A-PDF MERGER DEMO

Report n. 121

La metodologia statistica multilevel come strumento per lo studio delle interazioni tra il comportamento procreativo individuale e il contesto

G. Rivellini

Pisa, June 1997

LA METODOLOGIA STATISTICA *MULTILEVEL* COME POSSIBILE STRUMENTO PER LO STUDIO DELLE INTERAZIONI TRA IL COMPORTAMENTO PROCREATIVO INDIVIDUALE E IL CONTESTO

Giulia Rivellini

Dipartimento di Scienze Statistiche
Università degli Studi di Padova
E-mail : rivellini@iulm.it

ABSTRACT

La definizione dei meccanismi causali attraverso i quali un sistema di micro-variabili può essere collegato a macro-variabili, e viceversa, richiede ulteriori approfondimenti sia sul versante metodologico che su quello sostanziale. A tale proposito l'approccio di analisi ormai noto con il termine di *multilevel*, pur non risultando esaustivo della trattazione del problema, sembra offrire al ricercatore sociale uno strumento utile per indagare ed eventualmente tratteggiare in maniera più nitida i contorni di possibili forme di dipendenza tra i comportamenti procreativi individuali e gli «aggregati» cui l'individuo appartiene. Lo studio della variabilità delle relazioni di natura demografica nell'ambito di realtà contestuali tra loro diversamente connotate ha infatti recentemente condotto anche i demografi al dibattito sui metodi che favoriscono l'analisi degli effetti contestuali a livelli multipli della struttura sociale. Intuendo le potenzialità di questa nuova metodologia, il lavoro proposto nasce dal desiderio di riflettere sull'impiego di concetti e modelli predisposti per l'analisi di dati strutturati in maniera *nested*, tenendo conto esplicitamente di quelle che possono essere le esigenze applicative del demografo. Questo spiega la rivisitazione tecnica delle principali tipologie di modelli multilevel sia lineari che non realizzata con l'obiettivo di presentare l'insieme degli «utensili» tra i quali andare a scegliere quello più adeguato alle caratteristiche delle informazioni e delle variabili di analisi.

1. L'individuo e il contesto

La ricerca sociale si preoccupa spesso di mettere a fuoco le relazioni esistenti tra l'individuo e la società intesa nella sua accezione più generale. Gli individui interagiscono quotidianamente con il contesto sociale cui appartengono, sono influenzati dai gruppi sociali di riferimento, così come le proprietà di questi stessi gruppi vengono definite dagli individui che rientrano nel gruppo di appartenenza.

Il ruolo del gruppo è stato per lungo tempo un capitolo della teoria e della ricerca sociale: non a caso uno dei testi più classici della letteratura sociologica americana è intitolato *The Human Group* (Homans, 1950). Tra i temi di studio più rilevanti, accanto all'approfondimento delle dinamiche di funzionamento del gruppo stesso, compare con forza l'analisi dell'impatto che l'aggregato può esercitare sui singoli individui ad esso appartenenti. In questo caso specifico la ricerca è focalizzata sull'idea di gruppo come *contesto* interagente con gli individui che ad esso fanno riferimento. Ne deriva, quindi, che «the contextual analysis is the study of the role of the group context on actions and attitudes of individuals (Iversen, 1991).»

Per lo studioso di scienze sociali e conseguentemente anche per il demografo si impone allora una riflessione sul versante delle metodologie statistiche approntate per tale analisi contestuale, la quale necessita di informazioni sugli individui, così come sui gruppi cui essi appartengono.

Il caso tipico presenterà allora una variabile dipendente misurata sull'individuo e l'interesse del ricercatore sarà rivolto a cogliere e catturare gli effetti sul comportamento osservato generati sia dalle caratteristiche individuali, tanto quanto da quelle di gruppo. Più precisamente si tenderà a trovare qual è la forma del legame tra la variabile dipendente e le variabili esplicative associate sia all'individuo che al gruppo, quale il segno del legame e quale il tipo di legame stesso, se lineare o non lineare. Sarà da studiare se le carateristiche individuali e quelle contestuali agiscono contemporaneamente dando luogo ad effetti di interazione; insomma «as a long term goal, we seek to go beyond the statistical analysis and obtain a better understanding of the process by which individuals are affected by the group context (Iversen,1991)».

Ma riguardo al tipo di approccio da prediligere nel portare avanti un'analisi contestuale, non c'è sempre stata comunanza di opinioni. La letteratura dimostra, infatti , come il conflitto tra analisi ecologica e analisi individuale sia proceduto tra alterne vicende con una certa predominanza del secondo tipo di analisi, soprattutto per la vasta gamma di metodologie ad essa predisposta, oltre che per la ricchezza delle informazioni raccolte con indagini *ad hoc*. Se da un lato non si può pensare che il singolo possieda in sé tutte le determinanti che lo conducono a certe scelte (e quindi appare limitativo procedere esasperando al massimo il micro approccio), dall'altro il prediligere l'analisi ecologica, conferendo all'osservazione del comportamento medio dei gruppi un potere altamente

esplicativo della variabilità dei comportamenti individuali, porta inevitabilmente a scontrarsi con il problema più volte ribadito della fallacia ecologica[1].

