruppo le unità sono considerate perfettamente omogenee.
Analizzare variabili che appartengono a differenti livelli della gerarchia su un singolo e comune livello può risultare inadeguato e presentare degli inconvenienti, che diventano tanto più gravi quanto più la gerarchia è rilevante nella spiegazione del fenomeno analizzato. Da un lato, l'aggregazione comporta una sostanziale perdita di informazioni e, di conseguenza, l'analisi statistica perde precisione. Dall'altro, anche quando i dati vengono disaggregati, i test statistici ordinari considerano che i valori disaggregati siano, in genere, informazioni indipendenti provenienti dall'insieme della unità di basso livello. Invece, nelle situazioni in cui i dati sono gerarchicamente organizzati, i dati letti a livelli superiori non sono in genere indipendenti. Il comportamento degli individui è influenzato dal contesto sociale nel quale sono inseriti e le caratteristiche di un gruppo sono influenzate dagli individui che formano il gruppo stesso: gli individui e il contesto sociale nel quale vivono possono essere visti come un sistema gerarchico di individui e gruppi, nel quale gli individui e i gruppi stanno a livelli diversi. I test statistici tradizionali sono basati sull'assunto di indipendenza tra tutte le osservazioni, e se questa ipotesi risulta violata, le stime degli errori standard, calcolate attraverso le procedure statistiche convenzionali, sono distorte e, di conseguenza, i risultati che si ottengono possono apparire "impropriamente" significativi. Sul finire degli anni ′80, si assiste al tentativo di approdare ad un nuovo paradigma che, superando la dicotomia tra la dimensione macro (contestuale) e la dimensione micro (individuale), provi ad integrarle. Sempre negli stessi anni si sviluppano, dapprima in ambiti esclusivamente legati alla scienza dell'educazione (Goldstein, 1987; Raudenbush e Bryk, 1986; Aitkin e Longford, 1986), nuovi modelli statistici finalizzati all'analisi dei due livelli (micro e macro), al fine di superare la prospettiva riduzionista dal macro al micro ed agevolare l'integrazione tra le due prospettive analitiche: i modelli *multilevel*. Essi trovano quindi una giu-

stificazione nel risolvere le problematiche che si incontrano utilizzando dati a struttura complessa. Tra queste, ad esempio, una, di ragione concettuale, consiste nell'analizzare i dati ad un certo livello e formulare le conclusioni ad un altro livello (fallacia del livello decisionale). Questo tipo di errore può assumere sostanzialmente due forme (Pintaldi, 2003):

1. Atomistic Fallacy: problema in cui si incorre quando si formulano inferenze su un livello della gerarchia basandosi su analisi realizzate a un livello inferiore (Alker, 1969); si fanno ad esempio inferenze riguardanti associazioni a livello di gruppo mediante associazioni a livello individuale. In tal modo non si considera che i fattori che spiegano la variabilità tra individui all'interno dei gruppi non sono necessariamente gli stessi che spiegano la variabilità tra i gruppi (Hox, 1995), oppure non agiscono nel medesimo modo.
2. Ecological Fallacy: consiste nell'interpretare dati aggregati come se fossero dati individuali. Si fanno inferenze riguardanti il livello individuale sulla base dei dati inerenti il livello di gruppo, considerando cioè aggregazioni a livello del gruppo cui gli individui appartengono (Robinson, 1950); in tal modo si utilizza la correlazione tra variabili a livello di gruppo per fare affermazioni su relazioni di livello micro (Snijders e Bosker, 1999).

Unità di Analisi	Livello Inferenza	Tipo Fallacia
Gruppo	Individuale	ECOLOGICA
Individuo	Gruppo	ATOMISTICA
Individuo, esclusa rilevanza gruppo	Individuale	PSICOLOGICA
Gruppo, esclusa rilevanza individuo	Gruppo	SOCIOLOGICA

Tabella 1.7: Tipi di fallacia

Si è a lungo dibattuto se per dati con struttura di tipo gerarchico fosse da prediligere un approccio ecologico o un'analisi individuale: se da un lato non si può pensare che il singolo possieda in sè tutte le determinanti che lo conducono a certe scelte (e quindi appare limitativo procedere considerando il solo livello individuale), dall'altro il prediligere l'analisi ecologica, conferendo all'osservazione del comportamento medio dei gruppi un potere altamente esplicativo della variabilità dei componenti individuali, porta inevitabilmente a scontrarsi con il problema dell'errore dell'analisi ecologica (le relazioni tra gli aggregati si sono spesso rilevate inconsistenti, o addirittura opposte, nel momento dell'induzione sui comportamenti individuali). L'errore è prima di tutto sul piano logico ed è dovuto ad una omissione in fase di modellizzazione. Emerge, quindi, la necessità di utilizzare un modello statistico che tenga conto della non indipendenza delle osservazioni e che consenta, allo stesso tempo, di analizzare simultaneamente variabili classificate a diversi livelli della gerarchia.

1.2.1 I limiti dell'inferenza ecologica

Il primo ad evidenziare i possibili errori derivanti dall'analisi di una relazione riscontrata tra variabili rilevate su unità di livello d'aggregazione superiore ad unità di livello inferiore fu Thorndike (1939). E', tuttavia, con Robinson (1950) che il problema della *ecological fallacy* attira l'interesse della comunità scientifica, a tal punto che tuttora si è soliti parlare di paradosso di Robinson. Nello studio Robinson considera N individui caratterizzati da due proprietà X la razza e Y l'analfabetismo, entrambe variabili dicotomiche, con modalità $X = 0$ (bianco), $X = 1$ nero e $Y = 0$ (alfabetizzato), $Y = 1$ (analfabeta). Si considerano gli individui sia come unità distinte che suddivise in m sottogruppi, creati in base ai valori assunti da una terza variabile Z, l'area geografica, corrispondente a uno dei distretti oppure ad uno degli stati americani (Tabella 1.8).

i	X	Y	Z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	x_N	y_N	z_N

Tabella 1.8: In questo tipo di tabella le x_i e le y_i assumono valori $0, 1$ mentre le z_i uno dei possibili m valori

E' noto che

$$Cov(X, Y) = Cov_W(X, Y) + Cov_B(X, Y)$$

cioè la covarianza tra X e Y per le N unità è pari alla somma della covarianza entro i gruppi ($Cov_W(X, Y)ottenutacomemediadellecovarianzecalcolateneigruppi$) e dalla covarianza fra i gruppi, o covarianza ecologica, ($Cov_B(X, Y)ottenutacomecovarianzadellemediedeigrupp$) (o covarianza ecologica). Si osservi che le medie di gruppo altro non sono che le percentuali di neri all'interno dei gruppi (per la variabile X) e le percentuali di analfabeti (per la variabile Y). Per le correlazioni si avrà poi:

$$\rho_{X,Y} = \frac{Cov_W(X,Y) + Cov_B(X,Y)}{\sqrt{Var(X)Var(Y)}} =$$
$$= \frac{Cov_W(X,Y)}{\sqrt{Var_W(X)Var_W(Y)}}\sqrt{\frac{Var_W(X)Var_W(Y)}{Var(X)Var(Y)}} +$$
$$+\frac{Cov_B(X,Y)}{\sqrt{Var_B(X)Var_B(Y)}}\sqrt{\frac{Var_B(X)Var_B(Y)}{Var(X)Var(Y)}} =$$
$$= Corr_W\sqrt{\frac{Var_W(X)Var_W(Y)}{Var(X)Var(Y)}} + Corr_B\sqrt{\frac{Var_B(X)Var_B(Y)}{Var(X)Var(Y)}}$$

Robinson trovò che a una bassa correlazione individuale tra livello di analfabetismo e razza afro-americana (Tabella 1.9) (Corr(X,Y)=0,203) corrispondeva una correlazione ecologica molto più elevata ($Corr_B(X,Y) = 0,946$), come può evincersi dalla (Figura 1.12 e Figura 1.11) riprese dal lavoro originale e che dovrebbero essere più idoneamente realizzate con diagrammi a bolle analogamente alla Figura 1.10. Gli effetti delle modalità di raggruppamento e della covarianza all'interno dei gruppi interferiscono con la relazione tra X e Y calcolata a livello individuale.

		0	1	
		Bianco	Nero	Totale
0	Alfabetizzato	2406	1512	3918
1	Analfabeta	85574	7780	93354
	Totale	87980	9292	97272

Tabella 1.9: Correlazione individuale tra razza e analfabetismo per gli Stati Uniti d'America nel 1930.

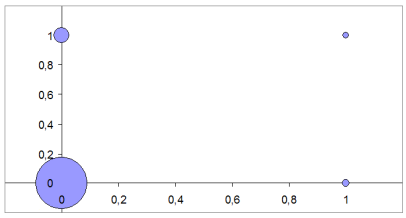

Figura 1.10: Diagramma a bolle che evidenzia la correlazione individuale

Analogamente, analizzando la relazione tra la percentuale di cittadini di razza afro-americana e il tasso di analfabetismo, considerando come gruppo

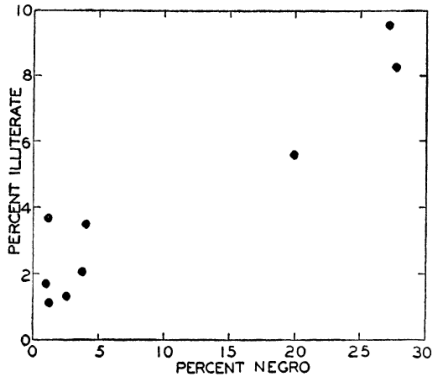

Figura 1.11: Grafico a dispersione che evidenzia la correlazione ecologica tra razza e analfabetismo in nove distretti Americani nel 1930

di aggregazione gli stati americani, si ottiene un valore di $+0,773$ (Figura 1.12). Ciò, sottolinea Robinson, indurrebbe un ricercatore a concludere che esiste una forte correlazione tra l'essere neri e l'essere analfabeti, conducendo ad un'interpretazione erronea del fenomeno.

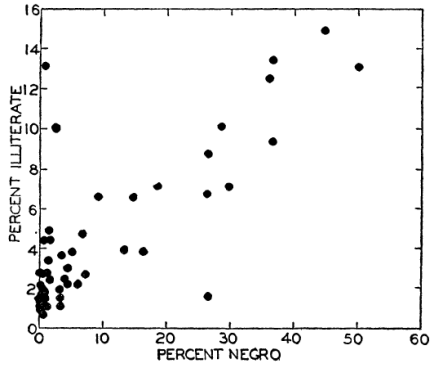

Figura 1.12: Grafico a dispersione che evidenzia la correlazione ecologica tra la razza e l'analfabetismo negli stati americani nel 1930

Anche considerando la relazione tra nativi stranieri e analfabetismo (Tabella 1.10) si ottiene una correlazione individuale pari a 0.118, mentre se si considera la stessa relazione calcolata a livello aggregato, cioè considerando come area di agrregazione gli stati americani oppure le nove divisioni in distretti (Figura 1.14) si ottengono rispettivamente $0,619$ e $0,526$.

Col suo contributo, Robinson evidenzia due punti fondamentali: la cor-

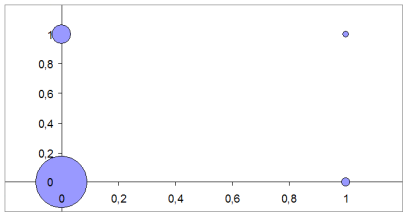

Figura 1.13: Diagramma a bolle che evidenzia la correlazione individuale tra nativi e analfabeti nel 1930 negli Stati Uniti

		0	1	
		Nativo	Straniero	Totale
0	Alfabetizzato	2614	1304	3918
1	Analfabeta	81441	11913	93354
	Totale	84055	13217	97272

Tabella 1.10: Correlazione individuale tra nativi e analfabeti nel 1930 negli Stati Uniti

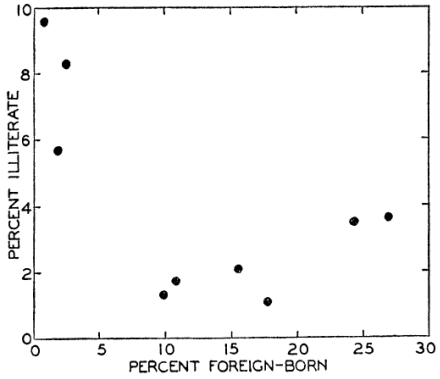

Figura 1.14: Grafico a dispersione che evidenzia la correlazione ecologica tra nati all'estero e analfabetismo in distretti americani nel 1930

relazione tra due variabili misurata a livello ecologico è molto diversa dalla correlazione misurata a livello individuale. All'aumentare del livello di aggregazione dell'unità, infatti, corrisponde, in genere, sia un incremento del coefficiente di correlazione all'interno delle unità di aggregazione sia una diminuzione del valore del rapporto di correlazione. L'intuizione di Robinson, confermata da altri autori (Yule e Kendall, 1950), se da una parte pone fine ad una pratica molto diffusa, ovvero quella di operare inferenze sulla base di correlazioni calcolate su livelli di analisi differenti (Stokes, 1969), dall'altra

pone numerosi problemi di tipo teorico e metodologico. Robinson evidenzia come due coefficienti di correlazione misurati a livelli di aggregazione diversi indicano, di fatto, relazioni diverse. In particolare date n tabelle a doppia entrata costruite per ciascuno dei gruppi considerati, la correlazione ecologica si basa solo sulle marginali di queste, non tenendo conto della distribuzione dei casi entro le celle, mentre quella individuale è calcolata sulla tabella relativa al totale dei casi, non facendo riferimento alla suddivisione degli individui per gruppo di appartenenza.
Volendo stabilire i criteri per cui è possibile procedere all'inferenza da un livello di aggregazione ad un altro si potrà considerare la relazione (Identità di Blalock):

$$\rho_{xy} = \rho_{xyW}\sqrt{(1-\eta^2_{yz})(1-\eta^2_{xz})} + \rho_{xyB}(\eta_{yz}\eta_{xz})$$

dove:

- ρ_{xy} = correlazione totale (a livello individuale)
- ρ_{xyW} = correlazione interna ai gruppi (media delle correlazioni di gruppo)
- ρ_{xyB} = correlazione tra gruppi (a livello ecologico)
- η^2_{xz} = rapporto di correlazione tra la variabile X e la variabile classificatoria Z
- η^2_{yz} = rapporto di correlazione tra la variabile Y e la variabile classificatoria Z

Correlazione individuale ed ecologica coincidono solo se c'è totale omogeneità entro i gruppi. Si può considerare una semplice simulazione in cui si evidenzia che è addirittura possibile che la correlazione tra due variabili a livello aggregato possa essere di segno opposto rispetto alla correlazione a livello individuale. Si considerino, a tale scopo, due variabili X e Y la cui relazione è monotona decrescente, condizionatamente a una variabile di contesto Z.

X	Y	Z
1	3	a
2	2	a
3	1	a
3	5	b
4	4	b
5	3	b
5	7	c
6	6	c
7	5	c

Tabella 1.11: dati

Se si considera tutta la popolazione disaggregata si evidenzia una buona correlazione positiva $\rho_{yx} = 0,60$ (vedi Figura 1.15).

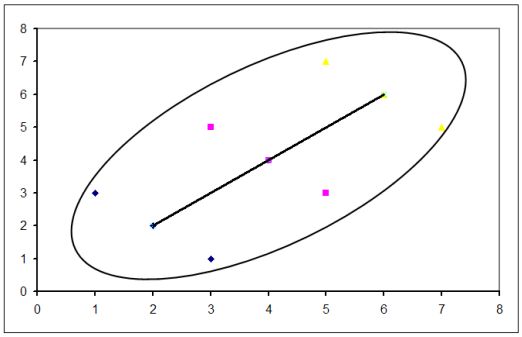

Figura 1.15: Considerando tutta la popolazione, si evidenzia una buona correlazione positiva tra X e Y

Analizzando invece i dati aggregati secondo le modalità di Z, la correlazione ecologica (between) è perfetta: $\rho_{yxB} = 1$.

Z	X	Y
a	2	2
b	4	4
c	6	6

Tabella 1.12: dati

Se invece si studia la relazione tra X e Y entro ogni singolo contesto, le tre ellissi vedi Fig. 1.16, evidenziano perfette correlazioni negative: $\rho_{YX}^{(a)} = \rho_{YX}^{(b)} = \rho_{YX}^{(c)} = -1$.

I rapporti di dipendenza di Pearson ($\eta_{YZ}^2 = 0,8 = \eta_{XZ}^2$) mostrano stretta dipendenza di X e Y dal contesto. Se si calcola ora la correlazione "entro" (ρ_{yxW}) come media ponderata delle correlazioni trovate nei singoli contesti, l'identità di Blalock trova piena conferma:

$$\rho_{yx} = \eta_{xz}\eta_{yz}\rho_{yxB} + \sqrt{(1-\eta_{xz}^2)}\sqrt{(1-\eta_{yz}^2)}\rho_{yxW} =$$
$$= \sqrt{0,8}\sqrt{0,8}(1) + \sqrt{0,2}\sqrt{0,2}(-1) = 0,60$$

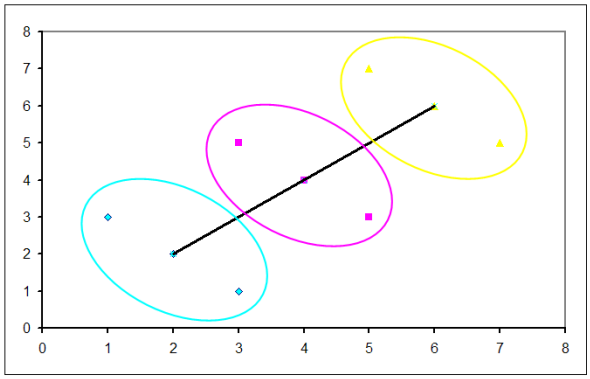

Figura 1.16: Relazione tra X e Y entro ogni singolo contesto.

In un recente lavoro, Guseo (2006) ha trattato il problema della *ecological fallacy* confrontando la correlazione parziale e la correlazione ecologica. Noto che l'analisi della varianza è uno strumento statistico che consente di individuare fonti separate di variabilità, dovute alla presenza di un legame di dipendenza esistente tra le variabili, si consideri un gruppo di N individui, caratterizzati da due variabili X e Y: le determinazioni distinte della variabile X si indicano con x_i o, più sinteticamente, attraverso il semplice deponente i, $i = 1, 2, ..., K_1$. Analogamente, le determinazioni distinte della variabile Y si indicano con il simbolo y_j o, per semplicità, con il deponente j, $j = 1, 2, ..., K_2$. K_1 e K_2 rappresentano la cardinalità delle determinazioni distinte delle marginali X e Y. Talvolta si è interessati allo studio diretto della covarianza tra le componenti di una variabile doppia (X, Y) tenendo sotto controllo gli effetti sulla relazione dovuti alla presenza di una variabile concomitante Z. Gli N membri possono quindi essere suddivisi in m sottogruppi in base al valore della variabile concomitante Z. Si consideri lo studio della correlazione parziale tra le variabili X e Y al netto del contributo lineare dovuto a Z. Occorre determinare la covarianza tra le variabili residuo (covarianza parziale) ottenute dopo aver eliminato il contributo lineare, secondo i minimi quadrati, delle variabili concomitanti. Nel caso di una sola variabile concomitante, la covarianza parziale tra X e Y al netto del contributo lineare di Z è

$$Cov_Z(X, Y) = \sigma_Z(X, Y) = M[(X - \widehat{a} - \widehat{b}Z)(Y - \widehat{c} - \widehat{d}Z)]$$

essendo $\widehat{a}$, $\widehat{b}$, $\widehat{c}$ e $\widehat{d}$ le stime delle relazioni lineari del contributo di Z rispettivamente su X e Y.

Se si aggiunge e toglie lo scostamento rispetto alle funzioni di regressione $\mu_X(Z)$ $\mu_Y(Z)$ si ha:

$$\begin{aligned}\sigma_Z(X,Y) &= M[(X+\mu_X(Z)-\mu_X(Z)-\hat{a}-\hat{b}Z)(Y+\mu_Y(Z)-\mu_Y(Z)-\hat{c}-\hat{d}Z)] = \\ &= M_Z\{\sigma_{XY}(Z)\} + M_Z[(\mu_X(Z)-\widehat{a}-\widehat{b}Z)(\mu_Y(Z)-\widehat{c}-\widehat{d}Z)] = \\ &= \sigma^*_{XY} +_L \sigma_{XY},\end{aligned}$$

cioè la covarianza parziale è costituita dalla somma di due addendi: la covarianza residua e il cosidetto *covariance lack of fit* ${}_L\sigma_{xy}$. Sfruttando la scomposizione a tre termini:

$$\begin{aligned}\sigma_{XY} &= M_Z\{\sigma_{XY}(Z)\} + M_Z[(\mu_X(Z)-\hat{a}-\hat{b}Z)(\mu_Y(Z)-\hat{c}-\hat{d}Z)] + \frac{\sigma_{XZ}\sigma_{YZ}}{\sigma^2_Z} = \\ &= \sigma^*_{XY} + \sigma_{XY} +_C \sigma_{XY}\end{aligned}$$

è immediato osservare che la covarianza parziale tra X e Y al netto del contributo dovuto a Z è pari a

$$\sigma_Z(X,Y) = \sigma_{XY} - \frac{\sigma_{XZ}\sigma_{YZ}}{\sigma^2_Z} = \sigma_{XY} -_C \sigma_{XY}$$

E' da notare che se almeno una delle due funzioni di regressione, $\mu_X(Z)$ o $\mu_Y(Z)$ è rettilinea, allora

$$\sigma(X,Y) = \sigma^*_{XY} + \sigma_{XY}$$

e quindi la covarianza parziale e quella ecologica non presentano l'addendo comune, ovvero *covariance lack of fit*.

Restano ora da individuare le corrispondenti varianze, precisamente:

$$\begin{aligned}\sigma^2_p(X) &= M[(X-\widehat{a}-\widehat{b}Z)^2] = \\ &= M[(X-\mu_X+\widehat{b}\mu_Z-\widehat{b}Z)^2] = \\ &= M[(X-\mu_X)-\widehat{b}(Z-\mu_Z)]^2 = \\ &= \sigma^2_X + \frac{\sigma^2_{XZ}}{\sigma^4_Z}\sigma^2_Z - 2\frac{\sigma^2_{XZ}}{\sigma^2_Z} = \\ &= \sigma^2_X - \frac{\sigma^2_{XZ}}{\sigma^2_Z} = \\ &= \sigma^2_X(1-\rho^2_{XZ})\end{aligned}$$

$$\sigma_p^2(Y) = M[(Y - \widehat{c} - \widehat{d}Z)^2] =$$
$$= \sigma_Y^2 - \frac{\sigma_{YZ}^2}{\sigma_Z^2} =$$
$$= \sigma_Y^2(1 - \rho_{YZ}^2)$$

Il quadrato del coefficiente di correlazione parziale tra X e Y, al netto del contributo lineare di Z è, pertanto,

$$\rho_Z^2(X,Y) = \frac{(\sigma_{XY}\sigma_Z^2 - \sigma_{XZ}\sigma_{YZ}\sigma)^2\sigma_Z^4}{(\sigma_Y^2\sigma_Z^2 - \sigma_{YZ}^2)(\sigma_X^2\sigma_Z^2 - \sigma_{XZ}^2)\sigma_Z^4} =$$
$$= \frac{(\rho_{XY} - \rho_{XZ}\rho_{YZ})^2}{(1 - \rho_{YZ}^2)(1 - \rho_{XZ}^2)}.$$

Tale risultato è ben noto in letteratura. Sotto l'ipotesi di correlazione parziale nulla, cioè $\rho_Z^2(X,Y) = 0$, si ha che la covarianza tra X e Y assume la forma

$$\sigma_{XY} = \frac{\sigma_{XZ} + \sigma_{YZ}}{\sigma_Z^2},$$

ovvero

$$\rho_{XY} = \rho_{XZ} + \rho_{YZ}.$$

Si evidenzia quindi che l'apporto informativo apparente di X e Y su σ_{XY} o su ρ_{XY} dipende esplicitamente dalla presenza dei legami non nulli di Z con X e di Z con Y. Il contributo informativo di X in un modello di regressione lineare che contiene già Z come esplicativa è allora irrilevante.
Si consideri il caso in cui la relazione tra i due caratteri Y e X sia monotona decrescente condizionatamente a Z, mentre al crescere di Z le relazioni siano monotone crescenti. Il grafico 1.17 ne costituisce un esempio.

I punti rappresentano le osservazioni individuali mentre i quadrati sono le medie condizionate ad uno specifico livello di Z, ovvero di coordinate $(\mu_X(Z), \mu_Y(Z))$, $Z = 1, 2, ..., K$. L'evidente covarianza ecologica positiva, $\sigma_{\overline{XY}} = Cov(\mu_X(Z), \mu_Y(Z)) > 0$ ed il corrispondente coefficiente di correlazione positivo, $\rho_{\overline{XY}} > 0$ possono dar luogo alla cosidetta fallacia ecologica, dovuta ad un errore di riferimento inferenziale. La relazione positiva tra $\overline{X}$ e $\overline{Y}$ nello spazio delle unità aggregate, contrasta con le relazioni condizionali rispetto a Z, riferite alle unità individuali, che presentano un evidente segno negativo.
La fallacia ecologica si presenta come un problema concreto se si procede con informazioni medie aggregate e si pretende di riferire la relazione globale alle unità individuali presenti all'interno dei sottogruppi.

Si è a lungo dibattuto se per dati con struttura di tipo gerarchico fosse da prediligere un approccio ecologico o un'analisi individuale: se da un lato

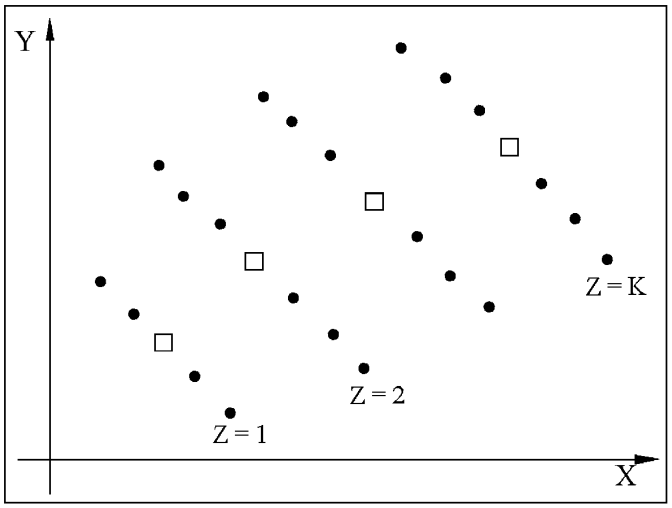

Figura 1.17: Relazioni locali e fallacia ecologica.

non si può pensare che il singolo possieda in sè tutte le determinanti che lo conducono a certe scelte (e quindi appare limitativo procedere considerando il solo livello individuale), dall'altro il prediligere l'analisi ecologica, conferendo all'osservazione del comportamento medio dei gruppi un potere totalmente esplicativo della variabilità dei componenti individuali, porta inevitabilmente a scontrarsi con il problema dell'errore dell'analisi ecologica (le relazioni tra gli aggregati si sono spesso rilevate inconsistenti, o addirittura opposte, nel momento dell'induzione sui comportamenti individuali). L'errore è prima di tutto sul piano logico ed è dovuto ad una omissione in fase di modellizzazione. Emerge, quindi, la necessità di utilizzare un modello statistico che tenga conto della non indipendenza delle osservazioni e che consenta, allo stesso tempo, di analizzare simultaneamente variabili classificate a diversi livelli della gerarchia. Per risolvere il problema connesso alle interazioni tra la sfera individuale e il contesto in cui l'individuo è inserito, che rappresenta il livello superiore, è necessario impostare analisi statistiche multilevel, avendo correttamente individuato la struttura dei dati.

1.2.2 Dati ecologici e dati individuali

I paradigmi interpretativi dei fenomeni sociali si sono mossi storicamente all'interno di una prospettiva dicotomica che tende a privilegiare alternativamente o le dimensioni micro o le dimensioni macro.

Sul rapporto e sulla relativa portata euristica delle informazioni raccolte a livello micro ed a livello macro è possibile delineare tre prospettive (Pintaldi, 2003). In primo luogo i dati ecologici non hanno un valore euristico in quanto non permettono la stima dei dati individuali. Secondo tale prospettiva, definita "riduzionista", i dati ecologici non hanno alcun ruolo nella ricerca sociale, se non quello di studiare, con i limiti connessi alla fallacia ecologica, le relazioni a livello individuale (Robinson, 1950).
L'approccio riduzionista, riducendo il livello d'analisi all'individuo, implica da un punto di vista tecnico, il ricorso a una serie di strumenti di rilevazione basati sul contatto diretto con il soggetto. Diventano fondamentali a questo fine strumenti di rilevazione "soggettiva" quali le indagini (*survey*, *exit-polls*) ed ogni altra tecnica in grado di rilevare opinioni ed atteggiamenti a livello individuale. Il ricorso a tali strumenti di rilevazione, tuttavia, pone una serie di problemi, inerenti in primo luogo l'affidabilità del dato, causati da fattori esterni e dalla possibilità di un repentino cambiamento d'opinione che potrebbero inficiare la qualità del dato. In secondo luogo, una rilevazione di tipo diretto comporta inevitabilmente un costo maggiore, determinato dalla numerosità campionaria necessaria a produrre risultati statisticamente significativi e anche dall'organizzazione delle modalità di rilevazione. Da un punto di vista teorico, inoltre, un approccio che fonda l'analisi esclusivamente sui dati individuali sottintende una definizione di struttura sociale quale entità neutrale. Il contesto, in altri termini, non assume rilevanza, dal momento che non svolge alcun ruolo nella scelta individuale, determinata esclusivamente da caratteristiche soggettive (Coleman, 1968). In una seconda prospettiva i dati ecologici presentano un valore in sè, in quanto forniscono delle informazioni differenti e complementari ai dati individuali.
Alla prima posizione si contrappone quella di chi ritiene che i dati ecologici presentino, infatti, un valore specifico, poichè forniscono delle informazioni differenti rispetto a quelle ottenute attraverso i dati individuali (Brown, 1995). Secondo questo punto di vista, alcuni fenomeni sociali assumono significato solo a livello aggregato. Alcuni autori distinguono tra il rischio di utilizzare come unità di analisi gli individui, mentre si studia un fenomeno che origina e si spiega in relazione al contesto in cui questi si trovano, e che può, quindi, essere spiegato concentrando l'attenzione sul particolare processo da cui origina (*individual-level fallacy*) ed il rischio di trarre conclusioni fallaci sul comportamento individuale, utilizzando variabili i cui valori hanno come referente un aggregato (*ecological fallacy*). Quest'ultimo approccio, si potrebbe definire "oggettivo" perchè si annulla l'incidenza dell'insieme dei fattori di disturbo dovuti al coinvolgimento.
Se il vantaggio di quest'ultimo approccio consiste nella possibilità di ottenere informazioni esaustive con un relativo basso impiego relativamente bassso di risorse, lo svantaggio è rappresentato ancora una volta dalla sostanziale impossibilità di inferire al livello individuale il comportamento riscontrato a livello aggregato.
D'altro canto, in alcune situazioni, i dati ecologici sono rilevanti in quanto a

certe condizioni possono essere impiegati per stimare i dati individuali; cioè si pone in luce il problema dell'integrazione tra dati individuali e aggregati (Lazarsfeld e Menzel, 1961; Goodman, 1953 e 1959; Blau, 1960; Davis et al., 1961). In questo contesto emerge lo studio sistematico delle relazioni tra effetti individuali e contestuali e, più specificatamente, assume rilevanza la necessità di risolvere il problema della fallacia ecologica.

1.2.3 Il contesto: un problema di definizione

I problemi legati alla relazione micro-macro non sono stati del tutto risolti. Se da una parte infatti, si è giunti alla costruzione di algoritmi sempre più sofisticati, tali da prevedere strutture gerarchiche incrociate e con appartenenza multipla (Goldstein, 1999), dall'altra rimangono irrisolti quei problemi di ordine concettuale (a partire dalla definizione di contesto) connessi al rapporto funzionale tra ambiente sociale e comportamento individuale, ovvero, tra legami relativi a unità situate a livelli gerarchici differenti.
A tal proposito è possibile individuare almeno tre definizioni di contesto (Zaccarin e Rivellini, 2002):

- Raggruppamento "naturale". Rappresenta il criterio di aggregazione più intuitivo e si può affermare che la modellistica di cui ci stiamo occupando nasce dalle riflessioni su questa modalità di raggruppamento. In questo caso la struttura gerarchica è intrinseca. I soggetti vengono naturalmente classificati come appartenenti ad un gruppo. E' il caso tipico degli alunni aggregati per classi, o di individui residenti nella stessa area geografica.
- Raggruppamento "ambientale" [3]. In questo caso la correlazione tra unità appartenenti allo stesso gruppo emerge da considerazioni di tipo teorico. Ci si trova dinnanzi ad una situazione in cui l'aggregazione non è di tipo geografico ma ambientale. In altri termini, si suppone che l'esposizione allo stesso ambiente (di lavoro, ad esempio) favorisca una comunanza di valori, atteggiamenti, comportamenti tali da costruire dei veri e propri gruppi. Alcuni autori hanno sottolineato l'importanza dell'appartenenza di classe in relazione ad esempio alle scelte politiche e/o elettorali (Andersen e Heath, 2002; Charnock, 1996);
- Raggruppamento "teorico". Ci si riferisce ad aggregazioni formulate sulla base dei costrutti teorici fondati su fattori di tipo economico, sociale e culturale. Sicuramente tra le tre tipologie è quella più problematica da indagare, proprio per l'incertezza dei confini.

Dal momento che il raggruppamento gioca un ruolo fondamentale nell'analisi dei comportamenti non sembra banale evidenziare i limiti concettuali ed i problemi metodologici ed interpretativi che tali definizioni presentano.

[3] le autrici utilizzano *working grouping* volendo probabilmente intendere un concetto simile a quello di classe

Innanzitutto vi è la questione dei confini. Alcuni gruppi presentano dei confini fissi e ben determinati. In questo caso l'individuo, o l'unità d'analisi gerarchicamente inferiore, può appartenervi oppure no. Non è prevista una situazione intermedia. E' il caso tipico dei raggruppamenti naturali: non si può appartenere a due comuni o a due province. Nell'ambito delle scienze sociali, tuttavia, quest'ultima condizione rappresenta l'eccezione piuttosto che la regola. Nella maggior parte dei casi, infatti, il ricercatore si trova dinnanzi a strutture di gruppo fluide, dai contorni sfocati, dai confini incerti ed indeterminabili e diventa fondamentale stabilire, non tanto se un'unità appartiene ad un raggruppamento, bensì "in che misura" vi appartiene. Un altro aspetto collegato alla definizione di contesto (anche se in senso etimologico) è quello relativo alla mobilità. Gli individui si muovono al di là dei confini stabiliti.

1.2.4 Relazione tra variabile di contesto e variabile dipendente

In vari ambiti disciplinari (sociologico, economico, demografico, sanitario etc.), si ha spesso a che fare con fenomeni a struttura gerarchica, in cui i dati si presentano a più livelli: individuale, familiare, territoriale, sociale. In queste circostanze bisogna procedere all'analisi di una relazione tra gli individui e la società. Gli individui interagiscono col contesto sociale cui appartengono, cioè i soggetti sono influenzati dalle caratteristiche dei gruppi di cui fanno parte e, a loro volta, le proprietà di questi gruppi risentono dell'influenza dei singoli individui. Matrici di dati che presentano una struttura gerarchica sono caratterizzate da relazioni tra variabili differenti ai differenti livelli (Figura 1.18).

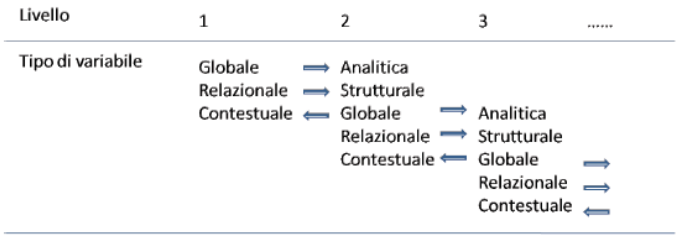

Figura 1.18: Schema semplificato della tipologia di variabili ai differenti livelli.

In simili circostanze, individui (unità) e gruppi (macro-unità) sono presi in considerazione come un sistema gerarchico, osservabile a differenti livelli; ciò conduce ad un'analisi dell'interazione tra le variabili che caratterizzano

gli individui, con quelle che caratterizzano i gruppi.
In una struttura gerarchica si possono evidenziare tre tipologie di effetti (Snijders e Boskers, 1999; Tacq, 1986) che potrebbero entrare in gioco, influenzando il comportamento delle unità di livello inferiore. Seguendo l'impostazione grafica di Tacq (1986), ben nota in letteratura e adottata anche da Snijders e Bosker, le figure che seguiranno adottano le seguenti convenzioni: la linea tratteggiata separa due livelli: al di sotto c'è il livello micro, al di sopra il livello macro; le lettere maiuscole servono ad indicare variabili misurate al livello macro, mentre quelle minuscole al livello micro; infine la freccia indica la presunta relazione causale.
In primo luogo si potrebbe essere interessati all'analisi della relazione tra una variabile indipendente x ed una dipendente y all'interno di un contesto (Figura 1.19). Questo concerne lo studio della relazione tra la variabile obiettivo y e i fattori che caratterizzano le unità di livelli; in tal caso non vi è un'influenza delle variabili di livello macro su quello inferiore.

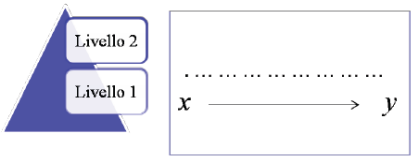

Figura 1.19: Relazione tra le variabili all'interno di un gruppo

In altri casi l'attenzione del ricercatore può focalizzarsi sul ruolo delle dimensioni contestuali. Si potrebbe essere interessati al ruolo che alcune variabili, misurate a livello macro, potrebbero assumere, nel condizionare il comportamento delle unità elementari, tenendo sotto controllo le variabili individuali (Figura 1.20). L'obiettivo è rilevare l'influenza che la variabile Z al livello macro, ha sulla variabile y, al livello micro, in cui vi è la presenza congiunta anche dell'effetto della variabile x, legata ad essa da un nesso di causalità. Qui lo scopo è verificare se le variabili del livello superiore sono in relazione con le variabili del livello inferiore. In questo caso, si possono presentare tre diversi tipi di relazione tra i due livelli: un primo in cui una variabile macro influisce sulle determinazioni della variabile micro; un secondo tipo, che deriva dal primo, in cui c'è una relazione tra la variabile macro e quella micro, dato l'effetto sulla variabile micro di un'altra variabile dello stesso livello; un terzo caso, che può essere considerato speculare rispetto ai tre precedenti, in cui è la variabile di livello inferiore ad avere effetto su una variabile di livello superiore.

La situazione più comune nei vari ambiti della ricerca sociale si verifica quando si suppone una interazione delle variabili tra differenti livelli. Un'ultima ipotesi, quindi, è quella in cui l'indagine si focalizza sulle interazioni tra

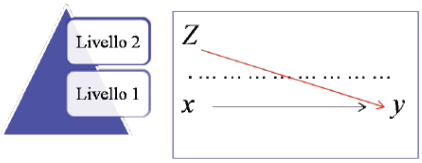

Figura 1.20: Relazione tra variabili di contesto e variabile dipendente

i due livelli (Figura 1.21). Questo è noto anche come "interazione *cross-level*", in cui la relazione tra una macro e una micro-variabile dipende da una diversa variabile di livello inferiore. In questa circostanza la relazione tra x e y dipende dall'influenza di Z.

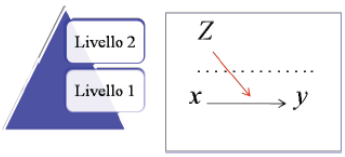

Figura 1.21: Effetto d'interazione

In questo caso si hanno due possibilità: o l'interazione tra variabili di primo livello è condizionata dalla variabile di contesto (Z) oppure l'interazione tra variabile di contesto (Z) e variabile dipendente (y) è condizionata dalla variabile di primo livello (x). Le strutture fin qui descritte, pur essendo le più frequenti, rappresentano solamente alcune delle possibili relazioni che legano le variabili di differenti livelli. Si può comunque definire un quadro generale dei fattori che possono influenzare le decisioni di un individuo: possono essere distinti in individuali, micro-contestuali e macro-contestuali. Si dicono contestuali tutti quei fattori propri dell'ambiente in cui l'individuo vive e che hanno un effetto sui risultati della sua azione. Volendo distinguere tra micro- e macro-contestuali, i primi possono riferirsi, considerando ad esempio uno studente, all'ambiente familiare, scolastico e socioeconomico della zona di residenza; i secondi, invece, riguardano un contesto più ampio, regionale o nazionale, e sono importanti soprattutto nelle comparazioni internazionali.

1.3 Considerazioni sulla struttura dei modelli complessi

Come è stato evidenziato in precedenza, la struttura da adottare dipende sia dal contenuto informativo dei dati che dalle finalità dell'analisi. Nelle

applicazioni accade spesso di disporre di dati organizzati in una struttura di tipo gerarchico, essendo gli stessi classificati, in via naturale o in modo funzionale all'analisi, in classi o gruppi, suscettibili a loro volta di essere ripartiti in sottogruppi e così via. La struttura gerarchica si dipana, dunque, in livelli successivi fino a pervenire alle cosidette "unità elementari", che si trovano al livello più basso della gerarchia e vengono dette unità di primo livello; in generale, a partire da esse, le unità che formano raggruppamenti intermedi vengono dette unità di $g - esimo$ livello con $g = 1, ..., G$ $(G > 1)$. Si dirà, in tal caso, che la struttura gerarchica è a G livelli.

Dati che presentano una struttura siffatta si prestano in generale all'applicazione delle consuete tecniche di analisi statistica multivariata, sia nella fase puramente esplorativa, sia nella analisi ed interpretazione, mediante modelli statistici, dei legami tra i fenomeni osservati. Così, per esempio, se lo scopo dell'analisi è lo studio della dipendenza, ed i dati seguono una struttura gerarchica, è opportuno, ai fini di uno studio più accurato, tenere in considerazione tale struttura e le eventuali ipotesi che si sono fatte: sulle fonti di variabilità oltre a quelle di dipendenza. In caso contrario, le conclusioni potrebbero risultare non adeguate.

L'esempio più classico è offerto dallo studio del rendimento scolastico (vedi Figura 1.22), dove le unità statistiche (gli studenti) sono raggruppate in scuole. Tali raggruppamenti, annidati, costituiscono la natura multilivello dei

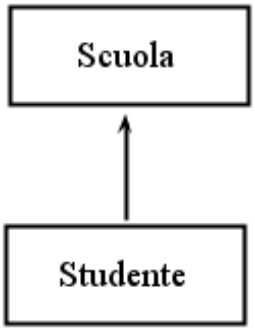

Figura 1.22: Rappresentazione di una struttura nested secondo Brown.

dati, caratterizzata dalla possibile elevata somiglianza delle unità statistiche all'interno dei gruppi. Infatti, come più volte osservato, una caratteristica dei dati strutturati in maniera gerarchica è che gli individui che fanno parte del medesimo gruppo sono più somiglianti fra loro, rispetto a quelli appartenenti a gruppi diversi. Di tale effetto si deve tenere opportunamente conto, in modo da utilizzare le procedure statistiche più opportune.
Uno dei primi esempi di utilizzo del modello multilevel è contenuto in *"Statistical modeling of data teaching styles (with discussion)"* di Aitkin, Anderson

e Hinde (1981), che uscì in risposta a un articolo di Bennet (1976). Questo, analizzando il comportamento di alcuni alunni di scuole elementari, giunse alla conclusione che i bambini esposti a uno stile di insegnamento "formale" compivano maggiori progressi rispetto agli altri bambini. I dati vennero analizzati utilizzando le tradizionali tecniche di regressione multipla e i risultati furono definiti statisticamente significativi. Aitkin, Anderson e Hinde dimostrarono invece che tenendo in considerazione l'effetto del raggruppamento dei bambini in classi le differenze scomparivano.
Sempre negli anni '80, oltre a Aitkin e Longford (1986), altri studiosi quali De Leeuw e Kreft (1986), Goldstein (1986), Mason, Wong e Entwistle (1984) e Raundenbush e Bryk (1986) proposero procedure di stima e software statistici per i modelli multilevel. Per questo motivo in anni relativamente recenti, grazie anche allo sviluppo delle possibilità di calcolo offerte dai nuovi dispositivi di elaborazione automatica dei dati (PC), la teoria e i metodi dei modelli lineari ad effetti misti, da una parte, e i modelli gerarchici, dall'altra, hanno conosciuto, oltre a sviluppi metodologici, anche una notevole diffusione in campo applicativo. E' stata poi fornita una sistematizzazione teorica dei modelli interpretativi per lo studio di dati strutturati, che ha preso il nome di Analisi Multilivello e, in modo analogo, i modelli che essa raccoglie vengono denominati Modelli Multilivello (Bryk e Raudenbush, 1992; Goldstein, 1995; Snijders e Bosker, 1999).

Con riferimento al Goldstein (1995) si può apprezzare quanto si sia sviluppata la metodologia multilevel. Nel presente lavoro si prenderanno in considerazione solo i modelli lineari, il cui uso è più comune nelle applicazioni, rispetto a quelli non lineari.
L'idea di base di un modello lineare a due livelli è molto semplice. Vengono definiti per i vari contesti (o gruppi di individui) modelli lineari diversi (detti di primo livello) che di solito si riferiscono alla stessa variabile risposta, le stesse variabili esplicative, ma diversi coefficienti di regressione. I suddetti modelli vengono collegati da un modello di livello superiore, in cui i coefficienti di regressione sono espressi in funzione delle variabili esplicative di secondo livello. L'idea di modelli di regressione distinti per ogni gruppo, seguiti da modelli in cui i coefficienti di regressione sono legati a variabili esplicative di secondo livello, non è però sufficiente per specificare un modello multilevel. E' necessario comprendere che esso implica un collegamento tra i modelli specificati ai diversi livelli: questo è il punto focale. Il tipo di integrazione più semplice si ha con i modelli di regressione a coefficienti casuali, per i quali i coefficienti di regressione di primo livello sono trattati come variabili casuali di secondo livello. Ciò significa che tali coefficienti sono originati da una distribuzione di probabilità. Assumere che i coefficienti di regressione siano variabili casuali, significa ritenere che ciascun gruppo costituisce un gruppo scelto a caso da una ipotetica popolazione di gruppi ed inoltre all'interno di ciascun gruppo, si assume che le unità statistiche rappresentino un campione casuale semplice estratto da una specifica popolazione ipotetica di unità di primo livello. L'aggiunta di covariate di secondo livello legate ai coefficienti

casuali, rende il modello ancora più generale. Inizialmente, si può considerare il caso di una sola variabile esplicativa. Per rendere il discorso concreto, si faccia riferimento, come specifico ambito di analisi, all'ambito scolastico. Naturalmente le considerazioni fatte possono essere estese a tutti gli altri campi in cui si manifestano dati a struttura gerarchica. Si considereranno nel seguito dei semplici esempi di analisi di regressione e di analisi della varianza (ANOVA) i cui corrispondenti modelli statistici possano essere visti come casi particolari dei modelli lineari gerarchici.
Si consideri, a tal proposito , la regressione dei risultati in matematica Y in funzione della variabile x, condizione socio-economica (SES); si pensi inizialmente ad un'unica scuola all'interno della quale vengono rilevati i risultati in matematica di un campione di n di studenti ($i = 1, ..., n$) (vedi Figura 1.23).