Recentemente si è largamente diffuso un nuovo tipo di approccio sia logico che metodologico rivolto ad individuare il punto di incontro tra le dimensioni micro e quelle macro, tentando di conciliare le due posizioni prima presentate.

Nel campo delle scienze sociali capita allora di riflettere sulla possibilità di costruire modelli statistici *multilevel* che includano più livelli di osservazione, quello relativo all'individuo e quello contestuale che può derivare sia da aggregazioni di individui che riferirsi a caratteristiche proprie dell'area cui l'individuo appartiene. Tutto questo per tentare di condurre analisi sui comportamenti individuali che non prescindano dal contesto in cui essi si formano (Racioppi, 1994). D'altro canto appare ormai condiviso il fatto che tra i metodi per l'analisi degli effetti contestuali, l'approccio *multilevel* possa offrire «non più un semplice sfondo alla centralità dell'individuo, ma un quadro ben più articolato in cui situazioni circostanti oggettive, norme, tradizioni, disponibilità e accessibilità di servizi e strutture interagiscono attivamente con l'azione sociale individuale (Borra-Racioppi, 1995)».

Altri contributi della letteratura italiana chiariscono la rilevanza concettuale di questo nuovo modo di interpretare i fenomeni che vedono come protagonista attivo l'individuo appartenente insieme ad altri ad un livello superiore di osservazione, quale può essere la famiglia, il gruppo sociale di riferimento, la realtà territoriale di residenza, o altre dimensioni contestuali che generano insieme all'individuo una struttura cosiddetta *nested* dei dati.

Le riflessioni dei ricercatori italiani vanno infatti dai problemi al confine con la filosofia (Micheli, 1995), a quelli metodologici dell'integrazione micro macro, con la proposta di alcuni modelli di analisi per strutture gerarchiche di dati (Borra-Racioppi, *op.cit.*), fino a mostrare con casi di ricerca la necessità concreta di sviluppare i rapporti tra la dimensione micro e quella macro, quando, per esempio si conclude che nelle analisi comparative internazionali i dati individuali non sono utili a capire le differenze di comportamento tra paesi (De Rose, 1995).

Lavori recenti accennano al problema della congruenza micro-macro, così come al modo di concepire l'effetto contesto, corrispondentemente a differenti logiche dell'azione individuale, aspetti concettuali non affatto trascurabili nella progettazione globale di una ricerca multilevel diretta a cogliere le interazioni tra le variabili che descrivono l'individuo e quelle che descrivono i gruppi sociali, generiche variabili di secondo livello. Allo stesso modo non dimenticano di ammonire circa la necessità di progettare disegni di indagine capaci di integrare già in fase di rilevazione i livelli oltre ad essere sensibili agli effetti contesto sia globali di status che di struttura (cfr. Micheli, *op. cit.*).

[1]Le relazioni tra gli aggregati si sono spesso rivelate inconsistenti e soprattutto opposte nel momento dell'induzione sui comportamenti individuali.

Guardando alla letteratura internazionale un buon punto di partenza per la ricerca sociologica è rappresentato dall'articolo di Mason *et al.* (1983) con il quale vengono poste le basi strutturali per la concettualizzazione e lo studio empirico dei problemi di analisi contestuale.

Il valore concettuale di un approccio cosiddetto *human ecological* per lo studio delle dinamiche di popolazione è messo in luce, invece, da un recente articolo di Namboodiri nel quale si sostiene che .."it is fruitful for demographers to think in human ecological[2] terms..." (Namboodiri, 1994). E tra le varie metodologie proposte per mettere a punto questo importante tipo di analisi, si sollecita anche l'uso della ricerca multilevel, strumento che a parere dell'autore non risulta ancora pienamente sfruttato da parte degli *human ecologists*.

Si può dire, inoltre, che l'importanza del contesto sia ormai completamente riconosciuta dalla ricerca demografica soprattutto riguardo alla necessità di produrre informazioni adatte (cfr. Pinnelli, 1995): se prima il dato a livello macro era dato esclusivamente dall'aggregazione di osservazioni tratte a livello individuale, a partire dalla World Fertility Survey (1979), in occasione della quale è stato elaborato un modulo di indagine per la rilevazione delle caratteristiche contestuali, lo stesso è divenuto oggetto di una rilevazione apposita, indipendente da quella effettuata sull'individuo. E sempre in occasione di questa indagine molteplici sono stati gli interventi con i quali è stata data voce all'importanza di quello che più volte viene citato come il *community level* (cfr. Casterline e Werner, 1985), sebbene manchi un preciso orientamento di ricerca nell'uso e nell'accostamento delle diverse categorie di contesto.