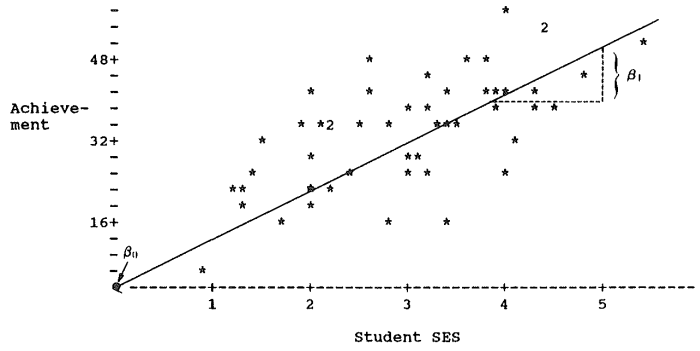

Figura 1.23: Scatterplot della relazione tra il risultato e SES in una ipotetica scuola

La nuvola dei punti può essere riassunta da una retta con intercetta β_0 e pendenza β_1. Una prima relazione ipotizzabile è data quindi dalla seguente equazione di regressione:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

in base alla quale la variabile dipendente osservata per il soggetto $i-esimo$, con $i = 1, ..., n$ è funzione lineare dello status socio economico della famiglia. La componente e_i definisce l'errore, casuale, associato all'individuo $i-esimo$. L'intercetta β_0 è definita come il risultato teorico in matematica di uno studente il cui SES è zero. La pendenza β_1 è l'incremento che si attende nel risultato in matematica quando aumenta un'unità di SES.
Tipicamente si assume che le e_i siano determinazioni di v.c. E_i normalmente distribuite con media nulla e stessa varianza pari a σ^2, ovvero $E_i \sim N(0, \sigma^2)$,

fra loro incorrelate. Risulta spesso comodo riscalare la variabile X, sottraendo la media $\overline{x}$ da ogni punteggio: $x_i - \overline{x}$ (vedi Figura 1.24). Se ora si interpretasse y con la retta di regressione in funzione di $x_i - \overline{x}$ si avrebbe che l'intercetta β_0 diviene pari alla media dei risultati in matematica mentre la pendenza rimane immutata.

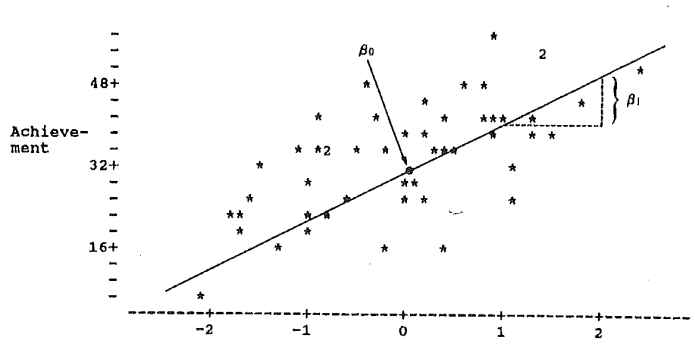

Figura 1.24: Scatterplot che mostra la relazione tra il risultato e SES (centrato)

Estendendo l'analisi al caso di due scuole (vedi Figura 1.25), le equazioni di regressione associate rispettivamente e separatamente alla scuola 1 e alla scuola 2 possono risultare del tipo:

$$y_{i1} = \beta_{01} + \beta_{11}(x_{i1} - \overline{x}_1) + r_{i1} \quad i = 1, ..., n_1$$

$$y_{i2} = \beta_{02} + \beta_{12}(x_{i2} - \overline{x}_2) + r_{i2} \quad i = 1, ..., n_2$$

I dati presi in considerazione indicano che le due scuole differiscono in due modi. Per prima cosa, la scuola 1 ha una media più alta della scuola 2, in quanto per le due intercette $\beta_{01} > \beta_{02}$. In secondo luogo, SES è meno predittivo del risultato in matematica nella scuola 1 rispetto alla scuola 2, come si evidenzia dal confronto tra le due pendenze $\beta_{11} < \beta_{12}$. Se si ipotizza che gli studenti siano stati assegnati casualmente alle due scuole, si potrebbe dire che la scuola 1 è in entrambi i casi più efficace e più giusta della scuola 2. La miglior efficacia è indicata dal valore medio più alto del livello di risultato nella scuola 1 ($\beta_{01} > \beta_{02}$). La miglior equità è indicata dalla pendenza più debole ($\beta_{11} < \beta_{12}$). Tuttavia, gli studenti non sono solitamente assegnati a caso nelle scuole, così molte interpretazioni degli effetti della scuola sono ingiustificate se non si considera la diversa composizione degli studenti.

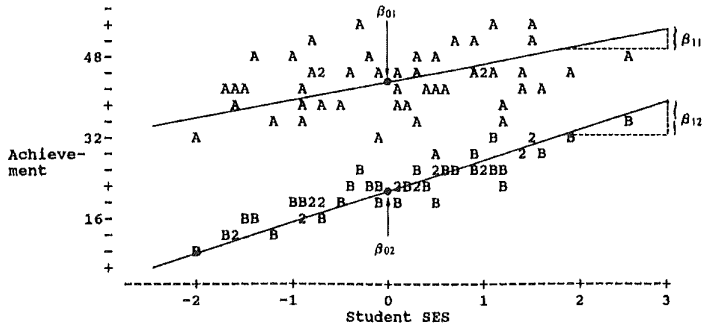

Figura 1.25: Scatterplot che mostra la relazione tra il risultato e SES in due ipotetiche scuole

Si ipotizzi ora di estendere la relazione studiata ad un'intera popolazione di scuole e di avere a disposizione solo un campione casuale di J scuole da detta popolazione. Non è pratico riassumere i dati con uno scatterplot per ogni scuola, ma è possibile comunque generalizzare l'equazione di regressione per la $j-esima$ scuola:

$$y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij} - \overline{x}_j) + e_{ij} \quad con \;\; i = 1, ..., n_j \quad e \quad j = 1, ..., J$$

dove per semplicità si assume che le e_{ij} siano determinazioni di v.c. E_{ij} normalmente distribuite con media nulla e varianza omogenea tra le scuole e tra gli studenti cioè $E_{ij} \sim N(0, \sigma^2)$. Per ogni scuola, l'efficacia e l'equità sono descritte da una coppia di valori (β_{0j}, β_{1j}). Quindi a differenza della regressione ordinaria sul campione di tutti gli studenti le cui prestazioni sono descritte da un unico modello, ciascuna unità $j-esima$ di secondo livello è caratterizzata da differente intercetta β_{0j} e differente coefficiente di regressione β_{1j}.
In tale situazione può risultare sensato e conveniente assumere che le intercette e le pendenze che caratterizzano la popolazione delle scuole abbiano una distribuzione normale bivariata (B_0, B_1), caratterizzata da 5 parametri:

$$E(B_0) = \gamma_0 \qquad E(B_1) = \gamma_1$$

$$Var(B_0) = \tau_{00} \qquad Var(B_1) = \tau_{11}$$

$$Cov(B_0, B_1) = \tau_{01}$$

che hanno il seguente significato

- γ_0 valore atteso per il risultato in matematica dell'intero sistema scolastico
- τ_{00} variabilità tra i valori medi delle scuole

- γ_1 è il valore atteso per il coefficiente angolare della generica scuola rappresentativa dell'intero sistema scolastico
- τ_{11} variabilità tra i coefficienti angolari delle scuole
- τ_{01} covarianza tra i coefficienti angolari e le intercette nell'intera popolazione di scuole

Un valore positivo di τ_{01} implica che le scuole con una media più alta tendono anche ad avere una pendenza positiva. Nelle applicazioni pratiche il vero valore dei parametri della popolazione $(\gamma_0, \gamma_1, \tau_{11}, \tau_{00}, \tau_{01})$ e i veri valori relativi alle singole scuole (means and slopes) (β_{0j} e β_{1j}), devono essere stimati tramite i dati.
Si consideri, ad esempio, lo scatterplot della relazione tra le stime $\hat{\beta}_{0j}$ e $\hat{\beta}_{1j}$ per un ipotetico campione di 200 scuole (vedi Figura 1.26).

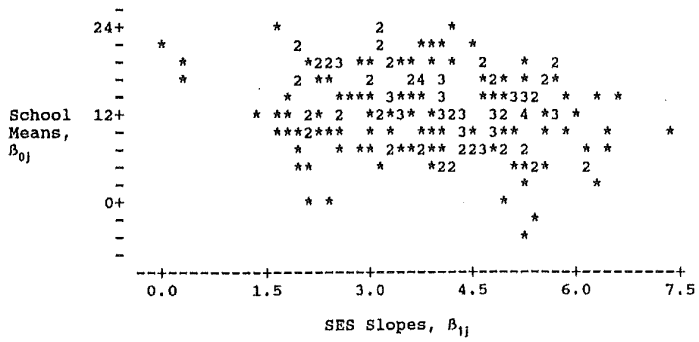

Figura 1.26: Scatterplot delle medie delle scuole (asse verticale) e pendenze SES (asse orizzontale) per 200 ipotetiche scuole

Da notare che c'è più dispersione tra i livelli medi che tra le pendenze, ovvero $\tau_{00} > \tau_{11}$. Si nota anche che i due effetti tendono ad essere correlati negativamente: le scuole con alta media di successo β_{0j} tendono ad avere una relazione debole SES-risultato β_{1j}. Simbolicamente $\tau_{01} < 0$.
La scuola efficace ed egalitaria (alta media dei risultati con un grande valore di β_{0j} e debole effetto SES, piccoli valori di β_{1j}) è quella,ad esempio, che si trova, sulla parte alta dello scatterplot, della precedente Figura 1.25, in cui i punti sono contrassegnati dal simbolo (A).
Un'ulteriore analisi di interesse potrebbe essere poi quella di considerare particolari raggruppamenti di scuole, quali ad esempio, le pubbliche e le private. La caratterizzazione dei gruppi può essere fatta introducendo opportune variabili esplicative di livello superiore. Nel caso di pubbliche e private si può considerare una semplice variabile indicatrice W che assume valore $w_j = 1$

per le scuole private e valore $w_j = 0$ per le scuole pubbliche. Coleman, Hoffer and Kilgore (1982) sostenevano che W è positivamente correlata con l'efficacia (le scuole private hanno una più alta media dei risultati rispetto alle scuole pubbliche) e negativamente legate alla pendenza (l'effetto SES sui risultati in matematica risulta minore nella scuola privata, rispetto alla scuola pubblica). In questo caso il modello è espresso attraverso le relazioni:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}$$

dove

- γ_{00} è la media dei risultati delle scuole pubbliche
- γ_{01} è la media della differenza dei risultati tra la scuola privata e pubblica (efficacia della scuola privata)
- γ_{10} è la media delle pendenze nelle scuole pubbliche
- γ_{11} è la differenza media nella pendenza SES-risultato tra la scuola privata e la scuola pubblica (vantaggio nell'equità della scuola privata)
- u_{0j} è l'effetto della scuola j sulla costante
- u_{1j} è l'effetto della scuola j sulla pendenza

Si assume che u_{0j} e u_{1j} siano determinazioni di variabili casuali U_0 e U_1 con media nulla e varianza rispettivamente τ_{00} e τ_{11} (rappresentano la variabilità di β_{0j} e β_{1j} al netto dell'effetto fisso di W) e covarianza τ_{01} e che siano indipendenti dalle componenti erratiche di primo livello e_{ij}. Si noti che i coefficienti di regressione γ non compaiono più con l'indice j; infatti essi sono le medie (delle corrispondenti v.c.); la variabilità è ora descritta dalle determinazioni u_{0j} u_{1j} che rappresentano la diversità tra le singole scuole. Sostituendo le equazioni

$$\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j}$$

e

$$\beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}$$

nell'equazione

$$y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij} - \overline{x}_{.j}) + e_{ij}$$

si ottiene l'equazione generale del modello gerarchico a due livelli:

$$y_{ij} = \gamma_{00} + \gamma_{01} w_j + \gamma_{10}(x_{ij} - \overline{x}_{.j}) + \gamma_{11} w_j (x_{ij} - \overline{x}_{.j}) + u_{0j} + u_{1j}(x_{ij} - \overline{x}_{.j}) + e_{ij}$$

Questa equazione si discosta però da quella tipica di un modello lineare, per la cui stima si usano in genere gli OLS, dove gli errori si ipotizzano essere indipendenti, normalmente distribuiti e con varianza costante. In questo modello invece gli errori hanno la forma più complessa

$$u_{0j} + u_{1j}(x_{ij} - \overline{x}_{.j}) + e_{ij}$$

Tali errori sono dipendenti dalle singole scuole, perchè le componenti u_{0j} e u_{1j} sono comuni ad ogni studente della scuola j-esima. Gli errori hanno anche una varianza diversa perchè la quantità

$$u_{0j} + u_{1j}(x_{ij} - \overline{x}_{.j})$$

dipende da u_{0j} e u_{1j} che variano tra le scuole, mentre il valore $(x_{ij} - \overline{x}_{.j})$ varia tra gli studenti. Solo nel caso particolare in cui u_{0j} e u_{1j} fossero nulle per ogni j l'equazione

$$Y_{ij} = \gamma_{00} + \gamma_{01} w_j + \gamma_{10}(x_{ij} - \overline{x}_{.j}) + \gamma_{11} w_j (x_{ij} - \overline{x}_{.j}) + u_{0j} + u_{1j}(x_{ij} - \overline{x}_{.j}) + e_{ij}$$

diverrebbe equivalente al modello di regressione OLS.

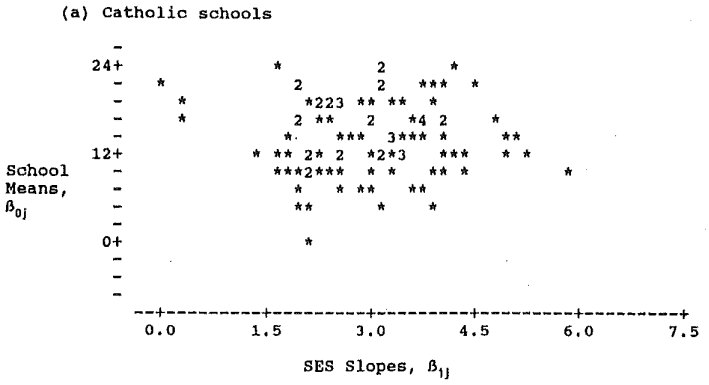

Figura 1.27: relazione tra media e pendenza (scuole cattoliche)

Nelle figure 1.27 e 1.28 si ha una rappresentazione grafica del modello specificato dalle equazioni

$$\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j}$$

e

$$\beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}$$

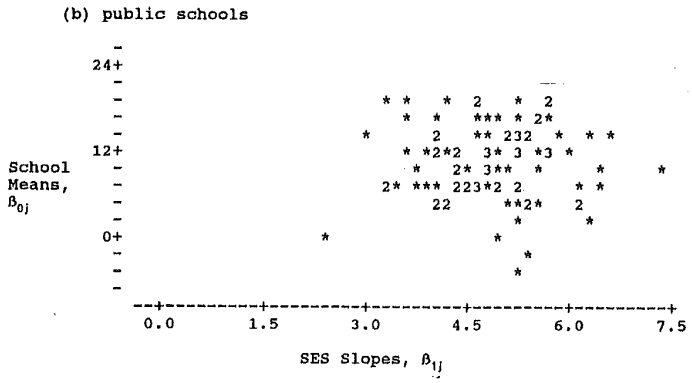

Figura 1.28: relazione tra media e pendenza (scuole pubbliche)

1.3.1 Effetti fissi e casuali

In letteratura si riscontrano pareri differenti (Gelman, Hill, 2007) al riguardo dell'utilizzo di procedure con effetti fissi o casuali. In genere si sostiene che gli effetti fissi sono più appropriati se interessano i coefficienti del livello dei gruppi, mentre gli effetti casuali quando si ha interesse nello studio dell'intera popolazione (si pensi ad esempio, al problema di stima della media generale della popolazione che può ottenersi come media, ovviamente pesata, delle medie parziali o di gruppo). Altro criterio è quello di utilizzare gli effetti fissi quando i gruppi osservati rappresentano tutti i possibili gruppi e utilizzare invece gli effetti casuali quando la popolazione contiene gruppi non sempre presenti nei dati. Gelman e Hill (2007) consigliano comunque di utilizzare sempre effetti casuali nei modelli multilivello.
La natura non fornisce fattori etichettati come fissi e come casuali. Ogni fattore può assumere l'una o l'altra caratteristica in funzione delle ipotesi specificate. La distinzione è basata sul modo in cui i livelli del fattore sono scelti. Un fattore si dice "fisso" quando i livelli da includere nell'esperimento sono definiti nel problema in esame e sono "fissati" dallo sperimentatore. Nel caso di un fattore fisso tutti i livelli rilevanti all'analisi sono inclusi nell'esperimento. Un fattore si dice "casuale" se i livelli inclusi nell'esperimento sono un sottoinsieme di quelli teoricamente possibili e la selezione avviene in modo casuale. Nel caso di un fattore casuale solo un campione di possibili livelli è incluso nell'esperimento. Un criterio utile per stabilire se un fattore è fisso o casuale consiste nel porre la domanda se l'eventuale ripetizione dello studio e quindi il riesame della stessa ipotesi può essere eseguito includendo esattamente gli stessi livelli del fattore considerato, potranno presentarsi livelli diversi. Vi sono quindi differenze sostanziali nel modo in cui sono definiti i livelli di un fattore fisso e quelli di un fattore casuale. Vi sono anche differenze sostanziali tra le due classi di fattori nel contenuto informativo e quindi

nelle conclusioni che possono essere tratte dalla analisi. Nel caso di un fattore fisso i livelli del trattamento sono misurati senza errore. Ciò deriva dal fatto che per un tale fattore tutti i livelli rilevanti all'analisi dell'ipotesi sono stati inclusi nell'esperimento. Le conclusioni dello studio sono tuttavia limitate ai soli livelli esaminati. Nel caso di un fattore casuale, invece, i livelli si presentano in modo casuale. Esiste quindi una varianza associata a ciascun effetto in conseguenza del fatto che l'esperimento include un campione di possibili livelli di fattore. In questo caso, tuttavia, le conclusioni dello studio sono generalizzabili a tutta la popolazione statistica da cui i livelli sono stati estratti. Se si considera ad esempio l'analisi del metodo di insegnamento nelle scuole e si è interessati al metodo di insegnamento sull'intera popolazione di tutte le scuole si sceglie casualmente un campione di scuole dall'intera popolazione. Questo campione risulta essere rappresentativo delle caratteristiche dell'intera popolazione e in questo caso i fattori sono da considerarsi casuali. Se si è invece interessati solo alle caratterestiche di un certo numero di scuole e le si considerano tutte nell'analisi, i fattori sono da considerarsi fissi (Goldstein, "Multilevel Statistical Models"). In sostanza è possibile per un fattore essere considerato fisso in alcune circostanze e casuale per altre.

In definitiva, quindi, i modelli multilivello possono essere considerati in due differenti modi: usando modelli con effetti fissi e modelli con effetti casuali. Quale di queste due situazioni è la più appropriata dipende dall'obbiettivo dell'analisi inferenziale, la natura dei gruppi e la natura specifica del campionamento. Per Snijders e Boskers (1999):

1. Se i gruppi sono considerati come specifiche entità e il ricercatore intende in primo luogo trarre conclusioni riguardanti ciascuno di questi specifici gruppi, allora è opportuno utilizzare un modello con effetti fissi;
2. Se i gruppi sono considerati un campione da una (reale o ipotetica) popolazione e il ricercatore vuole trarre conclusioni riguardanti la popolazione, allora il modello ad effetti casuali è il più appropriato.
3. Se il ricercatore vuole verificare gli effetti della variabile del livello del gruppo, si dovrebbe usare il modello ad effetti casuali, questo perchè il modello ad effetti fissi "mostra" già tutte le differenze tra i gruppi, e non spiega la variabilità tra i gruppi lasciando che sia spiegata dalla variabile di livello del gruppo.
4. Specialmente per gruppi relativamente piccoli, i modelli ad effetti casuali hanno particolari vantaggi, purchè le assunzioni sugli effetti siano ragionevoli. Il modello comprende le assunzioni di indipendenza e somiglianza al riguardo della distribuzione degli effetti dei gruppi u_{0j}. In maniera meno formale, gli effetti non spiegabili del gruppo sono governati da meccanismi abbastanza simili tra gruppo e gruppo e operano indipendentemente da gruppo a gruppo. I gruppi sono cioè considerati scambiabili.
5. Il modello con effetti casuali è spesso usato con l'aggiunta, ad esempio, dell'assunzione che gli effetti casuali siano normalmente distribuiti. Se questa ipotesi è poco attendibile, i risultati ottenuti con questa analisi risulterebbero non realistici.

Altre considerazioni riguardanti la scelta di effetti fissi o casuali sono trattate anche nel Searle (1992, sezione 1.4).

E' importante notare che le proprietà dei fattori fissi e dei fattori casuali possono essere combinate con opportuni disegni sperimentali misti, permettendo così l'analisi di situazioni più complesse.

1.3.2 Un esempio

Premettiamo un semplice esempio, riguardante la coltivazione dei pomodori, che prevede una prima fase di trapianto delle piantine e l'uso di fertilizzanti per favorire la crescita. Per verificare se l'uso di uno specifico fertilizzante influenza la crescita dei pomodori si è estratto un campione di 25 piante di pomodoro, a cui è stato assegnato casualmente uno dei due fertilizzanti (1=fertilizzante 1; 2=fertilizzante 2); l'altezza di ogni piantina di pomodoro è stata misurata al momento del trapianto x e dopo dieci giorni y. La risposta y è rappresentata dall'altezza delle piantine di pomodoro dopo essere state trapiantate e trattate.

L'analisi potrebbe essere effettuata mediante la consueta analisi della varianza ad un criterio di classificazione. In tal caso la procedura si basa sul confronto delle medie dei risultati osservati sui due gruppi di piantine alle quali sono stati somministrati i fertilizzanti in questione. Così procedendo, implicitamente si suppone che i soggetti presi in esame siano tutti caratterizzati dalla stessa altezza iniziale, mentre, si è detto, ciascuna piantina ha una sua altezza particolare. Si comprende allora come nella esemplificazione appena proposta, e in tutti i casi dove entrano in gioco attitudini non omogenee delle unità sperimentali, divenga necessario ipotizzare gli effetti di un ulteriore fattore, che viene chiamato, in questo caso, fattore aggiuntivo di disturbo. In generale ci si propone di confrontare, mediante il *test* F, due o più medie e arrivare a decidere se esista o meno l'influenza dei vari fattori. Essendo solo due i fertilizzanti considerati nell'esempio (si confrontano solo due medie campionarie) si otterrebbero gli stessi risultati utilizzando il *test* t *di Student*. Il *test* F consente di confrontare due o più gruppi, mentre il *test* t permette di confrontarne solamente due. Se si osserva l'esito dell'esecuzione di entrambe le procedure, il risultato suggerisce di prendere le medesime decisioni circa l'ipotesi nulla. In particolare, il valore di F impiegato nel confronto tra 2 medie, altro non è che il valore di t elevato al quadrato: $F = t^2$ (Kinnear e Gray, 2006).

Ricorrendo con i dati in esame all'ANOVA si ottengono i seguenti risultati: fissato un livello di significatività pari ad $\alpha = 0,05$, essendo $F = 1,65 < 4,28 = F_{0,95}$ valore critico del test, l'ipotesi dell'uguaglianza degli effetti provocati dai predetti due fertilizzanti può essere accettata.
Tuttavia, l'approccio appena descritto non tiene conto del fatto che ognuna delle piantine è caratterizzata da una propria altezza nel momento del

trapianto e pertanto queste non possono intendersi omogenee, come invece dovrebbe essere, per dare corretto significato ai risultati della suddetta analisi.

Accade, cioè, che i risultati, ad un più attento esame, si rivelino per così dire "inquinati" dagli effetti di un fattore di disturbo, non considerato in precedenza ma che, una volta riconosciuto come suscettibile di esercitare una qualche influenza, deve esser tenuto nel debito conto, se si vuole che l'esito dell'analisi probabilistica fornisca indicazioni corrette. Per depurare i valori osservati dai valori che esprimono l'effetto del fattore di disturbo è opportuno riconoscere che tra la variabile x, detta anche "covariata", esiste un legame con la variabile y oggetto di interesse. L'analisi in questi casi va sotto il nome di analisi della covarianza, nella quale si suppone che esista un legame, lineare, fra l'altezza iniziale prima del trattamento (x) e finale dopo il trattamento (y) delle piantine e che quindi y dipenda oltre che dal fertilizzante anche dall'altezza iniziale x.
L'analisi può poi essere preceduta dalla cosiddetta verifica d'ipotesi di parallelismo delle rette di regressione che interpretano per ciascun trattamento il predetto legame lineare: se le rette sono parallele il legame è simile nei due gruppi e può essere quindi espresso con un'unico parametro k, invece che da due distinti coefficienti angolari.

Nel nostro caso l'ipotesi di parallelismo delle rette di regressione può essere accettata (in quando non risulta significativo l'effetto di interazione). Si perviene infatti al seguente risultato (vedi Figura 1.29):

Tests of Between-Subjects Effects

Dependent Variable:y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Model	6,196E6	4	1549055,643	1469,067	,000	,996
Fertilizzante	51554,969	2	25777,485	24,446	,000	,700
x	15026,810	1	15026,810	14,251	,001	,404
Fertilizzante * x	5,435	1	5,435	,005	,943	,000
Error	22143,427	21	1054,449			
Total	6218366,000	25				

a. R Squared = ,996 (Adjusted R Squared = ,996)

Figura 1.29: Ancova.

fissato in $\alpha = 0,05$ il livello di significatività, l'ipotesi riguardante l'uguaglianza di effetti dei due fertilizzanti deve ora essere rifiutata. Il risultato ottenuto con l'impiego dell'analisi della covarianza è del tutto opposto a quello ottenuto con l'impiego della sola analisi della varianza, potendo concludere che, ritenendo valida l'influenza lineare della covariata x, i due fertilizzanti hanno una diversa inflenza sull'altezza delle piante di pomodoro dopo dieci giorni dal trapianto.

Il modello con dati a struttura gerarchica diviene in questo caso del tipo:

$$y_{ij} = \alpha_j + \beta x_{ij} + e_{ij}$$

dove le α_j rappresentano gli effetti dei fertilizzanti e β il coefficiente (unico) che esprime il legame lineare tra le altezze osservate a 10 giorni di distanza. Le opzioni possibili sono due:

1. assumere che $\alpha_1, ..., \alpha_J$ siano delle costanti da stimare (*effetti fissi*). Ciò equivale ad usare un modello di analisi della covarianza (ANCOVA).
2. assumere che $\alpha_1, ..., \alpha_J$ siano effetti *effetti casuali*, cioè realizzazioni di una variabile aleatoria $N(\alpha, \sigma_u^2)$ di cui ci si limita a stimare media e varianza.

La scelta tra le due opzioni dovrebbe basarsi sulla natura dei gruppi e sul fine dell'indagine. Si useranno:

- effetti fissi se nel campione sono presenti tutti i possibili gruppi (ad esempio, pazienti trattati con una serie di farmaci alternativi) oppure se sono presenti tutti i gruppi di interesse ai fini dell'analisi (ad esempio, se il fine è quello di studiare solo le scuole incluse nel campione senza voler estendere i risultati anche ad altre scuole);
- effetti casuali se nel campione sono presenti dei gruppi che rappresentano una popolazione di gruppi e il fine dell'analisi è quello di estendere i risultati a tale popolazione.

In teoria per gli effetti casuali, a seguito della usuale assunzione di normalità, la popolazione da cui provengono i gruppi[4] dovrebbe avere un numero infinito di elementi, anche se, in pratica questo requisito non è sempre soddisfatto.

1.4 Un modello generale e alcuni semplici sottomodelli

I modelli multilevel hanno in generale lo scopo di spiegare il legame tra una variabile dipendente e un insieme di variabili esplicative. In questi modelli la funzione è specificata in modo tale da considerare esplicitamente la struttura gerarchica dei dati, integrando l'analisi a livello individuale con quella a livello aggregato. Le assunzioni basilari che un modello di regressione lineare deve assicurare sono: linearità delle relazioni, normalità, omoschedasticità ed indipendenza degli errori. In un modello multilevel le prime due assunzioni sono in genere rispettate, mentre le successive, soprattutto l'indipendenza, in genere non lo sono. L'idea che sta alla base dei modelli multilevel è quella di considerare non un'unica equazione, per tutti i gruppi in cui può essere suddivisa la popolazione, ma un insieme di equazioni con parametri diversi per ogni gruppo di individui. Le possibili variabili esplicative sono legate ai

[4] un'interessante discussione sulle implicazioni della scelta del modello è contenuta nel classico articolo di Aitkin e Longford (1986).

diversi livelli della struttura gerarchica della popolazione.
Le condizioni operative che stanno alla base dell'analisi dei modelli multilevel sono le seguenti:

- un dataset strutturato gerarchicamente;
- una variabile dipendente misurata a livello più basso;
- delle variabili esplicative misurate sui diversi livelli della gerarchia.

La metodologia dei modelli multilivello può essere introdotta in modo conveniente, e senza perdita di generalità, con riferimento ad una struttura gerarchica dei dati a due soli livelli di raggruppamento. Più esattamente, si supponga che le singole unità di osservazione, nonchè elementari o di primo livello, siano aggregate in J gruppi di unità di secondo livello e si assuma che le unità elementari raggruppate entro il $j-esimo$ gruppo siano pari a n_j $(j = 1, ..., J)$. Sia Y la variabile oggetto di interesse osservata sulle unità elementari; lo scopo è di indagare in merito al legame di dipendenza che sussiste fra la stessa e una variabile esplicativa X. Si suppone, inoltre, che tale legame possa non mantenersi costante da gruppo a gruppo, ma vari in relazione, per esempio, all'azione di una variabile esplicativa che interviene al secondo livello. Il modello multilivello si propone di collegare con un'unica formulazione statistica modelli di regressione specificabili separatamente entro i diversi gruppi. Formalmente, la relazione fra X e Y viene espressa a livello del $j-esimo$ gruppo tramite il seguente modello:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij} \tag{1.3}$$

dove l'osservazione y_{ij} del fenomeno di interesse effettuata sulla $i-esima$ unità elementare entro l'unità $j-esima$ del secondo livello è generato da una variabile aleatoria y, x_{ij} è il valore assunto dalla variabile esplicativa X sulla stessa unità, mentre e_{ij} indica la componente casuale d'errore del modello $(i = 1, ..., n_j; j = 1, ..., J)$.
Si assume, inoltre, che gli errori e_{ij}, nel seguito detti "di primo livello", abbiano valore atteso nullo, varianza costante pari a σ^2 e siano fra loro incorrelati all'interno dello stesso gruppo così come fra gruppi diversi. Il modello definito dall'equazione (1.3) è detto modello di livello 1. Come si nota i parametri β_{0j} (intercetta relativa al gruppo $j-esimo$) e β_{1j} (coefficiente angolare della variabile X relativo al gruppo $j-esimo$) dipendono dall'indice j di gruppo. Con ciò si vuole indicare che, al variare del gruppo, le rette di regressione possono essere caratterizzate da diversa intercetta e/o da diversa pendenza, e quindi che la variabile X può esercitare un'influenza lineare diversa da gruppo a gruppo. Solitamente, la natura variabile dei parametri β_{0j} e β_{1j} viene a sua volta espressa mediante modelli di regressione. Questi ultimi possono prevedere la presenza di una variabile esplicativa W di secondo livello, che agisce con intensità differente da gruppo a gruppo, ma costante all'interno dello stesso gruppo $j = 1, ..., J$, secondo le relazioni:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j}$$

e

$$\beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}$$

tale modello è detto di livello 2 a parametri γ_{00}, γ_{10} e, se presenti, γ_{01}, γ_{11}, essendo indipendenti dalla struttura di gruppo, sono fissi. Le variabili U_{0j} e U_{1j} che generano le determinazioni u_{0j} e u_{1j} costituiscono, invece, la parte aleatoria dei rispettivi modelli e vengono dette effetti casuali o errori di secondo livello. Si assume che esse abbiano valore atteso nullo, ma non necessariamente varianze uguali, e in generale siano fra loro correlate.
Considerando l' equazione del modello al livello 1

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$$

e le equazioni del modello al livello 2

$$\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}$$

con le dovute sostituzioni si perviene alla definizione del cosiddetto modello combinato:

$$y_{ij} = \gamma_{00} + \gamma_{01} w_j + \gamma_{10} x_{ij} + \gamma_{11} w_j x_{ij} + u_{0j} + u_{1j} x_{ij} + e_{ij} \tag{1.4}$$

per $i = 1, ..., n_j$ e $j = 1, .., J$, nel quale:

- $\gamma_{00} + \gamma_{01} w_j + \gamma_{10} x_{ij} + \gamma_{11} w_j x_{ij}$ costituisce la parte deterministica del modello (γ_{00} è l'intercetta o la costante; γ_{01} rappresenta l'effetto della variabile esplicativa del livello 2; γ_{10} indica l'effetto dei predittori del livello 1; γ_{11} è l'effetto della interazione cross-level tra i predittori del livello 1 e quelli del livello 2), il prodotto $\gamma_{11} w_j x_{ij}$ indica l'interazione fra il primo e il secondo livello (interazioni cross-level).
- $u_{0j} + u_{1j} x_{ij} + e_{ij}$ costituisce la parte casuale del modello (u_{0j} sono i residui della variabilità del livello 2 sull'intercetta del livello 1 al netto della variabile esplicativa w_j; u_{1j} indica i residui della variabilità del livello 2 sulla pendenza del livello 1 al netto di x_{ij}; e_{ij} sono gli errori al livello 1 omettendo le variabili esplicative del primo livello, misurano gli errori in y_{ij} rappresentano la variabilità in y_{ij} attribuibile alle unità del livello 1).

La specificazione del modello (1.4) non è completa se non si esplicitano le assunzioni sulla parte residuale del modello $u_{0j} + u_{1j} x_{ij} + e_{ij}$.

Nell'analisi multilivello si considerano le seguenti assunzioni concernenti le componenti residuali :

1. $E(U_{0j}) = E(U_{1j}) = E(E_{ij}) = 0$; questo implica che non ci sono errori sistematici nei parametri e nel modello 1.

2. $Var(U_{0j}) = \tau_{00}$, $Var(U_{1j}) = \tau_{11}$, $Var(E_{ij}) = \sigma^2$; questo postula che gli errori del livello 1 e del livello 2 hanno varianza costante [5].
3. $Cov(U_{0j}, U_1) = \tau_{01}$; questo perchè gli errori dell'intercetta e della pendenza al livello 2 possono essere correlati. La covarianza τ_{01} cattura la relazione tra l'intercetta e la pendenza e, in generale, uno può sempre stimare questi termini (Snijders and Bosker 1999).
4. U_{0j} e U_{1j}; sono normalmente distribuite, come E_{ij}. [6] Considerando contemporaneamente le assunzioni (1) - (4) si ha che i residui di livello 2 sono descritti da una distribuzione normale bivariata con media nulla e matrice di varianza-covarianza:

$$\Sigma = \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}$$

   mentre i residui del livello-1 si distribuiscono con una distribuzione normale con media nulla e varianza σ^2.
5. $Cov(U_{0j},\ E_{ij}) = Cov(U_{1j},\ E_{ij}) = 0$ questo implica che gli errori della pendenza e dell'intercetta sono incorrelati con gli errori delle unità del primo livello al netto delle variabili dipendenti. Questa assunzione è necessaria per ottenere un modello ben identificato. Questo implica inoltre che gli errori del livello-1 abbiano varianza costante.

Si consideri l'espressione dei residui del modello multilivello:

$$\delta_{ij} = u_{0j} + u_{1j}x_{ij} + e_{ij}$$

essi costituiscono determinazioni di variabili casuali che indicheremo con Δ_{ij}, che non sono caratterizzate da varianza costante [7], infatti:

$$\begin{aligned} Var(\Delta_{ij}) &= E[(U_0 + U_1X_{ij} + E_j)^2] = \\ &= E[U_{0j}^2] + 2x_{ij}E[U_0, U_1] + x_{ij}^2E[U_1^2] + E[E_{ij}^2] = \\ &= \tau_{00} + 2x_{ij}\tau_{01} + x_{ij}^2\tau_{11} + \sigma^2 \end{aligned} \tag{1.5}$$

E' chiaro che $Var(\Delta_{ij})$, e quindi $Var(Y_{ij})$, è in parte una funzione dei predittori di livello-1, quindi Δ_{ij} ha una varianza non costante (sebbene U_0, U_1 e E_{ij} abbiano varianza costante per l'assunto (2)). Si avrà varianza costante solo nel caso in cui $U_1 = 0$, questo vuol dire che w_j definisce in maniera esatta

[5] Questa assunzione potrebbe essere rilassata per gli errori del livello 1 (si veda Browne et al. 2000; Snijders and Bosker 1999). E' nota anche un'applicazione in cui le unità del livello 2 sono caratterizzate da differenti strutture di varianza-covarianza (Thum 1997).

[6] Modelli per dati categoriali, conteggio (count), o dati di durata richiedono una specificazione differente per quanto concerne la distibuzione degli errori del livello 1

[7] per questa dimostrazione, bisogna considerare un'assunzione addizionale, cioè che $Cov(e_{ij}, e_{kl})$ per $i \neq j$, $k \neq l$. Questa assunzione è inclusa solamente per convenienza e potrebbe essere rilassata nelle applicazione di modelli multilivello per time series of pooled cross-sections (Goldstein 1995)

la differenza della pendenza rispetto ad x_{ij} al netto delle unità di livello-2. Si fa inoltre osservare che i residui dei modelli multilivel sono anche correlati, per le unità di livello-1 nested nelle unità di livello-2. Indicati con δ_{ij} e δ_{kj} due generici residui del livello-2, abbiamo infatti:

$$\begin{aligned} Cov(\Delta_{ij}, \Delta_{kj}) &= E[(U_0 + U_1 X_{ij} + E_{ij})(U_0 + U_1 X_{ij} + E_{ij})] = \\ &= E[U_0^2] + x_{ij}E[U_0, U_1] + x_{kj}E[U_0, U_1] + x_{ij}x_{kj}E[U_1^2] = \\ &= \tau_{00} + x_{ij}\tau_{01} + x_{kj}\tau_{01} + x_{ij}x_{kj}\tau_{11} \end{aligned} \tag{1.6}$$

Questa covarianza assumerà valore nullo nel caso in cui $U_0 = U_1 = 0$. Questo significa che le w_j definiscono in maniera esatta la variazione delle unità di livello-2 nel modello intercetta e pendenza di livello-1 . Dalla covarianza (1.6) si ottiene la cosiddetta correlazione intra-classe:

$$\rho = \frac{Cov[\Delta_{ij}, \Delta_{kj}]}{\sqrt{Var(\Delta_{ij})}\sqrt{Var(\Delta_{kj})}}$$

Questo coefficiente fornisce una misura dell'omogeneità all'interno di uno stesso gruppo, ma rappresenta anche la proporzione di varianza residua spiegata dal raggruppamento (Kreft e De Leeuw, 1998).
Il modello (1.4) è proposto nella sua formulazione più generale. Infatti, potrebbe non essere necessario specificare nel modello tutte le ipotesi relative a tutte le componenti casuali, così come potrebbe non essere necessario spiegare la variazione dei parametri β_{0j} e/o β_{1j} mediante la variabile esplicativa W, e neppure inserire la variabile esplicativa X, come accade nel modello di analisi della varianza ad effetti casuali. Ne discende che il modello può essere specificato nei modi più appropriati a seconda delle relazioni ipotizzate sulle variabili.

Si ricorda che dato un singolo predittore al livello-1 X_{ij} e un singolo predittore al livello-2 W_j il modello è dato da

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_j + u_{1j}$$

che è il più semplice modello lineare gerarchico. Quando alcuni termini di questo modello sono posti uguali a zero ci si riconduce a modelli ancora più semplici e familiari. Si cerca ora di illustrare la connessione con i più comuni metodi di analisi dati, sia per dimostrare il range di applicazione della modellistica multilevel, sia per evidenziare le loro connessioni con il modello più generale.

1.4.1 One-way ANOVA con effetti fissi

Nel caso in cui si considerino gli effetti fissi (i livelli del fattore sono scelti non casualmente dallo sperimentatore, ma consistono in tutti i possibili livelli presenti nella popolazione), il modello ANOVA può essere presentato come modello multilevel "degenere". Può infatti rappresentare un caso particolare di modello gerarchico nel quale l'equazione di primo livello risulta:

$$y_{ij} = \beta_{0j} + e_{ij}$$

dove $E_{ij} \sim N(0, \sigma^2)$, mentre l'equazione relativa all'unico coefficiente di regressione è definita dalla seguente riparametrizzazione

$$\beta_{0j} = \gamma_{00} + \alpha_j$$

dove γ_{00} rappresenta la media dei β_{0j} e quindi $\sum \alpha_j = 0$.

La y_{ij} è la $i-esima$ misurazione per il trattamento $j-esimo$, β_{0j} è l'effetto dovuto al trattamento $j-esimo$ ed e_{ij} è l'errore casuale ovvero la differenza dell'osservazione $i-esima$ del trattamento j dal suo valore di riferimento. L'ANOVA viene realizzata verificando, attraverso il confronto tra varianze cosiddette *between* e *within*, la significatività di almeno due effetti α_j.

1.4.2 One-way ANOVA con effetti casuali

Il modello lineare gerarchico più semplice risulta quello equivalente ad un modello ANOVA con effetti casuali, per il quale l'equazione di primo livello è ancora:

$$y_{ij} = \beta_{0j} + e_{ij}$$

dove β_{0j} rappresenta il risultato medio per la $j-esima$ unità di livello-2 e quindi $\beta_{0j} = \mu_{Y_j}$. L'equazione relativa all'unico coefficiente di regressione è definita da

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

dove γ_{00} rappresenta il valore medio della popolazione per la variabile risultato (media generale della variabile sotto osservazione) mentre u_{0j} è l'effetto casuale associato alla $j-esima$ unità di secondo livello, con media nulla e varianza pari a τ_{00} (ovvero l'effetto casuale dovuto all'appartenenza al $j-esimo$ gruppo). Si assume che ogni errore al livello-1 e_{ij} sia determinazione di una v.c. normalmente distribuita con media pari a zero e varianza costante e pari a σ^2. Si nota che questo modello predice il risultato per ogni unità al livello-1 con solo un parametro al livello-2, l'intercetta β_{0j}. In questo caso β_{0j} è la media delle unità di livello-2, cioè $\beta_{0j} = \mu_{Y_j}$. L'equazione di un modello a due livelli per una one-way ANOVA con effetti casuali, e con γ_{01} del modello

generale, posto uguale a zero, si ottiene sostituendo l'equazione

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

nell'equazione

$$Y_{ij} = B_0 + E_{ij}$$

Si ottiene così la formula per esprimere il modello:

$$Y_{ij} = \gamma_{00} + U_0 + E_{ij}$$

che è certamente un modello one-way ANOVA con media di popolazione γ_{00} con un effetto gruppo (livello-2) u_{0j}; e con un effetto soggetto (livello-1) e_{ij}. Questo modello, nella terminologia dei modelli multilivello, è definito *model intercept only* ossia senza variabili esplicative. Anche in questo modo la varianza totale di Y_{ij} è composta da due componenti: σ^2 che indica la varianza dei residui della variabile a livello individuale (varianza within) e τ_{00} che rappresenta la varianza dei residui della variabile a livello aggregato (varianza between). Infatti, dalla $y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$ e dalle assunzioni fatte in precedenza la variabilità totale può essere scomposta in due componenti:

$$Var(Y_{ij}) = Var(U_0 + E_{ij}) = \tau_{00} + \sigma^2$$

In realtà ci si trova alle prese con un modello di analisi della varianza ad effetti casuali, dove si assume che la componente u_{0j} vari casualmente tra le unità di secondo livello, mentre e_{ij} rappresenta la componente erratica associata all'$i-esima$ unità di primo livello. Dal momento che le osservazioni si assumono correlate positivamente all'interno di ciascun gruppo, a causa del fattore di appartenenza al gruppo, si avrà che:

$$Cov(Y_{ij}, Y_{i'j'}) = \begin{cases} 0 \; se \; j \neq j' \\ \tau \; se \; j = j' \; e \; se \; i \neq i' \end{cases}$$

In altri termini, σ^2 e τ rappresentano rispettivamente la varianza all'interno dei gruppi (within) e la varianza tra i gruppi (between). L'analisi della varianza tradizionale ad effetti fissi, come è noto, fa riferimento ad un numero fisso e noto di gruppi (all'interno dei quali si distribuiscono le osservazioni), tale da implicare l'esaustività della classificazione ed evitare ipotesi sulla distribuzione della popolazione di gruppi dalla quale, in caso di campionamento, i gruppi verrebbero estratti. L'ANOVA ad effetti casuali, di contro, in virtù dell'introduzione dell'effetto casuale, permette, sotto opportune ipotesi distribuzionali, di procedere all'inferenza.

L'ANOVA a effetti casuali è spesso utilizzata come analisi preliminare dei dati con struttura gerarchica, producendo in prima istanza una stima puntuale e un intervallo di confidenza per la media della popolazione γ_{00}.
Inoltre, un usuale indicatore associato alla ANOVA a effetti casuali è la cosiddetta correlazione intra-classe, cioè la correlazione esistente tra due individui

appartenenti allo stesso gruppo. Il coefficiente di correlazione intra-classe ρ è definito come:

$$\rho = \frac{\tau_{00}}{(\tau_{00} + \sigma^2)}$$

e varia tra 0 e 1; esso rappresenta una misura che giustifica il ricorso al modello gerarchico.Un valore del coefficiente molto basso, infatti, non segnalando la presenza di correlazione all'interno dei gruppi, suggerisce di evitare la modellizzazione a più livelli e di ricorrere ai tradizionali modelli regressivi ad un solo livello (Muthen e Satorra, 1995). All'aumentare del coefficiente di correlazione intraclasse aumenta il contributo esplicativo dovuto alla strutturazione gerarchica. Questo coefficiente fornisce una misura dell'omogeneità all'interno di uno stesso gruppo e rappresenta la proporzione di varianza residua spiegata dal raggruppamento; misura quindi la parte di variabilità dovuta all'effetto di raggruppamento e quella derivante dalla dipendenza tra osservazioni raggruppate in unità dello stesso livello. Purtroppo l'unica variabile esplicativa è quella relativa all'appartenenza ad un gruppo, per cui le analisi possono far emergere l'esistenza di forti differenze tra i gruppi, senza indicare però le cause di tale differenza.
Il passo successivo prevede allora l'inserimento nel modello di variabili predittive, sia appartenenti al livello individuale, sia a livello gerarchicamente superiore, per meglio comprendere la variabilità all'interno dei gruppi e tra i gruppi stessi. La specificazione dei modelli multilevel consente di modellare la variabilità dei coefficienti di regressione (sia intercetta, che coefficiente angolare) esistente tra le unità di secondo livello, prendendo in considerazione anche le variabili esplicative definite al secondo livello di analisi.

1.4.3 Means-as-Outcomes Model

Questo sottomodello è abbastanza simile al precedente, infatti il modello di primo livello è esattamente lo stesso:

$$y_{ij} = \beta_{0j} + e_{ij} \tag{1.7}$$

Nel modello di regressione *means-as-outcomes*, però, si considerano le medie di ciascun gruppo come risultato che deve essere previsto in funzione delle caratteristiche di gruppo; il modello di livello-2 è infatti:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j} \tag{1.8}$$

con predittore W. Sostituendo l'equazione (1.8) nella (1.7) si ottiene il modello:

$$y_{ij} = \gamma_{00} + \gamma_{01} w_j + u_{0j} + e_{ij}$$

Anche per questo modello la varianza dei risultati è rappresentata da

$$Var(Y_{ij}) = Var(U_0 + E_{ij}) = \tau_{00} + \sigma^2$$

Si noti che u_{0j} ha un significato diverso rispetto al modello precedente. Mentre prima la variabile casuale u_{0j} rappresentava la deviazione di ogni $j - esima$ media di gruppo dalla media generale, ora identifica il residuo

$$u_{0j} = \beta_{0j} - (\gamma_{00} + \gamma_{01} w_j).$$

Similmente, la varianza τ_{00} è ora la varianza residuale o condizionale di B_0, dopo aver tolto l'effetto di W. Si ha quindi che il coefficiente di correlazione intra-classe ρ, dato che $Var(\Delta_{ij}) = \tau_{00} + \sigma^2$ e che $Cov(\Delta_{ij}, \Delta_{kj}) = \tau_{00}$, è specificato utilizzando la solita equazione:

$$\rho = \frac{\tau_{00}}{(\tau_{00} + \sigma^2)}.$$

1.4.4 One-way ANCOVA con effetti casuali

Se si considera di nuovo il modello completo

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}$$

e poniamo pari a zero nel livello-2 i coefficienti γ_{01} e γ_{11} e gli effetti casuali u_{1j} per tutte le j, il modello risultante diventa un modello ANCOVA ad un fattore con effetti casuali e con un singolo predittore al livello-1 come covariata. Il modello al livello-1 rimane del tipo

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$$

con il predittore X_{ij}. Il modello al livello-2 diventa invece

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

E' da notare che il contributo di x_{ij} è trasformato attraverso un valore fisso per ciascuna unità del livello-2 come è indicato dall'equazione $\beta_{1j} = \gamma_{10}$.
Il modello completo risulta quindi essere:

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{0j} + e_{ij}$$

L'unica differenza tra questa equazione e il modello ANCOVA standard è l'effetto gruppo u_{0j} considerato come casuale piuttosto che fisso; come nell'ANCOVA γ_{10} è il coefficiente di regressione di Y_{ij} su X_{ij} nel gruppo j. Si ricorda anche che $Var(E_{ij}) = \sigma^2$ risulta essere anche la varianza residua dopo l'aggiustamento con la covariata del livello-1, X_{ij}. Un'estensione per il modello ANCOVA con effetti casuali permette l'introduzione delle covariate al livello-2. Per esempio, se il coefficiente γ_{01} è non nullo, il modello diventa

$$y_{ij} = \gamma_{00} + \gamma_{01} w_j + \gamma_{01} x_{ij} + u_{0j} + e_{ij}$$

e considera per il livello-2, le ulteriori covariate W_j, mentre controlla il livello-1 con le covariate X_{ij} e gli effetti casuali delle unità del livello-2 con u_{0j}. Anche il modello classico ANCOVA assume che l'effetto covariata γ_{10} sia costante per ogni gruppo.

1.4.5 Modello con tutti gli effetti casuali

Tutti i modelli discussi sopra sono esempi di modelli con sola intercetta casuale, ovvero il coefficiente dell'intercetta del livello-1 β_{0j}. La pendenza al livello-1 non è nemmeno considerata nella one-way ANOVA a effetti fissi o casuali o nella means as outcomes. Nel modello ANCOVA ad effetti casuali è casuale la sola intercetta, mentre il coefficiente β_{1j} costituisce la parte di effetti fissi (o addirittura costante nei vari gruppi). Una prima generalizzazione dei modelli lineari gerarchici più semplici include i casi in cui le pendenze del livello-1 sono ipotizzate variabili casuali legate alle unità di livello-2. Un primo caso è quello di un modello in cui sia l'intercetta che la pendenza sono assunte casuali, senza che, però, venga inserita nel processo di spiegazione del fenomeno alcuna variabile di secondo livello. In questo modello, detto a coefficienti casuali, sia l'intercetta del livello-1 sia le pendenze del livello-1 variano in maniera casuale. Il modello di livello-1 è

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$$

mentre il modello al livello-2 è ancora una semplificazione delle equazioni

$$\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}$$

in cui sia γ_{01} e γ_{11} sono ipotizzate nulle, ovvero:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Si nota che queste equazioni non prevedono i predittori. Il modello dopo le opportune sostituzioni risulta quindi essere rappresentato dall'equazione:

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij}$$

che implica che la risposta è funzione dell'equazione di regressione

$$\gamma_{00} + \gamma_{10}x_{ij}$$

più un errore casuale a tre componenti:

- u_{0j} effetto casuale dell'unità $j - esima$;
- $u_{1j}x_{ij}$ dove u_{1j} è l'effetto casuale dell'unità $j-esima$ sulla pendenza β_{1j};
- l'errore del livello-1 e_{ij}.

*1.4.6 Modelli a coefficienti variabili (*slope-as-outcomes models*)*

Osservando la parte casuale del modello a coefficienti casuali

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij}$$

si constata come esso permetta di stimare la variabilità dei coefficienti di regressione tra le unità di secondo livello. L'idea di fondo del modello a coefficienti variabili consiste invece nell'intervallo specificato a costruire e stimare i parametri di tanti modelli regressivi quanti sono i gruppi considerati. Il passo logicamente successivo è cercare di spiegare tale variabilità tramite le variabili esplicative di secondo livello. Nel modello a coefficienti variabili si suppone che tanto l'intercetta quanto la pendenza vari da un gruppo all'altro. Il modello a coefficienti variabili permette di stimare la variabilità nei coefficienti di regressione (sia l'intercetta, che la pendenza) attraverso le unità del livello-2. Sucessivamente si riesce a modellare questa varibilità.
Dato un predittore al livello-1 X_{ij} e un predittore al livello-2 W_j

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$

Il modello al livello-2 diventa

$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_j + u_{1j}$$

questo modello ha il pregio di strutturare i dati in senso gerarchico, di prevedere, pertanto, due livelli di analisi, ciascuno caratterizzato da una specifica

modellizzazione. Il modello *slope-as-outcomes*, inoltre non prende in considerazione eventuali aspetti comuni alle unità di secondo livello, che potrebbero non essere spiegati, ed eventuali variabili esplicative introdotte.

1.4.7 Modello con coefficienti angolari non casuali e legati ad una variabile di contesto

A volte può capitare che, dopo aver introdotto nel modello generale la covariata di secondo livello W, la varianza residuale di β_{1j} sia prossima a zero. Ovvero la varianza dei residui u_{1j} nell'equazione $\beta_{1j} = \gamma_{10} + \gamma_{11}w_j + u_{1j}$ sia trascurabile. Per ragioni legate all'efficienza statistica e alla stabilità computazionale, diviene conveniente porre $u_{1j} = 0$, (quindi $\tau_{11} = 0$ e $\tau_{01} = 0$). Pertanto i residui dell'equazione $\beta_{1j} = \gamma_{10} + \gamma_{11}w_j + u_{1j}$ sono fissati a zero e il modello del livello-2 per la pendenza diventa

$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_j$$

e questo modello con le equazioni

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j}$$

produce il modello

$$Y_{ij} = \gamma_{00} + \gamma_{10}w_j + \gamma_{10}x_{ij} + \gamma_{11}w_jx_{ij} + u_{0j} + e_{ij} \tag{1.9}$$

dove i coefficienti angolari variano da gruppo a gruppo ma la loro variazione non è più casuale. Precisamente, le pendenze β_{1j} sono una funzione di W. Il modello può essere visto come ulteriore esempio di modello ad intercetta casuale, in quanto β_{0j} è la sola componente che varia casualmente tra le unità di livello-2.

1.4.8 Ricapitolazione

Nei precedenti paragrafi si sono considerati semplici modelli lineari gerarchici con una sola variabile esplicativa di livello 1, X, e una sola variabile esplicativa al livello-2, W. In questo contesto il modello al livello-1 è caratterizzato da due parametri: l'intercetta e la pendenza. Al livello-2 ciascuno di questi possono essere ottenuti mediante una funzione di W e una componente accidentale. A partire poi dal modello completo si sono ottenuti sottomodelli,

ponendo uguali a zero alcuni parametri; tali sottomodelli risultano in genere utili nelle analisi preliminari di dati di tipo gerarchico.
I sottomodelli considerati sono classificabili come: modelli ad intercetta casuale (random intercept models) e modelli a pendenza casuale (randomly varying slope models). I modelli ad intercetta casuale si suddividono poi in:

1. One-way random effects ANOVA model
2. Means as outcomes model
3. One-way ANCOVA model
4. Modello con coefficienti angolari non casuali

In questi modelli le componenti di varianza sono: la varianza al livello-1, σ^2, e la varianza al livello-2, τ_{00}. Si può notare che nei modelli ANOVA e Means as outcomes model non esiste la pendenza nella relazione di livello-1. Nel modello ANCOVA viene considerata la pendenza al livello 1 ma è fissa per le unità del livello 2. Nei modelli non-randomly varying slopes la pendenza varia in base a una funzione di W senza l'aggiunta di componenti casuali. Per quanto riguarda i modelli:

1. Random Coefficients model
2. Slope and intercepts as outcomes model

questi presentano sia pendenza che intercetta variabili (ma non casualmente). Un'altra distinzione potrebbe poi essere fatta considerando i modelli che includono termini di cross-level interaction. I modelli combinati possono includere termini di cross-level interaction per predire la variazione nella pendenza. Questi termini appaiono in due dei sottomodelli analizzati:

1. intercepts and slope as outcomes model
2. modello con coefficienti angolari non casuali

Si possono quindi osservare, per il modello gerarchico lineare a due livelli, alcuni casi speciali:

1. *regressione ordinaria*. La variabilità tra i gruppi è nulla e quindi i coefficienti sono fissi (vedi Figura 1.30).

$$\tau_{00} = 0, \tau_{11} = 0 \Longrightarrow Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \varepsilon_{ij} \tag{1.10}$$

2. *intercetta casuale*. La varianza del coefficiente di regressione è nulla (e quindi anche la covarianza tra i coefficienti). Inoltre la varianza dell'intercetta non dipende da X (la centratura di X è infatti irrilevante). Le rette di regressione relative ai gruppi sono parallele ed è quindi possibile ordinare i gruppi (vedi Figura 1.31).

$$\tau_{00} > 0, \tau_{11} = 0 \Longrightarrow Y_{ij} = (\gamma_{00} + u_{0j}) + \gamma_{10} x_{ij} + \varepsilon_{ij} \tag{1.11}$$

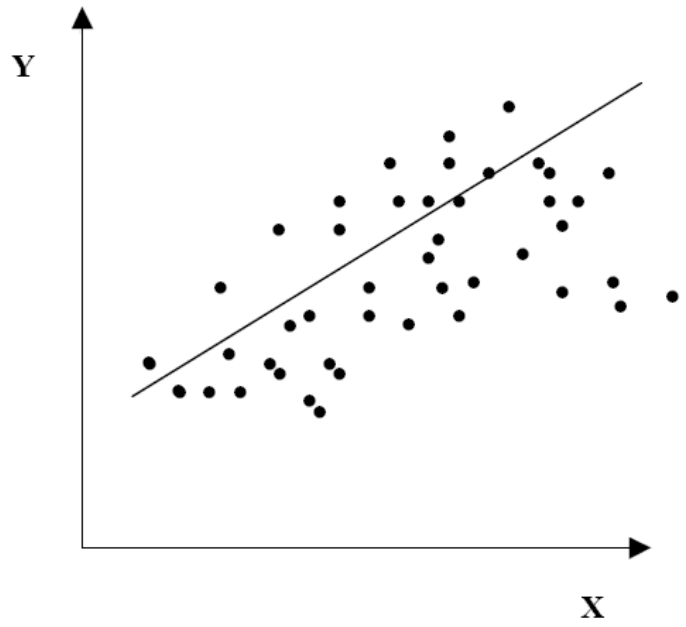

Figura 1.30: Regressione ordinaria.

3. *pendenza casuale.* La varianza dell'intercetta τ_{00} e la covarianza intercetta-pendenza (τ_{01}) si riferiscono alla variabile di contesto W e dipendono da W (vedi Figura 1.32). Poichè spesso l'origine di X è arbitraria è bene non vincolare a zero la covarianza. Non esiste un ordinamento univoco dei gruppi: l'ordinamento varia al variare del valore X considerato.

$$\tau_{00} > 0, \tau_{11} > 0 \Longrightarrow Y_{ij} = (\gamma_{00} + u_{0j}) + (\gamma_{10} + u_{1j})x_{ij} + \varepsilon_{ij} \qquad (1.12)$$

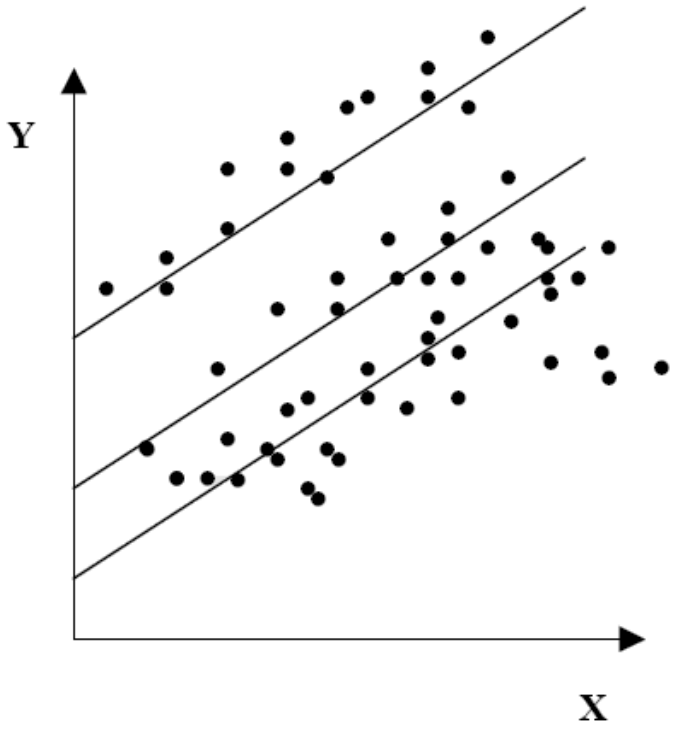

Figura 1.31: Intercetta casuale

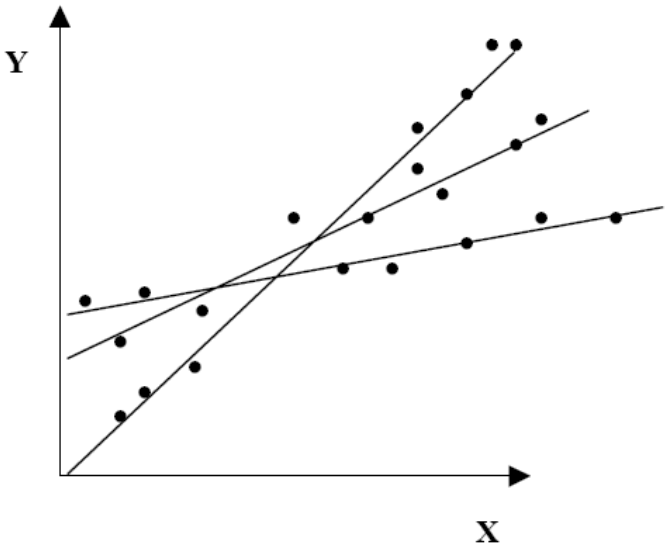

Figura 1.32: Pendenza casuale

Capitolo 2
Caratteristiche dei modelli lineari multilivel

Per affrontare il trattamento dei modelli descritti nel precedente capitolo, sono state sviluppate metodologie statistiche di natura parametrica che considerano la presenza delle gerarchie. I modelli multilevel infatti costituiscono lo strumento più adatto per trattare le informazioni presenti all'interno delle strutture gerarchiche: tengono conto, in maniera esplicita, sia della presenza di relazioni tra le variabili appartenenti ad uno specifico livello, sia delle relazioni tra i differenti livelli, considerando in tal modo l'effetto netto sulle unità e le interazioni presenti. In letteratura sono stati proposti diversi modelli di regressione multilevel: *random coefficient model*, *variance component model* e *hierarchical linear model*. Questi modelli, essendo basati su un approccio comune, formano la classe dei *multilevel regression model*. Essi partono dall'assunto che ci sia un dataset strutturato in maniera gerarchica, una sola variabile esplicativa misurata al livello più basso e almeno una variabile esplicativa ad ogni livello presente nella struttura. Trattandosi di dati gerarchicamente organizzati, le osservazioni individuali non sono indipendenti. I modelli di regressione ordinari assumono indipendenza fra le unità e quindi la presenza della correlazione intraclasse[1] costituisce una violazione di tale assunzione, dalle conseguenze potenzialmente molto gravi. Concettualmente si è soliti immaginare i modelli di regressione multilevel come sistemi gerarchici di equazioni di regressione. Anche nel prosieguo di questa trattazione si considereranno due soli livelli, senza comunque trascurare le possibili generalizzazioni. Una tipologia specifica delle tecniche multilevel riguarda i modelli ad intercetta casuale (*random intercept model*), oppure i modelli a pendenza casuale (*random slope*).

[1] Il *coefficiente di correlazione intraclasse* può essere definito come la porzione di variabilità attribuibile ai gruppi o, equivalentemente, come la correlazione fra due unità dello stesso gruppo. Naturalmente quando è presente un effetto casuale relativo ad una covariata tale coefficiente non può essere calcolato, ma il concetto che rappresenta continua ad essere valido.

2.1 Coefficiente di correlazione intraclasse

La correlazione intraclasse è una misura del grado di dipendenza degli individui: più gli individui condividono le esperienze comuni dovute alla vicinanza nel tempo e nello spazio, più sono simili. Il più alto livello di dipendenza può presentarsi, ad esempio, tra due osservazioni di gemelli monozigoti, oppure bambini nati e cresciuti nella stessa famiglia. Un altro esempio ben conosciuto di osservazioni dipendenti riguarda le "misure ripetute" sulla stessa persona. La caratteristica principale dell'analisi multilevel è costituita dal fatto che in genere, trattandosi di dati gerarchicamente organizzati, le osservazioni individuali non sono indipendenti. La correlazione media esistente tra individui appartenenti allo stesso gruppo viene detta *intra-class correlation*, generalmente indicata con il simbolo ρ; essa può essere spiegata in diversi modi; ad esempio, può anche essere definita come misura di omogeneità di un gruppo. A partire da Ronald Fisher, la correlazione intraclasse è stata considerata nel quadro dell'analisi della varianza (ANOVA) e, più recentemente, nel quadro dei modelli ad effetti casuali (*random effect*). Il coefficiente di correlazione intraclasse proposto da Fisher (1954) consiste nel rapporto tra la media dei prodotti degli scarti da μ (media generale) per tutte le $N_j(N_j - 1)$ coppie distinte che si possono formare con le N_j osservazioni

$$
1 \left\{ \begin{array}{|c|c|} \hline 1 & 2 \\ \hline 1 & 3 \\ \hline \vdots & \\ \hline 1 & N_i \\ \hline \end{array} \right\} N_i - 1
$$

$$
2 \left\{ \begin{array}{|c|c|} \hline 2 & 1 \\ \hline 2 & 3 \\ \hline \vdots & \\ \hline 2 & N_i \\ \hline \end{array} \right\} N_i - 1
$$

$$
\vdots
$$

$$
N_i \left\{ \begin{array}{|c|c|} \hline N_i & 1 \\ \hline N_i & 2 \\ \hline \vdots & \\ \hline N_i & N_i - 1 \\ \hline \end{array} \right\} N_i - 1
$$

contenente all'interno di ognuno dei K gruppi e il prodotto degli scarti quadratici delle N osservazioni che formano le coppie (e quindi la loro varianza) dove $N = \sum_{i=1}^{K} N_i$

$$\rho = \frac{\frac{\sum_j^K \sum_{i \neq i'}^{N_j} (X_{ij} - \mu)(X_{i'j} - \mu)}{N^*}}{\frac{\sum_i \sum_j^{N_j} (X_{ij} - \mu)^2}{N}}$$

dove

$$\mu = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \frac{x_{ij}}{N}$$

ed

$$N^* = \sum_{j}^{K} N_j (N_j - 1)$$

è il numero delle coppie distinte senza ripetizione che si possono formare dentro i K gruppi. Il numeratore e il denominatore, essendo medie di covarianze e varianze calcolate all'interno dei gruppi si possono denominare, rispettivamente, covarianza e varianza intra-gruppo. Una scrittura alternativa utile per comprendere la natura del coefficiente di correlazione intraclasse, è basata sulle distanze euclidee tra unità appartenenti allo stesso gruppo:

$$\rho = \frac{\sum_j^K \sum_{i \neq i'}^{N_j} (X_{ij} - X_{i'j})^2}{\sum_j (N_j - 1) \sum_i (X_{ij} - \mu)^2}$$

Si consideri ad esempio un data-set di N coppie di dati $(x_{n,1}, x_{n,2})$ per $n = 1, ..., N$. Il coefficiente di correlazione intraclasse proposto da Fisher (1954) è

$$\frac{1}{Ns^2} \sum Ns^2 \sum_{n=1}^{N} (x_{n,1} - \bar{x})(x_{n,2} - \bar{x})$$

dove

$$\bar{x} = \frac{1}{2N} \sum_{n=1}^{N} (x_{n,1} + x_{n,2})$$

$$s^2 = \frac{1}{2N} \{ \sum_{n=1}^{N} (x_{n,1} - \bar{x})^2 + \sum_{n=1}^{N} (x_{n,2} - \bar{x})^2 \}$$

Il coefficiente di correlazione intraclasse in questa accezione viene definito anche per gruppi con più di due valori. Ad esempio per gruppi con tre soggetti:

$$\bar{x} = \frac{1}{3N}\sum_{n=1}^{N}(x_{n,1} + x_{n,2} + x_{n,3})$$
$$s^2 = \frac{1}{3N}\{\sum_{n=1}^{N}(x_{n,1} - \bar{x})^2 + \sum_{n=1}^{N}(x_{n,2} - \bar{x})^2 + \sum_{n=1}^{N}(x_{n,3} - \bar{x})^2\}$$
$$\rho = \frac{1}{3Ns^2}\sum_{n=1}^{N}(x_{n,1} - \bar{x})(x_{n,2} - \bar{x}) + (x_{n,1} - \bar{x})(x_{n,3} - \bar{x}) + (x_{n,2} - \bar{x})(x_{n,3} - \bar{x})$$

Al crescere della numerosità di individui per gruppo ρ cresce rapidamente. Nella forma equivalente

$$\rho = \frac{K}{K-1}\frac{N^{-1}\sum_{n=1}^{N}(\bar{x}_n - \bar{x})^2}{s^2} - \frac{1}{K-1}$$

dove K è il numero di dati in ogni gruppo, $\bar{x}_n$ è la media dell'$n-esimo$ gruppo (Harris, 1913). Per K grande, questo coefficiente di correlazione intraclasse risulta:

$$\frac{N^{-1}\sum_{n=1}^{N}(\bar{x}_n - \bar{x})^2}{s^2}$$

che può essere allora interpretato come la frazione della varianza totale imputata alla varianza tra i gruppi.
Questo ICC e il coefficiente di correlazione intraclasse (Pearson) coincidono nel caso in cui i gruppi tendono all'infinito e la numerosità all'interno dei gruppi diverge. Con i dati organizzati in una struttura gerarchica a due livelli, l'*intra-class correlation* è definita quindi come proporzione di variabilità attribuibile ai gruppi o, equivalentemente, la correlazione fra due unità dello stesso gruppo. Se si è in presenza di correlazione intraclasse, come potrebbe succedere con questo tipo di dati, il presupposto della indipendenza delle osservazioni non è rispettato. Un effetto di tale violazione è l'incremento non controllabile della probabilità di commettere l'errore di prima specie (livello α), in letteratura associato proprio alla presenza della correlazione intraclasse. I test statistici tradizionali sono basati sull'assunto di indipendenza tra le osservazioni. Se questa ipotesi risulta violata, le stime degli errori standard prodotte dalle procedure convenzionali risultano distorte per difetto e, di conseguenza, i risultati che si ottengono potrebbero essere "impropriamente" significativi.

Al fine di esplicitare meglio il coefficiente di correlazione intraclasse è utile specificare meglio le sue caratteristiche. Tale misura parte dalla scomposizione della varianza totale in *within* (infra-gruppo) e *between* (inter-gruppi) (Snijders e Bosker, 1999):

$$Var(Y_{ij}) = \tau^2 + \sigma^2$$

Per giungere alla sua formulazione campionaria si considera un campione di numerosità N, si indica con J il numero totale di macro unità osservate e con n_j il numero delle micro unità nella $j-esima$ macro unità, quindi $N = \Sigma_j n_j$. Si considerino ad esempio, le ricerche effettuate nel campo dell'istruzione dove ci si propone di rilevare l'esistenza di differenze tra classi (gruppi di unità statistiche) di studenti (unità statistiche) sulla base di una certa misura individuale di risultato (Y), tenendo conto del fatto che le caratteristiche (X) degli studenti e quelle (Z) delle classi possono essere rilevanti nel determinare tale risultato (Aitkin e Longford, 1986; Goldstein e Spiegelhalter, 1996). La base logica di tali tecniche deriva dalla considerazione che il risultato individuale Y dipende sia da fattori riferibili all'unità statistica oggetto di studio (unità di primo livello), che da fattori riferibili al gruppo di appartenenza (unità statistica di secondo livello). Ciò che si può osservare (fattori osservabili) è rappresentato da una o più variabili X riferite all'unità di primo livello e da una o più variabili Z riferite all'unità di secondo livello. Invece, tutto ciò che non è osservabile (fattori non osservabili o non osservati) viene considerato come termine di errore (le variabili casuali ad esso abbinato vengono indicate con la lettera E, quando si fa riferimento all'unità statistica di primo livello, e con la lettera U nel caso delle unità di secondo livello).

Abbiamo allora, dal punto di vista campionario:

- La media della macro unità j:

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

- La media generale:

$$\overline{y} = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} y_{ij} = \frac{1}{N} \sum_{j=1}^{J} n_j \overline{y}_j$$

- La varianza *within* al gruppo j è data da:

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

La varianza *within* si può quindi interpretare come sintesi dei residui delle singole osservazioni all'interno delle macro unità. Per quanto riguarda la varianza *between*, il discorso si complica in riferimento alla dimensione delle

macro unità.
Per gruppi di uguale numerosità essa è definita come:

$$s^2_{between} = \frac{1}{J-1}\sum_{j=1}^{J}(\overline{y}_j - \overline{y})^2$$

per gruppi di diversa numerosità, il contributo dei vari gruppi deve essere pesato, quindi viene definita come:

$$s^2_{between} = \frac{1}{\widetilde{n}(J-1)}\sum_{j=1}^{J}(\overline{y}_j - \overline{y})^2$$

dove $\widetilde{n}$ è definito come:

$$\widetilde{n} = \frac{1}{J-1}\{N - \frac{\sum_j n_j^2}{N}\} = \overline{n} - \frac{s^2(n_j)}{J\overline{n}}$$

$\overline{n} = \frac{N}{J}$ è dimensione media delle macro unità e

$$s^2(n_j) = \frac{1}{J-1}\sum_{j=1}^{J}(n_j - \overline{n})^2$$

è la varianza della dimensione delle macro unità.
La varianza totale può, allora, essere scritta come una combinazione della varianza *within* e *between*:

$$Var(y_{ij}) = \frac{1}{(N-1)}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2 = \frac{N-J}{N-1}s^2_{within} + \frac{\widetilde{n}(J-1)}{N-1}s^2_{between}$$

Le complicazioni relative alla varianza *between* derivano dal fatto che i residui rispetto alle singole osservazioni, contribuiscono, benchè in misure minore, alla sua costruzione. In pratica non si conoscono le vere varianze *between* e *within*, ma è possibile stimarle attraverso i dati osservati.

Il valore atteso della variabile casuale varianze *within* è:

$$E(S^2_{within}) = \sigma^2$$

mentre il valore atteso della variabile casuale varianza *between* è:

$$E(S^2_{between}) = \tau^2 + \frac{\sigma^2}{\widetilde{n}}$$

Le loro stime $\hat{\sigma}^2$ e $\hat{\tau}^2$ possono essere calcolate come:

$$\widehat{\sigma}^2 = s^2_{within}$$

e

$$\widehat{\tau}^2 = s^2_{between} - \frac{s^2_{within}}{\widetilde{n}}$$

Supponendo, ad esempio, di avere un set di dati strutturato su due livelli, dove le micro unità rappresentano l'insieme delle n osservazioni (livello 1), raggruppate nei rispettivi gruppi di appartenenza (livello 2), si può dividere la variabilità totale in quella *within*, ossia all'interno dei gruppi, e quella *between*, ovvero tra i vari gruppi $Var(Y_{ij}) = \tau^2 + \sigma^2$.
In tale situazione il coefficiente di correlazione intraclasse ρ si può definire:

$$\rho(y_{ij}, y_{i'j}) = \frac{varianza\ popolazione\ tra\ macro\ unitá}{varianza\ totale} = \frac{\tau^2}{\tau^2 + \sigma^2}$$

Il parametro ρ è il coefficiente di correlazione intraclasse e indica la correlazione tra due individui dello stesso gruppo o anche la quota di variabilità totale a livello di gruppo. Nel caso in cui il coefficiente di correlazione è significativamente diverso da zero si può affermare che parte della variabilità è attribuibile ai gruppi, e che quindi, il macro livello influeza il micro.
Tale indice è stato proposto anche da Donner nel 1986, nella forma:

$$\rho(ICC) = \frac{var(tra\ le\ classi)}{var(tra\ le\ classi) + var(residua)} = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{residual}}$$

dove la classe è identificata con il suo livello medio o con lo scarto del livello medio rispetto alla media generale, spiegato dalla variabile moderatrice, espressione della gerarchia presente nei dati. Il coefficiente di correlazione intraclasse è usato per stimare la correlazione tra due unità dello stesso gruppo di appartenenza, per esempio due studenti nella stessa classe (Fisher 1925).

La correlazione intraclasse ha la carratteristica che media e varianza sono comuni a tutti i membri appartenenti al medesimo gruppo e, per un numero sufficientemente elevato di gruppi, fornisce la proporzione della varianza attribuibile alla differenza tra i gruppi. Viene anche usata, ad esempio, per la valutazione della coerenza o della riproducibilità delle misurazioni fatte da osservatori differenti sulle stesse quantità (Aitkin e Longford, 1986; Goldstein e Spiegelhalter, 1996).

2.2 Componenti di varianza e variabili esplicative

In un modello di regressione ordinario la varianza del termine di errore ha il significato di varianza residua, cioè di varianza non spiegata dai regressori. In genere l'inserimento di un nuova variabile comporta una riduzione della

varianza residua, la cui entità dipende dal suo potere esplicativo. La situazione è più complessa in un modello a componenti di varianza, nel quale la varianza non spiegata dai regressori viene scomposta in due parti: la componente *between* σ_u^2, ovvero la varianza non spiegata dai regressori e che è attribuibile agli effetti casuali, ovvero alla struttura gerarchica; la componente within σ_e^2, ovvero la varianza residua in senso stretto, che non è spiegata nè dai regressori, nè dall'appartenenza ai gruppi, ma che è legata alla variabilità individuale. L'effetto dell'inserimento di nuove variabili sulle componenti di varianza dipende dal tipo di variabile (Longford, 1993, pp. 29-30):

- variabile di contesto (livello 2): una variabile misurata a livello di gruppo che contribuisce a spiegare le differenze tra i gruppi e quindi a ridurre la componente *between*, mentre non ha nessun effetto sulla componente *within*;
- variabile individuale (livello 1): come è naturale attendersi, l'inserimento di una variabile individuale riduce la varianza *within*, ma la direzione del suo effetto sulla componente *between* non è determinabile a priori.

Bisogna pensare che la componente *between* è una misura del grado di eterogeneità dei gruppi non spiegata dai regressori e che l'inserimento di una nuova variabile individuale può sia aumentare che diminuire la misura di tale eterogeneità non spiegata. Consideriamo, ad esempio, uno studio sulla mortalità dei degenti di un insieme di ospedali (unità di livello 2) e supponiamo di inserire una variabile che misura la gravità dei pazienti. Se i pazienti più gravi sono ricoverati negli ospedali più qualificati, l'inserimento di tale variabile provoca un aumento della componente *between*, poichè porta alla luce un'eterogeneità che in precedenza era mascherata, in quanto non veniva tenuto conto del modo in cui i pazienti erano assegnati agli ospedali.

2.2.1 Vantaggi e limiti dei modelli gerarchici

Nei precedenti paragrafi sono state analizzate le caratteristiche principali dei modelli multilevel, tralasciando le possibili generalizzazioni a più livelli e la loro presentazione in forma matriciale; sono state evidenziate le ragioni del loro utilizzo e i limiti principali dei modelli di regressione classici.
Volendo riassumere quanto fin qui esposto, allo scopo di sottolineare le differenze principali tra i modelli di regressione lineare ed i modelli multilevel, si riportano in tabella 2.1 le assunzioni di base su cui essi sono fondati.

Per i dati a struttura gerarchica l'applicazione dell'analisi multilevel comporta i seguenti vantaggi:

- Trattamento della interdipendenza: l'idea generale che spiega questa caratteristica è che individui appartenenti alla medesima rete di relazioni siano più vicini o abbiano dei comportamenti tra loro più simili di

	regressione	multilevel
- linearità della relazione funzionale	sì	sì
- normalità distribuzionale	sì	sì
- omoschedasticità	sì	no
- indipendenza delle osservazioni	sì	no

Tabella 2.1: Confronto tra modelli parametrici

quanto non accada con individui appartenenti a reti di relazioni diverse (correlazione intra-classe).

- Scomposizione della struttura dell'errore (varianza) in una o più fonti di variabilità (una o più componenti), corrispondenti alle diverse unità di analisi (es. primo e secondo livello), riuscendo così ad esprimere anche la variabilità tra i gruppi.
- Le fonti di variabilità possono essere collegate a variabili esplicative relative a ciascun livello.

Inoltre i modelli multilivello consentono di:

- Eliminare la distorsione nella stima degli errori standard dei parametri;
- Stimare l'effetto del gruppo (*group effect*) scomponendo la variabilità in due componenti: quota interna ai gruppi (*within*) e tra gruppi (*between*);
- Introdurre variabili esplicative a livello di gruppo (*group-level predictors*) cercando così di dare una migliore descrizione della variabilità tra gruppi (*random effects model*);
- modellare gli effetti di interazione o *cross-level*.

Tra i principali limiti dell'analisi multilevel va sottolineato che:

- Nonostante il rigore metodologico di tali modelli c'è la necessità di sviluppare teorie che specificano a livello di gruppo e a livello individuale quali fattori possano congiuntamente configurare un determinato *outcome*, ad es. il supporto sociale.
- Come tutti i modelli statistici, anche i modelli multilevel necessariamente semplificano processi complessi. Un limite intrinseco che l'analisi multilevel condivide con gli altri metodi di regressione è il fatto di verificare separatamente gli effetti di più variabili.
- L'analisi multilevel non consente infine di abbracciare la complessa fenomenologia delle possibili relazioni tra variabili, poichè implica una struttura di regressione in cui una singola variabile dipende da un insieme di altre variabili.

Nel seguito vengono approfondite le specificazioni dei principali modelli multilivello e fornita l'espressione esplicita dell'ICC e la versione matriciale degli stessi, al fine di rendere più agevole la successiva trattazione delle procedure inferenziali.

2.3 Il modello ad intercetta casuale

Questo modello rappresenta un caso particolare del modello gerarchico lineare, conosciuto anche col nome di *Random Intercept Model* (Snijders e Bosker, 1999). Come nel classico modello di regressione lineare, si è in presenza di una variabile dipendente Y e di un set di predittori X, misurati al livello degli individui. In particolare, la formalizzazione del modello avviene nel seguente modo:

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + e_{ij}$$

dove y_{ij} rappresenta la variabile risposta, con l'indice i relativo agli individui e l'indice j relativo alle unità di secondo livello[2]. L'obiettivo è quello di stimare il valore atteso di y_{ij}, considerando l'effetto del predittore X sia a livello individuale sia a livello di gruppo. Si ipotizza che la variabile esplicativa sia caratterizzata da livelli medi differenti in ogni gruppo. Tale modello considera l'effetto gruppo del predittore attraverso le variazioni dell'intercetta. In altre parole, si stima un modello in cui il coefficiente di regressione è costante nei gruppi (parallelismo) e ciò che distingue gli stessi rispetto al predittore è la diversa intercetta. Gli e_{ij} sono gli errori a livello degli individui. L'intercetta variabile a livello di gruppo viene modellata come:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

dove γ_{00} rappresenta l'intercetta media tra tutti i gruppi, mentre u_{0j} rappresenta la componente aleatoria. In altre parole, la generica intercetta è la somma della media generale e dell'effetto casuale a livello di gruppo, che misura la deviazione rispetto alla media. Sostituendo quest'ultima equazione nella precedente si ottiene il modello completo:

$$y_{ij} = \gamma_{00} + \beta_1 x_{ij} + u_{0j} + e_{ij}$$

Nel modello così ottenuto, gli u_{0j} potrebbero essere considerati sia come parametri fissi, che come variabili casuali indipendenti ed identicamente distribuite. Il primo caso si ha quando i gruppi sono specificati a priori, riconducendosi quindi all'analisi della covarianza in cui la variabile di raggruppamento è un fattore fisso; nel secondo caso gli u_{0j} sono gli effetti casuali di gruppo non spiegati dalla regressione. Tale interpretazione porta alla definizione del *Random Intercept Model* in cui l'intercetta varia tra i gruppi in maniera casuale, poichè i gruppi sono considerati un campione estratto casualmente da una popolazione di gruppi.

Per comprendere come si giunge a questo modello, bisogna in realtà partire dal considerare il modello ANOVA ad effetti casuali, in cui le variabili esplicative (in genere indicate con i simboli X e Z) ai diversi livelli non compaiono (questo modello contiene solo i gruppi casuali e le variazioni casuali interne).

[2] In questo modello non compaiono variabili esplicative di secondo livello; l'effetto su di esso sarà specificato nei modelli *random slopes*

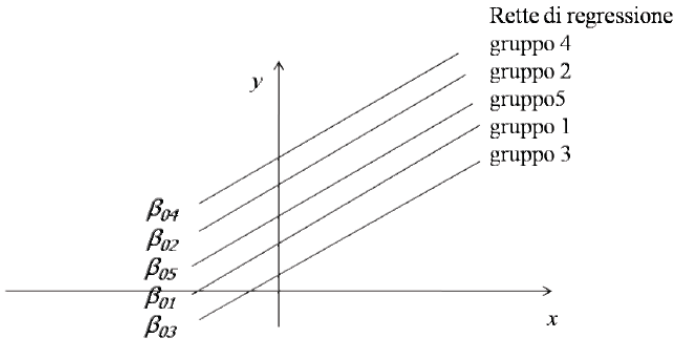

Figura 2.1: *Random intercept model*

Questo modello è definito *Empty Model*. Esso può essere espresso come un modello in cui la variabile dipendente è uguale alla somma della media generale γ_{00}, dell'effetto casuale a livello di gruppo u_{0j} e dell'effetto casuale a livello individuale e_{ij}. I gruppi con elevato u_{0j} avranno in media Y elevato, mentre i gruppi con basso u_{0j} avranno in media Y basso. Si assume poi che le variabili casuali u_{0j} e e_{ij} abbiano media 0 e siano mutuamente indipendenti. Tale modello permette, in questo modo, la partizione base della variabilità dei dati tra i due livelli. Anche nel modello generale $y_{ij} = \gamma_{00} + \beta_1 x_{ij} + u_{0j} + e_{ij}$ la varianza totale di Y può essere scomposta come la somma delle varianze a livello 1 e a livello 2 nel seguente modo:

$$Var(Y_{ij}) = Var(U_0) + Var(E_{ij}) = \tau_0^2 + \sigma^2$$

La covarianza tra due individui i e i' appartenenti allo stesso gruppo j è uguale alla varianza di u_{0j}:

$$Cov(Y_{ij}, Y_{i'j}) = Var(U_{0j}) = \tau_0^2$$

e la loro correlazione è

$$\rho(Y_{ij}, Y_{i'j}) = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}$$

Si ricorda che il parametro ρ è il coefficiente di correlazione intraclasse, e indica la correlazione tra due individui dello stesso gruppo o anche la quota di variabilità totale a livello di gruppo. Si può affermare, nell'ipotesi in cui il coefficiente di correlazione sia significativamente alto, che ha senso effettuare un'analisi multilevel in quanto buona parte della variabilità è attribuibile ai gruppi, e quindi il macro livello influenza il micro. A questo punto il successivo step è l'inclusione nel modello di eventuali variabili esplicative. Come nel classico modello di regressione lineare esse sono usate per spiegare parte

della variabilità della Y; nel caso specifico si riferisce alla variabilità sia del primo che del secondo livello. Se si considera una sola variabile X si ritrova il modello:

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{0j} + e_{ij}$$

Le assunzioni fondamentali sono che tutti gli errori u_{0j} e e_{ij} sono determinazioni di variabili casuali U_0 e E_{ij} mutuamente indipendenti, con medie nulle. Per u_{0j} e e_{ij} si assume che siano estratti da popolazioni distribuite normalmente e le loro varianze siano rispettivamente τ_0^2 e σ^2. La variabile casuale U_0 può essere vista come descrittiva degli errori a livello di gruppo, cioè come effetti di gruppo, non spiegati da X. Dal momento che gli errori casuali contengono quella parte di variabilità della variabile dipendente che non è considerata come funzione di variabili esplicative, si può affermare che questo modello contiene variabilità non spiegata a due livelli annidati. La partizione della variabilità non spiegata sui vari livelli è l'essenza dei modelli gerarchici ad effetti casuali. All'interno del modello, γ_{00} è sempre l'intercetta media dei gruppi e γ_{10} può essere visto come un coefficiente di regressione non standardizzato come nel modo usuale (infatti in tale equazione $\gamma_{10} = \beta_1$); cioè l'aumento unitario nel valore di X è associato con un aumento medio in Y di β_1 unità. La varianza residua condizionata al valore di X è:

$$Var(Y_{ij}|x_{ij}) = Var(U_0) + Var(E_{ij}) = \tau_0^2 + \sigma^2$$

mentre la covarianza tra due differenti individui i e i' nello stesso gruppo è ancora:

$$Cov(Y_{ij}, Y_{i'j}|x_{ij}, x_{i'j}) = Var(u_{ij}) = \tau_0^2$$

La frazione di variabilità residua ascrivibile al livello 1 è data da

$$\frac{\sigma^2}{(\sigma^2 + \tau_0^2)}$$

e per il livello 2 questa frazione è

$$\frac{\tau_0^2}{(\sigma^2 + \tau_0^2)}.$$

Della covarianza o correlazione tra due individui dello stesso gruppo, una parte può essere spiegata dai rispettivi valori di X, mentre l'altra parte non è spiegata. Questa è il cofficiente di correlazione intraclasse residuo:

$$\rho_I(Y|X) = \frac{\tau_0^2}{(\sigma^2 + \tau_0^2)}$$

Questo parametro è analogo all'usuale cofficiente di correlazione intraclasse, ma ora i parametri τ_0^2 e σ^2 sono riferiti alle varianze del modello

$$y_{ij} = \gamma_{00} + \beta_1 x_{ij} + u_{0j} + e_{ij},$$

che include gli effetti della variabile mentre prima erano riferiti alle varianze dell'*Empty Model*.

Quando il coefficiente di correlazione intraclasse è nullo (quando, ad esempio, u_{0j} è uguale a 0 per tutti i J gruppi) allora il raggruppamento è irrilevante per la variabile Y che condiziona X, e si può usare il normale modello di regressione lineare. Se il coefficiente di correlazione intraclasse residuo, o equivalentemente τ_0^2 è significativo, allora il modello lineare gerarchico risulta migliore di quello di regressione *Ordinary Least Squares* (OLS). Nel *Random Intercept Model*, i parametri da stimare sono quattro:

- i coefficienti di regressione γ_{00} e γ_{10} o (β_1);
- le componenti di varianza τ_0^2 e σ^2.

Ovviamente è possibile generalizzare il modello *Random Intercept Model* a più di due livelli.