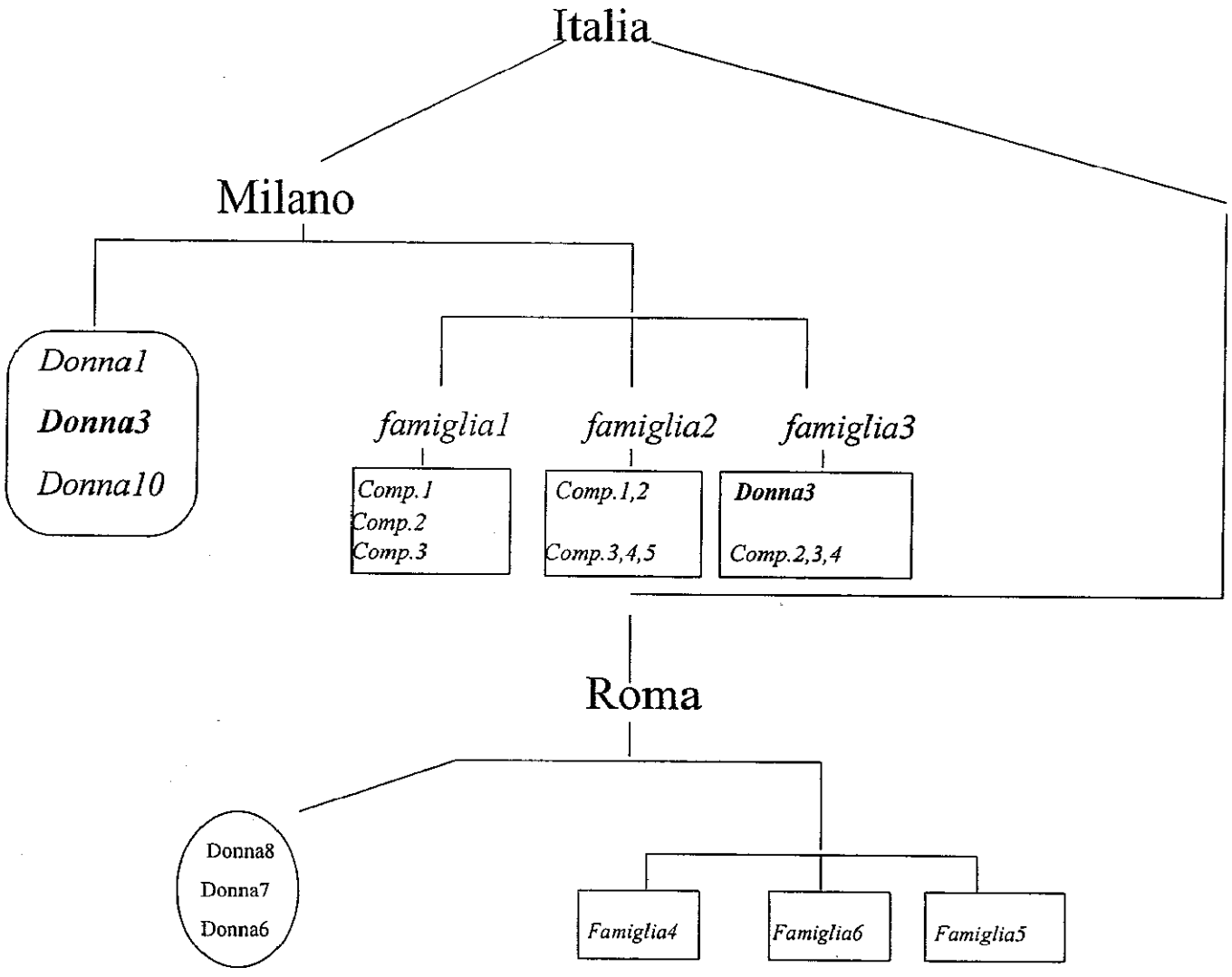
2. La struttura gerarchica delle osservazioni e il bisogno di tecniche di analisi multilevel

Questo tipo di problema connesso alle interazioni tra la sfera individuale e il contesto macro può essere in parte affrontato impostando analisi statistiche *multilevel*.

I modelli di analisi delle dipendenze a più livelli si sono infatti sviluppati negli ultimi dieci anni come un utile strumento per l'analisi di dati a struttura complessa. La letteratura più recente in campo socio-demografico mostra come sia divenuto oramai comune parlare di *multilevel modelling* quando si è di fronte a strutture gerarchiche dei dati.