2.4 Il modello completo a coefficienti casuali: Random slopes

Nei modelli ad intercetta casuale, i gruppi differiscono rispetto al valore medio della variabile dipendente: l'unico effetto casuale è cioè attribuibile all'intercetta. La relazione fra variabile dipendente e variabile esplicativa può tuttavia differire tra i gruppi in più modi: è possibile, ad esempio, che gli effetti dello stato socio-economico degli studenti di una scuola sul loro rendimento, sia più forte in alcune classi rispetto ad altre. Questo fenomeno, nell'analisi della covarianza, è conosciuto come eterogeneità della regressione fra i gruppi (non parallelismo); nei modelli gerarchici è noto come *random slopes*. Nella situazione appena descritta, la stima dei parametri di un modello multilevel può essere concettualmente distinta in due fasi successive. Nella prima fase, a livello degli individui, vengono adattati, all'interno di ciascun gruppo, modelli di regressione separati, al fine di predire la variabile risposta Y in funzione della variabile esplicativa X; nella seconda fase si introducono le variabili esplicative misurate a livello di gruppo, che descrivono la variazione dei coefficienti di regressione. Il modello in esame può essere specificato come segue:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$$

dove abbiamo che β_{0j} è la classica intercetta, β_{1j} è l'usuale coefficiente di regressione per la variabile esplicativa X, misurata sul livello degli individui, mentre e_{ij} rappresenta il termine d'errore. Come nel *random intercept model*, anche in questo caso la differenza rispetto al modello di regressione

non gerarchico consiste nel fatto che ogni gruppo possiede una diversa intercetta, β_{0j} ma ora anche un differente coefficiente di regressione β_{1j}. Inoltre, si assume che, all'interno di ciascun gruppo, gli errori al livello individuale siano indipendenti e normalmente distribuiti con media nulla e varianza σ^2, $E_{ij} \sim N(0; \sigma^2)$. A causa della variazione tra le unità di livello superiore, i coefficienti in esame prendono il nome di coefficienti casuali. Le macro-unità sono ancora viste come un campione proveniente da una più vasta popolazione di gruppi. A questo punto i coefficienti β_{0j} e β_{1j} del modello di regressione gerarchico, possono essere esplicitati in un coefficiente medio e una parte che risente della dipendenza dalle unità a livello superiore, ovvero:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$$

Anche in questo caso si assume che i termini di errore nelle equazioni di regressione a livello di gruppo u_{0j} e u_{1j}, spesso denominati macro-errori, siano normalmente distribuiti con media nulla e varianze τ_0^2 e τ_1^2, rispettivamente. Inoltre, si assume che i macro-errori siano indipendenti tra i gruppi e dagli errori di livello individuale e_{ij}; con σ_{u01}^2 viene indicata la covarianza tra i macro-errori u_{0j} e u_{1j}:

$$U_0 \sim N(0, \tau_0^2); \;\; U_1 \sim N(0, \tau_1^2); \;\; Cov(U_0; U_1) = \sigma_{u01}^2$$

Sostituendo le equazioni $\beta_{0j} = \gamma_{00} + \gamma_{01} z_j + u_{0j}$ e $\beta_{1j} = \gamma_{10} + \gamma_{11} z_j + u_{1j}$ nella equazione $y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$, il modello di regressione multilevel può essere scritto in un'unica equazione di regressione:

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} z_j + \gamma_{11} x_{ij} z_j + u_{1j} x_{ij} + u_{0j} + e_{ij}$$

Il termine $x_{ij} z_j$ è denominato cross-level interaction poichè risente dell'effetto moderante delle variabili esplicative misurate su differenti livelli della gerarchia come mostrato in figura 2.2.
La parte

$$[\gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} z_j + \gamma_{11} x_{ij} z_j]$$

viene denominata parte sistematica del modello, mentre la parte

$$[u_{1j} x_{ij} + u_{0j} + e_{ij}]$$

che contiene i termini casuali di errore, viene denominata parte aleatoria del modello. Essa costituisce una struttura complessa di errore e, come si può notare dalla formula, gli errori per le osservazioni all'interno delle macro unità sono correlati poichè u_{0j} e u_{1j} risultano comuni per le osservazioni che appartengono al medesimo gruppo. Il modello implica non solo che gli individui all'interno dello stesso gruppo hanno valori di Y correlati, ma anche che

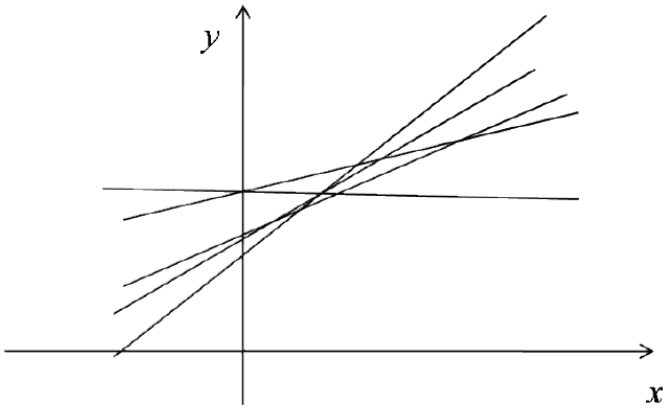

Figura 2.2: *Random slope model*

questa correlazione, così come la varianza di Y è dipendente dal valore di X, (il termine d'errore u_{1j} è connesso con x_{ij}). Da ciò deriva che l'errore totale sarà differente per differenti valori di X, situazione questa, che nei modelli di regressione ordinari, prende il nome di eteroschedasticità. Risultano pertanto violate le assunzioni di indipendenza e di omoschedasticità degli errori, su cui si basano i modelli di regressione ordinari. Attraverso l'equazione di regressione $y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}Z_j + \gamma_{11}x_{ij}Z_j + u_{1j}x_{ij} + u_{0j} + e_{ij}$ è, dunque, possibile stimare i coefficienti degli effetti fissi, degli effetti indipendenti delle variabili di secondo livello, di quelle di primo livello e la loro interazione. Il modello multilevel, inoltre, permette di quantificare la variabilità nei diversi livelli della gerarchia:

- variabilità entro gruppi, espressa dalla varianza σ^2;
- variabilità tra gruppi, espressa dalle varianze degli effetti casuali τ_0^2 e τ_1^2

Gli effetti stimati dal modello possono essere suddivisi in un primo insieme riguardante la *parte sistematica*, ovvero

- γ_{00} è l'intercetta: rappresenta il valore di Y qualora sia X che Z presentano valore zero
- γ_{01} è l'effetto del predittore del livello 2 (variabile esplicativa Z)
- γ_{10} è l'effetto del predittore del livello 1 (effetto di X su Y quando Z assume valore zero)
- γ_{11} è l'effetto dell'interazione tra i predittori del livello 1 e del livello 2

e in un secondo insieme riguardante la *parte aleatoria*, ovvero

- σ^2 varianza intra-classe (tra le unità di livello inferiore) controllando per l'effetto di X

- τ_0^2 varianza condizionata dell'intercetta rispetto a Z, (esprime la variabilità tra le macro unità per la parte relativa alla sola intercetta)
- τ_1^2 varianza condizionata del coefficiente di regressione rispetto a Z, (esprime la variabilità tra le macro unità per la parte legata all'effetto interazione)
- σ_{u01}^2 covarianza condizionata tra intercetta e coefficiente di regressione di primo livello.

Quando nel modello in esame si ha che la variabilità residua tra le unità di secondo livello relativa alle intercette e ai coefficienti di regressione risulta trascurabile, la parte casuale a livello macro risulta prossima allo zero; di conseguenza tendono a zero anche le stime delle varianze ad esse collegate τ_0^2 e τ_1^2. In una simile circostanza, il coefficiente di correlazione intraclasse è prossimo allo zero ed il modello di regressione multilevel si riduce ad un classico modello di regressione multipla, che include variabili indipendenti misurate indistintamente sia nel primo che nel secondo livello, poichè è inesistente la struttura gerarchica. In questa situazione, gli individui all'interno dei gruppi possono essere considerati indipendenti. Al contrario, l'esistenza di una variabilità significativa tra le intercette o tra i coefficienti di regressione, comporta la presenza di una elevata correlazione intraclasse e giustifica l'adozione del modello multilivello.

2.5 Design effect

Se i dati vengono disaggregati, tutte le unità di primo livello appartenenti allo stesso gruppo presenteranno gli stessi valori delle variabili di più alto livello. I test statistici ordinari trattano questi dati disaggregati come informazioni indipendenti. La dimensione campionaria di queste variabili dovrebbe però essere pari al numero delle unità di più alto livello. Usare come dimensione campionaria il numero dei casi disaggregati può in genere condurre a dei test di significatività che rifiutano l'ipotesi nulla più frequentemente rispetto al livello nominale α. L'importanza e le implicazioni di questo problema sono ben documentate nell'ambito degli studi di *sample survey*. Quando i dati hanno una struttura gerarchica, il campione che si estrae dalla popolazione è un campione a più stadi (ad esempio, l'estrazione di un campione di distretti scolastici, da ognuno dei quali si estrae poi un campione di scuole e così via). Nel campionamento a stadi si estraggono le macro-unità e, successivamente, si estraggono le unità all'interno del gruppo. In questo caso le probabilità di scelta sono note, ma non costanti. Un errore che frequentemente si commette è quello di ignorare la struttura dei dati e pretendere che le unità al livello più basso siano selezionate indipendentemente da quelle di livello superiore. In realtà, una volta selezionata l'unità primaria, aumentano le probabilità di scelta di un'unità secondaria appartenente a quel gruppo. Un

disegno campionario a stadi può essere descritto graficamente come in figura 2.3.

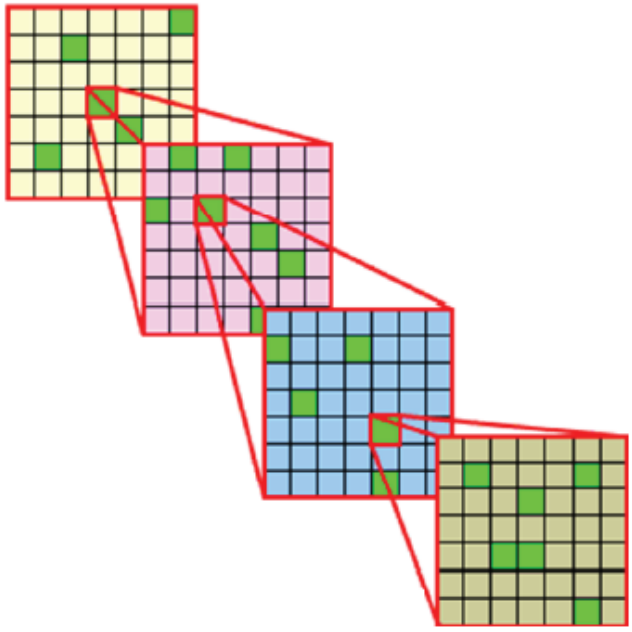

Figura 2.3: Il campionamento a stadi.

Le unità più scure sono quelle selezionate ad ogni livello, a partire dalle macro-unità in alto e, seguendo un percorso di tipo *top-down*, a cascata fino al arrivare al livello 1 delle unità elementari. E' preferibile utilizzare il campionamento a stadi in quanto i costi, per la fase di intervista o testing, sono fortemente ridotti se i soggetti da intervistare sono riconducibili a raggruppamenti geograficamente vicini o ad altri tipi di organizzazione in gruppi. Dopo la specificazione del disegno campionario, per lo studio di gerarchie, o in generale dei sistemi multilivello, è necessario distinguere le relazioni presenti tra le micro-unità e quelle tra le macro unità e le macro-micro unità. Inoltre, è bene sottolineare che i modelli statistici multilevel, per essere correttamente impiegati, sono necessariamente riferiti a un disegno campionario a più stadi (*multi-stage*).
Un campione su più stadi porta ad estrarre le unità iniziali dai livelli più elevati per poi procedere verso gli stadi via via più bassi. In campioni così realizzati le osservazioni individuali non risultano generalmente indipendenti. Inoltre, è importante tener conto del fatto che gli individui appartenenti ad uno stesso gruppo interagiscono tra loro e sono influenzati dal contesto socia-

le cui appartengono. Dal momento che non si può non valutare come e in che misura la struttura gerarchica delle osservazioni determini modificazioni nel comportamento della variabile dipendente di interesse, i dati devono essere elaborati come se si trattasse di un campione generato da più popolazioni caratterizzate da valori diversi dei parametri. In tale contesto, i modelli multilevel sono la risposta più adeguata, poichè sono strutturati per consentire l'analisi simultanea di variabili appartenenti a livelli differenti della gerarchia, includendo nel modello anche tutte le possibili forme di dipendenza.

Kish (1995) inoltre evidenzia gli effetti che questo tipo di campionamento ha sulla varianza campionaria. A tal proposito, egli definisce *design effect* (*deff*) il rapporto tra la varianza campionaria effettiva e la varianza campionaria che si otterrebbe con un campione casuale semplice della stessa numerosità. Quindi *deff* è il fattore per il quale quest'ultima varianza campionaria va moltiplicata per ottenere la varianza campionaria effettiva. Kish fornisce una formula per calcolare il *design effect* per un modello a due livelli nel caso di dati bilanciati:

$$deff = 1 + (n - 1)\rho \tag{2.1}$$

dove ρ è la già citata *intra-class correlation* e n è il numero (costante) di unità di primo livello all'interno di ogni unità di secondo livello. Nel caso di dati non bilanciati, se le numerosità dei gruppi non sono molto diseguali, si può usare la numerosità media come approssimazione. E' evidente che il *design effect* è pari a uno solo quando la *intra class correlation* è nulla o quando la numerosità dei gruppi è pari a 1. In tutti gli altri casi il *design effect* è più grande di 1 e quindi i test statistici standard tendono a sottostimare la varianza campionaria, comportando un errore di primo tipo α più elevato di quello nominale. I modelli multilevel risolvono questo problema incorporando un effetto casuale per ogni livello. La variabilità di questi effetti casuali è tenuta in considerazione quando vengono stimati gli errori standard. Usando la terminologia che compete all'ambito della *survey research*, queste stime degli errori standard tengono in considerazione l'effetto dell'*intra class correlation* (o del *design effect*) che si manifesta come risultato del campionamento adottato (Bryk e Raudenbush).
La dipendenza tra le osservazioni individuali può essere considerata come un fattore che "riduce" la numerosità campionaria effettiva. Considerando un campionamento a due stadi in cui tutti i gruppi sono costituiti dallo stesso numero di unità elementari, la numerosità campionaria effettiva n_{eff} può essere calcolata come segue (Kish, 1965):

$$n_{eff} = \frac{n}{1 + (n_{clus} - 1)\rho} \tag{2.2}$$

dove n è la numerosità campionaria totale, n_{clus} è la dimensione di ciascun gruppo e ρ è una opportuna misura della correlazione intraclasse[3]. Le cor-

[3] In letteratura sono stati proposti diversi coefficienti per la misura della correlazione

rezioni per gli effetti da disegno, come quella proposta da Kish, presentano due pesanti limiti. In primo luogo, la correlazione intraclasse varia al variare della variabile di interesse. In secondo luogo, i problemi relativi all'analisi di strutture gerarchiche sono in genere resi più complessi dalla presenza di variabili misurate su tutti i livelli della gerarchia. Emerge, quindi, la necessità di utilizzare un modello statistico che tenga conto della non indipendenza delle osservazioni e che consenta, allo stesso tempo, di analizzare simultaneamente variabili che provengono da diversi livelli della gerarchia [4].

I vantaggi che i modelli multilivello presentano, rispetto ai modelli di regressione classici, sono che essi consentono:

- di scomporre la varianza osservata in più fonti di variabilità riferite alle diverse unità di analisi. Tutto ciò non è fattibile con i metodi classici di regressione ordinaria dei minimi quadrati, attraverso i quali è possibile stimare una sola componente di varianza residuale, non essendo tenuta in considerazione la struttura gerarchica dei dati.
- di produrre errori standard dei coefficienti di regressione corretti (non sottostimati) e conseguentemente test di significatività più conservativi rispetto a quelli tradizionali ottenuti ignorando la presenza dei raggruppamenti.
- di misurare gli effetti delle interazioni *cross-level*. Queste ultime sono definite come interazioni tra variabli misurate a diversi livelli della gerarchia.

In letteratura sono trattati molti esempi. Nelle scienze dell'educazione e in sociologia, un esempio è costituito dalla "teoria dell'interazione attitudine-trattamento" (Cronbach e Webb (1975), Cronbach e Snow (1977)). Questa teoria postula che gli stili di insegnamento differiscono e che alcuni di questi sono più efficaci con studenti meno dotati, mentre altri con studenti più recettivi. Esiste quindi un effetto di interazione tra l'insegnante (variabile di secondo livello) e lo studente (variabile di primo livello). E' la possibilità di misurare gli effetti delle interazioni *cross-level* a costituire la caratteristica più apprezzata dei modelli multilivello nel campo delle ricerche educative, benchè non è stato dimostrato che questi modelli aiutino a scoprire interazioni che non possono essere scoperte con altri metodi (Kreft, 1996). Non è facile stabilire in quale misura essi abbiano contribuito allo sviluppo delle teorie *cross-level*. Va comunque sottolineato che è necessario disporre di dataset numerosi per individuare, se esistono, tali interazioni.

---

intraclasse. Tra i più importanti si segnalano quelli di Donner (1986), e Searle, Casella e McCulloch (1992).

[4] Per maggiori dettagli sull'argomento si faccia riferimento a Barcikowski (1981) e Cochran (1977)

2.6 Il modello multilivello lineare nella notazione matriciale

Al fine di illustrare le proprietà statistiche e le procedure di stima adatte per il modello più generale possibile è utile introdurre una notazione matriciale, facendo anche riferimento al seguente modello a due livelli

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + u_{0j} + u_{1j}z_{1ij} + u_{2j}z_{2ij} + e_{ij}.$$

Posto in generale:

- n = numero di unità elementari
- H = numero di livelli
- n_h = numero di unità di livello h $(h = 1, ..., H)$ (quindi $n_1 = n$)
- $n_{h(j)}$ = numero di unità elementari appartenenti alla $j - esima$ unità di livello h
- p = numero di parametri fissi
- q_h = numero di effetti casuali di livello h (dove $q_1 = 1$, componente e_{ij})

il modello multilivello lineare può scriversi nella forma matriciale:

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\beta + \mathbf{\Sigma}_{h=2}^{H}\mathbf{Z}_h\mathbf{u}_h + \epsilon = \\ &= \mathbf{X}\beta + \mathbf{\Sigma}_{h=1}^{H}\mathbf{Z}_h\mathbf{u}_h\end{aligned}$$

dove[5]

- $\underset{(n\times 1)}{\mathbf{y}}$ il vettore delle risposte;
- $\underset{(n\times p)}{\mathbf{X}}$ è la matrice delle variabili esplicative (effetti fissi);
- $\underset{(p\times 1)}{\beta}$ è il vettore dei parametri fissi;
- $\underset{(n\times(n_h q_h))}{\mathbf{Z}_h} = \oplus_{j=1}^{n_h}\{\underset{(n_{h(j)}\times q_h\times q_h)}{\mathbf{Z}_{h(j)}}\}$ è la matrice diagonale a blocchi relativa agli effetti casuali di livello h;
- $\underset{(n\times 1)}{\epsilon}$ è il vettore degli effetti casuali di primo livello

[5] Le dimensioni delle matrici sono riportate sotto le stesse nella forma (numero righe)×(numero colonne). Per quanto riguarda gli operatori matriciali, l'apice indica la matrice trasposta, $\otimes$ è il prodotto di Kronecker, mentre $\oplus$ è la *somma diretta* cioè $\oplus_{j=1}^{n}\mathbf{A}_j$ è la matrice diagonale a blocchi i cui blocchi sono, da sinistra verso destra, $\mathbf{A}_1, ..., \mathbf{A}_n$ (cfr. Searle et al., 1992, Appendix M)

- $\underset{(n_h q_h \times 1)}{\mathbf{u}} = (\underset{(1\times q_h)}{\mathbf{u}'_{h(1)}}, ..., \underset{(1\times q_h)}{\mathbf{u}'_{h(n_h)}})'$ è il vettore degli effetti casuali di livello h[6].

Indicheremo con $\mathbf{Y}$, $\mathbf{U}_h$ ed $\mathbf{E}$ i vettori relativi alle variabili casuali corrispondenti a $\mathbf{y}$, $\mathbf{u}_h$ ed ϵ .

Le ipotesi della parte casuale del modello sono le seguenti:

1. $E(\mathbf{U}_h) = \mathbf{0}$ gli effetti casuali hanno valore atteso nullo;
2. $Cov(\mathbf{U}_h, \mathbf{U}_{h'}) = \mathbf{0}$ per $h \neq h'$ gli effetti casuali relativi a unità appartenenti a livelli diversi sono incorrelati;
3. $Var(\mathbf{U}_h) = \mathbf{S}_h = \mathbf{I}_{n_h} \otimes \mathbf{\Omega}_h$ gli effetti casuali relativi a unità diverse appartenenti allo stesso livello sono incorrelati ed hanno la stessa matrice di covarianza.

Pertanto la matrice di covarianza di $\mathbf{Y}$ (condizionata a $\mathbf{X}$) è data da

$$\begin{aligned}\mathbf{V}_h &= Var(\sum_{h=1}^{H} \mathbf{Z}_h \mathbf{u}_h) = \\ &= \sum_{h=1}^{H} \mathbf{Z}_h \mathbf{S}_h \mathbf{Z}'_h = \\ &= \sum_{h=1}^{H} (\oplus_{j=1}^{n_h} \{\mathbf{Z}_{h(j)} \mathbf{\Omega}_h \mathbf{Z}'_{h(j)}\}) = \\ &= \sum_{h=1}^{H} \mathbf{V}_H^{(h)}\end{aligned}$$

dove $\mathbf{V}_H^{(h)}$ è il contributo degli effetti casuali di livello h della matrice di covarianza delle risposte in un modello a H livelli. Si noti che $\mathbf{V}_H^{(h)}$ è diagonale a blocchi, con blocchi corrispondenti alle unità di livello h. Per quanto riguarda la distribuzione degli effetti casuali, l'ipotesi usuale è quella di normalità, che risulta conveniente soprattutto in presenza di molti effetti casuali (Goldstein, 1995, p. 22 [7])

---

[6] Il termine di errore individuale è rappresentato da $\mathbf{u}_1$, che può essere pensato come effetto casuale di primo livello

[7] Quando l'ipotesi di normalità non è soddisfatta, gli stimatori dei parametri sono consistenti, ma non efficienti, mentre gli stimatori degli errori standard non sono consistenti (Goldstein,1995, p.22). L'ipotesi di normalità viene solitamente controllata per mezzo dei *diagrammi quantile-quantile* (Goldstein, 1995, p. 28; Longford, 1993, cap. 3).

2.7 Il modello multilivello lineare a due livelli nella notazione matriciale

A titolo esemplificativo si presenta ora la forma matriciale esplicita del modello multilivello a due livelli.
La rappresentazione per una singola unità risulta essere:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + u_{0j} + u_{1j}z_{1ij} + u_{2j}z_{2ij} + e_{ij}$$

mentre per l'intero campione risulta essere (con $H = 2$ livelli e $j = 2$ gruppi):

$$\begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \\ y_{12} \\ y_{22} \\ y_{32} \\ y_{42} \end{bmatrix} = \begin{bmatrix} 1\ x_{111}\ x_{211} \\ 1\ x_{121}\ x_{221} \\ 1\ x_{131}\ x_{231} \\ 1\ x_{112}\ x_{212} \\ 1\ x_{122}\ x_{222} \\ 1\ x_{132}\ x_{232} \\ 1\ x_{142}\ x_{242} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} 1\ z_{111}\ z_{211}\ 0\ 0\ 0 \\ 1\ z_{121}\ z_{221}\ 0\ 0\ 0 \\ 1\ z_{123}\ z_{231}\ 0\ 0\ 0 \\ 0\ 0\ 0\ 1\ z_{112}\ z_{212} \\ 0\ 0\ 0\ 1\ z_{122}\ z_{222} \\ 0\ 0\ 0\ 1\ z_{132}\ z_{232} \\ 0\ 0\ 0\ 1\ z_{142}\ z_{242} \end{bmatrix} \begin{bmatrix} u_{01} \\ u_{11} \\ u_{21} \\ u_{02} \\ u_{12} \\ u_{22} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{32} \\ \epsilon_{42} \end{bmatrix}$$

A livello aggregato (livello-2) per $y_{j(2)}$ risulta essere:

$$\begin{bmatrix} \mathbf{y}_{1(2)} \\ \mathbf{y}_{2(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{1(2)} \\ \mathbf{x}_{2(2)} \end{bmatrix} \beta + \begin{bmatrix} \mathbf{z}_{1(2)} & 0 \\ 0 & \mathbf{z}_{2(2)} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{1(2)} \\ \epsilon_{2(2)} \end{bmatrix}$$

La rappresentazione (livello-2) per le y_{ij} (in questo caso per il gruppo $j = 1$) risulta quindi essere:

$$\begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \end{bmatrix} = \begin{bmatrix} 1\ x_{11j}\ x_{21j} \\ 1\ x_{12j}\ x_{22j} \\ 1\ x_{13j}\ x_{23j} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} 1\ z_{11j}\ z_{21j} \\ 1\ z_{12j}\ z_{22j} \\ 1\ z_{12j}\ z_{23j} \end{bmatrix} \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} + \begin{bmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \epsilon_{3j} \end{bmatrix}$$

2.8 Stima dei parametri

La stima dei parametri di un modello multilivello lineare ha costituito per lungo tempo un problema proibitivo a causa della notevole mole di calcoli richiesta dagli algoritmi di stima. Fra i metodi proposti in letteratura si ricordano:

- Massima verosimiglianza (ML) (Harville, 1977; Longford, 1987);
- Massima verosimiglianza vincolata (REML) (Patterson e Thompson, 1971);

- Minimi quadrati generalizzati iterati (IGLS) (Goldstein, 1986);
- Minimi quadrati generalizzati iterati vincolati (RIGLS) (Goldstein, 1989);
- Algoritmo EM (Aitkin, 1981; Bryk e Raudenbush, 1992);
- Analisi bayesiana con metodi Markov Chain Monte Carlo (MCMC) (Gilks, 1996).

Per quanto riguarda la verosimiglianza, si osserva che l'ipotesi di normalità degli effetti casuali permette di determinare facilmente la distribuzione marginale della risposta. Con riferimento alla notazione matriciale del modello

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{\Sigma}_{h=1}^{H}\mathbf{Z}_h\mathbf{u}_h$$

si ottiene che la v.c. che genera le osservazioni è

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{V}_H(\theta))$$

dove θ è il vettore che raccoglie i parametri cosiddetti casuali contenuti nelle matrici $\mathbf{\Omega}_1, ..., \mathbf{\Omega}_H$. Pertanto il logaritmo naturale della verosimiglianza marginale è dato da

$$l(\beta, \theta|\mathbf{y}) = -\frac{1}{2}\{n\log(2\pi) + \log(\det \mathbf{V}_H(\theta)) + (\mathbf{y} - \mathbf{X}\beta)'\mathbf{V}_H^{-1}(\theta)(\mathbf{y} - \mathbf{X}\beta)\} \tag{2.3}$$

Il vettore delle derivate parziali prime della funzione di log-verosimiglianza è chiamato vettore di *scoring* $\mathbf{S}$. Se un punto di massimo $\widehat{\theta}$ della funzione di log-verosimiglianza si trova all'interno dello spazio parametrico $\mathbf{\Theta}$ e il vettore di *scoring* è definito nell'intorno di questo punto, allora:

$$\mathbf{S}(\mathbf{y}; \widehat{\theta}) = \mathbf{0}.$$

Un approccio standard delle stime di massima verosimiglianza è quello di trovare tutti i vettori $\widehat{\theta}$ per i quali $\mathbf{S}(\mathbf{y}; \widehat{\theta}) = \mathbf{0}$ e di esplorare il comportamento della funzione di log-verosimiglianza lungo la frontiera dello spazio parametrico e nei punti in cui il vettore di *scoring* non è definito. Solo in questi punti si possono trovare le stime di massima verosimiglianza.
Gli stimatori ora descritti possiedono delle proprietà desiderabili: sotto ipotesi generali, sono consistenti e asintoticamente efficienti; al crescere del campione, la loro distribuzione campionaria diventa approssimativamente normale. Inoltre, se si vuole stimare una funzione dei parametri, se si sostituisce al valore di detti parametri le stime di massima verosimiglianza, la funzione risultante è a sua volta uno stimatore di massima verosimiglianza.
Tuttavia è noto che, con questo metodo, nella stima delle componenti di varianza non si tiene conto dei gradi di libertà dovuti alla stima degli effetti fissi. Questo problema naturalmente si presenta nei modelli ad effetti casuali: nelle stime di massima verosimiglianza non viene fatta alcuna distinzione tra parametri noti oppure stimati. Una soluzione generale di questo problema, che si deve a Patterson e Thompson (1971), è basare le stime delle componen-

ti di varianza e covarianza non sulle osservazioni, ma su combinazioni lineari di queste ultime, scelte in modo tale da non contenere effetti fissi. Queste combinazioni lineari, che si indicheranno con $\mathbf{k}'\mathbf{y}$, risultano essere equivalenti ai residui ottenuti dopo aver stimato gli effetti fissi. Il vettore $\mathbf{k}'$ è scelto in modo tale che $\mathbf{k}'\mathbf{y} = \mathbf{k}'XZ\gamma + Xu + e$ non contenga termini fissi, cioè in modo che $\mathbf{k}'XZ\gamma = 0 \ \forall\gamma$, quindi $\mathbf{k}'XZ = 0$[8]. Harville (1977) chiama $\mathbf{k}'y$ *error contrast*: gli *error contrasts* formano uno spazio lineare di dimensione $[N - rango(XZ)]$.
Massimizzare la funzione di log-verosimiglianza per $\mathbf{y}$ porta alle stime *full maximum likelihood* (FML), mentre massimizzare tale funzione per l'insieme degli "*error contrasts*" porta alle stime *restricted maximum likelihood* (REML). I metodi FML e REML in genere producono risultati molto simili per quanto riguarda σ^2, ma abbastanza diverse per quanto concerne $\mathbf{V}$ matrice di varianze e covarianze. Per essere più precisi, la differenza delle stime dipende dalla numerosità delle unità di secondo livello, in quanto lo stimatore FML è pari a $\frac{J-F}{J}$ volte lo stimatore REML, dove F è la dimensione del vettore γ (Bryk e Raudenbush, 1992).
Il calcolo delle stime di massima verosimiglianza, siano esse FML o REML, implica complesse espressioni non lineari nei parametri. In queste situazioni le equazioni vengono risolte tramite procedure iterative, attraverso l'utilizzo di specifici software. La massimizzazione della (2.3) comporta quindi alcuni problemi computazionali, che sono stati risolti da Longford (1987), il quale ha proposto un algoritmo di massimizzazione di tipo *Fisher scoring*. Di seguito si illustrerà in dettaglio il metodo IGLS proposto da Goldstein (1986), poichè è quello implementato nel programma MLwiN (Goldstein *et. al*, 1998). In realtà, gli algoritmi *Fisher scoring* e IGLS sono formalmente equivalenti (Goldstein, 1995, p. 23).
Il metodo IGLS si basa sulla seguente osservazione: se i parametri fissi fossero noti si potrebbe usare il principio dei minimi quadrati generalizzati (GLS) per stimare i parametri casuali, e viceversa. Pertanto, partendo da una stima iniziale dei parametri fissi (ad esempio ottenuta con i minimi quadrati ordinari), l'algoritmo IGLS alterna la stima dei parametri casuali e fissi con il metodo GLS, fino a convergenza. Usando la notazione del modello

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{\Sigma}_{h=1}^{H}\mathbf{Z}_h\mathbf{u}_h$$

e scrivendo $\mathbf{V}$ in luogo di $\mathbf{V}_H$, i due passi dell'algoritmo IGLS possono essere formalizzati come segue:

1. Stima GLS dei parametri fissi
   Nota la matrice $\mathbf{V}$, lo stimatore GLS dei parametri fissi è

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{2.4}$$

   con

[8] Searle (1992)

$$Cov(\hat{\beta}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

2. Stima GLS dei parametri casuali
Noto il vettore β, lo stimatore GLS dei parametri casuali (inclusi nella matrice $\mathbf{V}$, in base alla

$$\begin{aligned}\mathbf{V}_h &= var(\sum_{h=1}^{H} \mathbf{Z}_h \mathbf{u}_h) = \\ &= \sum_{h=1}^{H} \mathbf{Z}_h \mathbf{S}_h \mathbf{Z}_h' = \\ &= \sum_{h=1}^{H} (\oplus_{j=1}^{n_h} \{\mathbf{Z}_{h(j)} \mathbf{\Omega}_h \mathbf{Z}_{h(j)}'\}) = \\ &= \sum_{h=1}^{H} \mathbf{V}_H^{(h)}\end{aligned}$$

può essere ottenuto come segue. Indichiamo con

$$\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\beta$$

il vettore dei residui, per i quali vale la relazione $E(\tilde{\mathbf{y}}\tilde{\mathbf{y}}') = \mathbf{V}$. Poi definiamo un nuovo vettore $\mathbf{y}^*$ tale che

$$\mathbf{y}^* = vec(\tilde{\mathbf{y}}\tilde{\mathbf{y}}') \tag{2.5}$$

dove *vec* è l'operatore che forma un vettore da una matrice impilando le sue colonne una sotto l'altra. Adesso, indicando i parametri casuali come vettore θ, è possibile scrivere un modello lineare per i parametri casuali:

$$E(\mathbf{Y}^*) = \mathbf{X}^*\theta$$

dove la matrice dei regressori $\mathbf{X}^*$ può essere determinata colonna per colonna in base alla seguente formula (Goldstein, 1986; Goldstein and Rasbash, 1992):

$$\mathbf{x}_k^* = vec(\frac{\partial \mathbf{V}}{\partial \theta_k}) = vec[\oplus_{j=1}^{n_{\overline{h}}} \{\mathbf{Z}_{\overline{h}(j)} (\frac{\partial \mathbf{\Omega}_{\overline{h}}}{\partial \theta_k}) \mathbf{Z}_{\overline{h}(j)}'\}]$$

dove $\mathbf{x}_k^*$ è la $k-ma$ colonna di $\mathbf{X}^*$, mentre θ_k è il $k-mo$ elemento di θ che assumiamo essere un effetto casuale appartenente al livello arbitrario $\overline{h}$. Il modello lineare consente di usare il metodo GLS per stimare gli effetti casuali:

$$\hat{\theta} = (\mathbf{X}^{*'}\mathbf{V}^{*-1}\mathbf{X}^{*})^{-1}\mathbf{X}^{*'}\mathbf{V}^{*-1}\mathbf{y}^{*} \tag{2.6}$$

dove $\mathbf{V}^{*} = \mathbf{V} \oplus \mathbf{V}$. Si noti che $\mathbf{V}^{*}$ non è esattamente la matrice di covarianza di $\mathbf{y}^{*}$, la quale è singolare e quindi non può essere usata nella stima GLS[9]. Si dimostra poi (Goldstein and Rasbash, 1992) che

$$Cov(\hat{\theta}) = 2(\mathbf{X}^{*'}\mathbf{V}^{*-1}\mathbf{X}^{*})^{-1}$$

L'algoritmo IGLS itera tra la (2.4) e la (2.6) fino a convergenza, usando di volta in volta le stime correnti dei parametri fissi e casuali (le stime iniziali dei parametri fissi sono solitamente ottenute con i minimi quadrati ordinari). Goldstein (1986) dimostra che, sotto ipotesi di normalità, le stime così ottenute sono di massima verosimiglianza. In assenza di normalità, lo stimatore IGLS è comunque consistente, anche se non pienamente efficiente; tuttavia il corrispondente stimatore degli errori standard non è più consistente (Goldstein, 1995, p. 22). Lo stimatore IGLS in generale è distorto e ciò può costituire un problema nei campioni di piccola numerosità. Pertanto è utile disporre anche di uno stimatore non distorto, che può essere ricavato apportando una piccola modifica alla procedura IGLS. Infatti, il passo dell'algoritmo deputato alla stima dei parametri casuali si basa sulla relazione

$$E[(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)'] = \mathbf{V}$$

Tuttavia tale relazione non è più vera se si sostituisce β con il suo stimatore GLS $\widehat{\beta}$, poichè in tal caso

$$E[(\mathbf{Y} - X\hat{\beta})(\mathbf{Y} - \mathbf{X}\hat{\beta})'] = \mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{'} \tag{2.7}$$

Per correggere questo errore si può sommare il termine $\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'$ a $\widetilde{\mathbf{y}}$ prima di calcolare $\mathbf{y}^{*}$. In tal modo lo stimatore IGLS diviene corretto e viene indicato con l'acronimo RIGLS (Goldstein, 1989)[10]. Comunque, anche nei campioni di piccola numerosità, la scelta fra IGLS e RIGLS non è ovvia, poichè alcuni studi di simulazione mostrano che la correttezza del metodo vincolato viene pagata con una minore efficienza e che non esistono linee guida per risolvere tale conflitto in favore di un metodo o dell'altro (Kreft and De Leeuw, 1998, par. 5.4).

[9] La matrice di covarianza di $\mathbf{y}^{*}$ è $\mathbf{V}^{*}(\mathbf{I}+\mathbf{S}_N)$, dove $\mathbf{S}_N$ è la cosidetta *vec permutation matrix* (Searle et al., 1992, par. 12.3)

[10] Nell'acronimo RIGLS la R sta per Restricted. L'origine di tale termine va ricercata nel fatto che lo stimatore RIGLS è equivalente allo stimatore di massima verosimiglianza vincolata (REML).

2.9 Stima degli effetti casuali (o residui)

Anche l'uso di un modello multilivello presuppone la stima degli effetti casuali (o residui). In realtà gli effetti casuali sono variabili aleatorie per cui ciò che si stima è la realizzazione di tali variabili aleatorie nei vari gruppi. In un modello di regressione classico i termini di errore (che si riferiscono ad un unico livello) sono usualmente stimati attraverso i residui di regressione. Invece in un modello multilivello i residui $\mathbf{y} - \mathbf{X}\beta$, detti residui totali stimati, devono essere opportunamente scomposti nelle loro componenti di primo, secondo, . . ., $H-mo$ livello. Supponendo per il momento noti tutti i parametri del modello, i residui di livello h possono essere stimati per mezzo del loro valore atteso, condizionato ai residui totali veri $\widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\beta$,

$$\hat{\mathbf{u}}_h = E(\mathbf{U}_h|\tilde{\mathbf{y}}) \tag{2.8}$$

Se si assume una distribuzione normale degli effetti casuali, il valore atteso (2.8) può essere calcolato tramite la distribuzione a posteriori $\mathbf{u}_h|\tilde{\mathbf{y}}$. In generale, in assenza di ipotesi distribuzionali specifiche (come nel caso dello stimatore IGLS), si può usare una semplice regressione lineare di $\mathbf{u}_h$ su $\tilde{\mathbf{y}}$ (Goldstein, 1995, app. 2.2). Poiché $cov(\tilde{\mathbf{y}}, \mathbf{u}_h) = \mathbf{R}_h$, dove

$$\mathbf{R}_h = \oplus_{j=1}^{n_h}\{\mathbf{Z}_{h(j)}\boldsymbol{\Omega}_h\}$$

dalla regressione abbiamo

$$\hat{\mathbf{u}}_h = \mathbf{R}_h^{'}\mathbf{V}^{-1}\tilde{\mathbf{y}} \tag{2.9}$$

Sostituendo ai parametri incogniti il loro valore stimato, si ottiene uno stimatore consistente degli effetti casuali che, nell'ipotesi di normalità, coincide con lo stimatore bayesiano empirico. La sua matrice di covarianza, data la (2.7), è

$$Var(\hat{\mathbf{U}}_h) = \mathbf{R}_h^{'}\mathbf{V}^{-1}(\mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}')\mathbf{V}^{-1}\mathbf{R}_h \tag{2.10}$$

La (2.10) è nota come matrice di covarianza non condizionata e i relativi errori standard vengono detti diagnostici, poichè vengono usati per standardizzare i residui ai fini diagnostici (ad esempio per tracciare il diagramma quantile-quantile per la verifica dell'ipotesi di normalità). Tuttavia, se il fine è quello di fare inferenza sul valore assunto dagli effetti casuali (ad esempio costruendo un intervallo di confidenza) è opportuno usare la *matrice di covarianza condizionata*

$$\mathbf{S}_h\mathbf{R}_h^{'}\mathbf{V}^{-1}(\mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}')\mathbf{V}^{-1}\mathbf{R}_h \tag{2.11}$$

che può essere ottenuta come errore quadratico medio della regressione di $\mathbf{u}_h$ su $\tilde{\mathbf{y}}$, oppure, nell'ipotesi di normalità, come varianza della distribuzione a posteriori $\mathbf{u}_h|\tilde{\mathbf{y}}$. Gli errori standard ottenuti dalla (2.11) si dicono compara-

tivi, in quanto spesso vengono usati nei confronti fra effetti casuali relativi a unità diverse. Si noti che nel calcolo della (2.10) e della (2.11) si tiene conto della variabilità campionaria dei coefficienti fissi, ma non di quella degli effetti casuali. Pertanto, in campioni di piccola numerosità, può essere opportuno stimare tali matrici di covarianza con procedure di tipo bootstrap (Goldstein, 1995, par. 3.5).

2.10 L'effetto *shrinkage*

Al fine di mostrare le proprietà dello stimatore dei residui (2.9) è utile considerare il seguente modello a componenti di varianza:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$$

$$\begin{cases} \beta_{0j} = \beta_0 + u_j \\ U_j \sim N(0, \sigma_u^2) \\ E_{ij} \sim N(0, \sigma_e^2) \end{cases} \tag{2.12}$$

Si ha $\mathbf{\Omega}_1 = \sigma_e^2$; $\mathbf{\Omega}_2 = \sigma_u^2$; $\mathbf{Z}_{1(j)} = 1$ per ogni $j = 1, ..., n_1$; $\mathbf{Z}_{2(j)} = \mathbf{1}_{n_{2(j)}}$ per ogni $j = 1, ..., n_2$ ($\mathbf{1}_k$ indica il vettore unitario di lunghezza k). Pertanto dalla (2.9) si ricava

$$\hat{u}_j = s(n_j, \tau) \cdot \check{y}_j, \tag{2.13}$$

dove

- $\check{y}_j = (\bar{y}_j - \hat{\alpha} - \hat{\beta}\bar{x}_j)$ è il residuo stimato medio del $j - mo$ gruppo ($\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$ e analogamente $\bar{x}_j$)
- $s(n_j, \tau) = \frac{1}{1 + \frac{1}{n_j \tau}}$ è il cosiddetto *shrinkage factor*
- $\tau = \frac{\sigma_u^2}{\sigma_e^2}$ è il rapporto delle componenti di varianza).

Lo shrinkage factor, compreso fra 0 e 1, comprime il residuo stimato medio, in modo differenziato a seconda della numerosità del gruppo n_j e del rapporto fra le componenti di varianza τ. In particolare, lo *shrinkage* sarà più forte nei gruppi poco numerosi che in quelli molto numerosi; inoltre, a parità di numerosità, lo *shrinkage* sarà più forte quando la componente di varianza *between* è piccola rispetto a quella *within*. Lo *shrinkage* rende più affidabile la stima degli effetti casuali, poichè tende a riportare verso lo zero (cioè verso la media degli effetti casuali nella popolazione) la stima relativa ai gruppi poco numerosi, che contengono cioè poca informazione per la stima dell'effetto casuale. D'altra parte lo *shrinkage* ha delle conseguenze indesiderate quando si vogliano confrontare due gruppi sulla base dei residui stimati, poichè può accadere che un gruppo con un elevato valore dell'effetto casuale

ma di scarsa numerosità abbia lo stesso residuo stimato di un gruppo con un piccolo valore dell'effetto casuale ma di grande numerosità. L'effetto dello *shrinkage* si ripercuote anche sulla stima di α_j nella

$$y_{ij} = \beta_{0j} + \beta x_{ij} + e_{ij}$$

$$\begin{cases} \beta_{0j} = \beta_0 + u_j = \\ U_j \sim N(0, \sigma_u^2) = \\ E_{ij} \sim N(0, \sigma_e^2) \end{cases}$$

Infatti

$$\begin{aligned} \hat{\beta}_{0j} &= \hat{\beta}_0 + \hat{u}_j \\ &= \hat{\beta}_0 + s(n_j, \tau) \cdot \check{y}_j \\ &= (1 - s(n_j, \tau))\hat{\beta}_0 + s(n_j, \tau)(\bar{y}_j - \hat{\beta}_1 \bar{x}_j) \end{aligned}$$

per cui $\hat{\alpha}_j$ risulta un valore intermedio tra $\hat{\alpha}$ (stima del coefficiente medio nella popolazione) e $\bar{y}_j - \hat{\beta}\bar{x}_j$ (stima relativa al $j - mo$ gruppo). Questa proprietà viene indicata con il termine *borrowing strenght*. I concetti di *shrinkage* e *borrowing strenght*, che abbiamo illustrato per il modello a componenti di varianza, valgono in generale (Bryk e Raudenbush, 1992) e rappresentano uno degli aspetti più caratteristici dell'analisi multilivello.

Capitolo 3
Caratteristica dei modelli multilevel per dati politomici

I modelli multilivello sono stati inizialmente concepiti per lo studio di variabili quantitative e con specificazione lineare del valore atteso. Il florido sviluppo di questa classe di modelli può essere spiegato in parte dal loro vasto campo di applicazione e in parte dalla relativa semplicità della trattazione matematica e dell'interpretazione statistica. Tuttavia le esigenze della ricerca scientifica, soprattutto in ambito sociale e biomedico, hanno spinto verso un'estensione dei modelli multilivello, per poter includere specificazioni non lineari del valore atteso della risposta. In particolare, una specificazione non lineare del valore atteso si rende necessaria quando la risposta è di tipo qualitativo, come accade di frequente nelle indagini in ambito sociale. In questo paragrafo si concentrerà l'attenzione sui modelli lineari generalizzati multilivello, che costituiscono la scelta più conveniente per l'analisi di dati categorici con struttura gerarchica. Dopo un'introduzione generale, si esamineranno in dettaglio i modelli per dati binari, politomici, ordinali e di sopravvivenza in tempo discreto, concludendo con la descrizione di alcune specifiche procedure di stima.