Un problema di analisi *multilevel* riguarda infatti popolazioni con una struttura gerarchica delle osservazioni e un esempio è fornito dallo schema 1, nel quale si ipotizza che esista una gerarchia definita dalla nazione, dal comune, dalla famiglia e da ultimo dall'individuo. All'interno di un

[2] «Human ecology is a specialization of ecology, tailored to suit the characteristic features of human population», dove *ecologia* concerne tutto quello che concerne con le popolazioni interagenti oltre che tra loro, anche con il loro ambiente, e non il comporetamento e la biologia delle forme individuali di vita.



SCHEMA 1: Un esempio di struttura gerarchica delle osservazioni

comune sono residenti un certo numero di individui che occupano il livello più basso nella gerarchia, anche nel caso in cui tra il comune e l'individuo sia interposto il livello della famiglia di appartenenza.

Analogamente si può ipotizzare che le donne residenti in uno stesso comune adottino un comportamento procreativo significativamente diverso da quello rilevato su donne viventi in un altro comune distinto dal primo per il valore assunto da indicatori di struttura della popolazione, dei costumi coniugali e familiari, di ricchezza o di occupazione così come di istruzione.

Tutti i casi di studio caratterizzati dal presentare un evidente struttura gerarchica delle osservazioni conducono allora il ricercatore - sia esso demografo, sociologo, economista o epidemiologo - a compiere analisi su sistemi in cui le unità statistiche elementari, rappresentanti il primo livello, costituiscono dei gruppi, i quali a loro volta definiscono unità statistiche ad un livello gerarchicamente superiore.

Un campione a 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 formulati le osservazioni individuali non risultano generalmente indipendenti. Gli scolari di una stessa scuola tendono, infatti, ad essere simili l'uno all'altro a causa di processi di selezione (alcune scuole attraggono individui appartenenti ad una medesima classe sociale, per esempio) o a causa della comune storia che condividono vivendo nella medesima realtà scolastica.

Oltre al problema della dipendenza delle osservazioni che caratterizza fortemente le popolazioni con struttura complessa delle osservazioni, è importante anche tenere conto del fatto che gli individui appartenenti ad un gruppo interagiscono e conseguentemente sono influenzati dal contesto sociale cui appartengono. 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 osservati in queste particolari situazioni non possono allora essere elaborati come un campione proveniente da una singola popolazione, ma come generati da più popolazioni con diversi valori dei parametri (Borra e Racioppi, *op. cit.*).

Si intuisce quindi l'importanza di individuare una categoria di modelli statistici che permetta di considerare tale organizzazione gerarchica, valutando come questa particolare struttura possa influenzare la variabile oggetto di studio osservata al livello più basso della gerarchia (Racioppi, 1993).

Di fronte a queste esigenze di analisi i modelli *multilevel* si presentano attualmente come la risposta metodologica forse più adeguata: essi sono infatti strutturati per analizzare simultaneamente variabili appartenenti a livelli differenti della gerarchia, usando dei modelli statistici che includano tutte le possibili forme di dipendenza. Senza dimenticare che il desiderio di ricongiungere le dimensioni micro e macro, superando il conflitto tra l'analisi ecologica e l'analisi individuale,

sfociava proprio nella possibilità di costruire dei modelli statistici che includessero più livelli di informazione, tra cui sicuramente quello individuale e quello contestuale.

E non è difficile percepire l'eco di questo generale consenso scorrendo la più recente letteratura applicativa prodotta sia a livello internazionale che nazionale. Infatti se fino alla fine degli anni '80 l'utilizzo della categoria dei modelli *multilevel* sembrava essere limitato principalmente agli studi di scienza dell'educazione, dove gli alunni appartengono alle classi, le classi a loro volta formano nel loro insieme la scuola, e le scuole rientrano in distretti amministrativi[3], negli ultimi sei sette anni la letteratura appare ricca di lavori costruiti intorno all'adozione della modellistica *multilevel* volta a spiegare fenomeni rientranti nel campo dell'epidemiologia, della macro economia (Blien *et al.*, 1994), della demografia con particolare riguardo alle interazioni tra il credo religioso e l'uso dei metodi anticoncezionali (Amin *et al.*, 1996) o a variabili dicotomiche relative alla probabilità di avere un secondo o un terzo figlio per la donna italiana (Zaccarin, 1995), o dell'utilizzo dei servizi sanitari (Cislaghi *et al.*, 1996), così come dello stato di salute della popolazione (Galletti, 1996).

Un' ampia gamma di discipline sembra allora interessata a comprendere il ruolo che il contesto ricopre nella determinazione delle scelte e dei comportamenti individuali, un contesto che può essere dato dall'area territoriale di residenza, così come dalla famiglia di appartenenza, dall'azienda in cui si lavora, dal villaggio in cui una donna vive quotidianamente o ancora dall' *area naturale* di appartenenza (Billari,Rivellini, 1996).