3.1 Definizione e interpretazione

Prima di parlare di modelli multilivello non lineari è opportuno soffermarci su alcune proprietà degli analoghi modelli lineari che a prima vista sembrano ovvie, ma che in realtà sono fondamentali per capire le implicazioni della non linearità. A fini illustrativi si considera il seguente modello lineare a due livelli[1]:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij}$$

[1] Rispetto alla notazione precedentemente usata si è sostituito γ_{00} con β_{0j} e γ_{01} con β_{1j}. Si ricorda inoltre che "iid" ste per "indipendenti e identicamente distribuiti" e che il simbolo $\perp$ indica, in presenza di normalità, indipendenza stocastica.

dove le e_{ij}, u_{0j} e u_{1j} sono determinazioni di variabili casuali

$$E_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$$

$$\begin{bmatrix} U_{0j} \\ U_{1j} \end{bmatrix} \overset{iid}{\sim} N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01}^2 & \sigma_{u1}^2 \end{bmatrix})$$

Una conseguenza di queste ipotesi distribuzionali è che sono normali sia la distribuzione della risposta condizionata agli effetti casuali, sia la sua distribuzione marginale:

$$Y_{ij}|u_{0j}, u_{1j} \sim N(\beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij}, \sigma_e^2)$$

$$Y_{ij} \sim N(\beta_{0j} + \beta_{1j}x_{ij}, \sigma_e^2 + (\sigma_{u0}^2 + \sigma_{u1}^2 x_{ij}^2 + 2\sigma_{u01}x_{ij})).$$

Osservando la distribuzione marginale notiamo che rispetto ad un modello di regressione classico la struttura di covarianza è più complessa, ma i valori attesi sono identici; in altre parole, gli effetti casuali modificano solo la struttura di covarianza. Come si vedrà tra breve, ciò non è più vero nei modelli non lineari. Un'altra considerazione riguarda l'effetto delle covariate sulla risposta, infatti si ha

$$\frac{\partial}{\partial x_{ij}} E(Y_{ij}|u_{0j}, u_{1j}) = \beta_{1j} + u_{1j} \;\; (effetto \;\; medio \;\; nel \;\; gruppo \;\; j)$$

$$\frac{\partial}{\partial x_{ij}} E(Y_{ij}) = \beta_{1j} \;\; (effetto \;\; medio \;\; nella \;\; popolazione),$$

per cui l'effetto di X su Y dipende dal gruppo, ma nel gruppo medio (cioè, quello per il quale $u_{0j} = u_{1j} = 0$) tale effetto coincide con quello medio della popolazione. Anche questa fondamentale proprietà è peculiare dei modelli lineari. Dopo queste osservazioni si prende in esame la classe più importante di modelli multilivello non lineari, cioè i modelli lineari generalizzati multilivello generalizzati, che, per una struttura a due livelli, possono essere definiti come segue:

1. in ogni gruppo le risposte sono indipendenti condizionatamente agli effetti casuali del gruppo e seguono un modello lineare generalizzato (GLM, Generalised Linear Model: McCullagh e Nelder, 1989);
2. gli effetti casuali dei vari gruppi sono un campione casuale semplice da una distribuzione comune multivariata (solitamente gaussiana).

Questo modo di definire il modello non consente, in generale, di determinare in forma analitica la distribuzione marginale, salvo il caso del modello lineare con distribuzione normale. Negli altri casi è possibile definire direttamente la distribuzione marginale (Diggle et al., 1994), ma, come si vedrà, ciò risulta

meno conveniente per gli sviluppi teorici. Per quanto riguarda la specificazione della distribuzione condizionata, l'uso dei GLM costituisce la regola, poiché tali modelli possiedono un elevato grado di generalità e poggiano su fondamenti teorici ormai consolidati. Anche l'assunzione di normalità degli effetti casuali è largamente diffusa, sebbene non manchino proposte in senso contrario[2].

In un GLM multilivello le risposte sono marginalmente indipendenti tra gruppi diversi, mentre all'interno di uno stesso gruppo l'indipendenza non è marginale, ma è condizionata agli effetti casuali. Ciò significa che la dipendenza esistente tra le risposte di un certo gruppo è interamente attribuibile agli effetti casuali, cioè a quei fattori non osservabili comuni a tutte le unità del gruppo. Formalmente, la versione GLM multilivello del modello lineare

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij}$$

comporta le seguenti implicazioni:

1. dati (u_{0j}, u_{1j}), le risposte $y_{1j}, ..., y_{n_j j}$ sono mutuamente indipendenti e seguono un GLM con densità

$$f(y_{ij}|u_{0j}, u_{1j}) = exp\{\frac{[y_{ij}\theta_{ij} - \psi(\theta_{ij})]}{\phi} + c(y_{ij}, \phi)\},$$

dove θ_{ij} è il parametro naturale, ϕ è il parametro di dispersione, mentre $\psi(\cdot)$ e $c(\cdot)$ sono funzioni note; il valore atteso e la varianza condizionati soddisfano:

$$\mu^u_{ij} = E(Y_{ij}|u_{0j}, u_{1j}) = g^{-1}(\beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij}),$$

$$\upsilon^u_{ij} = Var(Y_{ij}|u_{0j}, u_{1j}) = \upsilon(\mu^u_{ij})\phi,$$

dove $g^{-1}(\cdot)$ è l'inversa della funzione link $g(\cdot)$, mentre $\upsilon(\cdot)$ è la funzione di varianza;
2. gli effetti casuali $\{(u_{0j}, u_{1j}) : j = 1, ..., J\}$ sono un campione casuale semplice da una distribuzione multivariata, solitamente gaussiana:

$$\begin{bmatrix} U_{0j} \\ U_{1j} \end{bmatrix} \overset{iid}{\sim} N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix})$$

[2] Generalmente l'assunzione di normalità viene abbandonata nel caso di presenza di effetti casuali con distribuzione discreta (McDonald, 1994) oppure quando si voglia ottenere la distribuzione della risposta nella cosiddetta forma chiusa (Conaway, 1990). L'assunzione di normalità resta comunque la scelta classica, poichè è conveniente da un punto di vista teorico ed è difficilmente confutabile dall'evidenza empirica, in quanto gli effetti casuali sono quantità non osservabili. Inoltre, Gibbons et al. (1994) e Gibbons e Hedeker (1997) mostrano che, nelle applicazioni da loro discusse, la scelta di una distribuzione uniforme degli effetti casuali conduce a risultati del tutto simili a quelli ottenibili con la distribuzione normale.

A differenza del modello lineare, nel GLM multilivello non compare il termine di errore di primo livello, che viene implicitamente specificato con la distribuzione della risposta. L'unico parametro relativo alla variabilità di primo livello è il parametro di dispersione ϕ, che però in alcune importanti distribuzioni (es. Binomiale, Poisson) risulta fissato a priori[3]. Si noti inoltre che, ad eccezione del modello lineare normale, la varianza

$$\upsilon_{ij}^{u} = Var(Y_{ij}|u_{0j}, u_{1j}) = \upsilon(\mu_{ij}^{u})\phi,$$

dipende dal valore atteso condizionato e quindi dalla realizzazione degli effetti casuali. Il valore atteso marginale della risposta è dato da

$$E(Y_{ij}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g^{-1}(\beta_{0j} + \beta_{1j}x_{ij} + u_0 + u_1 x_{ij})\varphi(u_0, u_1)du_0 du_1,$$

dove $\varphi(\cdot, \cdot)$ è la densità di probabilità degli effetti casuali. Tale valore atteso è diverso da quello condizionato relativo al gruppo medio

$$E(Y_{ij}|u_{0j} = 0, u_{1j} = 0) = g^{-1}(\beta_{0j} + \beta_{1j}x_{ij})$$

Una conseguenza di questo fatto è che:

$$\frac{\partial}{\partial x_{ij}} E(Y_{ij}|u_{0j} = 0, u_{1j} = 0) \neq \frac{\partial}{\partial x_{ij}} E(Y_{ij})$$

cioè, l'effetto di x su y nel gruppo medio non coincide con l'effetto medio nella popolazione. Questa osservazione è fondamentale per interpretare correttamente il coefficiente β, che ha un significato diverso rispetto al β di un analogo modello senza effetti casuali. A questo proposito giova ricordare che in letteratura sono presenti due approcci alternativi per l'analisi di dati gerarchici (Zeger et al., 1988; Goldestein e Rasbash, 1996):

1. L'approccio *unit specific*, o condizionato (nel quale rientrano i modelli multilivello), che consiste nel rendere esplicita l'influenza della gerarchia per mezzo degli effetti casuali, specificando la distribuzione della risposta in modo condizionato; in questi modelli i coefficienti si riferiscono all'effetto delle covariate per ogni dato gruppo (unità di secondo livello).
2. L'approccio *population average* o marginale, che si basa sulla specificazione della distribuzione marginale della risposta, considerando la correlazione generata dalla gerarchia come un fattore di disturbo; in questi

[3] Se si effettua la stima con un metodo di quasi-verosimiglianza (Wedderburn, 1974) il parametro ϕ può diventare comunque oggetto di stima, qualora si voglia modellare una extra-variabilità (Williams, 1982). Tuttavia nei GLM multilivello il problema dell'extravariabilità è più raro che nei GLM ordinari, poichè gli effetti casuali contribuiscono a modellare in modo migliore la variabilità. Sul ruolo dell'extra-variabilita nei modelli multilivello cfr. Goldstein (1995), pp. 98-99.

modelli i coefficienti si riferiscono all'effetto medio delle covariate nella popolazione.

I coefficienti di regressione nei due casi possono essere diversi. Ad esempio, in un modello logit ad intercetta casuale i coefficienti di regressione del modello *unit specific* $(\beta_1^{US}, ..., \beta_p^{US})$ e del modello *population average* $(\beta_1^{PA}, ..., \beta_p^{PA})$ sono tali che (Neuhaus et al., 1991):

1. $|\beta_k^{PA}| \leq |\beta_k^{US}|$ per ogni $k = 1, ..., p$
2. l'uguaglianza vale solo se $\beta_k^{US} = 0$
3. la differenza fra β_k^{PA} e β_k^{US} aumenta all'aumentare della varianza dell'effetto casuale

La scelta fra i due approcci è dettata dalle finalità dell'indagine: se la struttura gerarchica ha un interesse specifico è opportuno usare un modello *unit specific*, altrimenti si possono usare entrambi. In effetti, il modello *unit specific* è più generale, poichè può essere usato anche per studiare l'effetto medio delle covariate nella popolazione. L'unica difficoltà è che il valore atteso marginale,

$$E(Y_{ij}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g^{-1}(\alpha + \beta x_{ij} + u_0 + u_1 x_{ij}) \varphi(u_0, u_1) du_0 du_1,$$

è dato da un integrale che spesso non ha soluzione analitica: tuttavia il problema può essere facilmente risolto per mezzo di un'approssimazione analitica o di una simulazione Monte Carlo (Goldstein, 1995, par. 5.3).

3.2 Modelli per dati binari

Quando la risposta è binaria, cioè $y_{ij} \in \{0, 1\}$, il valore atteso $E(Y_{ij})$ coincide con la probabilità di successo $P\{Y_{ij} = 1\}$, e ciò vale anche condizionatamente agli effetti casuali. Pertanto, il GLM multilivello viene solitamente scritto sostituendo $\mu_{ij}^u = E(Y_{ij}|u_{0j}, u_{1j})$ con[4]

$$\pi_{ij} = P\{Y_{ij} = 1|u_{0j}, u_{1j}\}.$$

I modelli lineari generalizzati (GLM) costituiscono un'ampia classe di modelli statistici nei quali l'effetto delle variabili esplicative sulla risposta viene modellato attraverso la specificazione del predittore lineare, della funzione link e della funzione di varianza[5]. Nel modello lineare classico la funzione *link* è la

[4] Nel seguito, per non appesantire la notazione, si ometterà di indicare esplicitamente la dipendenza della probabilità di successo degli effetti casuali, scrivendo π_{ij} in luogo di π_{ij}^u.

[5] A differenza di quanto accadeva per il modello lineare (distribuzione Normale), in questo contesto la varianza è tipicamente funzione del valore atteso

funzione identità, nei GLM invece la funzione *link* può essere una qualunque funzione monotona e differenziabile. Attraverso la funzione *link* il predittore lineare viene messo in relazione (per mezzo di una funzione g) con il valore atteso riferito alla variabile dipendente, cioè con π. Varie sono le funzioni *link* comunemente usate e spesso la loro scelta è arbitraria. Solitamente, nelle applicazioni poca importanza è data alla selezione della funzione link e quasi sempre la scelta ricade sui link canonici, quanto alla funzione *link* del GLM, le tre scelte più comuni sono

- *probit*: $g(\pi) = \Phi^{-1}(\pi)$ dove è $\Phi(\cdot)$ la funzione di ripartizione della distribuzione normale standard;
- *logit*: $g(\pi) = \log(\frac{\pi}{1-\pi})$ la cui inversa $g^{-1}(x) = \frac{1}{1+\exp(-x)}$ è la funzione di ripartizione della distribuzione logistica standard;
- *complementary log-log*: $g(\pi) = \log[-\log(1-\pi)]$, la cui inversa $g^{-1}(x) = 1-\exp[-\exp(x)]$ è la funzione di ripartizione di una distribuzione di tipo "extreme-value".

Si ricorda che la distribuzione logistica standard ha media nulla, varianza $\frac{\pi^2}{3}$ e una forma molto simile a quella di una normale di pari varianza, rispetto alla quale, però, ha le code leggermente più "pesanti". Pertanto i risultati ottenibili con i link probit e logit sono praticamente identici, a meno che non si abbiano probabilità molto vicine a 0 oppure a 1. In effetti la scelta solitamente dipende dall'impostazione teorica che sottende il modello: il logit, essendo il link canonico, semplifica le proprietà del modello (McCullagh e Nelder, 1989) e, inoltre, ha il vantaggio di consentire un'interpretazione dei risultati in termini di *odds ratio* (Agresti, 1990); d'altra parte, come si vedrà di seguito, il probit rappresenta la scelta più naturale nel caso di un modello a soglia con variabile latente (Winship e Mare, 1983). Il link *complementary log-log* si distingue dagli altri due per la sua asimmetria e per la varianza della relativa distribuzione "extreme-value", che è pari a $\frac{\pi^2}{6}$. Questo link trova importanti applicazioni nei modelli per dati ordinali e per dati di sopravvivenza in tempo discreto.

3.2.1 Versione con variabile latente e soglia

I modelli a soglia sono interessanti, perchè consentono di derivare certe proprietà del modello non lineare da quelle del modello lineare; inoltre possono essere facilmente estesi al caso di variabili di risposta ordinali. Usando la notazione del GLM multilivello, un modello multilivello a soglia per dati binari viene costruito definendo una variabile latente (non osservabile) y^*_{ij} che segue un modello lineare a due livelli

$$y^*_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij}$$

dove i termini di errore di primo livello sono ipotizzati determinazioni di una normale, una logistica o una distribuzione di tipo "extreme-value"[6]. Anche se y^*_{ij} non è direttamente osservabile, si suppone poi che sia possibile sapere se la sua realizzazione supera oppure no un certo valore, detto soglia, che solitamente viene posto uguale a zero[7], e definiamo la variabile binaria osservabile y_{ij} come

$$Y_{ij} = I\{Y^*_{ij} > 0\}$$

dove $I\{\cdot\}$ è la funzione indicatrice che vale 1 quando l'evento in parentesi è vero. Pertanto, usando la notazione introdotta con la

$$\pi_{ij} = P\{Y_{ij} = 1|u_{0j}, u_{1j}\}.$$

si ottiene

$$\pi_{ij} = P\{Y^*_{ij} > 0|u_{0j}, u_{1j}\} = g^{-1}(\beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij}),$$

ovvero

$$g(\pi_{ij}) = \beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij}$$

dove, a seconda della distribuzione del termine di errore di primo livello, la funzione di *link* g è *probit*, *logit* oppure *complementary log-log*. Nel caso del modello probit è possibile calcolare in forma chiusa il valore atteso marginale della risposta. Infatti, poichè

$$Y^*_{ij} \sim N(\beta_{0j} + \beta_{1j}x_{ij}, 1 + (\sigma^2_{u0} + \sigma^2_{u1}x^2_{ij} + 2\sigma_{u01}x_{ij}))$$

si ottiene

$$E(Y_{ij}) = P\{Y^*_{ij} > 0\} = \Phi(\frac{\beta_{0j} + \beta_{1j}x_{ij}}{[1 + (\sigma^2_{u0} + \sigma^2_{u1}x^2_{ij} + 2\sigma_{u01}x_{ij})]^{\frac{1}{2}}})$$

da cui risulta evidente che i coefficienti del modello marginale sono attenuati rispetto a quelli del modello multilivello[8]. Un vantaggio della specificazione tramite variabile latente riguarda il calcolo del coefficiente di correlazione intraclasse. Infatti, con l'eccezione del modello lineare normale, in un GLM

[6] Tutte queste distribuzioni vengono assunte nella forma standard, cioè con la varianza fissata (1 nel caso della normale, $\frac{\pi^2}{3}$ nel caso della logistica, $\frac{\pi^2}{6}$ nel caso della "extreme-value"). La scelta arbitraria della varianza del termine di errore non causa una perdita di generalità poichè è una condizione necessaria per l'identificabilità del modello (Winship e Mare, 1983; Hedeker e Gibbons, 1994).

[7] Per l'identificabilità del modello è necessario porre un vincolo sulla soglia oppure sulla costante della variabile latente. L'opzione più comune è quella di porre a zero la soglia (Winship e Mare, 1983; Hedeker e Gibbons, 1994).

[8] Per il modello *logit* il calcolo del valore atteso marginale in forma chiusa non è possibile, ma Zeger et al. (1988) hanno derivato una formula approssimata che conferma il fenomeno dell'attenuazione.

multilivello il coefficiente di correlazione intraclasse non è definito nemmeno quando è presente un unico effetto casuale sull'intercetta, poichè, per la dipendenza dalla media, la varianza marginale della risposta non è costante. Tuttavia è possibile calcolare il coefficiente sulla variabile latente: infatti, se

$$y^*_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + e_{ij}$$

il coefficiente di correlazione intraclasse è

$$\frac{\sigma^2_{u0}}{\sigma^2_e + \sigma^2_{u0}}$$

dove σ^2_e è fissato a 1 nel modello *probit*, $\frac{\pi^2}{3}$ nel modello *logit* e $\frac{\pi^2}{6}$ nel modello *complementary log-log*.

3.3 Modelli per dati politomici

I dati si dicono politomici quando le risposte appartengono ad un insieme non ordinato di $m > 2$ categorie, ad esempio "bianco", "nero", "rosso". Per l'individuo i del gruppo j, la risposta politomica può essere rappresentata da un vettore multinomiale (codifica disgiuntiva completa):

$$\mathbf{y}'_{ij} = (y^{(1)}_{ij}, ..., y^{(m)}_{ij}),$$

dove $y^{(s)}_{ij} \in \{0, 1\}$ è una variabile di Bernoulli che vale 1 quando l'osservazione cade nella categoria s $(s = 1, ..., m)$. Poichè le categorie sono mutuamente esclusive, si ha $\sum_{s=1}^{m} y^{(s)}_{ij} = 1$. Analogamente al caso binario poniamo

$$\pi^{(s)}_{ij} = P\{Y^{(s)}_{ij} = 1 | u_{0j}, u_{1j}\}, \quad s = 1, ..., m$$

dove u_{0j} e u_{1j} sono, come al solito, gli effetti casuali relativi all'intercetta e al coefficiente angolare. Le probabilità

$$\pi^{(s)}_{ij} = P\{y^{(s)}_{ij} = 1 | u_{0j}, u_{1j}\}, \quad s = 1, ..., m$$

sono legate dal vincolo $\sum_{s=1}^{m} \pi^{(s)}_{ij} = 1$. Poichè le variabili che compongono il vettore $\mathbf{y}'_{ij}$ sono linearmente dipendenti, una di esse deve necessariamente essere esclusa. Convenzionalmente la variabile esclusa è $y^{(m)}_{ij}$, dove m è indicata come categoria di base e scelta arbitrariamente. In questo modo il vettore $(y^{(1)}_{ij}, ..., y^{(m-1)}_{ij})$ ha una distribuzione multinomiale con matrice di covarianza non singolare, i cui elementi sono

$$Var(Y^{(s)}_{ij}) = \pi^{(s)}_{ij}(1 - \pi^{(s)}_{ij}) \quad s = 1, ..., m - 1$$

$$Cov(Y_{ij}^{(s)}, Y_{ij}^{(r)}) = -\pi_{ij}^{(s)}\pi_{ij}^{(r)} \quad s \neq r.$$

Il modello più comune per l'analisi di dati politomici si basa sul *link logit multivariato* (Fahrmeir e Tutz, 1994). Nel caso di m categorie, si ottengono $m-1$ modelli logistici, ognuno dei quali confronta la probabilità di ognuna delle $m-1$ categorie scelte con quella di base (quella esclusa, per comodità la $m-esima$):

$$\log(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(m)}}) = \beta_{0j}^{(s)} + \beta_{1j}^{(s)}x_{ij} + u_{0j}^{(s)} + u_{1j}^{(s)}x_{ij} \quad s = 1, ..., m-1$$

Si noti che gli effetti fissi e gli effetti casuali sono specifici di ogni equazione e che i parametri casuali modellano non solo la variabilità degli effetti casuali all'interno delle singole equazioni, ma anche la variabilità degli effetti casuali appartenenti a equazioni diverse. Scrivendo il secondo membro delle equazioni

$$\log(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(m)}}) = \beta_{0j}^{(s)} + \beta_{1j}^{(s)}x_{ij} + u_{0j}^{(s)} + u_{1j}^{(s)}x_{ij} \quad s = 1, ..., m-1$$

come $\eta_{ij}^{(s)}$, le probabilità delle singole categorie sono date da (Fahrmeir e Tutz, 1994)

$$\pi_{ij}^{(s)} = \frac{\exp(\eta_{ij}^{(s)})}{1 + \sum_{h=1}^{m-1}\exp(\eta_{ij}^{(h)})}$$

$$\pi_{ij}^{(m)} = \frac{\exp(\eta_{ij}^{(s)})}{1 + \sum_{h=1}^{m-1}\exp(\eta_{ij}^{(h)})}$$

con $s = 1, ..., m-1$.

I parametri del modello logit multivariato possono essere facilmente interpretati osservando che, per due categorie arbitrarie diverse da quella di base, si ha

$$\begin{aligned}\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(r)}} &= \exp(\eta_{ij}^{(s)} - \eta_{ij}^{(r)}) = \\ &= \exp(\beta_{0j}^{(s)} - \beta_{0j}^{(r)})\exp((\beta_{1j}^{(s)} - \beta_{1j}^{(r)})x_{ij}) \times \\ &\quad \exp(u_{0j}^{(s)} - u_{0j}^{(r)})\exp((u_{1j}^{(s)} - u_{1j}^{(r)})x_{ij}).\end{aligned}$$

3.4 Modelli per dati ordinali

I dati ordinali sono caratterizzati dal fatto che le risposte appartengono ad un insieme ordinato di categorie, ad esempio "basso", "medio", "alto". Solitamente le categorie vengono contrassegnate con i numeri naturali; tuttavia la numerazione delle categorie è una semplice convenzione, da non confondere con l'assegnazione di punteggi da utilizzare nei modelli[9]. L'interesse per i dati ordinali deriva dalla loro ampia diffusione, considerando che possono rientrare in questa categoria anche i dati cosiddetti di sopravvivenza in tempo discreto. I dati ordinali possono essere rappresentati in modo del tutto analogo a quello dei dati politomici, cioè assumendo per ogni unità statistica un vettore multinomiale

$$y_{ij}^{'} = (y_{ij}^{(1)}, ..., y_{ij}^{(m)}),$$

dove $y_{ij}^{(s)} \in \{0, 1\}$ è una variabile di Bernoulli che vale 1 quando l'osservazione cade nella categoria s ($s = 1, ..., m$). Come nel caso politomico poniamo

$$\pi_{ij}^{(s)} = P\{Y_{ij}^{(s)} = 1 | u_{0j}, u_{1j}\} \quad s = 1, ..., m$$

con il vincolo $\sum_{s=1}^{m} \pi_{ij}^{(s)} = 1$.
L'ordinamento delle categorie, che differenzia i dati ordinali da quelli politomici, può essere tenuto in considerazione basando i modelli sulle variabili cumulate:

$$z_{ij}^{(s)} = \sum_{l=1}^{s} y_{ij}^{(l)} \quad s = 1, ..., m-1$$

l'ultima variabile cumulata, $z_{ij}^{(m)}$, è sempre uguale a 1 e quindi non viene presa in considerazione, non contenendo alcuna informazione.
Le variabili cumulate hanno valore atteso pari a

$$E(Z_{ij}^{(s)}) = \sum_{l=1}^{s} \pi_{ij}^{(l)} = \gamma_{ij}^{(s)} \quad s = 1, ..., m-1$$

dove $\gamma_{ij}^{(s)}$ è la probabilità che la risposta cada in una categoria non superiore a s. Inoltre, assumendo una distribuzione multinomiale delle risposte, le variabili cumulate hanno varianze e covarianze pari a:

$$Var(Z_{ij}^{(s)}) = \gamma_{ij}^{(s)}(1 - \gamma_{ij}^{(s)}) \quad s = 1, ..., m-1$$

$$Cov(Z_{ij}^{(s)}, Z_{ij}^{(r)}) = \gamma_{ij}^{(s)}(1 - \gamma_{ij}^{(r)}) \quad s \leq r$$

[9] Si eviterà di fare ricorso ai modelli basati sull'assegnazione di punteggi, che sono suscettibili di numerose critiche (cfr. Agresti, 1984)

I due modelli fondamentali per l'analisi dei dati ordinali, basati sulle probabilità cumulate, sono il modello con il *link logit*, detto *proportional odds*,

$$\log(\frac{\gamma_{ij}^{(s)}}{1-\gamma_{ij}^{(s)}}) = \beta_{0j}^{(s)} + \beta_{1j}x_{ij} + u_{0j} + u_{1jx_{ij}} \qquad s = 1, ..., m-1 \tag{3.1}$$

e il modello con il link *complementary log-log*, detto *proportional hazards*,

$$\log[-\log(1-\gamma_{ij}^{(s)})] = \beta_{0j}^{(s)} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij} \qquad s = 1, ..., m-1 \tag{3.2}$$

In entrambe le specificazioni il coefficiente β è identico per tutti gli s, mentre le intercette sono ordinate in modo non decrescente, $\beta_{0j}^{(1)} \leq \beta_{0j}^{(2)} \leq ... \leq \beta_{0j}^{(m-1)}$. Gli effetti casuali sono ipotizzati comuni a tutti gli s, anche se questa non e un'assunzione necessaria. Il modello (3.1) viene detto *proportional odds* perchè il rapporto fra gli *odds* di due individui non dipende da s:

$$\frac{\frac{\gamma_{ij}^{(s)}}{1-\gamma_{ij}^{(s)}}}{\frac{\gamma_{i'j'}^{(s)}}{1-\gamma_{i'j'}^{(s)}}} = \exp[\beta_{1j}(x_{ij} - x_{i'j'}) + (u_{0j} - u_{0j'}) + (u_{1j}x_{ij} - u_{1j'}x_{i'j'})].$$

Invece il modello (3.2) è noto come *proportional hazards* perché, pensando i dati ordinali come tempi di sopravvivenza, rappresenta una versione discreta del modello a rischi proporzionali di Cox che si ottiene raggruppando le osservazioni in intervalli. Questa versione discreta, dovuta a McCullagh (1980), si aggiunge alla versione discreta di Prentice e Gloeckler (1978).
Nel prossimo paragrafo si introdurrà l'analisi di sopravvivenza e si discuteranno brevemente le proprietà delle versioni discrete del modello di Cox. Per il momento ci si limita ad osservare che nel modello (3.2) la funzione di sopravvivenza discreta è data da

$$1 - \gamma_{ij}^{(s)} = \{\exp[-\exp(\beta_{0j}^{(s)})]\}^{\exp(\beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij})}$$

dove $[-\exp(\beta_{0j}^{(s)})]$ è la funzione di sopravvivenza di base, cioè relativa ad un individuo con covariate ed effetti casuali nulli. Inoltre il *rischio* o *hazard* al tempo s è dato da

$$\frac{\gamma_{ij}^{(s)} - \gamma_{ij}^{(s-1)}}{1 - \gamma_{ij}^{(s-1)}} = 1 - \exp\{[\exp(\beta_{0j}^{(s-1)}) - \exp(\beta_{0j}^{(s)})]\exp(\beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij})\}.$$

3.4.1 Versione con variabile latente e soglie

I modelli (3.1) e (3.2) possono essere derivati anche per mezzo di un modello a soglia, in modo del tutto analogo a quanto visto per dati binari. Data la variabile latente

$$y^*_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij}$$

e un insieme di valori di soglia $-\infty = \nu_0 \leq \nu_1 \leq \nu_2 \leq ... \leq \nu_{m-1} \leq \nu_m = \infty$, ponendo

$$y^{(s)}_{ij} = I\{\nu_{s-1} < y^*_{ij} \leq \nu_s\} \quad s = 1, ..., m$$

si ottiene

$$\begin{aligned}\pi^{(s)}_{ij} &= P\{\nu_{s-1} < y^*_{ij} \leq \nu_s | u_{0j}, u_{1j}\} = \\ &= P\{y^*_{ij} \leq \nu_s | u_{0j}, u_{1j}\} - P\{y^*_{ij} \leq \nu_{s-1} | u_{0j}, u_{1j}\} = \\ &= g^{-1}(\nu_s - (\beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij})) + \\ &\quad -g^{-1}(\nu_{s-1} - (\beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij}))\end{aligned}$$

dove, a seconda della distribuzione del termine di errore di primo livello, il link g è il *logit* oppure il *complementary log-log*. In termini di probabilità cumulate ciò equivale a

$$\gamma^{(s)}_{ij} = g^{-1}(\nu_s - (\beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij})) \quad s = 1, ..., m-1$$

ovvero

$$g(\gamma^{(s)}_{ij}) = \nu_s - (\beta_{0j} + \beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij}) \quad s = 1, ..., m-1 \qquad (3.3)$$

Il modello (3.3), a parte una differenza nella parametrizzazione, è lo stesso delle equazioni (3.1) e (3.2), cioè è il modello di *proportional odds* se g è il link *logit* ed è il modello *proportional hazards* se g è il link *complementary log-log*. A proposito di parametrizzazioni si osserva che:

1. Il modello (3.3) necessita di un vincolo, poichè ha un parametro di troppo per l'intercetta. Nei modelli a soglia il vincolo di solito è $\nu_1 = 0$ (Hedeker e Gibbons, 1994), altrimenti si può porre $\beta_{0j} = 0$, ottenendo così la stessa parametrizzazione dei modelli (3.1) e (3.2).
2. Il coefficiente β_{1j} e gli effetti casuali hanno il segno invertito rispetto alle equazioni (3.1) e (3.2). Infatti, una covariata che ha un effetto positivo sulla variabile latente (nel senso che un incremento di X determina un

incremento di $E(Y^*)$) ha, allo stesso tempo, un effetto negativo sulle probabilità cumulate, come risulta evidente dalla definizione del modello a soglia.

3.5 Modelli per dati di sopravvivenza in tempo discreto

L'analisi statistica dei tempi di sopravvivenza è stata oggetto, negli ultimi anni, di un rinnovato interesse da parte di vari settori della ricerca. L'elemento essenziale per cui un problema si colloca nell'ambito di tale analisi è costituito dalla presenza di una variabile aleatoria a valori non negativi che descrive il tempo di accadimento di uno o più eventi di interesse. E' proprio la genericità estrema con cui tale evento è definito che consente di includere situazioni della più diversa natura, interessando vari campi di studio. Si presentano ora i principali modelli dell'analisi di sopravvivenza in tempo discreto, esaminandone poi l'estensione al caso multilivello.
I dati di sopravvivenza scaturiscono da indagini di tipo longitudinale, finalizzate all'osservazione del tempo intercorrente fra due eventi, il secondo dei quali viene convenzionalmente chiamato morte[10]. L'oggetto di interesse è dunque il tempo di attesa T, che, a seconda dei casi, si ipotizza essere una variabile aleatoria continua o discreta. Tuttavia l'osservazione di T per tutti gli individui del campione è generalmente impossibile, sia perché ciò può richiedere un tempo di osservazione estremamente lungo e non pianificabile, sia perché alcuni individui vengono osservati prima di aver sperimentato l'evento di interesse. Questo fenomeno, tipico dei dati di sopravvivenza, è noto con il termine di censura a destra[11]. Pertanto i dati di sopravvivenza sono solitamente costituiti da coppie di variabili aleatorie (X, δ), dove X è il tempo osservato e δ è un indicatore che vale 1 se l'osservazione si è conclusa con l'evento di interesse e 0 se si è conclusa con la censura. I metodi dell'analisi di sopravvivenza fanno inferenza su T a partire dalle osservazioni su (X, δ). Essi sono usualmente basati sulla funzione di sopravvivenza $S(t)$ e sulla funzione di rischio o *hazard*, che, nel caso di una variabile aleatoria continua T, sono definite rispettivamente da

[10] I termini sopravvivenza e morte traggono origine dalle indagini demografiche e mediche in cui l'evento finale è la morte della persona (l'evento iniziale può essere la nascita, la diagnosi di una certa malattia ecc.). Naturalmente l'evento finale può essere di qualunque tipo e può avere connotati positivi (ad esempio, trovare lavoro), ma la terminologia in uso è quella relativa alle indagini sulla sopravvivenza.

[11] Nei dati di sopravvivenza esiste un'ampia casistica di osservazioni incomplete, di cui la censura a destra rappresenta il caso di gran lunga più frequente (Kalbfleisch e Prentice, 1980). Un'ipotesi fondamentale che sta alla base dell'analisi di sopravvivenza è che il meccanismo di censura sia indipendente dal processo che governa il succedersi degli eventi: questa è una condizione necessaria per poter riferire le conclusioni dell'analisi al tempo sottostante T, che non è direttamente osservabile per tutti gli individui (Kalbfleish e Prentice, 1980).

$$S(t) = P(T > t)$$

$$\lambda(t) = lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Dunque $S(t)$ è la probabilità di sopravvivere oltre il tempo t, mentre $\lambda(t)$, che assume valori nell'intervallo $[0, \infty)$, è il rischio istantaneo di morte al tempo t per un individuo sopravvissuto fino a quell'istante. Le funzioni di sopravvivenza e di rischio non sono altro che modi alternativi di caratterizzare la distribuzione di T, e risultano utili per la definizione dei modelli e l'interpretazione dei risultati. Si può passare da una funzione all'altra per mezzo delle seguenti relazioni (Kalbfleisch e Prentice, 1980):

$$\lambda(t) = -\frac{\partial}{\partial t} \log S(t)$$

$$S(t) = \exp\{-\int_0^t \lambda(s) ds\}$$

Se la variabile aleatoria T è discreta, la definizione della funzione di sopravvivenza rimane invariata, mentre quella della funzione di rischio diviene

$$S(t) = \prod_{s=1}^{t} (1 - \lambda(s))$$

Inoltre, mentre in tempo continuo $P(T = t) = 0$, in tempo discreto si ha

$$P(T = t) = \lambda(t) S(t-1) = \lambda(t) \prod_{s=1}^{t-1} (1 - \lambda(s))$$

3.5.1 Alcuni modelli classici

Consideriamo innanzitutto il modello a rischi proporzionali di Cox (Cox, 1972), ovvero il modello in tempo continuo più ampiamente usato, che rappresenta un punto di partenza per gli sviluppi di altri modelli in tempo discreto. Dato un campione casuale di individui $i = 1, ..., n$, il modello di Cox si basa sulla seguente specificazione della funzione di rischio:

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t) \exp(\mathbf{x}_i' \beta_\mathbf{1}) \quad t \in [0, \infty)$$

dove $\mathbf{x}_i$ è un vettore di covariate fisse[12] per l'individuo i, β_1 è un vettore di parametri e $\lambda_0(\cdot)$ è una funzione non specificata, detta *funzione di rischio di*

[12] Nell'analisi di sopravvivenza il termine *covariata fissa* viene usato in contrapposizione al termine *covariata tempo-dipendente*, per indicare che la covariata assume,

base, che rappresenta l'andamento del rischio per un individuo con $\mathbf{x}_i = \mathbf{0}$. Il modello è semiparametrico poichè, nonostante la presenza di un vettore di parametri, la distribuzione del tempo T non è completamente specificata[13]: proprio questa è la maggiore virtù del modello di Cox, che permette di studiare il rischio relativo fra gli individui senza bisogno di fare troppe ipotesi sulla distribuzione di T. Il modello viene detto a rischi proporzionali perchè fra le funzioni di rischio di due generici individui esiste un rapporto di proporzionalità costante nel tempo:

$$\frac{\lambda(t|\mathbf{x}_i)}{\lambda(t|\mathbf{x}_{i'})} = \exp[(\mathbf{x}_i - \mathbf{x}_{i'})'\beta_1] \quad \textit{per ogni } t\ [0, \infty).$$

Il modello di Cox viene usualmente specificato per mezzo della funzione di rischio, ma, data la relazione

$$S(t) = \exp\{-\int_0^t \lambda(s)ds\}$$

può anche essere visto in termini della funzione di sopravvivenza:

$$S(t|\mathbf{x}_i) = S_0(t)^{\exp(\mathbf{x}_i'\beta_1)}$$

dove $S_0(t)$ è la funzione di sopravvivenza di base, cioè per un individuo con $\mathbf{x}_i = \mathbf{0}$. Inoltre il modello di Cox può essere esteso al caso di covariate tempo-dipendenti: in tal caso però, oltre ad avere problemi teorici e computazionali, si perde la proprietà di rischi proporzionali (Kalbfleisch e Prentice, 1980).

Mentre il tempo di per sé è continuo, la sua misurazione avviene necessariamente ad intervalli discreti. Quando tali intervalli non sono abbastanza piccoli l'uso dei modelli in tempo continuo presenta dei seri problemi dovuti alla presenza di ties, cioè individui che sperimentano l'evento di interesse nello stesso intervallo di misura[14]. Inoltre ci sono delle situazioni in cui l'evento di interesse può verificarsi solo in determinati momenti, per cui il tempo deve considerarsi discreto (si pensi a indagini riguardanti le dichiarazioni dei redditi). Pertanto in molte applicazioni è opportuno, se non necessario, ricorrere a modelli in tempo discreto. I modelli di sopravvivenza in tempo discreto possono essere definiti seguendo due diversi approcci (Allison, 1982): il primo consiste nel trattare il tempo come se fosse effettivamente discreto (Myers et al., 1973), mentre il secondo assume l'esistenza di un modello sottostante in tempo continuo con osservazioni raccolte in determinati intervalli temporali (Holford, 1976). Nell'ambito del primo approccio il modello più largamente

---

per ogni individuo, un unico valore che non cambia nel tempo.

[13] La distribuzione del tempo T può essere completamente specificata solo specificando la funzione di rischio di base $\lambda_0(\cdot)$: ad esempio, se tale funzione è costante, T ha una distribuzione esponenziale.

[14] In un modello in tempo continuo una parità è un evento di probabilità nulla e quindi la presenza di molte parità rende il modello inadeguato.

usato si basa sul link *logit*. Dato un campione casuale di individui $i = 1, ..., n$ e indicando con k l'ultimo tempo osservato nel campione si ha (Cox, 1972)

$$\log[\frac{\lambda(t|\mathbf{x}_{it})}{1-\lambda(t|\mathbf{x}_{it})}] = \beta_{0t} + \mathbf{x}'_{it}\beta_1 \quad t = 1, ..., k,$$

dove $\mathbf{x}_{it}$ è un vettore di covariate per l'individuo i al tempo t. I parametri $(\beta_{01}, ..., \beta_{0k})$ modellano il rischio di base, svolgendo un ruolo analogo a quello della funzione $\lambda_0(\cdot)$ nel modello di Cox, mentre i parametri del vettore β_1 misurano l'effetto delle covariate sul logit del rischio. Per due individui i e i' si ha

$$\frac{\frac{\lambda(t|\mathbf{x}_{it})}{1-\lambda(t|\mathbf{x}_{it})}}{\frac{\lambda(t|\mathbf{x}_{i't})}{1-\lambda(t|\mathbf{x}_{i't})}} = \exp[(\mathbf{x}_{it} - \mathbf{x}_{i't})'\beta_1]$$

per cui, se le covariate sono fisse (cioè $\mathbf{x}_{it} = \mathbf{x}_i$ per ogni t), gli *odds* dei rischi sono proporzionali. Se invece si assume che i dati siano generati da un modello di Cox, il corrispondente modello in tempo discreto per osservazioni raggruppate in intervalli si basa sul link *complementary log-log* (Prentice e Gloeckler, 1978):

$$\log[-\log(1 - \lambda(t|\mathbf{x}_{it}))] = \beta_{0t} + \mathbf{x}'_{it}\beta_1 \quad t = 1, ..., k$$

dove il vettore di parametri β è identico a quello del modello di Cox sottostante[15]. Si noti che nella versione discreta la funzione di sopravvivenza ha la stessa specificazione del modello di Cox: infatti dalla

$$S(t) = \prod_{s=1}^{t}(1 - \lambda(s))$$

si ottiene

$$S(t) = \exp[-\sum_{s=1}^{t}\exp(\beta_{0s})\exp(\mathbf{x}'_{is}\beta_1)]$$

che, nel caso di covariate fisse, fornisce:

$$S(t|\mathbf{x}_i) = S_0(t)^{\exp(\mathbf{x}'_i\beta_1)}$$

[15] Per questo motivo nel modello con link *complementary log-log*, a differenza del modello con link *logit*, il vettore di parametri β è invariante rispetto alla suddivisione del tempo in intervalli (Allison, 1982). Comunque, ai fini pratici, la differenza fra i due modelli è spesso irrilevante e si annulla quando la partizione del tempo in intervalli è molto fine, poichè il modello con link *logit* converge al modello di Cox al tendere a zero dell'ampiezza degli intervalli (Thompson, 1977).

dove $S_0(t) = \exp[-\sum_{s=1}^{t} \exp(\beta_s)]$ è la funzione di sopravvivenza di base. Si noti, però, che nella versione discreta i rischi non sono proporzionali[16]. La specificazione

$$S(t|\mathbf{x}_i) = S_0(t)^{\exp(\mathbf{x}_i'\beta_1)}$$

della funzione di sopravvivenza caratterizza anche la versione discreta proposta da McCullagh (1980) che si è introdotta nel paragrafo precedente direttamente per il caso multilivello [17]. In entrambe le versioni discrete il vettore di parametri β è identico a quello del modello di Cox sottostante (in sostanza le due versioni discrete differiscono solo per la parametrizzazione della funzione di sopravvivenza di base: cfr. Laara e Matthews, 1985). A differenza dei modelli in tempo continuo, nei modelli in tempo discreto l'inclusione di covariate tempo-dipendenti è del tutto naturale. Ciò risulta utile anche quando i dati contengono esclusivamente covariate fisse: infatti, il modo più semplice per consentire ad una covariata fissa di avere un effetto variabile nel tempo è quello di costruire una covariata tempo-dipendente fittizia definita dall'interazione fra il tempo e la covariata fissa di interesse. Sia nel modello *logit* che in quello *complementary log-log* si possono imporre delle restrizioni sui parametri β_t che risultano particolarmente utili quando l'insieme di tali parametri sia molto numeroso. Ad esempio, Mantel e Hankey (1978) propongono una specificazione attraverso un polinomio in t:

$$\beta_{0t} = \sum_{r=0}^{R} \delta_r t^r$$

per cui i parametri $(\beta_{01}, ..., \beta_{0k})$ vengono sostituiti dai parametri $(\delta_0, ..., \delta_R)$.

3.5.2 Rappresentazione per mezzo di variabili indicatrici

Prima di passare alla versione multilivello di questi modelli è opportuno esaminare brevemente una rappresentazione alternativa dei dati di sopravvivenza, utile a fini computazionali, che verrà poi estesa alla versione multilivello. Per cominciare osserviamo che, per un qualsiasi modello di sopravvivenza in tempo discreto, la verosimiglianza è

$$L = \prod_{i=1}^{n} \{[P(T_i = t_i)]^{\delta_i} [P(T_i > t_i)]^{1-\delta_i}\}$$

[16] In tempo discreto la proporzionalità dei rischi è resa impossibile dal fatto che i rischi sono compresi nell'intervallo $[0, 1]$. Quello che si può imporre è la proporzionalità degli *odds* dei rischi, come accade nel modello con il link *logit*.

[17] vedi formula $1 - \gamma_{ij}^{s} = \{\exp[\exp(\beta_{0j}^{(s)})]\}^{\exp(\beta_{1j}x_{ij} + u_{0j} + u_{1j}x_{ij})}$

dove t_i è il tempo osservato per l'individuo i e δ_i è l'indicatore di non censura[18]. Pertanto, in base alla

$$S(t) = \prod_{s=1}^{t}(1 - \lambda(s))$$

e alla

$$P(T = t) = \lambda(t)S(t-1) = \lambda(t)\prod_{s=1}^{t-1}(1 - \lambda(s))$$

la verosimiglianza può essere scritta in termini di *hazard* nel seguente modo:

$$L = \prod_{i=1}^{n}\{[\lambda(t_i|\mathbf{x}_{it_i})]^{\delta_i}[1 - \lambda(t_i|\mathbf{x}_i t_i)]^{1-\delta_i}\prod_{s=1}^{t_i-1}[1 - \lambda(s|\mathbf{x}_{is})]\} \tag{3.4}$$

Adesso, seguendo un'idea di Brown (1975), si definisce un'insieme di variabili *dummy* y_{is} tali che $y_{is} = 1$ se e solo se l'individuo i sperimenta l'evento di interesse al tempo s $(s = 1, 2, ..., t_i)$. In questo modo, per ogni individuo del campione, la coppia (t_i, δ_i) viene sostituita da un vettore $(y_{i1}, y_{i2}, ..., y_{it_i})$ che assume i valori $(0, 0, ..., 0, 1)$ se $\delta = 1$ oppure $(0, 0, ..., 0, 0)$ se $\delta = 0$. Pertanto la verosimiglianza (3.4) può riscriversi come

$$L = \prod_{i=1}^{n}\prod_{s=1}^{t_i}[\lambda(s|\mathbf{x}_{is})]^{y_{is}}[1 - \lambda(s|\mathbf{x}_{is})]^{1-\mathbf{y}_{is}} \tag{3.5}$$

La (3.7) non è altro che la verosimiglianza di un campione casuale

$$\{y_{is} : i = 1, ..., n; s = 1, 2, ..., t_i\} \tag{3.6}$$

di variabili casuali Bernoulli con probabilità di successo

$$P(y_{is}|\mathbf{x}_{is}) = \lambda(s|\mathbf{x}_{is}). \tag{3.7}$$

Pertanto i modelli di sopravvivenza in tempo discreto possono essere visti come modelli di regressione per dati binari applicati ad un campione esteso che si ottiene sostituendo ad ogni record i contributi relativi alle singole unità temporali in cui l'individuo è stato osservato. Pertanto la stima dei parametri dei modelli

$$\log[\frac{\lambda(t|\mathbf{x}_{it})}{1 - \lambda(t|\mathbf{x}_{it})}] = \alpha_t + \mathbf{x}'_{it}\beta \quad t = 1, ..., k,$$

e

[18] In $L = \prod_{i=1}^{n}\{[P(T_i = t_i)]^{\delta_i}[P(T_i > t_i)]^{1-\delta_i}\}$ si assume che un individuo censurato venga osservato fino all'unità temporale t_i inclusa. Nel caso di dati osservati ad intervalli ciò equivale ad assumere che la censura intervenga al termine dell'intervallo. Sulle implicazioni di questa assunzione cfr. Allison (1982), p.71.

$$\log[-\log(1-\lambda(t|\mathbf{x}_{it}))] = \alpha_t + \mathbf{x}'_{it}\beta \quad t = 1, ..., k$$

può essere effettuata applicando al campione opportunamente esteso la procedura di stima relativa ad un modello di regressione per dati binari con link *logit* o *complementary log-log*[19]. Come risulta evidente dalla (3.7) le osservazioni del campione esteso sono statisticamente indipendenti, anche quelle che si riferiscono a tempi diversi di uno stesso individuo. Questa indipendenza può apparire strana, ma è necessaria per garantire l'equivalenza dei modelli. In effetti abbandonare l'ipotesi di indipendenza per le osservazioni relative ad un individuo significa ammettere la presenza di *eterogeneità non osservabile* (Allison, 1982, p. 82).

3.5.3 Versione multilivello

I modelli di sopravvivenza in tempo discreto presentati possono essere estesi al caso multilivello. Si consideri una struttura gerarchica a due livelli, indicando con i deponenti ij l'individuo i del gruppo j ($i = 1, ..., n_j; j = 1, ..., J$). Una semplice versione multilivello dei modelli

$$\log[\frac{\lambda(t|\mathbf{x}_{it})}{1-\lambda(t|\mathbf{x}_{it})}] = \beta_{0t} + \mathbf{x}'_{it}\beta_1 \quad t = 1, ..., k, \tag{3.8}$$

e

$$\log[-\log(1-\lambda(t|\mathbf{x}_{it}))] = \beta_{0t} + \mathbf{x}'_{it}\beta_1 \quad t = 1, ..., k \tag{3.9}$$

è data da

$$g(\lambda(t|\mathbf{x}_{ijt}, u_{0j})) = \beta_{0t} + x'_{ijt}\beta_1 + u_{0j} \quad t = 1, ..., k \tag{3.10}$$

dove $g(\cdot)$ è la funzione *logit* o *complementary log-log*, mentre $u_{0j} \overset{iid}{\sim} N(0, \sigma^2_{u0})$. In questo modello l'effetto casuale provoca una traslazione della funzione di rischio di base nella scala indotta dalla trasformazione $g(\cdot)$. Il modello (3.10) gode delle proprietà dei modelli (3.8) e (3.9) condizionatamente agli effetti casuali, ma non marginalmente (si ricordi la distinzione fra modelli *unit-specific* e *polulation-average* delineata in precedenza). Ad esempio, nel caso di link *logit* e covariate fisse si ha

$$\frac{\frac{\lambda(t|\mathbf{x}_{ij}, u_{0j})}{1-\lambda(t|\mathbf{x}_{ij}, u_{0j})}}{\frac{\lambda(t|\mathbf{x}_{i'j'}, u_{0j'})}{1-\lambda(t|\mathbf{x}_{i'j'}, u_{0j'})}} = \exp[(\mathbf{x}_{ij} - \mathbf{x}_{i'j'})\beta_1 + (u_{0j} - u_{0j'})]$$

[19] Ovviamente il limite di questa strategia sta nella numerosità del campione esteso, che può essere enorme nel caso che gli individui vengano osservati per una lunga sequenza di tempi. D'altra parte, quando il tempo sottostante è continuo e gli intervalli di osservazione sono sufficientemente piccoli si possono usare direttamente i modelli in tempo continuo.

per cui, condizionatamente agli effetti casuali, gli *odds* dei rischi sono proporzionali[20]. Il modello (3.10) può includere anche dei coefficienti casuali per modellare un effetto delle covariate differenziato nei gruppi. Ad esempio, se è presente una sola covariata ed il suo coefficiente è casuale, il modello (3.10) diviene

$$g(\lambda(t|x_{ijt}, u_{0j}, u_{1j})) = \beta_{0t} + \beta_1 x_{ijt} + u_{0j} + u_{1j} x_{ijt} \quad t = 1, ..., k \qquad (3.11)$$

con

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \overset{iid}{\sim} N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix})$$

Le considerazioni sull'equivalenza fra un modello di sopravvivenza in tempo discreto e un modello per dati binari applicato ad un campione opportunamente esteso possono essere ripetute anche nel caso multilivello, trattando tutte le variabili condizionatamente agli effetti casuali. Pertanto un modello come il (3.10) o il (3.11) è equivalente ad un modello multilivello per dati binari (con link *logit* o *complementary log-log*) applicato al campione esteso che si ottiene sostituendo ad ogni record i contributi relativi alle singole unità temporali in cui l'individuo è stato osservato. L'unica accortezza riguarda la definizione della struttura gerarchica. Infatti il campione esteso ha, formalmente, una struttura a tre livelli, in cui le osservazioni sulle singole unità temporali di un individuo costituiscono il primo livello, gli individui il secondo livello e i gruppi il terzo livello. Tuttavia l'equivalenza di cui si discute richiede l'eliminazione del secondo livello, quello relativo agli individui, poichè si assume che tutte le osservazioni sulle singole unità temporali, anche quelle riferite ad uno stesso individuo, siano indipendenti condizionatamente agli effetti casuali relativi al gruppo. Pertanto nella specificazione del modello la variabilità di secondo livello deve essere vincolata a zero, oppure la struttura gerarchica deve essere ridotta a due livelli accorpando il primo e il secondo livello. Per analogia a quanto discusso nel caso dei modelli ordinari, l'inclusione di effetti casuali a livello di individuo comporterebbe una eterogeneità non osservabile, nel senso che due individui con le stesse covariate ed appartenenti allo stesso gruppo avrebbero rischi diversi.

3.6 Stima

La stima dei parametri dei GLM multilivello è un problema piuttosto complesso, innanzitutto perché la verosimiglianza marginale in generale non è

[20] In questo caso, incidentalmente, anche gli *odds* marginali sono proporzionali, poiché in un modello *logit* a intercetta casuale i parametri dei modelli marginale e condizionato sono legati da un fattore di proporzionalità (Zeger et al., 1988).

esprimibile in forma chiusa. Infatti tali modelli sono definiti condizionatamente agli effetti casuali e quindi è immediatamente determinabile solo la verosimiglianza condizionata, dalla quale si ottiene poi quella marginale integrando rispetto alla distribuzione degli effetti casuali: in generale questa operazione di integrazione non è fattibile per via analitica, salvo il caso di distribuzioni coniugate (Lee e Nelder, 1996). Il caso più importante di coniugatezza si ha quando sia la distribuzione condizionata della risposta che la distribuzione degli effetti casuali sono normali, cosa che accade nel modello multilivello lineare. Negli altri casi per sfruttare la coniugatezza occorre di volta in volta assumere per gli effetti casuali una distribuzione coniugata con quella della risposta, il che spesso costituisce un vincolo inaccettabile (Longford, 1996). I numerosi metodi proposti in letteratura rappresentano possibili soluzioni al problema della stima in presenza di una verosimiglianza non esprimibile in forma chiusa. Fra i principali metodi ricordiamo:

- Massima verosimiglianza marginale con integrazione numerica di Gauss-Hermite (Anderson e Aitkin, 1985; Hedeker e Gibbons, 1994; Rampichini, 1994);
- Metodi di quasi-verosimiglianza, fra cui Quasi-Verosimiglianza Marginale (Marginal Quasi-Likelihood, MQL) e Quasi-Verosimiglianza Penalizzata (Penalized Quasi-Likelihood, PQL) (Goldstein, 1991; Breslow e Clayton, 1993; Goldstein e Rasbash, 1996);
- Metodi basati sulle equazioni di stima generalizzate (Liang e Zeger, 1986);
- Metodi bayesiani basati su simulazioni di tipo Markov Chain Monte Carlo (MCMC) (Zeger e Karim, 1991);
- Metodi classici basati su simulazioni (Mealli e Rampichini, 1999; Calzolari et al., 1999).

Descriveremo in dettaglio i seguenti metodi:

1. Massima verosimiglianza marginale con integrazione numerica di Gauss-Hermite, implementato nei programmi MIXOR (Hedeker e Gibbons, 1996) e MIXNO (Hedeker, 1998);
2. MQL e PQL, implementati nel programma MLwiN (Goldstein et al., 1998).

3.6.1 Massima verosimiglianza marginale con integrazione numerica di Gauss-Hermite

Consideriamo un modello multilivello per dati ordinali nella specificazione basata su variabile latente e soglie. Questo tipo di modello consente di illustrare in modo naturale il metodo di stima, includendo inoltre come caso particolare i modelli a risposta binaria. Successivamente, si farà cenno all'estensione ai modelli a risposta politomica.
Per semplificare la notazione si scrive la varibile latente come

$$y_{ij}^{*} = z_{ij} + e_{ij}$$

dove

$$z_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij}.$$

Assumiamo m categorie e un insieme di valori di soglia

$$-\infty = \nu_0 \leq \nu_1 \leq \nu_2 \leq ... \leq \nu_{m-1} \leq \nu_m = \infty$$

in cui la soglia ν_1 viene posta uguale a 0 per motivi di identificabilità. Pertanto, la probabilità di una generica categoria s è

$$\pi_{ij}^{(s)} = g^{-1}(\nu_s - z_{ij}) - g^{-1}(\nu_{s-1} - z_{ij})$$

dove g è la funzione link (probit, logit o complementary log-log). Raccogliendo i parametri liberi di soglia nel vettore $\nu' = (\nu_2, ..., \nu_{m-1})$ e ponendo

$$\mathbf{y}_j' = (y_{1j}, ..., y_{n_j j}), \quad u_j' = (u_{0j}, u_{1j}),$$

la verosimiglianza condizionata agli effetti casuali può scriversi, relativamente al $j - mo$ gruppo, nel seguente modo:

$$L(\beta_0, \beta_1, \sigma_{u0}^2, \sigma_{u1}^2, \sigma_{u01}, \nu | \mathbf{y}_j, \mathbf{u}_j) = \prod_{i=1}^{n_j} \prod_{s=1}^{m} [g^{-1}(\nu_s - z_{ij}) - g^{-1}(\nu_{s-1} - z_{ij})]^{d_{ijs}}$$

dove d_{ijs} è un indicatore che vale 1 se e solo se $y_{ij} = s$. Adesso si considera la scomposizione di Cholesky della matrice di covarianza degli effetti casuali, indicando con Ψ quella matrice sottotriangolare tale che

$$\mathbf{\Psi}\mathbf{\Psi}' = \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix}$$

Ciò consente di riparametrizzare il modello, sostituendo $\mathbf{u}_j$ con $\mathbf{\Psi}\mathbf{w}_j$. Assumendo per $\mathbf{u}_j$ una distribuzione normale multivariata, segue che $\mathbf{w}_j$ ha una distribuzione normale multivariata standard[21]. Perciò i tre parametri casuali $(\sigma_{u0}^2, \sigma_{u1}^2, \sigma)$ vengono sostituiti dai tre parametri di $\mathbf{\Psi}$, che indichiamo con il vettore ψ.

Tenuto conto della riparametrizzazione, la verosimiglianza marginale del $j - mo$ gruppo è data da

[21] Questa riparametrizzazione è utile per l'implementazione dell'integrazione numerica. Tuttavia presenta anche il vantaggio di consentire una stima più stabile dei parametri casuali nel caso che questi siano prossimi a zero (infatti la scomposizione di Cholesky è una versione matriciale della radice quadrata).

$$L(\beta_0, \beta_1, \psi, \nu|\mathbf{y}_j) = \int_{-\infty}^{+\infty} L(\beta_0, \beta_1, \psi, \nu|\mathbf{y}_j, \mathbf{w})\phi(\mathbf{w})d\mathbf{w} \tag{3.12}$$

dove $\phi(\cdot)$ denota la densità della distribuzione normale multivariata standard. Poichè le osservazioni relative a gruppi diversi sono marginalmente indipendenti, la verosimiglianza complessiva è data dal prodotto di verosimiglianze come la (3.12) e quindi il suo logaritmo naturale è esprimibile come una somma di contributi, uno per ogni gruppo:

$$\log L = \sum_{j=1}^{J} \log L(\beta_0, \beta_1, \psi, \nu|\mathbf{y}_j) \tag{3.13}$$

dove $L = L(\beta_0, \beta_1, \psi, \nu|\mathbf{y}_1, ..., \mathbf{y}_J)$. La log-verosimiglianza marginale (3.13) può essere massimizzata con il classico algoritmo *Fisher Scoring*: indicando con θ il vettore di tutti i parametri del modello e con θ_t il valore che esso assume alla $t - ma$ iterazione, si ha

$$\theta_{t+1} = \theta_t + [E(-\frac{\partial^2 \log L}{\partial\theta\partial\theta'}|_{\theta=\theta_t})]^{-1}[\frac{\partial \log L}{\partial\theta|_{\theta=\theta_t}}]$$

dove la matrice di informazione attesa è data da (Hedeker e Gibbons, 1994)

$$E(-\frac{\partial^2 \log L}{\partial\theta\partial\theta'}) = \sum_{j=1}^{J}(L(\theta|\mathbf{y}_j))^{-2}(\frac{\partial L(\theta|\mathbf{y}_j)}{\partial\theta})(\frac{\partial L(\theta|\mathbf{y}_j)}{\partial\theta})'$$

Le espressioni delle derivate di $\log L$ rispetto ai vari tipi di parametro sono riportate in Hedeker e Gibbons (1994). Ognuna di queste derivate include un integrale rispetto alla densità della distribuzione normale multivariata standard. Poichè l'algoritmo *Fisher Scoring* prevede il calcolo, ad ogni iterazione, del valore delle derivate di $\log L$ nel punto θ_t, si rende necessario approssimare in qualche modo gli integrali presenti, poichè tali integrali non sono risolvibili per via analitica. Una soluzione semplice ed efficace consiste nell'integrazione numerica secondo il metodo di quadratura di Gauss-Hermite che consiste nell'approssimare l'integrale con la somma ponderata dei valori della funzione integranda calcolati in una serie di punti, detti punti di quadratura. Nel caso di un integrale ad una dimensione si ha

$$\int_{-\infty}^{+\infty} f(s)\phi(s)ds \simeq \sum_{q=1}^{Q} f(x_p)p_{x_q}$$

dove $f(\cdot)$ è una funzione generica, $\phi(\cdot)$ è la densità della distribuzione normale univariata standard, Q è il numero di punti di quadratura e $(x_q, p_{x_q}) : q = 1, ..., Q$ sono, rispettivamente, i punti di quadratura e i pesi associati, che vengono scelti in base a criteri di ottimalità (Stroud e Sechrest, 1966). La scelta fondamentale riguarda il valore di Q, al crescere del quale aumenta la bontà dell'approssimazione, ma anche la mole di calcoli. In genere valori compresi

tra 5 e 10 rappresentano un buon compromesso. Nel caso di un integrale a r dimensioni ogni punto di quadratura diviene un vettore $r - dimensionale$

$$\mathbf{x}_q = (x_{q1}, ..., x_{qr})$$

il cui peso (scalare) associato è dato dal prodotto dei corrispondenti pesi univariati:

$$p_{x_q} = \prod_{h=1}^{r} p_{x_{qh}}$$

Poichè i punti di quadratura $r - dimensionali$ si ottengono incrociando in tutti i modi possibili i punti unidimensionali, si ha un totale di Q^r punti. Ciò costituisce il limite di questa procedura, perchè al crescere di r la mole di calcoli diviene presto insostenibile, anche se Hedeker e Gibbons (1994) suggeriscono che al crescere di r si può comunque ridurre Q senza incidere troppo sulla bontà dell'approssimazione[22]. La quadratura di Gauss-Hermite viene usata per approssimare gli integrali che compaiono nelle espressioni delle derivate della log-verosimiglianza che servono per implementare l'algoritmo Fisher Scoring. In tal caso la dimensione degli integrali è pari al numero di effetti casuali presenti nel modello, per cui la quantità di calcoli rimane accettabile solo per modelli relativamente semplici. La quadratura consente inoltre di approssimare la log-verosimiglianza marginale: infatti dalle (3.12) e (3.13) si ottiene[23]

$$\begin{aligned} \log L &= \sum_{j=1}^{J} \log L(\beta_0, \beta_1, \psi, \nu | \mathbf{y}_j) \\ &= \sum_{j=1}^{J} \log \sum_{q=1}^{Q^r} L(\beta_0, \beta_1, \psi, \nu | \mathbf{y}_j, \mathbf{x}_q) p_{x_q} \end{aligned}$$

La log-verosimiglianza così calcolata può essere usata, nel modo convenzionale, per il test χ^2 del rapporto di verosimiglianza (Hedeker e Gibbons, 1994)[24]. Ciò costituisce un vantaggio del presente metodo di stima rispetto ai metodi di quasi-verosimiglianza di cui si parlerà a breve, per i quali

[22] Ad esempio, in un'applicazione con $r = 5$ può talora essere sufficiente usare $Q = 3$, limitando così il numero totale di punti a $3^5 = 243$ (si noti che già con $Q = 5$ si avrebbero ben 3125 punti).

[23] Il modello che si è preso ad esempio ha due effetti casuali, per cui $r = 2$.

[24] Nel caso si voglia verificare l'ipotesi che una componente di varianza sia nulla, bisogna tener presente che tale ipotesi cade sulla frontiera dello spazio parametrico e che, quindi, la distribuzione asintotica del rapporto di verosimiglianza non è χ^2 con 1 grado di libertà (Self e Liang, 1987). Il problema può essere risolto considerando uno spazio parametrico esteso in cui la componente di varianza può assumere anche valori negativi (Longford, 1993, p. 172; in effetti, la maggior parte degli algoritmi

non esiste un'approssimazione affidabile della verosimiglianza. Il programma MIXOR (Hedeker e Gibbons, 1996) usa la procedura appena descritta per la stima dei parametri di modelli a risposta ordinale a due livelli, con quattro possibili link (*probit*, *logit*, *log-log*, *complementary log-log*). La possibilità di scegliere il numero di punti di quadratura consente di ottenere il livello di approssimazione desiderato. Il programma MIXNO (Hedeker, 1998) estende la procedura al modello logistico politomico a due livelli. In tal caso, poichè gli effetti casuali sono diversi per ogni equazione (si veda infatti $\log(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(m)}}) = \beta_0^{(s)} + \beta_1^{(s)} x_{ij} + u_{0j}^{(s)} + u_{1j}^{(s)} x_{ij}$ con $s = 1, ..., m-1$), la riparametrizzazione del modello avviene ponendo

$$\mathbf{u}_j^{(s)} = \mathbf{\Psi}^{(s)} \mathbf{w}_j \quad s = 1, ..., m-1$$

dove $\mathbf{\Psi}^{(s)}$ è la matrice sotto-triangolare che si ottiene dalla scomposizione di Cholesky della matrice di covarianza degli effetti casuali relativi alla categoria s, mentre $\mathbf{w}_j$ è un vettore aleatorio con distribuzione normale multivariata standard.

3.6.2 Quasi-Verosimiglianza Marginale (MQL) e Penalizzata (PQL)

I metodi di quasi-verosimiglianza MQL e PQL si basano su un'approssimazione lineare del modello di cui si vogliono stimare i parametri, in modo tale da poter usare gli algoritmi IGLS e RIGLS per i modelli lineari. Per illustrare i metodi MQL e PQL si considera il caso di dati binari a due livelli, scrivendo il modello come

$$\begin{aligned} y_{ij} &= \pi_{ij} + e_{ij} \\ &= h(\beta_0 + \beta_1 x_{ij} + u_{0j}) + e_{ij} \end{aligned}$$

dove $h(\cdot) = g^{-1}(\cdot)$ è l'inversa della funzione link e e_{ij} è il termine di errore di primo livello, con valore atteso nullo e varianza $\pi_{ij}(1 - \pi_{ij})$. La rappresentazione (3.14) richiama alla mente il modello multilivello lineare, ma esistono due differenze fondamentali:

1. la funzione h in generale non è lineare;

---

consente una stima non vincolata dei parametri casuali). In alternativa, l'ipotesi di componente di varianza nulla può essere verificata per mezzo di un test *score* (Lin, 1997).

2. il termine di errore di primo livello non è indipendente dall'effetto casuale, poichè la sua varianza dipende da π_{ij} e quindi da u_{0j} .

Data un generica funzione di due variabili $h(\eta+\theta)$, consideriamo la seguente approssimazione in serie di Taylor intorno al punto $(\eta_0 + \theta_0)$, arrestando lo sviluppo al primo termine per η e al secondo termine per θ:

$$h(\eta+\theta) \simeq h(\eta_0+\theta_0) + h'h(\eta_0+\theta_0)\cdot(\eta - eta_0) \\ +h'(\eta_0+\theta_0)\cdot(\theta-\theta_0) + \frac{1}{2}h''(\eta_0+\theta_0)\cdot(\theta-\theta_0)^2$$

dove l'apice indica la derivata. Questa approssimazione può essere applicata al modello (3.14) ponendo η e θ uguali rispettivamente alla parte fissa e alla parte casuale del predittore lineare:

$$\begin{cases} \eta = \beta_{0t+1} + \beta_{1t+1}x_{ij} \\ \theta = u_{0j,t+1} \end{cases}$$

dove l'indice $t+1$ significa che stiamo considerando i valori da stimare alla iterazione $t+1$. I possibili metodi di stima differiscono in base al punto $(\eta_0+\theta_0)$ dell'approssimazione e in base alla presenza o meno del termine di secondo ordine. Per quanto riguarda il punto $(\eta_0+\theta_0)$, due possibili scelte sono

1. Marginal Quasi-Likelihood (MQL)

$$\begin{cases} \eta_0 = \hat{\beta}_{0t} + \hat{\beta}_{1t}x_{ij} \\ \theta_0 = 0 \end{cases}$$

2. Penalized Quasi-Likelihood (PQL)

$$\begin{cases} \eta_0 = \hat{\beta}_{0t} + \hat{\beta}_{1t}x_{ij} \\ \theta_0 = \hat{u}_{0j,t} \end{cases}$$

dove la notazione $\hat{\beta}_{0t}$ indica il valore stimato alla iterazione t. Inoltre i due metodi possono essere implementati limitatamente ai termini di primo ordine dell'espressione (3.14), oppure includendo anche il termine di secondo ordine: pertanto sia MQL che PQL vengono distinti in termini di primo ordine e di secondo ordine. In seguito indicheremo con MQL1 e MQL2 i metodi MQL rispettivamente di primo e secondo ordine, e analogo significato avranno PQL1 e PQL2. In primo luogo si esamina in dettaglio la stima con PQL2 (Goldstein e Rasbash, 1996). Si indica con $\hat{h}_{ij,t}$ la quantità $h(\hat{\alpha}_t+\hat{\beta}_t x_{ij}+\hat{u}_{0j,t})$ e analogo significato si attribuisce a $\hat{h}'_{ij,t}$ e $\hat{h}''_{ij,t}$. In base alla (3.14), il modello (3.14) può essere approssimato alla iterazione $t+1$ da

$$y_{ij} = \hat{h}_{ij,t} + \hat{h}'_{ij,t} \cdot (\hat{\beta}_{0t+1} + \hat{\beta}_{1t+1} x_{ij} - \hat{\beta}_{0t} - \hat{\beta}_{1t} x_{ij}) + \\ + \hat{h}'_{ij,t} \cdot (\hat{u}_{0j,t+1} - \hat{u}_{0j,t}) + \frac{1}{2} \hat{h}''_{ij,t} \cdot (u_{0j,t+1} - \hat{u}_{0j,t})^2 + e_{ij}$$

Ora sostituiamo il termine quadratico con il suo valore atteso, $\frac{1}{2}\hat{h}''_{ij,t} \cdot \hat{c}_{j,t}$, dove $\hat{c}_{j,t}$ è la stima della varianza condizionata del residuo all'iterazione t. Poi si scriva e_{ij} come $\hat{z}_{ij,t} e^{\sharp}_{ij}$, dove

$$\hat{z}_{ij,t} = \sqrt{\hat{h}_{ij,t} \cdot (1 - \hat{h}_{ij,t})}$$

mentre $e^{\sharp}_{ij}$ è un termine di errore di media nulla e varianza unitaria[25]. La varianza di $e^{\sharp}_{ij}$ è dunque fissata e non è oggetto di stima, a meno che non si voglia stimare una componente di extravariabilità binomiale (Goldstein, 1995, pp. 98-99).
Pertanto il modello (3.14) può essere scritto come:

$$y^{\sharp}_{ij,t} = \pi^{\sharp}_{ij,t} + z^{\sharp}_{ij,t} e^{\sharp}_{ij}, \tag{3.14}$$

dove

$$y^{\sharp}_{ij,t} = y_{ij} - \hat{h}_{ij,t} + \hat{h}'_{ij,t} \cdot (\hat{\beta}_{0t} + \hat{\beta}_{1t} x_{ij} + \hat{u}_{0j,t}) - \frac{1}{2} \hat{h}''_{ij,t} \cdot \hat{c}_{j,t}$$

$$\pi^{\sharp}_{ij,t} = \hat{h}'_{ij,t} \cdot (\beta_{0t+1} + \beta_{1t+1} x_{ij} + u_{0j,t+1}).$$

La risposta modificata $y^{\sharp}_{ij,t}$ si ottiene sottraendo dalla risposta originaria y_{ij} un *offset* calcolato sulla base delle stime all'iterazione t, mentre il valore atteso modificato lineare $\pi^{\sharp}_{ij,t}$ non è altro che il predittore lineare moltiplicato per la costante $\hat{h}'_{ij,t}$ determinata all'iterazione t. Dunque, noti i risultati dell'iterazione t, il modello (3.14) è un modello multilivello lineare che può essere stimato con l'algoritmo IGLS o RIGLS. Riassumendo, ogni iterazione è composta dai seguenti passi:

1. calcolo delle variabili modificate (dette anche variabili di lavoro) presenti nel modello (3.14), basandosi sulle stime ottenute all'iterazione precedente;
2. stima dei parametri casuali e fissi per mezzo di un'iterazione dell'algoritmo IGLS o RIGLS;
3. stima degli effetti casuali e della loro varianza condizionata secondo la procedura per il modello lineare.

[25] La sostituzione di e_{ij} con $z_{ij,t} e^{\sharp}_{ij}$ ha il fine di eliminare la dipendenza dell'errore di primo livello dall'effetto casuale, scaricando la dipendenza di e_{ij} da π_{ij} su una covariata fittizia.

Le iterazioni vengono ripetute fino a convergenza. L'uso dei nomi Quasi-Verosimiglianza Marginale e Quasi-Verosimiglianza Penalizzata è divenuto comune in seguito ad un fondamentale articolo di Breslow e Clayton (1993). Il termine Quasi-Verosimiglianza sta ad indicare che questi metodi si basano solo sui valori attesi e sulla funzione di varianza, senza specificare l'intera distribuzione; il termine Marginale si riferisce al fatto che la relativa procedura approssima il modello multilivello con il modello marginale; infine, il termine Penalizzata è motivato dall'analogia della relativa procedura con la Quasi-Verosimiglianza Penalizzata usata da Green (1987) nell'ambito della regressione semiparametrica. Goldstein (1991) ha inizialmente proposto il metodo MQL, che si è ben presto rivelato insufficiente per una stima accurata (Breslow e Clayton, 1993; Rodriguez e Goldman, 1995). Le deficienze del metodo MQL vanno ricercate nel fatto che i parametri del modello marginale usato per l'approssimazione sono sistematicamente minori (in valore assoluto) di quelli del modello multilivello, come dimostra il risultato di Neuhaus et al. (1991). Dunque il metodo MQL produce, per i parametri fissi, stime distorte verso il basso, con una distorsione che cresce con la varianza degli effetti casuali. Le simulazioni Monte Carlo hanno inoltre mostrato forti distorsioni verso il basso delle stime degli effetti casuali[26]. Successivamente Goldstein e Rasbash (1996) hanno messo a punto il metodo PQL, mostrando, per via di simulazioni, che funziona assai meglio del metodo MQL, producendo stime solo leggermente distorte verso il basso. Naturalmente il metodo PQL è più complesso, poiché prevede l'uso dei residui ad ogni iterazione, e quindi presenta maggiori problemi di convergenza e maggiore variabilità degli stimatori. Per quanto riguarda la scelta fra le versioni PQL1 e PQL2, le simulazioni mostrano che PQL2 ha una performance leggermente migliore e qualche problema di convergenza in più. Le stime ottenute con i metodi MQL e PQL possono eventualmente essere corrette per mezzo delle procedure di bootstrap parametrico iterato implementate in MLwiN (Goldstein et al., 1998, cap. 7). Tuttavia occorre sottolineare che non è comunque possibile ottenere approssimazioni affidabili della verosimiglianza, per cui l'uso del test del rapporto di verosimiglianza è precluso. I metodi MQL e PQL, che abbiamo illustrato nel caso di dati binari a due livelli, possono essere estesi all'intera classe dei GLM multilivello con un numero arbitrario di livelli. In particolare, Goldstein (1995, p. 105) descrive le modifiche necessarie per la stima dei modelli per dati politomici e ordinali.

[26] Nei GLM per dati binari è difficile dare delle indicazioni sul tasso di distorsione, poichè ci sono molti fattori che entrano in gioco. Particolare importanza sembrano rivestire la struttura gerarchica (numero di gruppi e numerosità dei gruppi) e la variabilità della risposta nei gruppi. La situazione più sfavorevole si verifica quando i gruppi sono pochi ed al loro interno hanno risposte con poca variabilità.

3.7 Software per l'analisi

Gli algoritmi descritti in precedenza sono impiegati nei più importanti software per l'analisi multilivello; nel tempo infatti si sono via via sviluppati software specifici che tengono in considerazione le speciali proprietà dei modelli multilevel. Tuttavia possono essere utilizzati tutti i programmi riguardanti i modelli lineari misti, dato che i modelli multilevel lineari appartengono a questa classe, anche se sono dei casi speciali.
Si descrivono ora i programmi disponibili, con alcune indicazione degli algoritmi utilizzati, i punti forza e di debolezza. Alcuni di questi programmi possono essere parti di pacchetti statistici come ad esempio SAS.

- MLwin (Goldstein, 2003). E' l'ultimo di una lunga serie di programmi ML sviluppati dal Multilevel Project dell'Institute of Education di Londra; le precedenti versioni, in ambiente DOS, sono state soppiantate da MLwin che si interfaccia con Windows. Esiste un'ampia documentazione riguardante il programma Mlwin. Gli algoritmi utilizzati sono l'IGLS e il RIGLS.
- HLM (Raudenbush, Bryk, Cheong e Congdon 2004). La storia di HLM è simile a quella di MLwin. Inizialmente venne sviluppato un programma per l'analisi a due livelli, in un secondo momento un programma a tre livelli e infine una versione per Windows. Il software è fortemente legato al libro " Hierarchical linear models" di Anthony Bryk e Stephen Raudenbush (un legame anche più stretto di quello tra Harvey Goldstein e MLwin). Il pacchetto HLM gioca un ruolo importante nel campo della statistica educativa: negli Stati Uniti è stato adottato come software ufficiale per l'analisi multilevel in ambito scolastico.
  Per default, HLM/2L fornisce stime RML mentre HLM/3L stime FML ("2L" e "3L" stanno rispettivamente per "modelli a due livelli" e "modelli a tre livelli"). Attualmente i due programmi possono formire entrambe le stime. La più vecchia versione di HLM utilizzava l'algoritmo EM che, però, come già precisato, ha lo svantaggio di essere molto lento. Ora è possibile accelerare la convergenza attraverso l'algoritmo di Fisher scoring.
- R. In R sono disponibili due librerie tramite cui adattare modelli random e misti(mixed effects models), modelli in cui compaiano sia fattori fissi sia fattori random. La prima, che fa parte della distribuzione standard, è *nlme* (si veda Pinheiro, Bates, 2000) per una descrizione molto approfondita). La seconda, più recente e versatile, è *lme*4 (Bates, 2005), la quale non è contenuta nell'installazione standard e deve essere scaricata separatamente, assieme alla libreria *Matrix* da cui dipende. Per modelli lineari a effetti misti la maggiore differenza tra le due implementazioni è che le routine della librera *lme*4 sono in grado di trattare efficientemente modelli con effetti random incrociati o parzialmente incrociati e non solo modelli con effetti random annidati. E' disponibile una libreria specifica

per i modelli multilivello, la libreria *multilevel*, la quale però risulta meno flessibile delle librerie precedenti.

- Stata (GLLAMM). Pacchetto esterno di STATA, *gllamm* (*Generalized Linear Latent and Mixed Models*) è stato sviluppato da Skrondal, Rabe-Hesketh e Pickles (2004). I GLLAMM sono una classe di modelli per variabili latenti multilivello utilizzabili con vari tipi di variabili risposta: continue, conteggi, dati di durata, dicotomiche e dati categoriali. Le variabili latenti, o effetti casuali possono avere una distribuzione discreta o normale multivariata. Esempi di modelli appartenenti a questa classe sono: i modelli lineari generalizzati multilivello, i modelli fattoriali multilivello, i modelli a classi latenti e i modelli a equazioni strutturali multilivello. Alcuni esempi che utilizzano *gllamm*, scaricabile dal sito www.gllamm.org, sono disponibili nel sito: http://www.gllamm.org/examples.html.
  Sostanzialmente all'interno del "pacchetto" sono forniti due programmi principali, uno dedicato alla stima vera e propria dei modelli *gllamm* e uno dedicato esclusivamente alle operazioni "post-stima" *gllapred*. Per la sintassi completa con tutte le molteplici opzioni di *gllamm* si rinvia al manuale di Gllamm realizzato dai creatori del pacchetto stesso (Skrondal, Rabe-Hesketh e Pickles A., 2004).
- PROC MIXED: pacchetto di SAS. PROC MIXED si occupa dell'analisi dei modelli misti.
- SPSS - PASW MIXED- per modello generale lineare misto. Supporta i seguenti tipi di modelli:
  - Modello ANOVA ad effetti fissi
  - Disegni a blocchi completamente randomizzati
  - Disegni di esperimento Split-Plot
  - Modello ad effetti puramente casuali
  - Modello a coefficienti casuali
  - L'analisi multilivello
  - L'analisi delle misure ripetute
  - Analisi delle misure ripetute con covariate dipendenti dalla scala temporale
- Mplus (Muthen e Muthen 1998 - 2006) Mplus offre una vasta scelta di modelli, stimatori e algoritmi. Il programma ha un'interfaccia di facile utilizzo e una chiara visualizzazione grafica sia dei dati che dei risultati delle analisi. Mplus permette l'analisi sia di dati annidati sia di dati longitudinali, sia ad un livello sia multilivello, dati con o senza eterogenità e con dati mancanti. Le analisi possono essere eseguite sia per variabili continue, sia censurate, sia per dati binari, sia per dati categorici (ordinali o nominali), e anche per combinazioni di questi tipi di variabili. Inoltre, *Mplus* offre una vasta gamma di studi di simulazione Monte Carlo, dove i dati possono essere generati ed analizzati secondo uno dei modelli inclusi nel programma. Il quadro di modellazione *Mplus* riprende il tema unificante delle variabili latenti. Le generalità del quadro di modellazione

Mplus deriva dall'uso esclusivo di entrambe le variabili latenti continue e categoriali.

Capitolo 4
Una proposta alternativa per le componenti erratiche

La metodologia dei modelli multilivello consente l'analisi di dati organizzati in una struttura di tipo gerarchico, ossia di dati raggruppati. Nella teoria "classica" si introduce ai fini inferenziali l'ipotesi distributiva normale per le componenti d'errore. Nonostante le numerose proprietà di questa distribuzione un aspetto critico è rappresentato dalla limitatezza delle forme che può assumere la funzione di densità. Tutto ciò è in contrasto con i molteplici andamenti che si possono trovare in natura e diventa un limite quando il tutto viene riportato nell'ambito dell'analisi dei dati reali. Quindi in alcuni casi tale assunzione può rilevarsi troppo restrittiva. Si propone, quale alternativa alla normale, la distribuzione Skew-Normal (SN) (Azzalini, Dalla Valle, 1996). La classe di distribuzioni normali asimmetriche, denotata con il simbolo SN, è una famiglia di densità di probabilità che generalizza la legge normale introducendo una possibile asimmetria. E' importante notare che, a differenza di altre proposte presenti in letteratura, la famiglia SN consente di passare dal caso simmetrico a quello asimmetrico con continuità, modificando il solo valore di un parametro. Da un punto di vista teorico, la classe normale asimmetrica, ha il vantaggio di essere matematicamente trattabile e di godere di un buon numero di proprietà tipiche della distribuzione normale. La prima analisi sistematica della classe normale asimmetrica nel caso scalare è stata effettuata da Azzalini (1985, 1986); successivamente Azzalini, Dalla Valle (1996) hanno introdotto la versione multidimensionale della normale asimmetrica. Di seguito viene quindi messa in luce l'utilità di impiego di questa distribuzione nell'ambito dell'analisi multilivello, e si discutono i principali problemi legati alla stima dei parametri. Nelle applicazioni accade spesso di disporre di dati organizzati in una struttura di tipo gerarchico, essendo gli stessi disposti, in via naturale o in modo funzionale all'analisi, in classi o gruppi, suscettibili a loro volta di essere ripartiti in sottogruppi e così via. Spesso nella realtà però la forma delle distribuzioni è non simmetrica. Si consideri ad esempio la distribuzione del peso di persone, alcune delle quali sovrappeso, mentre la maggior parte hanno un peso normale, questa distribuzione potrebbe quindi essere asimmetrica a destra.

Già in economia e nel marketing si è proposto quale alternativa alla normale la distribuzione lognormale. Meijer e Rouwendal (2006) hanno confrontato quanto trattato in letteratura e comparato, come alternative alla distribuzione normale, le distribuzioni lognormale e Gamma; nel loro lavoro sulla base dell'adattamento e dell'interpretabilità, le distribuzioni lognormale e Gamma si sono dimostrate più performanti, rispetto alla normale.
Si vuole ora presentare una possibile estensione del modello multilivello, considerando un'ulteriore tipo di componente aleatoria, che garantisca:

- l'inclusione della normale come sottocaso proprio
- un'ampia escursione di indici di asimmetria e curtosi
- di preservare alcune proprietà della normale
- la trattabilità matematica delle alternative.

Non è detto che tutte queste proprietà desiderate si possano conseguire appieno e simultaneamente con una singola formulazione. La distribuzione normale asimmetrica risulta essere un modello flessibile e con maggior capacità di adattamento a fenomeni osservati per via campionaria. Permette inoltre trattabilità matematica e maggior interpretabilità rispetto al problema sostanziale. Permette di estendere la classe normale per renderla più flessibile. Questo approccio presenta forti e naturali legami con varie tematiche in ambito applicativo. L'asimmetria infatti sembra nascere da una diversa reazione a sollecitazioni esterne. Dati che presentano una struttura siffatta richiedono particolari cautele nell'applicazione delle consuete tecniche di analisi statistica multivariata, sia nella fase puramente esplorativa, sia nella descrizione ed interpretazione mediante modelli statistici, dei legami fra i fenomeni osservati. Così per esempio se lo scopo dell'analisi è lo studio della dipendenza di un fenomeno da altri fenomeni ed i dati seguono una struttura gerarchica, è opportuno ai fini di un'analisi più accurata tenere in considerazione tale struttura, in caso contrario infatti, il modello potrebbe non risultare adeguato. Tutto quanto si dirà si riferisce a distribuzioni continue, e quindi dotate di funzione di densità. Si discuterà di una particolare formulazione, altri approcci in uso corrente, con particolare rilevanza in ambito multidimensionale, sono i cosidetti "miscugli" di distribuzioni (normali) e le copule. Un aspetto da sottolineare è che la famiglia delle distribuzioni normali risulterà un elemento proprio "centrale" della nuova costruzione, e non un caso limite. Questo risulta essere in accordo con la percezione comune della famiglia normale rispetto alle distribuzioni osservate empiricamente. Un livello di estensione è costituito dalla famiglia di densità simmetriche perturbate (Azzalini 2005, Genton 2004), le quali si ottengono partendo da una qualunque densità simmetrica multidimensionale e applicando a questa un fattore di ponderazione che può essere scelto con ampio margine di manovra, dovendo rispettare solo poche e semplici condizioni. La disponibilità di questo tipo di costruzione ha aperto la strada verso la costruzione di classi di distribuzioni che consentono non solo la regolazione della loro asimmetria, ma anche la curtosi e altri elementi.

4.1 Skew-Normal aspetti generali

"Il compito della statistica (. . .) non consiste solo nel determinare la legge di dipendenza dei diversi valori ed esprimerla con pochi numeri, ma anche nel fornire un aiuto allo studioso che vuole cercare le cause della variazione. . . . le curve teoriche studiate dal PEARSON e dall'EDGEWORTH (. . .) mentre danno con molta approssimazione la legge di variazione, a mio avviso sono difettose in quanto (. . .) nulla ci fanno sapere sulla legge di dipendenza, quasi nulla sulla relazione con la curva normale. Io penso che miglior aiuto per lo studioso potrebbero essere delle equazioni che supponessero una perturbazione della variabilità normale per opera di cause esterne" (Fernando de Helguero, 1908). Anche se tecnicamente la prima espressione formale della distribuzione normale asimmetrica risale addietro nel tempo (Birnbaum, 1950), tuttavia il punto di avvio della ricerca in questo ambito va collegato al riconoscimento del ruolo autonomo della distribuzione stessa come estensione della famiglia delle distribuzioni normali, specificamente con l'introduzione di un parametro di regolazione dell'asimmetria (Azzalini, 1985). La costruzione della versione multidimensionale della distribuzione normale asimmetrica (Azzalini e Dalla Valle, 1996) ha infatti contribuito alla crescita di interesse in questo ambito di ricerca.

La classe delle normali asimmetriche multivariate, introdotta da Azzalini e Dalla Valle nel 1996, include le distribuzioni normali multivariate e permette di modellare sia l'asimmetria che la curtosi. Le normali asimmetriche possono essere generate in diversi modi, che ne motivano l'applicazione nella medicina e nelle assicurazioni. I momenti, i cumulanti ed i principali indici di sintesi hanno una semplice forma analitica. La classe delle distribuzioni normali asimmetriche è chiusa rispetto a trasformazioni lineari e presenta notevoli proprietà di invarianza. Può essere ulteriormente estesa attraverso la classe delle distribuzioni normali asimmetriche generalizzate, a cui appartengono alcune trasformazioni di normali inverse comunemente usate.

La normale asimmetrica (Azzalini, 1985) oltre agli usuali parametri di posizione e scala, prevede l'introduzione di un terzo parametro che ne regola l'asimmetria. Essa gode di buone proprietà dal lato matematico-probabilistico e non mancano molti risultati utili in ambito inferenziale. I primi studi sulle distribuzioni delle probabilità asimmetriche appaiono alla fine del diciannovesimo secolo. Edgeworth (come citato nel precedente lavoro del de Helguero) per primo esamina il problema di adattamento di distribuzioni ai dati. Qualche anno dopo Karl Pearson discute (1894; 1895) l'ottenimento di densità asimmetriche dal miscuglio di due curve normali. Uno statistico italiano, Fernando de Helguero, al IV Congresso Internazionale dei Matematici di Roma del 1909, introduce una nuova famiglia di distribuzioni asimmetriche. L'idea che de Helguero propone riguarda una modifica delle distribuzioni di probabilità simmetriche tramite perturbazione per poter descrivere quei fenomeni in cui si osservano campioni selezionati. Il punto di partenza degli studi moderni sulle distribuzioni asimmetriche può identificarsi nel lavoro di Azzalini

(1985), dove si trova il seguente utile risultato.

Lemma 4.1. *Se f_0 è una funzione di densità di probabilità unidimensionale simmetrica in 0, G_0 una funzione di ripartizione derivabile tale che G_0' esiste ed è simmetrica attorno a 0, e w una funzione dispari, allora*

$$f(z) = 2f_0(z)G_0\{w(z)\} \qquad (z \in R) \tag{4.1}$$

è una funzione di densità.

Dimostrazione. Sia Y una v.c. con densità f_0 e X una v.c. con distribuzione G_0, indipendente da Y. E' immediato verificare che $w(X)$ ha distribuzione simmetrica, e quindi lo stesso vale per $X - w(Y)$ tale per cui

$$\frac{1}{2}P\{X - w(Y) \leq 0\} = E\{P\{X \leq w(Y)|Y = y\}\} = \int_R f_0(y)G_0\{w(y)\}dy$$

Si chiami f_0 distribuzione base, e $G(x) = G_0\{w(x)\}$ fattore di perturbazione (della simmetria). La

$$f(z) = 2f_0(z)G_0\{w(z)\}$$

fornisce un meccanismo semplice e generale per produrre una gran varietà di distribuzioni *perturbate* a partire da quella di base.
Il supporto di f è al più quello di f_0. Il risultato non richiede che il supporto di f_0 sia l'intera retta. Se si pone $G(x) = G_0\{w(x)\}$, allora $G(x) \geq 0$ e $G(x) + G(-x) = 1$; si può reimpostare il risultato precedente in termini di una $G(x)$ avente tali proprietà. In generale $G(x)$ non è una funzione di ripartizione. Siccome $w(x) \equiv 0$ è una funzione dispari, per la quale risulta che $G(x) = \frac{1}{2}$, allora l'insieme delle funzioni $f(\cdot)$ di tipo $f(z) = 2f_0(z)G_0\{w(z)\}$ include $f_0(\cdot)$.

Questo lemma generale garantisce la costruzione di un'intera famiglia di distribuzioni asimmetriche a partire, dalla perturbazione tramite $G_0\{w(z)\}$, di una funzione di densità simmetrica f_0. La nuova famiglia di funzioni f include f_0 per $w(z) = 0$. Da questa idea nasce un nutrito insieme di distribuzioni asimmetriche: si pensi all'articolo di Arnold e Beaver (2000) nel quale viene proposta una generalizzazione in cui la curva normale viene sostituita da distribuzioni con code più pesanti, come per esempio, la distribuzione multivariata asimmetrica di Cauchy.

4.2 La distribuzione normale asimmetrica

Si definisce ora la distribuzione di probabilità normale asimmetrica facendo riferimento al lemma precedente.

Siano $f_0 = \varphi$ e $G = \Phi$ la funzione di densità e di ripartizione di una normale standardizzata. Allora la densità

$$\phi(z;\alpha) = 2\varphi(z)\Phi(\alpha z) \quad (-\infty < z < \infty) \tag{4.2}$$

è chiamata normale asimmetrica con parametro di forma α e si indicherà, per seguire l'usuale notazione in uso in letteratura, che $Z \sim SN(\alpha)$.