Se da un lato sul versante applicativo iniziano a susseguirsi a ritmi abbastanza sostenuti svariati contributi e dall'altro il versante metodologico raccoglie diversi interventi che illustrano le problematiche da affrontare per impostare una corretta analisi contestuale, manca ancora qualche utensile nella cassetta degli attrezzi del demografo.

Sarà allora obiettivo dei successivi paragrafi proporre una riflessione sulle migliorie apportate dall'analisi *multilevel* rispetto alle tecniche di regressione ordinaria e di analisi della varianza, ripercorrendo quelle che sono le categorie principali di modelli *multilevel* lineari e non lineari. Tutto questo perseguendo l'obiettivo primario di valutare quanto questa metodologia sia in grado di rispondere alle esigenze di ricerca proprie del demografo.

[3] Un' eccezione è rappresentata da un articolo del 1988 apparso su *Enviroment and Planning* (Davies *et al.*, 1988) , nel quale le tecniche multilevel forniscono un supporto nella comprensione dei differenziali salariali osservati tra cinquantasette aziende appartenenti alla *Engineering Employers Federetation* (EEF).

3. Metodi di analisi multilevel *versus* metodi di regressione ordinaria

Perchè il ricercatore sia spinto ad apprendere nuove tecniche di analisi e ad interpretare in virtù di queste i risultati dei suoi studi, magari compiendo anche grandi sforzi di comprensione, è necessario che siano a lui ben chiare quali sono le novità apportate rispetto alle metodologie più tradizionali e quindi più note.

Per comprendere quali sono i vantaggi dell'analisi *multilevel* nei confronti delle tecniche di regressione ordinaria si può partire dalla riflessione condotta su tre degli obiettivi generali di ricerca preposti alle tecniche di analisi *multilevel* (con particolare riguardo ai modelli *random coefficients*, che in seguito verranno illustrati in maniera più dettagliata). Questi possono essere sintetizzati nel seguente modo (Raudenbush, 1992) :

- apportare migliorie nelle stime degli effetti osservati tra le unità individuali
- modellare gli effetti di interazione o *cross level*
- suddividere le componenti di varianza -covarianza

Vediamoli più da vicino uno per volta. Nel primo caso l'obiettivo è quello di generare stime migliori per un modello di regressione relativo a variabili dipendenti osservate su individui facenti parti di una unità di livello gerarchicamente superiore, «*borrowing strenght*» dal fatto che simili stime esistono per altre unità di livello superiore all'individuo.

Seguiamo l'applicazione di *Braun et al.* (1983) preoccupati di capire se il punteggio standardizzato ottenuto alla somministrazione di un test poteva essere adeguato per la selezione di studenti ammessi a frequentare università commerciali, considerando anche il gruppo delle minoranze nere. Dal momento che la maggior parte dei candidati sono bianchi, le loro informazioni dominano completamente il procedimento di stima. Di fronte a tale ostacolo la stima di equazioni separate per i candidati neri in ciascuna delle scuole considerate potrebbe rappresentare una soluzione, se non fosse per il fatto che la maggior parte delle scuole hanno un basso numero di studenti neri e che quindi offrono una scarsità di informazioni su cui sviluppare attendibili procedimenti di stima. Nel caso specifico riportato su 59 scuole 14 non hanno alcun studente nero, mentre 20 presentano una percentuale del 30%. Sviluppare equazioni predittive in queste 59 scuole mediante le tecniche di regressione ordinaria risulterebbe arduo. Inoltre anche se le 25 scuole rimanenti presentassero dati sufficienti a realizzare stime di equazioni separate, il campione di minoranze risulterebbe ancora esiguo. Alternativamente i dati potrebbero essere raggruppati tra le scuole ignorando la struttura *nested* degli studenti inseriti in scuole con caratteristiche diverse. In questo caso, dal momento che è

molto più probabile che le minoranze si trovino in alcune scuole piuttosto che in altre si potrebbero generare delle distorsioni nel procedimento di stima dei coefficienti. Gli autori individuano allora la risposta ai loro quesiti usando un modello lineare gerarchico: «by borrowing strength from the entire ensemble of data, they were able to efficiently utilize all of the available information to provide each school with separate prediction equations for whites and minorities. The estimator for each school was actually a weighted composite of the information from that school and the relations that exist in the overall sample».

Se ne deduce allora come il concetto del *borrowing of strenght* sia estremamente rilevante per due tipi di situazioni. Quando la presenza di dati sparsi costituisce un problema per l'analisi di comportamenti associati a membri di gruppi rappresentati in maniera non adeguata dai dati a disposizione (*sparse data*) insieme al caso in cui, nel modellare situazioni reali, i dati fanno riferimento a gruppi omogenei di individui (un' omogeneità che figura in un elevato coefficiente di correlazione intraclasse). In tal caso non si può accettare l'assunzione di indipendenza delle osservazioni, ipotesi basilare per le tradizionali tecniche statistiche.

Limitandosi alle analisi classiche realizzabili con i modelli di regressione lineari, si deve assicurare il rispetto delle quattro assunzioni di base, quali la linearità, la normalità, l'omochedasticità e l'indipendenza. Se le prime sono facilmente trasferibili in un problema di analisi di dati a struttura gerarchica, non altrettanto vale per le ultime due e soprattutto per l'assunzione di indipendenza. L'idea generale che spiega questo atteggiamento è che gli individui appartenenti al medesimo gruppo siano più vicini o abbiano dei comportamenti tra loro più simili di quanto non accada con gli individui appartenenti a gruppi diversi. Gli studenti frequentanti classi diverse possono essere indipendenti, mentre quelli che appartengono alla medesima classe condividono molte più variabili significative per esempio per lo studio del processo di apprendimento (preparazione dell'insegnante, livello culturale dei genitori degli studenti, livello intellettivo medio della classe). Alcune di queste variabili non saranno osservabili, il che significa che svaniranno nella componente erratica del modello lineare, causando così una correlazione tra i fattori di disturbo, concetto che può essere formalizzato usando dei modelli a componenti di varianza; le parti erratiche saranno allora definite da un fattore individuale e da uno di gruppo[4].

[4] Si può supporre in questa situazione che ciascuno dei gruppi sia caratterizzato da un differente modello di regressione definito, nel caso più semplice, da una propria intercetta e un proprio coefficiente angolare. Dal momento che i gruppi sono per la maggior parte delle volte campionati si può fare l'assunzione che i coefficienti di regressione siano il frutto anch'essi di un campionamento casuale realizzato da una popolazione di intercette e di coefficienti angolari. Questo modo di procedere come vedremo definisce un modello *random coefficients* ; se le stesse assunzioni vengono fatte limitatamente all'intercetta si parlerà invece di modelli a componenti di varianza.

In accordo a quanto affermano Bryk e Raudenbush (1992), tra gli altri obiettivi di ricerca associati alla modellistica *multilevel* compare senza ombria di dubbio «the formulation and testing of the hypotheses about cross.level interactions», aspetto che, come si è già avuto modo di ribadire, risulta alquanto appetibile nel campo delle scienze sociali, ove è importante attuare un collegamento tra la teoria e la pratica, tra l'individuo e il contesto di vita familiare, residenziale, lavorativo, ecc..

Un terzo uso della modellistica *multilevel* è infine associato alla stima delle componenti di varianza e covarianza con dati *nested*. Questa metodologia, a differenza della regressione ordinaria, permette infatti di suddividere la varianza totale osservata sulla variabile dipendente in una quota di varianza *within* e una *between*. Uno degli aspetti caratterizzanti i metodi di analisi più semplici quali la regressione ordinaria ai minimi quadrati (OLS) è rappresentato dal fatto che non vengono presi in considerazione i raggruppamenti delle osservazioni tra unità di analisi diverse. In tal modo diviene impossibile adottare un procedimento di decomposizione della varianza osservata entro una o più fonti di variabilità. Prendendo come esempio dati osservati su gruppi di famiglie le sorgenti di variabilità possono essere generate sia da variabili relative alla famiglia interamente considerata, così come da informazioni rilevate sui singoli membri o sul grado di reciprocità delle realazioni. Nel metodo OLS è presente solo una componente di varianza residuale : introducendo nel modello variabili esplicative il ricercatore è spinto ad ottenere un valore via via minore per questa componente. Nei modelli di analisi *multilevel* si tenta di ottenere valori bassi per tutte le componenti di variabilità osservate. Idealmente l'inserimento di variabili esplicative definite al livello più basso di analisi (caratteristiche individuali) diminuirà la componente corrispettiva di varianza osservata al primo livello var(e_{ij}); se le caratteristiche familiari differiscono tra le n famiglie considerate, la loro inclusione porterà anche ad una diminuzione della componente var(u_{0j}): l'inserimento di variabili esplicative osservate al livello familiare porterà ad un abbassamento della componente var(u_{0j}); l'inserimento di variabili di interazione tra X, variabile osservata al primo livello e Z, variabile ipoteticamente associata alle caratteristiche delle unità di livello superiore porterà ad una diminuzione della componente var(u_{1j}). E' certo che tutto questo non accade necessariamente: si parlerà, infatti, più correttamente di proporzione di varianza spiegata. Si noti che quest'ultimo tipo di interazione è realizzata tra variabili *individual*-level e *family*-level: tali interazioni prendono il nome di *cross level* interazione.