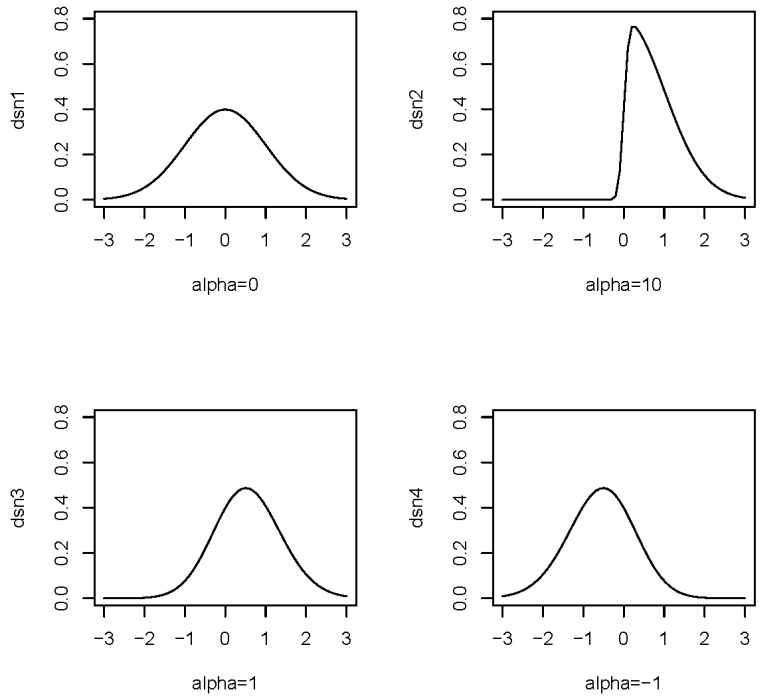

Figura 4.1: Grafico della funzione di densità di una $SN(\alpha)$ per alcuni valori di α

4.2.1 Famiglia di posizione e scala

Serve introdurre la famiglia di trasformazioni

$$Y = \xi + \omega Z \quad (\xi \in R, \omega \in R^+)$$

che produce la famiglia di distribuzioni aventi densità di probabilità in x pari a

$$\frac{2}{\omega}\varphi(\frac{x-\xi}{\omega}\Phi(\alpha\frac{x-\xi}{\omega})) \quad (-\infty < x < \infty)$$

dove ξ rappresenta il parametro di posizione e ω quello di scala. Scriveremo che $Y \sim SN(\xi, \omega^2, \alpha)$.

4.2.2 Momenti

La funzione generatrice dei momenti per una $SN(\xi, \omega^2, \alpha)$ è facilmente ottenibile sfruttando alcuni risultati ben noti:

Lemma 4.2. *(Complemento al quadrato). Se A è una matrice $k \times k$ simmetrica definita positiva e b un vettore $k \times 1$, allora*

$$\int_{R^k} \frac{1}{(2\pi)^{\frac{k}{2}}} \exp\{-\frac{1}{2}(y^T Ay - 2b^T y)\}dy = \frac{\exp\{\frac{1}{2}(b^T A^{-1} b)\}}{|A|^{\frac{1}{2}}}$$

dove dy sta per $dy_1...dy_k$

Lemma 4.3. *Se $U \sim N(0,1)$ e a, $b \in R$ allora:*

$$E[\Phi(a + bU)] = \Phi(\frac{a}{\sqrt{1+b^2}})$$

si veda Zacks(1981).

La funzione generatrice dei momenti per una $SN(\xi, \omega^2, \alpha)$ è data da

$$M(t) = 2\exp(\xi t + \frac{\omega^2 t^2}{2})\Phi(\delta\omega t)$$

dove $\delta = \frac{\alpha}{1+\alpha^2} \in (-1, 1)$, e la funzione generatrice dei cumulanti:

$$K(t) = \log M(t) = \xi t + \frac{\omega^2 t^2}{2} + \log[2\Phi(\delta\omega t)]$$

dove

$$\zeta_0(x) = \log\{2\Phi(x)\}$$

e in generale

$$\zeta_m(x) = \frac{d^m}{dx^m}\zeta_0(x) \quad (m = 1, 2, ...)$$

Derivando la funzione generatrice dei cumulanti otteniamo:

$$E[Y] = \xi + \omega\mu_z$$

$$Var[Y] = \omega^2(1 - \mu_z^2)$$

$$\gamma_1 = \frac{4-\pi}{2}\frac{\mu_z^3}{(1-\mu_z^2)^{\frac{3}{2}}}$$

$$\gamma_2 = 2(\pi - 3)\frac{\mu_z^4}{(1-\mu_z^2)^2}$$

dove $\mu_z = \sqrt{\frac{2}{\pi}\delta}$ mentre γ_1 e γ_2 sono il terzo e il quarto cumulante standardizzato, noti anche rispettivamente come il coefficiente di asimmetria e di curtosi.
E' da notare che gli indici γ_1 e γ_2 hanno un campo di variazione limitato. In particolare $|\gamma_1| \leq 0,995$ circa, mentre $0 \leq \gamma_2 \leq 0,869$. Questo implica un limite della normale asimmetrica nel rappresentare il comportamento a livello di asimmetria e curtosi per campioni di dati particolarmente asimmetrici o con pronunciata curtosi.

4.2.3 Parametrizzazione centrata

Per ovviare ai problemi legati all'inferenza sui parametri (ξ, ω^2, α) appena menzionati, Azzalini (1985) propone una riparametrizzazione del modello. Partendo dall'identità

$$Y = \xi + \omega Z = \mu_z + \sigma_z Z_0$$

dove Z ha una distribuzione $SN(\alpha)$ del tipo $\phi(z;\alpha) = 2\varphi(z)\Phi(\alpha z)$ e posto:

$$\mu_z = E(Z) = b\delta$$

e

$$\sigma_z^2 = Var(Z) = 1 - \mu_z^2$$

inoltre si definisce $Z_0 = \frac{1}{\sigma_z}(Z - \mu_z)$. La densità di Y sarà

$$\phi(z;\xi,\omega,\alpha) = \frac{2}{\omega}\phi(\frac{y-\xi}{\omega})\Phi(\alpha\frac{y-\xi}{\omega}) \quad (-\infty < z < \infty)$$

e si scriverà che $Y \sim SN(\xi, \omega^2, \alpha)$.
La parametrizzazione alternativa data da $(\mu, \sigma^2, \gamma_1)$ le cui espressioni esplicite

in termini dei parametri originali sono date da $E[Y] = \xi + \omega\mu_z$ $Var[Y] = \omega^2(1-\mu_z^2)$ $\gamma_1 = \frac{4-\pi}{2}\frac{\mu_z^3}{(1-\mu_z^2)^{\frac{3}{2}}}$.

Questa parametrizzazione è nota in letteratura con il nome di parametrizzazione centrata (o CP utilizzando il relativo acronimo dalla lingua inglese) in quanto viene introdotta a partire dalla variabile centrata Z_0, mentre la parametrizzazione fin qui utilizzata è chiamata parametrizzazione diretta (o DP). L'utilizzo della parametrizzazione centrata offre sicuramente numerosi benefici. Da un lato semantico, i nuovi parametri hanno un significato chiaro, più intuitivo e famigliare. Come nel modello normale, infatti, μ e σ^2 rappresentano esattamente la media e la varianza della distribuzione, mentre il parametro γ_1, rappresentando l'indice di asimmetria della distribuzione, risulta più logicamente collegato a questa rispetto al parametro α. Da un lato più pratico e inferenziale, la parametrizzazione centrata elimina talune caratteristiche che rendevano difficoltose le operazioni d'inferenza sulla normale asimmetrica.

Anche il problema della singolarità della matrice viene risolto tramite la parametrizzazione centrata, rendendo quindi possibile applicare gli usuali metodi dell'inferenza asintotica. Infatti i risultati sulla distribuzione asintotica delle stime di massima verosimiglianza affermano che, in condizioni di regolarità del problema di stima, dato un campione di numerosità n:

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1})$$

dove θ in questo caso indica il vettore dei parametri $\theta = (\mu, \sigma^2, \gamma)$ e dove $I(\theta)$ è la matrice d'informazione attesa per i parametri appena descritti.

4.2.4 Proprietà

In Azzalini (1985) sono state introdotte alcune importanti proprietà in cui risultano evidenti i legami tra la distribuzione normale e la distribuzione normale asimmetrica.

1. Se $\alpha = 0$ ci si riconduce alla densità di una $N(0,1)$;
2. se $X \sim SN(\alpha)$, allora $-X \sim SN(-\alpha)$;
3. se $\alpha \to \infty$ ci si riconduce alla mezza normale, $2\phi(x)$ per $z \geq 0$ ovvero

$$\lim_{\alpha\to\infty} \phi(z,\alpha)\frac{2}{\sqrt{\pi}}\exp\{-\frac{1}{2}z^2\} \quad z > 0$$

4. se $X \sim SN(\alpha)$, allora $Z \sim \chi_1^2$
5. per un fissato α la $\phi(z;\alpha) = 2\phi(z)\Phi(\alpha z)$ è fortemente unimodale, ossia $log(\phi(z;\alpha))$ è una funzione concava di z.

La distribuzione normale risulta quindi un caso interno a questa famiglia di distribuzioni.

Altre proprietà inerenti la distribuzione normale asimmetrica, ma di livello più approfondito si possono trovare in Azzalini (2005).

4.2.5 Generazione

La densità $\phi(z;\alpha) = 2\varphi(z)\Phi(\alpha z)$ è stata introdotta a partire dalla densità normale standardardizzata, applicando su di essa una perturbazione che da luogo all'asimmetria. Esistono diversi metodi per costruire una variabile aleatoria normale asimmetrica. Questi metodi hanno una importanza sia teorica che pratica in quanto consentono una facile generazione di valori pseudo-casuali da una legge normale asimmetrica. Si vogliono ora descrivere alcuni meccanismi reali che danno luogo a questo tipo di distribuzione.

4.2.5.1 Convoluzioni di normali

Un altro meccanismo di generazione di dati aventi distribuzione normale asimmetrica è presentato in Azzalini (1986).

Lemma 4.4. *Consideriamo due variabili casuali normali standard indipendenti, U_0 e U_1, e una costante $|\delta| < 1$. Se definiamo: $Z = \delta|U_0| + \sqrt{1-\delta^2}U_1$, allora $Z \sim SN(\frac{\delta}{\sqrt{1-\delta^2}})$*

Questo meccanismo di generazione compare nei lavori di Weinstein (1964), Aigner (1964) e Andel e al.(1984).

4.2.5.2 Rappresentazione stocastica

Se Z è una v.c. con densità $f(z) = 2f_0(z)G_0\{w(z)\}$, valgono le rappresentazioni

$$Z = (Y|X \leq w(Y)) \quad Z = \begin{cases} Y & se\ X \leq w(Y) \\ -Y & altrimenti \end{cases}$$

Invarianza distributiva rispetto alla perturbazione
Un corollario della forma di rappresentazione stocastica, cioè

$$Z = \begin{cases} Y & se\ X \leq w(Y) \\ -Y & altrimenti \end{cases}$$

è il seguente. Se $t(\cdot)$ è una funzione pari, allora segue che

$$t(Z) \stackrel{d}{=} t(Y)$$

dove $\stackrel{d}{=}$ indica uguaglianza in distribuzione.

4.2.5.3 Troncamento di una normale bivariata

Si supponga che $f(\widetilde{v}, z)$ sia una normale bivariata con vettore delle medie pari a $(0,0)^T$ e matrice di varianza con diagonale unitaria e correlazione pari a ρ. Si supponga inoltre di osservare un campione censurato in una delle due marginali, ossia di osservare solamente (v, z) se $v \geq 0$. La densità della variabile che osserviamo sarà quindi:

$$f(v,z) = \begin{cases} 2\widetilde{f}(v,z) \;\; per\ v \geq 0 \\ 0 \quad altrimenti \end{cases}$$

se marginalizziamo rispetto alla variabile z, integrando $f(v, z)$ sul dominio di v, otteniamo esattamente che:

$$Z \sim SN(\frac{\rho}{\sqrt{1-\rho^2}})$$

In contesti di selezione del personale o di ammissione a scuole o enti di formazione in cui è previsto un test d'ingresso di tipo attitudinale, l'osservazione del punteggio ottenuto in test successivi, è evidentemente correlata al test d'ingresso. Questo tipo di meccanismo ben si presta ad essere modellato pensando al troncamento di una distribuzione bivariata come quello appena descritto. Si veda a riguardo Birnbaum (1950) e Arnold et al. (1993). Si può estendere questa idea di campione selezionato, trattando la possibilità di effettuare il troncamento in un qualsiasi punto del dominio della funzione. Si ammetta la possibilità di troncamento latente in un qualsiasi punto del dominio u. Si supponga che $\widetilde{f}(v, z)$ sia la funzione di densità di una normale bivariata a media zero e varianza unitaria con correlazione pari a δ. Si immagini di osservare un campione selezionato, ossia solamente gli individui che superano una certa soglia (diciamo $-\tau$, per motivi di notazione che si chiariranno in seguito) per quanto riguarda v. La densità congiunta diventa quindi

$$f(v,z) = \begin{cases} \frac{\widetilde{f}(v,z)}{1-\Phi(-\tau)} \quad per\ v \geq -\tau \\ 0 \;\; altrimenti \end{cases}$$

Il denominatore $1 - \Phi(-\tau)$ ha la funzione di costante di normalizzazione. Se interesse dello studio è la sola marginale di z, marginalizzando rispetto a v:

$$\begin{aligned}\int_{-\tau}^{+\infty} f(v,z)dz &= \frac{1}{1-\Phi(-\tau)}\int_{-\tau}^{+\infty}\frac{1}{2\pi\sqrt{1-\delta^2}}\exp\{-\frac{1}{2}(\frac{v^2-2\delta vz+z^2}{(1-\delta^2)})\}dv \\ &= \frac{1}{\Phi(-\tau)}\frac{1}{\sqrt{1-\delta^2}}\int_{-\tau}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(\frac{v^2-2\delta vz+\delta^2z^2}{(1-\delta^2)})\} \\ &\quad \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(\frac{z^2-\delta^2z^2}{(1-\delta^2)})\}dv \\ &= \frac{1}{\Phi(-\tau)}\frac{1}{\sqrt{1-\delta^2}}\int_{-\tau}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(\frac{v-\delta z}{\sqrt{1-\delta^2}})^2\}dv\end{aligned}$$

Effettuando la sostituzione $u = \frac{v-\delta_z}{\sqrt{1-\delta^2}}$ e ponendo $\alpha = \frac{\delta}{\sqrt[n]{\sqrt{1-\delta^2}}}$, $\alpha_0(\tau) = \tau\sqrt{1+\alpha^2}$ si ottiene:

$$f_{ESN}(z;\alpha,\tau) = \frac{1}{\Phi(\tau)\phi(z)\Phi(\alpha_0(\tau)+\alpha z)}$$

densità di una distribuzione normale asimmetrica estesa, con parametro di forma α e parametro di troncamento τ indicata con $z \sim ESN(\alpha,\tau)$. Come per la normale asimmetrica si ammettono variazioni di posizione e scala. Siano $\xi \in R$ e $w \in R^+$; allora definita $Y = \xi + wZ$ la densità di Y sarà

$$f_{ESN}(z;\xi,w^2,\alpha,\tau) = \frac{1}{w\Phi(\tau)}\phi(\frac{y-\xi}{w})\Phi(\alpha_0(\tau)+\alpha\frac{y-\xi}{w})$$

e si scriverà che $Y \sim ESN(\xi,w^2,\alpha,\tau)$. Si noti che quando $\alpha = 0$ ci si riconduce alla distribuzione normale, per qualsiasi valore del parametro τ e che, quando $\tau = 0$, ci si riconduce alla distribuzione $\phi(z;\alpha) = 2\varphi(z)\Phi(\alpha z)$.

4.2.5.4 Distribuzione condizionata di una SN_2

Sia Y una variabile casuale normale asimmetrica bidimensionale del tipo $2\phi_d((y-\xi);\Omega)\Phi(\alpha^T w^{-1}(y-\xi))$ con $\Omega = w\Omega_z w$, ossia $Y \sim SN_2(\widetilde{\xi},\widetilde{\Omega},\widetilde{\alpha})$ e

$$\widetilde{\Omega} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22}\end{pmatrix}$$

Se si osserva una delle marginali, si pensi $Y_1 = y_1$, si ricava la distribuzione condizionata di $Y_2|Y_1 = y_1$. Siano:

$$\begin{aligned}
\xi &= \xi_2 + w_{21} w_{11}^{-1}(y_1 - \xi_1) \\
w &= w_{22} - w_{21} w_{11}^{-1} w_{12} \\
\alpha &= \frac{\alpha_1 + \frac{w_{12}}{\sqrt{w_{11} w_{22}}\alpha_2}}{(1 + \alpha_2^T w_{22}^{-1} \alpha_2)} \frac{1}{2} \\
\tau &= \alpha \sqrt{w_{11}^{-1}}(y_1 - \xi_1)
\end{aligned}$$

dove $\overline{w} = \frac{w}{w_{22}}$, per la densità condizionata risulta che

$$(Y_2 | Y_1 = y_1) \sim ESN(\xi, w, \alpha, \tau)$$

come mostrato in Azzalini e Dalla Valle (1996) e successivamente in Azzalini e Capitanio (1999).

4.3 Normale asimmetrica multivariata

Il fronte di maggior impatto della distribuzione SN è tuttavia costituito dal contesto multidimensionale, il cui interesse è decollato peraltro solo dopo il lavoro di Azzalini e Capitanio (1999) che ha introdotto una nuova parametrizzazione più chiara e connessa al caso scalare, tecnicamente equivalente alla prima (Azzalini e Dalla Valle, 1996) ed evidenziando le potenzialità della nuova classe di distribuzioni, sia in termini di proprietà formali che come fruibilità applicativa. Successivamente Arellano-Valle e Azzalini (2008) hanno proposto un'efficace soluzione per porre rimedio ad alcune problematiche di tipo inferenziale, come la singolarità della matrice di informazione attesa per α nullo.

Si introduce ora la funzione di densità della distribuzione multivariata, ma prima risulta necessario introdurre un concetto più ampio rispetto alla simmetria, ovvero la simmetria centrale.

Definition 4.5 (Simmetria centrale).
La v.c. $d-dimensionale$ Y è detta possedere simmetria centrale rispetto a un punto $\xi \in R^d$ se

$$Y - \xi \stackrel{d}{=} \xi - Y$$

Dato che si tratteranno le v.c. continue, è opportuno sottolineare che la definizione di simmetria centrale implica che la densità di probabilità f_0 corrispondente soddisfi la condizione $f_0(x - \xi) = f_0(\xi - x)$, per tutti i punti $x \in R^d$. Ad esempio la v.c. $N_d(\xi, \Omega)$ è centralmente simmetrica rispetto a ξ.

Lemma 4.6. *Sia f_0 una funzione di densità associata ad una v.c. d-dimensionale a simmetria centrale attorno a 0, G_0 una funzione di ripartizione derivabile*

tale che G_0' è simmetrica attorno a 0, e w una funzione da R^d a R dispari. Allora:

$$f(z) = 2f_0(z)G_0\{w(z)\} \quad z \in R^d$$

è una funzione di densità.

Dimostrazione. La dimostrazione è analoga a quella del caso scalare. Sia Y una v.c. con densità f_0 e X una v.c. con distribuzione G_0, indipendente da Y. Si indichi con A un insieme di *Borel* della retta reale, con $-A$ l'insieme che si ottiene invertendo il segno di tutti gli elementi di A, e si ponga $W = w(Y)$. Visto che $Y \stackrel{d}{=} -Y$, si può scrivere:

$$P\{W \in -A\} = P\{-W \in A\} = P\{w(-Y) \in A\} = Pw(Y) \in A$$

e questo implica che la simmetria rispetto a zero di W. Si ha quindi:

$$\frac{1}{2} = P\{X - W \leq 0\} = E\{P\{X \leq w(Y)|Y = y\}\} = \int_{R^d} f_0(y)G_0\{w(y)\}dy.$$

Una variabile casuale d-dimensionale Z, ha distribuzione normale asimmetrica d-dimensionale, se ha una funzione di densità del tipo:

$$\phi(\mathbf{z}; \mathbf{\Omega_z}; \alpha) = 2\varphi_{\mathbf{d}}(\mathbf{z}; \mathbf{\Omega_z})\mathbf{\Phi}(\alpha^{\mathbf{T}}\mathbf{z}) \tag{4.3}$$

dove α è un vettore $d - dimensionale \in R^d$, $\varphi(\mathbf{z}; \mathbf{\Omega_z})$ è la densità di una variabile casuale normale $d - dimensionale$ con vettore delle medie nullo e matrice di correlazione $\mathbf{\Omega_z}$ calcolata in $\mathbf{z} \in R^d$ e $\mathbf{\Phi}(.)$ è la funzione di ripartizione di una normale standard, in quanto le sue distribuzioni marginali hanno parametro di posizione pari a zero e parametro di scala pari a 1.

Diremo che $\mathbf{Z} \sim SN_d(\mathbf{\Omega}, \alpha)$. Generalizzando l'espressione

$$\phi(\mathbf{z}; \mathbf{\Omega_z}; \alpha) = 2\varphi_d(\mathbf{z}; \mathbf{\Omega_z})\Phi(\alpha^{\mathbf{T}}\mathbf{z})$$

con l'introduzione dei parametri di posizione e scala, abbiamo che

$$Y = \xi_d + \omega Z$$

dove adesso ξ_d è un vettore d-dimensionale e $\omega = diag(\omega_1, \omega_2, ..., \omega_d)$. La funzione di densità di Y sarà:

$$2\varphi_d((y - \xi); \Omega)\Phi(\alpha^T\omega^{-1}(y - \xi)) \tag{4.4}$$

con $\Omega = \omega\Omega_z\omega$ matrice di varianza e covarianza. Si indicherà con $Y \sim SN_k(\xi, \Omega, \alpha)$ la funzione di densità di Y.

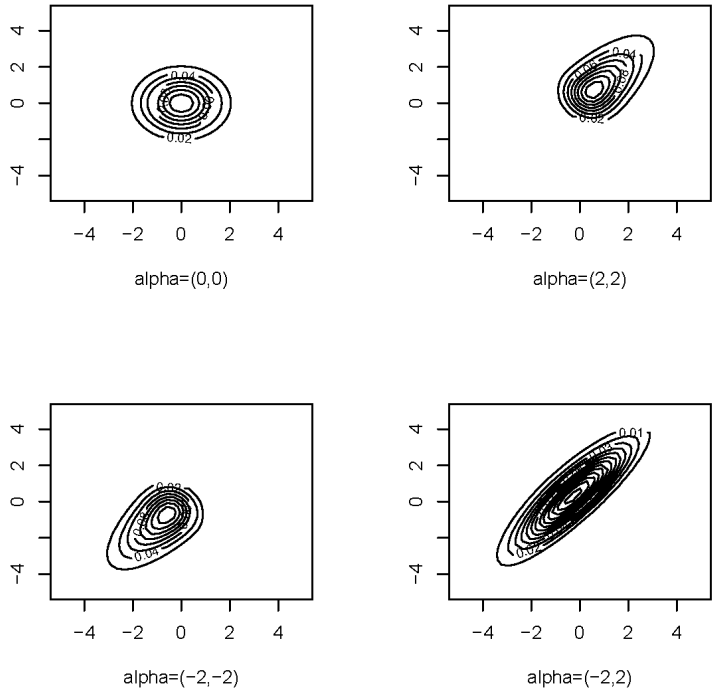

Figura 4.2: Grafico delle curve di livello di una distribuzione doppia Skew Normal per alcune scelte di α

4.3.1 Momenti

La funzione generatrice dei momenti per una $SN_k(\xi, \Omega, \alpha)$ è ottenibile sfruttando alcuni risultati ben noti, tra cui il seguente Lemma nell'estensione multidimensionale:

Lemma 4.7. *Se* $U \sim N_k(0, \Omega)$ *allora:*

$$E[\Phi(u + u^T U)] = \Phi(\frac{u}{\sqrt{1 + u^T \Omega u}})$$

per ogni u *scalare e* $v \in R^k$.

Utilizzando questi risultati si ottiene la funzione generatrice dei momenti:

$$M(t) = 2 \exp(\xi^T t + \frac{1}{2} t^T \Omega t) \Phi(\delta^T \omega t) \quad t \in R^k$$

dove

$$\delta = \frac{1}{(1+\alpha^T\overline{\Omega}\alpha)^{\frac{1}{2}}}\overline{\Omega}\alpha$$

e $\overline{\Omega}$ come per il caso scalare.

Derivando la seguente funzione generatrice dei cumulanti:

$$K(t) = t^T\xi + \frac{1}{2}t^T\Omega t + \zeta_0(\delta^T\omega t)$$

si ottiene

$$E[Y] = \xi + \omega\mu_z$$

$$Var[Y] = \Omega - \omega\mu_z\mu_z^T\omega$$

dove $\mu_z = \sqrt{\frac{2}{\pi}}\delta$ è il valore atteso della v.c.$Z = \omega^{-1}(Y-\xi) \sim SN_k(0, \overline{\Omega}, \alpha)$.

4.3.2 Parametrizzazione centrata

Come nel caso scalare, anche in quello multidimensionale viene proposta una parametrizzazione alternativa alla parametrizzazione diretta o DP (*direct parametrization*), per la quale la matrice di informazione attesa risulta singolare per alcune combinazioni dei parametri. Nell'articolo di Arellano-Valle e Azzalini (2008) viene presentata la parametrizzazione centrata o CP (*centred parametrization*), per la quale viene dimostrata la non singolarità della matrice di informazione attesa di Fisher. Per definire la CP si introducono nuovamente alcune espressioni seguendo lo schema di Azzalini e Capitanio (1999). Si definisce la v.c. "normalizzata"

$$Z = \omega^{-1}(Y-\xi) \sim SN_k(0, \overline{\Omega}, \alpha)$$

dove $\overline{\omega}$ e δ sono rispettivamente tali che (come per il caso scalare)

$$E[Z] = b\delta = \mu_z$$

$$Var[Z] = \overline{\omega} - \mu_z\mu_z^T = \overline{\omega} - b^2\delta\delta^T$$

dove $\sigma_z = diag(\sigma_{z,1}, ..., \sigma_{z,k})$, i cui termini sono la deviazione standard di Σ_z tale che il $j-esimo$ termine di $\sigma_{z,j} = (1-b^2\delta_{z,j}^2)^{\frac{1}{2}}$, per $j = (1, ..., k)$ e $b = \sqrt{\frac{2}{\pi}}$.

Ulteriori e importanti risultati ottenuti da Arellano-Valle e Azzalini (2008) sono le matrici di informazione osservata e attesa del Fisher per entrambe le parametrizzazioni, CP e DP.

4.3.3 Generazione

Come il caso scalare, anche quello multidimensionale si caratterizza per le diverse possibilità di generazione della famiglia SN_k.

4.3.4 Convoluzione di normali

Presa un v.c. multidimensionale Z tale che ogni sua componente si distribuisca come un SN, allora è naturale definire la distribuzione congiunta di Z una normale asimmetrica multidimensionale. Consideriamo una v.c. k-*dimensionale* $Y = (Y_1, ..., Y_k)^T$ con marginali standardizzate, indipendenti da una v.c. $Y_0 \sim N(0, 1)$:

$$\begin{pmatrix} Y_0 \\ Y \end{pmatrix} \sim N_{1+k}\{0, \begin{pmatrix} 1 & 0 \\ 0 & \Psi \end{pmatrix}\}$$

dove Ψ è matrice di correlazione $k \times k$. Si definisce

$$Z_j = \delta_j |Y_0| + (1 - \delta_j^2)^{\frac{1}{2}} Y_j \quad j = (1, ..., k)$$

dove $\delta_j \in (-1, 1)$. Quindi sfruttando il Lemma precedente risulta che

$$Z_j \sim SN(\frac{\delta_j}{\sqrt{1 - \delta_j^2}})$$

.
Allora è possibile scrivere che

$$SN \sim SN_k(0, \overline{\Omega}, \alpha)$$

. Per la relazione Ψ e $\overline{\Omega}$

Allora possiamo scrivere che

4.3.5 Metodo per condizionamento

Sia Y_0 una v.c. scalare e Y_1 una v.c. k-*dimensionale*, tale che

$$Y = \begin{pmatrix} Y_0 \\ Y_1 \end{pmatrix} \sim N_{1+k}(0, \Omega^*)$$

dove

$$\Omega^* = \begin{pmatrix} 1 & \delta^T \\ \delta & \overline{\Omega} \end{pmatrix}$$

e matrice di correlazione $k \times k$. Allora

$$Z = (Y_1 | Y_0 > 0) \sim SN_k(0, \overline{\Omega}, \alpha)$$

.

4.3.6 Proprietà

Molte proprietà della distribuzione SN semplice si estendono direttamente al caso multiplo. Valgono infatti i seguenti risultati

1. $\varphi_d(x; \Omega, 0) = \varphi_d(x; \Omega)$;
2. se $Z \sim SN(0, \overline{\Omega}, \alpha)$, allora $-XZ \sim SN(0, \overline{\Omega}, -\alpha)$;
3. $Z^T \overline{\Omega}^{-1} Z \sim \chi^2_d$ qualunque sia α.

In base alla rappresentazione del Lemma precedente, la v.c. $Y \sim SN_d(\xi, \Omega, \alpha)$ costituisce una perturbazione della distribuzione $N_d(\xi, \Omega)$. Mentre per il caso $d = 1$ l'effetto del parametro α sulla forma della densità è chiaramente identificato, per $d > 1$ la forma della densità è regolata congiuntamente ad α e da Ω.

4.3.7 Distribuzione normale asimmetrica k-dimensionale inversa

Una variabile casuale Z ha una $k - esima$ distribuzione inversa normale asimmetrica con parametri $\lambda_1, ..., \lambda_k \in R$ (si denota con $Z \sim SN_k(\lambda_1, ..., \lambda_k)$ per $k = 1, 2, ...$ se la sua funzione di densità è data da

$$\varphi^*_k(z) = c^*_k \phi(z)(1 - \Phi^*_{k-1}(\lambda_k z)) \tag{4.5}$$

dove $\Phi_{k-1}(.)$ è la funzione di distribuzione della distribuzione

$$RSN_{k-1}(\lambda_1, ..., \lambda_k)$$

$\Phi_0(.)$ è la funzione di distribuzione normale standard e

$$1/c_k^* = \int_{-\infty}^{\infty} \phi(z)(1 - \Phi_{k-1}^*(\lambda_k z))dz$$

4.4 Distribuzione normale asimmetrica chiusa

Un vettore casuale $Z = (Z_1, ..., Z_p)'$ ha una distribuzione normale asimmetrica chiusa con parametri $\mu \in R^p$, $\nu \in R^q$, $D \in R^p x R^q$, $\Delta \in R^p \times R^q$, scritta come $Z \sim CSN(\mu, \Sigma, D, \nu, \Delta)$ se la sua funzione di densità è

$$f(z) = \frac{\Phi_q(D(z-\mu); \nu, \Delta)}{\Phi_q(o_q; \nu, \Delta + D\Sigma D')} \phi_p(z; \mu, \Sigma) \tag{4.6}$$

per $z \in R^p$ dove $\phi_p(.; \xi, \Omega)$ e $\Phi_p(.; \xi, \Omega)$ sono rispettivamente la densità e la funzione di distribuzione della distribuzione $N_p(\xi, \Omega)$, Σ e Δ sono matrici definite positive, e $0_p = (0, ..., 0) \in R^q$

4.5 Distribuzione ellittica asimmetrica multivariata

Azzalini e Dalla Valle presentarono una teoria generale sulla versione multivariata di una distribuzione normale asimmetrica multivariata. Il loro articolo propone differenti metodi per generare distribuzioni SN. Esistono poi estensioni a distribuzioni ellittiche asimmetriche multivariate. Consideriamo un metodo condizionato per formare una nuova classe di distribuzioni. Consideriamo $X = (X_1, X_2, ..., X_k)^T$ un vettore casuale. Sia $X^* = (X_0, X^T)^T$ un vettore casuale $(k+1)$ dimensioni, tale che $X^* \sim El_{k+1}(\mu^*, \sum)$, dove $\mu^* = (0, \mu)$, $\mu = (\mu_1, ..., \mu_k)^T$ e Σ ha la forma:

$$\Sigma = \begin{pmatrix} 1 & \delta^T \\ \delta & \overline{\Omega} \end{pmatrix}$$

dove $\delta = (\delta_1, ..., \delta_k)^T$. Qui Ω è la matrice scalare associata al vettore X. Si afferma che che il vettore $Y = [X|X_0 > 0]$ ha una distribuzione ellittica asimmetrica e si indica con $Y \sim SA_k(\mu, \Omega, \delta)$, dove δ è il parametro di asimmetria. Se la densità di un vettore casuale X^* esiste e $P(X^* = 0) = 0$, allora la funzione di distribuzione di Y sarà:

$$f_Y(y) = 2 f_{g^k}(y) F_{g_{q(y)}}(\lambda^T(y - \mu))$$

dove $f_{g^k}(\cdot)$ è la funzione di distribuzione della $El_k(\mu, \Omega, g^{(k)})$ e $F_{g_{q(x)}}$ è la funzione di densità di $El_k(0, 1, g^{(k)})$ e $F_{g_{q(x)}}$, con

$$\lambda^T = \frac{\delta^T \Omega^{-1}}{(1 - \delta^T \Omega^{-1} \delta)^{\frac{1}{2}}}$$

$$g^{(k)(u)} = \frac{2\pi^{\frac{k}{2}}}{\Gamma(\frac{k}{2})} \int_0^\infty g^{(k+1)}(r^2+u) r^{k-1} dr \quad u \leq 0$$

$$g_{q(y)}(u) = \frac{g^{(k+1)}(u+q(y))}{g^{(k)}(q(y))}$$

e $q(y) = (y-\mu)^T \Omega^{-1}(y-\mu)$. In questo caso, si denota $Y \sim SE_k(\mu, \Omega, \delta; g^{(k+1)})$, dove $g^{(k+1)}$ è la funzione generatrice data in $g(k)(u)$ con k sostituito da $(k+1)$. Da λ^T e dal determinante positivo della matrice Σ, segue che δ e Ω devono soddisfare la condizione $\delta^T \Omega^{-1} \delta < 1$. Quindi

$$f_Y(y) = 2|\Omega|^{\frac{1}{2}} \int_{-\infty}^{\lambda^T(y-\mu)} g^{(k+1)}(r^2 + (y-\mu)^T \Omega^{-1}(y-\mu)) dr.$$

Capitolo 5
Simulazioni

Questo capitolo è interamente dedicato alle simulazioni eseguite allo scopo di confrontare le tecniche di stima sotto varie ipotesi di modello e studiarne quindi la robustezza, e osservare il comportamento dell'ICC in assenza di normalità. Per ottenere tutti i risultati sono stati utilizzati alcuni package del software R tra cui *lme*4, *nlme*, *lattice*, *sn*.

5.1 Scelta della dimensione campionaria

L'aumento della dimensione del campione rende più accurate le stime dei parametri e dei loro errori standard.
Kreft (1996) suggerisce come "*rule of thumb*" la cosiddetta "regola 30/30" secondo la quale per ottenere delle buone stime è opportuno avere un campione di almeno 30 gruppi ognuno dei quali costituito da almeno 30 individui.
Secondo Hox (1998) questa regola è valida se si è primariamente interessati alla stima degli effetti fissi; se l'interesse si concentra sulle interazioni *cross level*, il numero di gruppi dovrebbe essere ancora più elevato, risultando più appropriata una regola 50/20: circa 50 gruppi con circa 20 individui per gruppo; se l'attenzione, invece, ricade sulle componenti di varianza e covarianza, il numero dei gruppi deve essere considerevolmente più elevato: almeno 100 gruppi con 10 individui per gruppo.
Queste "*rules of thumb*" tengono conto del fatto che la raccolta dei dati implica il sostenimento dei relativi costi; pertanto se il numero dei gruppi aumenta, il numero di unità diminuisce. Bisogna però sottolineare che i costi per unità campionaria non sono costanti: intervistare 50 operai in una fabbrica piuttosto che 10 in 5 fabbriche diverse non comporta il sostenimento dei medesimi costi.
Snijders e Bosker (1994) discutono il problema della scelta della dimensione campionaria nel caso di un modello a due livelli tenendo in considerazione i costi per il reperimento dei dati.

Considerando quanto noto in letteratura, nelle simulazioni si è stabilito di generare campioni di 10, 30, 50 gruppi di 5, 10, 30 soggetti. I dati sono stati generati secondo un modello multilivello a due livelli, seguendo la procedura di Pinheiro e Bates (1998, 2000), sia con errori distribuiti secondo una v.c. normale, sia con errori distribuiti secondo una v.c *skew-normal* (con parametro di asimmetria α posto a vari livelli).

5.2 Simulazione ICC

In base a quanto noto in teoria si è ritenuto di generare un campione di 100 gruppi di 50 soggetti ciascuno secondo il modello multilivello a due livelli con una variabile esplicativa per livello. Il modello è stato dapprima generato considerando gli errori distribuiti secondo la v.c. normale; successivamente secondo la v.c. *skew normal*. Utilizzando un metodo sia grafico, che boostrap, si è voluto verificare per via empirica se la presenza di gruppi nei dati, generati con errori distribuiti normalmente e come una SN, veniva rilevata con la stessa efficacia. Un modo per misurare l'influenza dei gruppi è quello di confrontare la distribuzione delle medie dei gruppi, con quella di pseudo gruppi di individui assegnati casualmente. Se tutte le prime coincidono con le seconde non risulta esserci evidenza empirica degli effetti di gruppo. Se solo una o due di queste sono chiaramente diverse allora l'ICC non evidenzia empiricamente la differenza tra gruppi, ma solo la presenza di gruppi anomali.

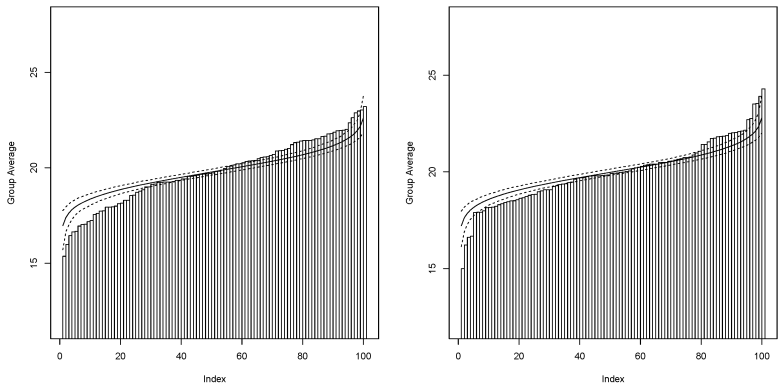

Figura 5.1: Simulazione ICC per modello multilivello con errori distribuiti normalmente (grafico a sinistra) e come una SN (grafico a destra)

Nel grafico 5.1 con tratto più evidente è indicata la distribuzione delle me-

die dei gruppi generati, mentre le linee tratteggiate ne indicano l'intervallo di confidenza bootstrap per le medie vere. Il grafico di sinistra in Fig. 5.1 riporta la situazione dei dati generati con il modello multilivello con errori distribuiti secondo una v.c. normale, mentre quello di destra con errori SN. L'ICC relativo al modello di sinistra è pari a 0,0282 mentre quello relativo al modello di destra è pari a 0,0245.
Sia graficamente che numericamente si nota che vi è minor evidenza nell'identificare la effettiva presenza di gruppi nel caso di modelli multilivello con errori distribuiti come una v.c. SN rispetto ai modelli classici con errori distribuiti normalmente.
La violazione dell'assunzione di osservazioni indipendenti porta ad una dimensione reale degli errori del *I tipo* più alta rispetto al valore nominale del 5% percento nominale (Barcikovski, R.S. 1981).
Nel suo articolo Barcikovski indica che a parità di numerosità di gruppo n_j, il valore di α aumenta all'aumentare del valore di ρ come dettagliato in tabella (5.1)

	ρ			
n_j	0,00	0,01	0,05	0,20
10	0,05	0,06	0,11	0,28
25	0,05	0,08	0,19	0,46
50	0,05	0,11	0,30	0,59
100	0,05	0,17	0,43	0,70

Tabella 5.1: Barcikovski, R.S. (1981) One Way ANOVA (2 gruppi)

Ne deriva che, se non si considera in maniera adeguata la presenza di ICC nei dati da analizzare, gli standard error dei parametri stimati risulteranno sottostimati.

5.3 Stima dei parametri: simulazioni e robustezza

Vengono ora riportati i risultati delle simulazioni per ogni combinazione possibile di parametri in ordine crescente di numerosità J dei gruppi. In parentesi viene riportato il vero valore del parametro (stimato con il metodo REML). Di seguito sono riportate le tabelle relative al modello generato con errori distribuiti secondo una v.c. normale.

Nelle successive tabelle si riportano i risultati delle simulazioni con dati generati considerando 10-30-50 gruppi di 5-10-30 soggetti ciascuno, ma con errori distribuiti secondo una v.c. *skew-normal*. Nelle simulazioni sono stati considerati diversi valori del parametro α di asimmetria. Dato il gran numero di combinazioni possibili si è ritenuto opportuno, sia in termini di spazio uti-

Normale $J = 10$						
	$n_j = 5$		$n_j = 10$		$n_j = 30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,01	0,441	20,00	0,39	20,01	0,40
$\gamma_{10}(1)$	1,00	0,43	1,01	0,59	1,01	0,24
$\gamma_{01}(1)$	1,01	0,44	1,01	0,40	1,00	0,40
$\gamma_{11}(8)$	8,01	0,42	7,98	0,59	8,01	0,24
$\sigma^2(1)$	0,73		0,86		0,95	
$\tau_{00}(1)$	0,87		0,91		0,95	
$\tau_{11}(1)$	0,86		0,89		0,94	
$\tau_{01}(0,77)$	0,70		0,73		0,72	

Normale $J = 30$						
	$n_j = 5$		$n_j = 10$		$n_j = 30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,21	20,00	0,21	20,01	0,19
$\gamma_{10}(1)$	1,01	0,26	1,00	0,20	1,00	0,20
$\gamma_{01}(1)$	0,99	0,23	1,00	0,21	1,01	0,20
$\gamma_{11}(8)$	8,01	0,26	8,01	0,20	8,00	0,21
$\sigma^2(1)$	0,76		0,84		0,94	
$\tau_{00}(1)$	0,84		0,91		0,97	
$\tau_{11}(1)$	0,80		0,90		0,97	
$\tau_{01}(0,77)$	0,72		0,75		0,76	

Normale $J = 50$						
	$n_j = 5$		$n_j = 10$		$n_j = 30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,16	19,95	0,15	20,00	0,15
$\gamma_{10}(1)$	1,00	0,16	1,00	0,15	1,00	0,15
$\gamma_{01}(1)$	1,01	0,16	1,00	0,15	1,00	0,15
$\gamma_{11}(8)$	8,00	0,17	8,00	0,15	8,00	0,15
$\sigma^2(1)$	0,75		0,84		0,94	
$\tau_{00}(1)$	0,84		0,91		0,97	
$\tau_{11}(1)$	0,79		0,91		0,96	
$\tau_{01}(0,77)$	0,70		0,75		0,77	

lizzato che di comprensione, riportare solo alcuni dei risultati ottenuti. Come per le simulazioni precedenti le stime sono state calcolate secondo REML.

Nelle simulazioni entrambi i modelli sono stati stimati supponendo che la parte casuale del modello si distribuisca come una v.c. normale, anche nel caso in cui il modello era stato generato con errori distribuiti con una v.c. *skew-normal*. In entrambi i casi le stime degli effetti fissi sono molto simili, così come per la stima di σ^2. Risultano molto diverse invece le stime di Σ. Quando l'ipotesi di normalità non è soddisfatta, gli stimatori dei parametri risultano infatti consistenti, ma non efficienti, mentre gli stimatori degli errori standard non sono consistenti (Goldstein, 1995, pag. 22).