Sintetizzando si può allora affermare che l'analisi *multilevel* offre il quadro metodologico adeguato cui fare riferimento quando l'assenza di un numero sufficiente di osservazioni all'interno delle unità di secondo livello impedisce di stimare modelli che contemplino solo effetti fissi. L'analisi *multilevel* tenta proprio di risolvere questo problema, dal momento che l'obiettivo non è quello di stimare la varianza dei coefficienti relativi a singole rette di regressione diverse per singola unità di livello

superiore a quello individuale. (se fosse così allora un coefficiente di regressione calcolato relativamente ad un comune con poche osservazioni porterebbe ad una distorsione delle stime. Il multilevel modelling ha tra gli obiettivi quello di dire qualcosa circa la popolazione generale delle provincie o dei comuni o meglio sulla popolazione delle unità di livello gerarchicamente superiore, tenendo in considerazione la struttura nested delle osservazioni. Considerato che l'interesse del ricercatore è volto a stimare la componente erratica su un'unica distribuzione e cioè quella dei coefficienti, a questo procedimento di stima contribuiscono tutte le osservazioni di tutte le unità di livello gerarchicamente superiore, superando così il problema dei dati sparsi.

L'approccio *multilevel* costituisce inoltre uno strumento basato sulla scomposizione della struttura dell'errore nelle sue componenti corrispondenti alle diverse unità di analisi, riuscendo così ad esprimere anche la variabilità tra i gruppi.

La complessa struttura della popolazione in esame evidenzia infine come gli usuali modelli di regressione siano inadeguati. La categoria dei modelli *multilevel* consente di tener conto dell'effetto legato al raggruppamento delle osservazioni in aree geografiche relativamente omogenee al loro interno per quanto attiene al tipo di comportamento o fenomeno indagato. L'effetto del raggruppamenteo, quindi, non è visto come un parametro di disturbo, ma come parte integrante della struttura della popolazione e in virtù di questo deve essere adeguatamente inserito nel modello in modo da contribuire alla comprensione delle relazioni tra le variabili oggetto di studio.

In linea generale donne appartenenti ad uno stesso gruppo precedentemente definito presentano correlazione intra-gruppo positiva: questo significa che i valori delle variabili rilevate tendono a essere più simili tra di loro di quanto non sarebbe accaduto con un raggruppamento casuale. Questa è una caratteristica della struttura della popolazione dovuta in parte all'influenza di ogni donna sulle altre donne del gruppo e in parte al fatto che le donne appartenenti allo stesso gruppo sperimentano esperienze comuni, sono immerse cioè in una situazione di contesto culturale, politico, sociale e economico a loro comune (Angeli, Rampichini, Salvini, 1996)

4. I modelli di regressione multilevel linerari: il punto di partenza per un percorso graduale di comprensione

4.1 Un atteggiamento di fondo

I *multilevel models* godono di un generale consenso rispetto al problema che si presentava in passato relativo alla trattazione di modelli con strutture gerarchiche dei dati.

Recenti sviluppi nella teoria della stima dei modelli lineari (Aitkin e Longford, 1986; Goldstein, 1989) hanno reso possibile la specificazione, la stima efficiente e la verifica di modelli per strutture gerarchiche. Tale struttura appare come una proprietà intrinseca del sistema oggetto di studio e l'utilizzo di modelli statistici per la sua descrizione è motivato dalla struttura stessa, indipendentemente dalla procedura di campionamento che ha generato i dati. I modelli di regressione *multilevel* sono quindi costruiti per tenere conto esplicitamente della struttura gerarchica della popolazione. In particolare il modello consente di includere variabili esplicative sia a livello individuale che di gruppo, avendo come obiettivo quello di descrivere sia la relazione tra variabili esplicative e variabile di risposta che la variabilità tra i gruppi (Angeli *et al.*)

Dopo circa quindici anni di studi e ricerche sono comparsi però alcuni aspetti che spiegano come mai tali modelli non rispondano perfettamente alle promesse che inizialmente erano state formulate. Questa strumentazione statistica non va quindi considerata come «the promised panacea in this type of research», sebbene rimanga salda la convinzione che se il ricercatore è interessato a riflettere sulle componenti di (co)varianza, i modelli a coefficienti casuali per l'analisi di dati nested risultano essere ancora una buona scelta (Kreft, 1996). Senza trascurare il fatto che «knowing that model selection is based on the theory or research question to which it is applied, and not only on the type of data and the way the data is collected, implies that a statistical model cannot be optimal in general, but only in specific research situations (Kreft,1996)»

Ma provando a porsi nell'ottica del ricercatore sociale intenzionato ad utilizzare questa metodologia statistica esclusivamente in modo applicativo sebbene non risulti ancora esperto di *mixed-effects models* o di *random-effects models*[5], difficilmente si coglie dalla letteratura finora esistente un quadro generale ed immediato delle diverse categorie di modelli multilevel costruito appositamente per la trattazione e risoluzione di problemi relativi alle dinamiche di popolazione.