$SN\ \alpha(2,2)\ \alpha(0)\ J=10$						
	$n_j=5$		$n_j=10$		$n_j=30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,03	20,00	0,20	20,00	0,02
$\gamma_{10}(1)$	1,00	0,16	1,00	0,31	1,00	0,16
$\gamma_{01}(1)$	1,00	0,09	1,00	0,22	1,00	0,04
$\gamma_{11}(8)$	8,00	0,17	8,00	0,31	8,00	0,16
$\sigma^2(1)$	0,73		0,83		0,94	
$\tau_{00}(1)$	0,83		0,91		0,97	
$\tau_{11}(1)$	0,78		0,89		0,96	
$\tau_{01}(0,77)$	0,63		0,66		0,67	

$SN\ \alpha(2,2)\ \alpha(0)\ J=30$						
	$n_j=5$		$n_j=10$		$n_j=30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,03	20,00	0,09	20,00	0,03
$\gamma_{10}(1)$	1,00	0,17	0,99	0,22	1,00	0,20
$\gamma_{01}(1)$	1,00	0,09	1,00	0,12	1,00	0,04
$\gamma_{11}(8)$	7,99	0,18	7,99	0,21	8,00	0,20
$\sigma^2(1)$	0,71		0,84		0,94	
$\tau_{00}(1)$	0,85		0,90		0,97	
$\tau_{11}(1)$	0,85		0,88		0,97	
$\tau_{01}(0,77)$	0,64		0,65		0,67	

$SN\ \alpha(2,2)\ \alpha(0)\ J=50$						
	$n_j=5$		$n_j=10$		$n_j=30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,03	20,00	0,05	20,00	0,01
$\gamma_{10}(1)$	1,00	0,15	1,00	0,10	1,01	0,14
$\gamma_{01}(1)$	1,00	0,08	1,00	0,07	1,01	0,03
$\gamma_{11}(8)$	8,01	0,15	8,00	0,18	8,00	0,14
$\sigma^2(1)$	0,72		0,84		0,94	
$\tau_{00}(1)$	0,83		0,90		0,97	
$\tau_{11}(1)$	0,79		0,89		0,97	
$\tau_{01}(0,77)$	0,62		0,65		0,67	

$SN\ \alpha(2,2)\ \alpha(2)\ J=10$						
	$n_j=5$		$n_j=10$		$n_j=30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,06	20,00	0,09	20,00	0,03
$\gamma_{10}(1)$	1,00	0,37	1,01	0,22	0,99	0,33
$\gamma_{01}(1)$	0,99	0,21	1,01	0,13	1,00	0,06
$\gamma_{11}(8)$	8,00	0,41	8,00	0,22	8,00	0,33
$\sigma^2(1)$	0,74		0,84		0,94	
$\tau_{00}(1)$	0,82		0,89		0,96	
$\tau_{11}(1)$	0,79		0,90		0,96	
$\tau_{01}(0,77)$	0,56		0,60		0,65	

$SN\ \alpha(2,2)\ \alpha(2)\ J=30$						
	$n_j=5$		$n_j=10$		$n_j=30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,07	20,00	0,04	20,00	0,02
$\gamma_{10}(1)$	0,99	0,23	0,99	0,24	1,00	0,27
$\gamma_{01}(1)$	1,00	0,12	1,00	0,08	1,00	0,04
$\gamma_{11}(8)$	8,01	0,24	7,99	0,23	8,00	0,27
$\sigma^2(1)$	0,73		0,84		0,94	
$\tau_{00}(1)$	0,83		0,91		0,96	
$\tau_{11}(1)$	0,80		0,89		0,96	
$\tau_{01}(0,77)$	0,61		0,64		0,66	

$SN\ \alpha(2,2)\ \alpha(2)\ J=50$						
	$n_j=5$		$n_j=10$		$n_j=30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,03	20,00	0,02	20,00	0,02
$\gamma_{10}(1)$	1,00	0,16	1,00	0,17	0,99	0,15
$\gamma_{01}(1)$	1,00	0,08	1,00	0,05	1,00	0,03
$\gamma_{11}(8)$	8,00	0,15	8,01	0,17	8,00	0,15
$\sigma^2(1)$	0,73		0,83		0,94	
$\tau_{00}(1)$	0,83		0,91		0,97	
$\tau_{11}(1)$	0,81		0,91		0,96	
$\tau_{01}(0,77)$	0,64		0,66		0,67	

$SN\ \alpha(10,10)\ \alpha(10)\ J=10$						
	$n_j=5$		$n_j=10$		$n_j=30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,12	20,00	0,05	20,00	0,19
$\gamma_{10}(1)$	0,98	0,37	1,00	0,30	0,99	0,44
$\gamma_{01}(1)$	0,99	0,23	1,00	0,11	1,00	0,21
$\gamma_{11}(8)$	7,99	0,41	8,00	0,30	7,99	0,46
$\sigma^2(1)$	0,74		0,84		0,94	
$\tau_{00}(1)$	0,86		0,90		0,95	
$\tau_{11}(1)$	0,83		0,92		0,95	
$\tau_{01}(0,77)$	0,56		0,62		0,59	

$SN\ \alpha(10,10)\ \alpha(10)\ J=30$						
	$n_j=5$		$n_j=10$		$n_j=30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,06	20,00	0,04	20,00	0,02
$\gamma_{10}(1)$	1,00	0,18	1,01	0,18	1,00	0,18
$\gamma_{01}(1)$	1,00	0,11	1,04	0,07	1,00	0,04
$\gamma_{11}(8)$	8,01	0,19	8,00	0,19	8,00	0,17
$\sigma^2(1)$	0,73		0,84		0,94	
$\tau_{00}(1)$	0,82		0,90		0,97	
$\tau_{11}(1)$	0,81		0,91		0,97	
$\tau_{01}(0,77)$	0,60		0,63		0,64	

$SN\ \alpha(10,10)\ \alpha(10)\ J=50$						
	$n_j=5$		$n_j=10$		$n_j=30$	
PARAMETRI	media	var	media	var	media	var
$\gamma_{00}(20)$	20,00	0,04	20,00	0,03	20,00	0,01
$\gamma_{10}(1)$	1,00	0,19	1,00	0,15	1,00	0,16
$\gamma_{01}(1)$	1,00	0,09	1,00	0,06	1,00	0,03
$\gamma_{11}(8)$	8,00	0,20	7,99	0,15	7,99	0,17
$\sigma^2(1)$	0,73		0,83		0,94	
$\tau_{00}(1)$	0,91		0,90		0,96	
$\tau_{11}(1)$	0,75		0,88		0,97	
$\tau_{01}(0,77)$	0,58		0,63		0,64	

5.4 Cenni sulle misure di adattamento del modello multilivello

Diverse statistiche possono essere utilizzate per valutare l'adattamento dei modelli. La stima dei parametri nei modelli multivel (intercetta, coefficienti di regressione e componenti della varianza) viene generalmente realizzata attraverso il metodo della Massima Verosimiglianza. Questo metodo produce stime asintoticamente efficienti e consistenti. Inoltre, con grandi campioni, le stime di Massima Verosimiglianza sono generalmente robuste rispetto a leggere violazioni dell'assunzione di normalità distributiva degli errori. Questo viene confermato anche dagli studi simulativi effettuati per studiare la robustezza delle stime in assenza di normalità, in particolare con distribuzione SN.

5.4.1 Principali test d'ipotesi nei modelli multilivello

5.4.2 Test di Wald

Uno dei test più utilizzati per la verifica di ipotesi nei modelli di regressione multilevel è il test di Wald, in cui la statistica test, Z, viene calcolata rapportando la stima puntuale del parametro di interesse all'errore standard della stima stessa. La distribuzione di riferimento per la statistica Z è la normale standardizzata. Il test di Wald si basa sull'assunto che i parametri sottoposti a verifica di ipotesi abbiano una distribuzione campionaria normale, con una varianza campionaria che può essere stimata a partire dalla matrice di informazione. Come discusso da Fears et al. (1996), in situazioni particolari, la statistica di Wald non risulta adatta per test sulle componenti della varianza, soprattutto nei casi in cui queste non siano prossime allo zero o nei casi in cui la numerosità campionaria sia molto ridotta. Si precisa, inoltre, che gli errori standard utilizzati per la costruzione del test sono di natura

asintotica, pertanto sono effettivamente validi solo per grandi campioni[1]. E' opportuno precisare che nelle regressioni multilivello, la numerosità campionaria, rilevante per i coefficienti di regressione e le componenti della varianza di secondo livello, è costituita dal numero dei gruppi che, generalmente, non è molto elevato.

5.4.3 Deviance Test

Il *deviance test* o anche *likelihood ratio test* si basa sul principio che, quando i parametri di un modello statistico sono stimati attraverso il metodo *maximum likelihood* (ML), la stima fornisce la *likelihood*. La devianza è definita come: $-2\log(L)$, dove per verosimiglianza si intende il valore della funzione di verosimiglianza alla convergenza. I modelli con devianza inferiore presentano un miglior adattamento ai dati. Se due modelli sono annidati, ovvero se un modello specifico può essere derivato da un modello più generale rimuovendo uno o più parametri dal modello generale, è possibile confrontare statisticamente i due modelli utilizzando le loro devianze. Infatti, la differenza tra le devianze di due modelli annidati, sotto l'ipotesi nulla di equivalenza tra i due modelli, si distribuisce come un $Chi-quadrato$ con gradi di libertà pari alla differenza nel numero dei parametri stimati dai due modelli. Questa proprietà può essere utilizzata per verificare l'ipotesi che l'adattamento ai dati del modello più generale sia significativamente superiore all'adattamento ai dati del modello specifico.

In genere non si considera direttamente il valore della devianza, ma le differenze nelle devianze di diversi modelli applicati agli stessi dati, ad esempio:

- M_0 è il modello con m_0 parametri e devianze D_0;
- M_1 è il modello con m_1 parametri e devianze D_1

Il test considerato sarà:

$$D_0 - D_1 = -2\ln L_0 + 2\ln L_1$$

dove L_0 è il modello nullo (modello ad intercetta casuale).

L'ipotesi nulla sarà:

$$H_0 : D_0 - D_1 = 0$$

sotto la quale $D_0 - D_1 \sim \chi^2$ *con* $m_1 - m_0$ *gradi di liberta'*.

[1] Non si conosce con precisione quale sia la numerosità campionaria sufficiente affinchè gli errori standard possano essere considerati accurati; per gli approfondimenti su questo tema si rimanda agli studi di simulazione condotti da Van der Leeden et al. (1997).

Questo test può essere applicato sia alla parte fissa che alla parte casuale del modello. Se la devianza è stata calcolata in base al criterio di stima *Residual ML*, si possono effettuare confronti solo tra modelli che presentano stessa parte fissa e differiscono solo nella parte casuale. Il test Chi-quadrato delle devianze è asintoticamente equivalente al test di Wald, ma risulta più efficace di questo per sottoporre a verifica di ipotesi le componenti di varianza. I test eseguiti sulle componenti della varianza, sia nel caso del test di Wald che nel caso del test Chi-quadrato delle devianze, sono "ad una coda" dal momento che le varianze non possono essere negative (Berkhof e Snijders, 2001).

5.4.4 Akaike Information Criterion

Un indice di adattamento finalizzato al confronto di modelli non annidati è costituito dal criterio di informazione di Akaike (*Akaike's Information Criterion* - AIC). Questo indice è basato sul principio di parsimonia, secondo il quale i modelli semplici sono da preferire rispetto ai modelli complessi; pertanto aggiusta il confronto statistico tra i modelli attraverso il numero dei parametri stimati. Nel caso dei modelli di regressione multilivello il criterio di informazione di Akaike assume la seguente forma:

$$AIC = d + 2q$$

dove d è la devianza del modello e q è il numero di parametri stimati. L'utilizzo di questo indice presuppone che i modelli posti a confronto facciano riferimento allo stesso data-set, utilizzando lo stesso metodo per la stima dei parametri. Un indice di adattamento piuttosto simile all'AIC è costituito dal criterio di informazione di Schwarz (*Schwartz's Bayesian Information Criterion* - BIC) che è dato da:

$$BIC = d + q\log(N)$$

dove N rappresenta il numero di osservazioni.
Dal momento che i dati gerarchicamente organizzati presentano, ai vari livelli, differenti numerosità campionarie, nelle analisi multilivello si preferisce utilizzare l'indice AIC poichè è molto più semplice da calcolare. In generale, dato un certo numero di modelli da confrontare, saranno da preferire quelli che presentano i valori più bassi di AIC o BIC.

5.5 Confronto tra i metodi di stima

Per quanto concerne i parametri fissi, tutte le simulazioni riportano che le stime OLS, GLS e di massima verosimiglianza (FML e RML) sono non distorte. L'unica differenza riguarda l'efficienza: le stime OLS sono meno efficienti, ma solo per dataset non grandi. La potenza del test di Wald per la significatività dei coefficienti fissi di primo livello dipende dalla numerosità del campione totale; quella per la significatività degli effetti di più alto livello e delle interazioni *cross level* dipende soprattutto dal numero dei gruppi. E' noto che le stime OLS degli errori standard di γ sono distorte per difetto. Lo studio di Barcikowski (1981) mostra, almeno per l'analisi della varianza, in che modo i risultati sono distorti in presenza di *intra class correlation*. le simulazioni indicano che quando le assunzioni di normalità e elevata numerosità del campione non sono soddisfatte, le stime di massima verosimiglianza risultano ancora non distorte, mentre gli errori standard sono distorti per difetto.

Per i parametri casuali il confronto fra le stime ML e REML evidenzia delle differenze. Le stime REML sono meno distorte, ma anche meno efficienti. Il trade-off tra i due metodi è molto complicato e gli autori delle simulazioni non sono stati in grado di affermare quando è meglio utilizzare un metodo piuttosto che l'altro. In contrasto con le conclusioni raggiunte per i parametri fissi, le iterazioni migliorano le stime degli effetti casuali.

5.6 Modello scelto dopo la simulazione

Si illustrerà ora la scelta del modello in base alla valutazione dell'accostamento dello stesso ai dati. Questi saranno generati mediante un modello multilivello, a due livelli, con una variabile esplicativa per ciascun livello (modello $M5$). Questi dati saranno stimati ipotizzando modelli differenti ($M0$, $M1$, $M2$, $M4$) dal modello con cui si sono generati, si veda a tal proposito la tabella 5.2. In questa tabella i dati sono stati generati mediante il modello $M5$ modello multilivello, a due livelli, con una variabile esplicativa per ciascun livello e con errori aventi distribuzione normale. Sempre con riferimento alla tabella 5.2 con il suffisso ML vengono indicate le stime del modello mediante il metodo della Massima Verosimiglianza, mentre $M0$, $M1$, $M2$ e $M4$ sono stati stimati mediante il $REML$.

Nella tabella 5.3 i dati sono stati generati mediante il modello $SNM5$, modello multilivello a due livelli, con una variabile esplicativa per ciascun livello e con errori aventi distribuzione SN. Anche qui il il suffisso ML vengono indicate le stime del modello mediante il metodo della Massima Verosimiglianza mentre $SNM0$, $SNM1$, $SNM2$ e $SNM4$ sono stati stimati mediante il $REML$.

I casi possibili posso essere numerosi, si sono riportate solo la tabella (5.2) e la tabella (5.3) a titolo esemplificativo. Come nei restanti casi, qui non

	AIC	BIC	$-logL$
M0	374.02	379.76	184.01
M1	375.82	383.46	183.91
M1ML	375.82	383.46	183.91
M2	216.79	228.26	102.40
M2ML	216.72	228.20	102.36
M4	181.18	194.57	83.59
M4ML	180.89	194.27	83.44
M5	**172.05**	**187.34**	**78.02**
M5ML	**171.61**	**186.91**	**77.81**

Tabella 5.2: Stima del modello $M5$ con errori distribuiti come una v.c. normale

	AIC	BIC	$-logL$
SNM0	375.59	381.32	184.79
SNM1	375.56	383.21	183.78
SNM1ML	375.56	383.21	183.78
SNM2	231.93	243.40	109.96
SNM2ML	231.86	243.33	109.93
SNM4	197.41	210.80	91.71
SNM4ML	197.15	210.53	91.57
SNM5	**194.45**	**209.75**	**89.23**
SNM5ML	**194.15**	**209.44**	**89.07**

Tabella 5.3: Stima del modello $SNM5$ con errori distribuiti come una v.c. SN

riportati, si osserva una minore evidenza empirica della falsità del modello nel caso si consideri un modello $SNM5$, modello generato con errori distribuiti come una SN. Il modello migliore risulta essere quello con AIC, BIC, $-LogL$ inferiore.

Considerando il modello con errori distribuiti secondo una v.c. normale si nota in fatti che:

	AIC	BIC	$-LogL$
M0	374.02	379.76	184.01
M5ML	171.61	186.91	77.81

Tabella 5.4: Confronto indici di adattamento del modello multilivello $M5$ con errori distribuiti come una v.c. normale

Mentre se gli errori si distribuiscono come una v.c. SN $\alpha(2,2)$ al secondo livello, e al primo livello si distribuiscono come una v.c. $\alpha(1)$ si nota che:

	AIC	BIC	$-LogL$
M0	375.59	381.32	184.79
M5ML	194.15	209.44	89.07

Osservando anche quando gli errori si distribuiscono come una v.c. SN $\alpha(10, 10)$al secondo livello, mentre al primo livello si distribuiscono come una v.c. $\alpha(10)$ si nota che:

	AIC	BIC	$-LogL$
M0	377.10	382.83	185.584
M5ML	181.07	196.36	82.533

Quindi come evidenziato da questi esempi, nel caso in cui cade la normalità si ha minore evidenza della falsità del modello. Infatti l'AIC, il BIC e il $-LogL$ risultano più alti nei modelli con errori SN rispetto a quelli con errori normali.

Capitolo 6
Conclusioni

La metodologia dei modelli multilivello consente l'analisi di dati organizzati in una struttura di tipo gerarchico, ossia di dati raggruppati. Nella teoria "classica" si introduce ai fini inferenziali l'ipotesi distributiva normale per le componenti d'errore. Tuttavia, in alcuni casi tale assunzione può rilevarsi troppo restrittiva. Nel presente lavoro è proposta, quale alternativa alla normale, la distribuzione *skew-normal* e una sua generalizzazione nel caso multidimensionale. Si è messa in luce l'utilità di impiego di queste distribuzioni nell'ambito dell'analisi multilivello, e si sono evidenziati i principali problemi legati alla stima dei parametri.

Se la stima del coefficiente di correlazione intraclasse mostra un grado di omogeneità tra i gruppi non trascurabile, non è sufficiente utilizzare un semplice modello ANOVA a effetti casuali, la cui caratteristica principale è quella di individuare la quota di variabilità attribuibile ai diversi livelli della gerarchia. Con il modello cosiddetto nullo non si è in grado di spiegare quali sono i fattori che determinano le differenze nei vari gruppi. Per questo motivo il modello deve essere adattato introducendo variabili esplicative sia a livello individuale che di gruppo. Per quanto riguarda le covariate che si introducono nel modello multilivello si ritiene opportuno sottolieneare che:

- il modello non deve contenere un numero eccessivo di variabili esplicative in quanto la correlazione tra le stesse renderbbe le stime dei parametri non affidabili, cioè tali per cui piccoli cambiamenti nel modello o nei dati campionari potrebbero comportare grandi cambiamenti nei risultati.
- le variabili esplicative non devono essere scelte a caso, ma sulla base della conoscenza del problema (è quindi sempre necessaria un'analisi esplorativa preliminare) e della teoria.

Il modello multilevel inoltre si distingue dal modello di regressione lineare in quanto oltre alle componenti fisse il modello si caratterizza per la presenza e la complessità della parte casuale. In un modello di regressione ai minimi quadrati ordinari è presente una sola componente di varianza: introducendo nel modello le variabili esplicative si può solo verificare la porzione di quell'u-

nica componente di variabilità che viene spiegata. In un modello multilevel invece la variabilità osservata è scomponibile in più fonti e, quando si introducono covariate, si può verificare in che modo esse abbiano influito sulla riduzione della variabilità ai vari livelli. L'introduzione di variabili esplicative di primo livello può comportare la riduzione della varianza di entrambi i livelli; di solito ci si attende che la riduzione della varianza maggiore si abbia al livello più basso, ma non sempre questo succede.

La caratteristica che va maggiormente apprezzata della modellistica multilivello riguarda la possibilità di un'analisi più accurata della variabilità, resa possibile da una definizione più articolata (e più rispondente al vero) della parte casuale del modello rispetto a un modello di regressione ordinaria. L'altro vantaggio è la stima più corretta dell'errore standard delle stime dei coefficienti fissi: la base per calcolare tali errori è sempre la parte casuale del modello. Se essa non rispecchia la realtà, come nel caso di un modello OLS, in cui si assume che gli errori siano omoschedastici e incorrelati, allora le stime risultano poco accurate, soprattutto quando il grado di omogeneità dei gruppi, misurato dal coefficiente di correlazione intraclasse, è molto elevato. La modellistica multilivello costituisce una soluzione al problema della scelta dell'appropriato livello di analisi: è sbagliato porsi il problema della scelta tra "analisi individuale" e "analisi ecologica", perchè a tutti i livelli sono presenti degli effetti strutturali e una variazione casuale. Quindi la modellistica multilevel è necessaria per rappresentare esplicitamente queste caratteristiche.

Con riguardo alla parte aleatoria del modello multilivello, nel presente lavoro si è ritenuto opportuno utilizzare la SN per la necessità di considerare alcune distribuzioni che potessero assumere i più svariati andamenti, adattandosi in modo più appropriato alle situazioni presenti in natura. Queste distribuzioni comprendono al loro interno anche la distribuzione normale. Quest'ultima può sempre essere tenuta in considerazione come ipotetica distribuzione dei dati e utilizzarla a ragion veduta.

Dal punto di vista inferenziale l'introduzione del parametro α, di asimmetria, ha permesso da un lato un migliore approccio all'analisi, potendo "manipolare", anche se non direttamente, la simmetria della distribuzione.
Gli algoritmi utilizzati dai più conosciuti package statistici utilizzano tuttavia metodi di stima che presuppongono errori, ad entrambi i livelli del modello multilevel, distribuiti secondo una v.c. normale. Come è risultato dalle simulazioni, in presenza di errori distribuiti secondo una SN si denota una difficoltà nel riconoscimento del vero modello sottostante alla generazione dei dati. Risulta anche di maggior difficoltà il riconoscimento della presenza di ICC col conseguente aumento degli errori di *I tipo*. Con riferimento alla stima dei veri parametri che caratterizzano il modello, risulta problematica la stima di varianze e covarianze Σ degli errori di secondo livello, per modelli generati con errori SN, soprattutto se J è molto piccolo. Inoltre, quando l'ipotesi di normalità non è soddisfatta, gli stimatori dei parametri, pur essendo consistenti, risultano non efficienti, mentre gli stimatori degli errori standard

non sono consistenti (Goldstein, 1995 pag 22).

Tra gli obiettivi ed i futuri sviluppi degli argomenti trattati vi è da approfondire il metodo di stima dei parametri di un modello multilivello con gli errori di entrambi i livelli distribuiti come una SN. A tale scopo occorre innanzitutto superare il problema critico della non chiusura della classe SN rispetto alle convoluzioni. Questo sarebbe possibile considerando la formulazione generale *SUN/Closed SN* (*Unified Skew Normal*) cioè una formulazione generale che unisce le formulazioni alternative della *Skew Normal*, ma chiusa rispetto a convoluzioni.

Bibliografia

[1] Aitkin M., Anderson D., Hinde J., (1981) *Statistical modelling of data on teaching style (con discussione)*, J.R. Statist. Soc. A, 144, 419-461.

[2] Aitkin M., Longford N. T. (1986) *Statistical modeling issues in school effectiveness studies (con discussione)*, J.R. Statist. Soc. A, 149, 1-43.

[3] Alker H. R., (1969) *A typology of fallacies*, in M. Dongan e S. Rokkan, Eds. Quantitative ecological analysis, in The Social Science, Cambridge Ma. M.I.T. Press.

[4] Andersen R., Aitkin M. (1985) *Variance components models with binary response: interviewer variability*, Journal of the Royal Statistical Society, B, Vol. 47, 203-210.

[5] Andersen R., Heath A. (2002) *Class matters. The persisting effects of contextual social class on individual voting in Britain*, In European Sociological Review, Vol. 18, 1964-97.

[6] Anderson T. W., (1984) *Estimating Linear Statistical Relationships*, The Annals of Statistics, 12, 1-45.

[7] Arellano Valle R. B., Azzalini A. (2006) *On the unification of families of skew-normal distributions* Scand. J. Statist., 33, 561-574.

[8] Arellano Valle R. B., Azzalini A. (2008) *The centred parametrization for the multivariate skew-normal distribution* J. Multivariate Anal., 99, 1362-1382.

[9] Arellano Valle R. B., Bolfarine H., Lachos V. H. (2005a) *Skew normal linear mixed models* Journal of data Science, 3, 415-438.

[10] Arellano Valle R. B., Branco M. D., Genton M. G. (2006) *A unified view on skewed distributions arising from selections* Canad. J. Statist., 34, 581-601.

[11] Arellano Valle R. B., Genton M. G., (2005) *On fundamental skew distributions* J. Multivariate Anal., 96, 93-116.

[12] Arellano Valle R. B., Gomez H. W., Quintana F. A. (2004) *A new class of skew-normal distributions* Communications in Statistics: Theory and Methods, 33, 1465-1480.

[13] Arellano Valle R. B., Gomez H. W., Quintana F. A. (2005b) *Statistical inference for a general class of asymmetric distributions* J.Statist. Plann. Inference, 128, 427-443.

[14] Arellano Valle R. B., Ozan S., Bolfarine H., Lachos V. H., (2005c) *Skew-normal measurement error models* J. Multivariate Anal., 96, 265-281.

[15] Arnold B. C., Beaver R. J., (2000c) *Some skewed multivariate distributions* Amer. J. of Mathematical and Management Sciences, 20, 27-38.

[16] Arnold B. C., Cox D., Bottai M., Robins J., (1993) *The non truncated marginal of a truncated bivariate normal distribution* Psychometrika, 58, 471-488.

[17] Azzalini A., (1985) *A class of distributions which includes the normal ones*, Scand. J. Statist., 12, 171-178.

[18] Azzalini A., (1986) *Further results on a class of distributions which includes the normal ones*, Statistica XLVI, 199-208.

[19] Azzalini A., (2005) *The skew-normal distribution and related multivariate families*, Scand. J. Statist., 32(2), 159-188.

[20] Azzalini A., Capitano A., (1999) *Statistical applications of the multivariate skew-normal distribution*, Scand. J. Statist., 61, 579-602.

[21] Azzalini A., Chiogna M., (2004) *Some results on the stress-strength model for skew-normal variates*, Metron LXII, 315-326.

[22] Azzalini A., Dalla Valle A., (1996) *The multivariate skew-normal distribution*, Biometrika, 83, 715-726.

[23] Azzalini A., (2006) *Some recent developments in the theory of distributions and their applications*, Atti della XLIII Riunione della Società Italiana di Statistica, 51-64.

[24] Berkhof J., Snijders T. A. B., (2001) *Variance component testing in multilevel models*, Journal of Educational and Behavioral Statistics, 26, 133-152.

[25] Birnbaum Z. W., (1950) *Effect of linear truncation on a multinormal population*, Ann. Math. Statist., 21, 272-279.

[26] Capitano A., Azzalini A., Stanghellini E., (2003) *Graphical models for skew-normal variates*, Scand. J. Statist., 30, 129-144.

[27] Chiogna M., (1998) *Graphical models for skew-normal variates*, J. Ital. Statist. Soc., 7, 1-13.

[28] Barcikowski R. S., (1981) *Statistical power with group mean as the unit of analysis*, Journal of the Educational Statistics, 6 (3), 267-285.

[29] Bates D., (2005) *Fitting linear mixed models in R. R News*, R News, 5(1):27-30, 2005.

[30] Bennet N., (1976) *Teaching styles and pupil progress*, London, Open Books.

[31] Birnbaum Z. W., (1950) *Effect of linear truncation on a multinormal population*, Ann. Math. Statist., 21, 272Ű279.

[32] Boyd L. H., Iverson G. R., (1979) *Contextual analysis: Concepts and statistical techniques*, Belmont, CA: Wadsworth.

[33] Breslow N. E., Clayton D. G., (1993) *Approximate inference in generalised linear mixed models* , Journal of the American Statistical Association, Vol 88, 9-25.

[34] Bryk W. J., Draper D., Goldstein H., Rasbash J. (2000) *Bayesian and Likehood Methods for Fitting Multilevel Modeling* , Computational Statistics and Data Analysis, Vol 39, No 2, 203-225.

[35] Bryk A. S., Raudenbush S. W., (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods*, SAGE publications, Newbury Park CA.

[36] Burstein L., (1980) *The Analysis of Multilevel Data in Educational Research and Evaluation*, Review of Research in Education, 8, 158-233.

[37] Burstein L., Kim, K.S. e Delandshere G., (1989) *Multilevel investigations of systematically varying slopes: issues, alternatives and consequences.*, In Bock R.D., Multilevel Analysis of educational data, Academic Press, New York.

[38] Charnock D., (1996) *Class and voting in the 1996 Australian Federal Election*, Electoral Studies, Vol. 16, 3.

[39] Chiogna M., (1998) *Some results on the scalar skew-normal distribution*, J. Ital. Statist. Soc, 7, 1-13.

[40] Cochran W. G., (1977) *Sampling Techniques*, Wiley, New York.

[41] Conaway M. R., (1990) *A random effects model for binary data*, Biometrics, 46, pp. 317-328.

[42] Dalla Valle A., (2007) *A test for the hypothesis of skew-normality in a population*, J. Statist. Comput. Simul., 77, 63-77.

[43] de Helguero F., (1908) *Sulla rappresentazione analitica delle curve abnormali*, Atti del IV Congresso Internazionale dei Matematici, Vol.III (sez. III-b), Ed. G. Castelnuovo. R. Accademia dei Lincei, Roma.

[44] De Leeuw J., Kreft I. G. G., (1995) *Questioning multilevel models*, Journal of Educational and Behavioral Statistics, 200, 171-189.

[45] De Leeuw J., Kreft I. G. G., (1986) *Random Coefficients Models for Multilevel Analisys*, Journal of Educational Statistics, 11, 57-86.

[46] Diggle P. J., Liang K.Y. e Zeger S.L. (1994) *Analysis of Longitudinal Data*, Claredon Press, Oxford.

[47] Donner A., Koval J. J. (1980) *The estimation of intraclass correlation in the analysis of family data*, Biometrics, 36, 19-25.

[48] Donner A., (1986) *A review of inference procedures for the intraclass correlation coefficient in the one-way random effect model*, International Statistical Review, 54, 67-82.

[49] Donner A., Koval J.J., (1980) *The large sample variance of an intraclass correlation*, Biometrika, 67, 719-722.

[50] Draper D., (1995) *Inference and hierarchical modeling in the social sciences (with discussion)*, Journal of Educational and Behavioral Statistics, 20, 115-147.

[51] Efron B., (1988) *Logistic Regression, survival analysis and Kaplan-Meier curve*, Journal of the American Statistical Association, 83, 414-425.

[52] Fahrmeir L. e Tutz G. (1994) *Multivariate statistical modelling based on generalized linear models*, Springer-Verlag, New York.

[53] Fienberg S., (1980) *The analysis of Cross-Classified Categorical Data*, Cambridge, MIT Press.

[54] Fisher R. A., (1921) *On the "probable error" of a coefficient of correlation deduced from a small sample*, Metron, 1, 3-32.

[55] Fisher R. A., (1954) *Statistical Methods for Research Workers (Twelfth ed.)*,Oliver and Boyd, http://psychclassics.yorku.ca/Fisher/Methods/.

[56] Galtung J., (1961) *Theory and methods of social research*, Columbia University Press, New York.

[57] Gelman A., (2005) *Analysis of Variance: Why It is More Important than Ever*, The Annals of Statistics, 33, 1-31.

[58] Genton M. G., (2004a) *Skew-elliptical distributions and their applications: a journey beyond normality*, Chapman and Hall/CRC.

[59] Gibbons R. D. e Hedeker D., (1997) *Random Effects Probit and Logistic Regression Models for Three-Level Data*, Biometrics, 53, 1527-1535.

[60] Gibbons R. D., Hedeker D., Charles S.C. e Frisch P., (1994) *A Random Effects Probit Model for Predicting Medical Malpractice Claims*, Journal of the American Statistical Association, 89, 760-767.

[61] Gilks W. R., Richaardson S. e Spiegelhalter D.J., (1996) *Markov Chain Monte Carlo in Practice*, Chapman e hall, Londra.

[62] Goldstein H., (1986) *Multilevel Mixed Linear Model Analysis using Iterative Generalized Least Squares*, Biometrika, 73, 43-56.

[63] Goldstein H., (1987) *Multilevel Covariance Component Models*, Biometrika, 74, 4300-431.

[64] Goldstein H., (1989) *Restricted unbiased iterative generalised least squares estimation*, Biometrika, 76, 622-623.

[65] Goldstein H., (1991) *Nonlinear multilevel models with an application to discrete response data*, Biometrika, 78, 45-51.

[66] Goldstein H., (1992) *Commentary: Better Ways to Compare Schools?*, Journal of Educational Statistics, 16, 89-91.

[67] Goldstein H., (1992) *Hierarchical Data Modeling in the Social Sciences*, Journal of Educational and Behavioral Statistics, 20, 201-204.

[68] Goldstein H., (1992) *Hierarchical Data Modeling in the Social Sciences*, Journal of Educational and Behavioral Statistics, 20, 201-204.

[69] Goldstein H. e Healy M.J.R., (1994) *The graphical presentation of a collection of means*, Journal of the Royal Statistical Society, A, 158, 175-177.

[70] Goldstein H. e Rasbash J., (1996) *Improved approximations for multilevels models with binary responses*, Journal of the Royal Statistical Society, A, 159, 505-513.

[71] Goldstein H., Rabash J., Plewis I., Draper D., Browne W., Yang M., Woodhouse G. e Healy M.J.R., (1998) *A User's Guide to MLwin*, Institute of Education, Londra.

[72] Ghosh P., Branco M. D., Chakraborty H., (2006) *Bivariate random effect model using skew-normal distribution with application to HIV-RNA*, Statist. Med., 26, 1255-1267.

[73] Gonzales Farias G., Dominguez-Molina J. A., Gupta A. K., (2004a) *Additive properties of skew normal random vectors*, J. Statist. Plann. Inference, 126, 521-534.

[74] Green P. J., (1987) *Penalized Likelihood for General Semi-Parametric Regression Models*, International Statistical Review, 55, 245-259.

[75] Gupta A. K., Chen T., (2001) *Goodness-of-fit tests for the skew-normal distribution*, Commun.Statist. - Simulation and Computation, 30, 907-930.

[76] Gupta A. K., Gonzales-Farias G., Dominguez-Molina J. A., (2004a) *A multivariate skew normal distribution*, J. Multivariate Anal., 89, 181-190.

[77] Gupta A. K., Huang W. J., (2002) *Quadratic forms in skew normal variates*, J. Math. Anal. Appl., 273, 558-564.

[78] Gupta A. K., Nguyen T. T., Sanqui J. A. T., (2004b) *Characterization of the skew-normal distribution*, Ann. Inst. Statist. Math., 351-360.

[79] Gupta A. K., (2004) *Generalized skew normal model*, Test, 13, 501-524.

[80] Gomez H. W., Salinas H. S., Bolfarine H., (2006) *Generalized skew-normal model: properties and inference*, Statistics, 40, 495-505.

[81] Gomez H. W., Venegas O., Bolfarine H., (2007) *Skew symmetric distributions generated by distribution function of the normal distribution*, Environmetrics, 18, 395-407.

[82] Guseo R., (2010) *Partial ecological correlation: a common three-term covariance decomposition*, Stat Methods Appl, 19, 31-46.

[83] Guseo R., (2006) *Statistica Terza edizione*, Padova, CEDAM.

[84] Hale G., (1977) *On Use of ANOVA in Developmental Research*, Child Development, 48, 1101-1106.

[85] Hardin J. W., Hilbe J. M., (2007) *Generalized linear models and extensions*, Stata Press.

[86] Harris A., (1913) *On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large*, Biometrika, 9, 446-472.

[87] Harville D. A., (1977) *Maximum Likelihood approaches to variance component estimation and related problems*, Journal of the American Statistical Association, 72, 320-340.

[88] Hedeker D. e Gibbons R.D., (1994) *A random effects ordinal regression model for multilevel analysis*, Biometrics, 50, 933-944.

[89] Hedeker D., Siddiqui O. e Hu F.B., (1999) *Random Effects Regression Analysis of Correlated Grouped Time Survival Data*, Statistics in Medicine, vol. 30,250-259.

[90] Henze N., (1986) *A probabilistic representation of the skew-normal distribution*, Scand. J. Statist., 13, 271-275.

[91] Hox J.J., (1995) *Applied Multilevel Analysis*, TT-Publikaties, Amsterdam.

[92] Hox J.J., (1995) *Multilevel Analysis: Techniques and Applications*, Erlbaum, New Jersey.

[93] Hox J.J., *Multilevel Modeling: When and Why*, I.Balderjahn, R. Mathar, M. Schader (Eds.) *Classification, data analysis, and data highways*, New York, Springer Verlag, 147-154.

[94] Jamalizadeh A., Behboodian J., Balakrishnan N., (2008) *On order statistics from bivariate skew-normal and skew t distributions*, Statist. Probab. Lett.

[95] Johnson N. L., Kotz S., Read C. B., (1988) *Skew normal distributions*, Encyclopedia of Statistical Sciences, 8, 507-507, Wiley, New York.

[96] Kalbfleish J.D. e Prentice R.L. (1980), *The Statistical Analysis of failure Time Data*, Wiley, New York.

[97] Kennet O., MacGrow S., Wong P. (1996), *Forming inferences about some intraclass correlation coefficients*, Psychological Methods, 1, 30-46.

[98] Kollo T., Traat I., (2001), *On the multivariate skew normal distribution*, in Revista de Estatistica, vol. II of Edicao Especial, 231-232, Portugal.

[99] Kreft Ita G. G., (1996) *"Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies*, Multilevel Models Project at the Institute of Education, University of London.

[100] Kreft Ita G. G., De Leeuw J., (1995) *Questioning Multilevel Models*, Journal of Educational and Behavioral Statistics, Vol 20, No. 2, 171-189.

[101] Kreft Ita G. G., De Leeuw J., Aiken L.S., (1996) *The Effect of Different Forms of Centering in Hierarchical Linear Models*, Multivariate Behavioral Research, 30, 1-21.

[102] Kreft Ita G. G., De Leeuw J. (1998) *Introducing Multilevel Modeling*, Sage, London.

[103] Kreft Ita G. G., De Leeuw J. (1986) *Random Coefficient Models for Multilevel Analysis*, Journal of Educational Statistics, 11, 57-85.

[104] De Leeuw J., Meijer E., (2008) *Handbook of Quantitative Multilevel Analysis*, Springer.

[105] Laara E. e Matthews J. N. S., (1985) *The equivalence of two models for ordinal data*, Biometrika, 72, 206-207.

[106] Lachos V. H., Gosh P., Arellano Valle R. B., (2010) *Likelihood based inference for skew-normal independent linear mixed models*, Statistica Sinica, 20, 303-322.

[107] Laird N., Ware J., (1982) *Random effects models for longitudinal data*, Biometrics, 38, 963-974.

[108] Langford I. e Lewis T., (1998) *Outliers in multilevel data*, Journal of the Royal Statistical Society, A, 161, 121-160.

[109] Lazarsfeld P. F., Menzel H., (1961) *On the relation between individual and collective properties*, Ed. A. Etzioni *Complex organizations: A sociological reader*, New York.

[110] Lee Y. e Nelder J. A., (1996) *Hierarchical generalized linear models (con discussione)*, Journal of the Royal Statistical Society, B, 57, 619-678.

[111] Liang K. Y. e Zeger S. L., (1986) *Longitudinal data analysis using generalized linear models*, Biometrika, 73, 45-51.

[112] Lin T. I., Lee S. Y. (2007) *Finite mixture modelling using the skew normal distribution*, Statistica Sinica, 17, 909-927.

[113] Lin X., (1997) *Variance Component Testing in Generalized Linear Models with Random Effects*, Biometrika, 84, 309-326.

[114] Little R. J., (1998) *Missing Data*, In Encyclopedia of Biostatistics, 2622-2635, Wiley.

[115] Liu J., Dey D. K., (2008) *Skew Random effects in multilevel binomial models: an alternative to non parametric approach*, Sattistical Modelling, 8, 221-241.

[116] Longford N. T., (1987) *A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects*, Biometrika, 74, 817-827.

[117] Longford N. T., (1996) *Discussione dell'articolo di Lee e Nelder (1986)*, Journal of the Royal Statistical Society, B, 57, 619-678.

[118] Longford N., (1993) *Random Coefficient Models*, Oxford, Clarendon Press.

[119] Loperfido N., (2001) *Quadratic form of skew-normal random vectors*, Statist. Probab. Lett., 54, 381-387.

[120] Mantel N., Hankey B., (1978) *A logistic regression analysis of response-time data where the hazard function is time dependent*, Communications in Statistics - Theory and Methods, A7, 333-347.

[121] Mardia K. V., (1970) *Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies*, Sankhya, 36, 115-128.

[122] Mardia K. V., Kent J. T., Bibby J. M., (1974) *Multivariate analysis*, London: Academic Press.

[123] Mare R. D., (1980) *Social Background and school continuation decision*, Journal of the American Statistical Association, 75, 295-305.

[124] Mason W. M., Wong G.Y., Entwisle B., (1984) *Contextual analysis through the multilevel linear model*, In S. Leinhardt (Ed), *Sociological methodology*, San Francisco, No 3, 271-284.

[125] Ma Y., Genton M. G., (2004) *Flexible class of skew-symmetric distributions*, Scand. J. Statist., 31, 459-468.

[126] McCullagh P. e Nelder J. A., (1989) *Generalized linear models (2nd edition)*, Chapman and Hall, Londra.

[127] McDonald R. P., (1994) *Two random effects models for multivariate binary data*, Biometrics, 50, 164-172.

[128] Meijer E., Rouwendal J. (2006) *Measuring welfare effects in models with random coefficients*, Journal of Applied Econometrics, 21, 227-244.

[129] Moerbeek Mirjam, Gerard J.P., (2000) *Design Issue for Experiments in Multilevel Populations*, Journal of Educational Behavioral Statistics, Vol 25, No 3, 271-284.

[130] Myers M., Hankey B. F. e Mantel N., (1973) *A logistic-exponential model for use with response-time data involving regressor variabiles*, Biometrics, 29, 257-269.

[131] Muthen L. K., Muthen B., (1994) *Multilevel covariance structure analysis*, Sociological Methods and Research, Vol 22, 376-398.

[132] Nelson L., (1964) *The sum of values from a normal and a truncated normal distribution*, Technometrics, Vol 4, 469-471.

[133] Patterson H. D., Thompson R., (1971) *Recovery of inter-block information when block sizes are unequal*, Biometrika, Vol 58, 545-554.

[134] Pearson K., (1894) *Contributions to the mathematical theory of evolution*, Phil. Trans. Royal. Soc. London, A, 185, 71.

[135] Pearson K., (1894) *Contributions to the mathematical theory of evolution II*, Phil. Trans. Royal. Soc. London, A, 186, 343.

[136] Pewsey A., (2000a) *Problems of inference for Azzalini's skew normal distribution*, Journal of Applied Statistics, 27, 859-770.

[137] Pfeffermann D., Skinner C. J., Holmes D., Goldstein H. e Rasbash J., (1997) *Weighting for unequal selection probabilities in multilevel models*, Journal of the Royal Statistical Society, B, 60, 23-40.

[138] Pinheiro, J.C., and Bates, D.M., (2000) *Mixed-Effects Models in S and S-PLUS*, Springer.

[139] Pintaldi F., (2003) *I dati ecologici nella ricerca sociale*, Carocci, Roma.

[140] Plewis I., Fielding A., (2003) *What is Multi-Level Modelling for? A Critical Response to Gorard*, British Journal of Educational Studies, 51, 408-419.

[141] Prentice R. L. e Gloeckler L.A., (1978) *Regression analysis of grouped survival data with application to breast cancer data*, Biometrics, 34, 57-67.

[142] Raudenbush S. W. e Willms J.D. (1995) *The estimation of school effects*, Journal of Educational and Behavioral Statistics, 20.

[143] Ringdal K., (1992) *Recent Developments in: Methods for Multilevel Analysis*, Acta Sociologica, Vol 35, No 3, 235-243.

[144] Robinson W. S., (1950) *Ecological Correlations and the Behavior of Individuals*, American Sociological Review, 15, 351-357.

[145] Rabash J., Yang M., Woodhouse G. e Goldstein H., (1995) *Mln command reference*, Institute of education, Londra.

[146] Rampichini C., Mealli F., (1999) *Estimating binary multilevel models through indirect inference*, Computational Statistics and Data Analysis, 29, 313-324.

[147] Reinhold Muller, Petra Buttner , (1994) *A critical discussion of intraclass correlation coefficients*, Statistics in Medicine, 13, 2465-2476.

[148] Rodriguez G., Goldman L. (1995) *An assessment of estimation procedures for multilevel models with binary responses*, Journal of the Royal Statistical Society, A, 159, 73-89.

[149] Rubin D. B., (1987) *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

[150] Salinas H. S., Arellano Valle R. B., Gomez H. W., (2007) *The extended skew-exponential power distribution and its derivation*, Commun. Statist. - Theory and Methods, 36, 1673-1689.

[151] Searle S. R., Casella G., McCulloch C. E. (1992) *Variance components*, Wiley, New York.

[152] Skinner C. J., (1989) *Domain means, regression and multivariate analysis. In Analysis of Complex Surveys*, (ed. Skinner C. J., Holt D., Smith T.M.F.), Wiley, New York.

[153] Snijders T.A.B., Bosker R.J., (1999) *Multilivel analysis: An introduction to basic and advanced multilevel modeling*, SAGE, Londra.

[154] Snijders T.A.B., Bosker R.J., (1994) *Standard Errors and Sampling Sizes for Two-Level Research*, Journal of Educational Statistics, 18, 237-261.

[155] Spanos A., (2005) *Where do statistical models come from? Revisiting the Problem of Statistical Model Specification*, 53, Institute of Mathematical Statistics.

[156] Stanish W., Taylor N., (1983) *Estimation of the intraclass correlation coefficient for the analysis of covariance model*, The American Statistician, 37, 221-224.

[157] Steenbergen M.R., Bradford S.J., (2002) *Modeling Multilevel Data Structures*, Journal of Educational Statistics, Vol 46, No 1, 218-237.

[158] Skrondal A., Rabe-Hesketh S., (2004) *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*, Boca Raton, Chapman and Hall.

[159] Skrondal A., Rabe-Hesketh S., Pickles A., (2004) *GLLAMM Manual*, http://www.bepress.com/ucbbiostat/paper160/.

[160] Swamy P. A. V. B., (1970) *Efficient Inference in a Random Coefficient Regression Model*, Econometrica, 38, 311-323.

[161] Swamy P. A. V. B., Tavlas G. S. (1995) *Random Coefficient Models: Theory and Applications*, Journal of Economic Surveys, 9, 165-196.

[162] Tate R. L., Wongbundhit Y., (1983) *Random versus nonrandom coefficient models for multilevel analysis*, Journal of Educational Statistics, 8, 103-120.

[163] Thum Y. M., (1997) *Hierarchical Linear Models for Multivariate Outcomes*, Journal of Educational and Behavioral Statistics, 22, 77-88.

[164] Umbach D., (2006) *Some moment relationships for multivariate skew-symmetric distributions*, Statist. Probab. Lett., 76, 507-512.

[165] Van den Eeden P., Huttner H. J. M., (1982) *Multilevel research*, Current Sociology, 30, 1-181.

[166] Verbeke L., Lesaffre E., (1997) *The effect of misspecifying the random-effects distributions in linear mixed models for longitudinal data*, Computational Statistic and Data Analysis, 23, 241-556.

[167] Wedderburn R. W. M., (1974) *Quasi-likelihood functions, generalized linear models and the Gauss-Newton method*, Biometrika, 61, 439-447.

[168] Wilk M. B., Kempthorne O., (1955) *Fixed, Mixed, and Random Models*, Journal of American Statistical Association, 50, 1144-1167.

[169] Williams D. A., (1982) *Extra-binomial variation in logistic linear models*, Applied Statistics, 31, 144-148.

[170] Winship C., Mare R. D., (1983) *Structural equations and path analysis for discrete data*, American Journal of Sociology, 89, 54-110.

[171] Woodhouse G., Yang M., Goldstein H., (1996) *Adjusting for Measurement Error In Multilevel Analysis*, Journal of the Royal Statistical Society, 159, 201-212.

[172] Yang M., (1997) *Multilevels models for multiple category responses - a simulation*, Multilevel Modelling Newsletter, vol 9, n.1, 10-16.

[173] Zaccarin S., Rivellini G., (2002) *Multilevel analysis in social research: an application of a cross-classified model*, Statistical Methods and Applications, Vol. 3, 97-.

[174] Zeger S.L., Liang K.Y. e Albert P.S. (1988) *Models for longitudinal data: a generalised estimating equation approach*, Biometrics, 44, 1049-1060.

[175] Zeger S.L. e Karim M.R., (1991) *Generalised linear models with random effects: a Gibbs Sampling approach*, Journal ofthe American Statistical Society, 86, 79-102.