Potrebbe allora ritornare utile una rivisitazione tecnica, anche se non troppo, che mostri le caratteristiche distintive di questo approccio metodologico, ponendo attenzione alle esigenze del demografo in quanto studioso di fenomeni coinvolgenti l'individuo interagente con una struttura sociale spesso complessa e articolata.

Si impone allora una riflessione sull'impiego di concetti, metodi e modelli per l'analisi degli effetti contestuali, rispetto ad un obiettivo di comprensione dei fattori esplicativi della fecondità.

Tralasciando il valore del manuale, la cui utilità rimane indiscussa per qualsiasi tipo di studio metodologico, questioni quali la diversa terminologia con cui spesso vengono spiegati i medesimi concetti, l'applicabilità a campi disciplinari svariati, la genericità del termine *multilevel*, così come la

[5] Mentre nella ricerca sociologica si utilizza principalmente la terminologia ormai più in voga di *multilevel linear models*, (cf. Goldstein, 1987, Mason et al., 1983), nelle applicazioni di biometria sono molto più comuni i termini *mixed effects models* a *random effects models* (Laird e Ware, 1982).

non trascurabile rapidità con cui si sono susseguiti negli ultimi dieci anni lavori metodologici volto a trattare il problema del comportamento dei parametri in un modello *random coefficients* e lavori applicativi, possono facilmente complicare la fase di studio e comprensione preliminare affrontata da colui che intende utilizzare questa strumentazione statistica con obiettivi meramente applicativi. Potrebbe allora facilitare il lavoro L'avere a disposizione uno strumento che possa assomigliare quasi ad un dizionario o thesaurus da tenere sulla scrivania nella fase di specificazione del modello e quindi di scelta della metodologia di analisi, faciliterebbe senza dubbio il lavoro del ricercatore.

Afferma molto saggiamente Blalock (1972) che «the manipulation of statistical formulas is no substitute for knowing what one is doing», suggerendo una sorta di linea interpretativa su cui si snoda la presentazione di quelli che sono i principali modelli statistici multilevel. Questi infatti rivelano indubbiamente le loro potenzialità e la loro utilità nel campo delle scienze sociali solo nella misura in cui vengono compresi a fondo ed applicati in maniera graduale. Lo stesso Goldstein insegna - chiosando l'introduzione alla più recente versione del manuale con una frase sulla quale è necessario riflettere prima di accingersi ad un qualsiasi utilizzo di tale metodologia, molto semplice, ma allo stesso modo ricca di significato «..multilevel models are tools to be used with care and understanding...(Goldstein, 1995)». L'insieme dei modelli *multilevel* mostra la sua piena potenzialità e significatività di utilizzo nella misura in cui vengono applicati gradatamente e capiti in profondità, altrimenti il vantaggio offerto rispetto ai modelli di regressione ordinaria difficilmente viene colto.

Ecco perché questa parte è stata scritta avendo in mente chiaramente che tali modelli devono essere usati dal ricercatore sociale esclusivamente per una migliore comprensione dei fenomeni osservati nella realtà sociale, non sempre così facilmente riproducibili tramite modelli matematici o statistici, pur complessi o nuovi che siano.

Per avere un'idea delle numerose categorie rientranti nella generica definizione di modelli *multilevel* è sufficiente scorrere l'indice del testo di Goldstein.

Inizialmente nati nel campo della scienza dell'educazione, dalla letteratura recente appare sempre più nitidamente il tentativo tuttora *in fieri* di trasferire questo impianto metodologico al maggior numero possibile di problematiche statistiche. L'obiettivo del presente lavoro non è certo quello di trattare approfonditamente tutte queste categorie dal momento che esula dall'intento generale che ha mosso la ricerca. Ampio spazio sarà dato quindi a quei modelli che si trovano ad essere più vicini alle problematiche di studio socio-demografico ; la presentazione sommaria delle altre categorie va letta nella direzione di mostrare quali possono essere le potenzialità di analisi di questa strumentazione, oltre che di fare luce sui possibili ambiti di applicazione per la disciplina demografica o per altri versanti disciplinari che ora non possono più ritenersi staccati dallo studio dei fenomeni relativi alla popolazione. Va aggiunto, inoltre, che «models for multilevel analysis cannot be a universal panacea (Goldstein, 1995)» In alcune circostanze, laddove è presente una lieve complessità strutturale,

difficilmente questi modelli possono rivelarsi strettamente necessari e i tradizionali modelli di regressione ad un unico livello possono essere più che sufficienti, sia per l'analisi che per la presentazione dei risultati. D'altra parte analisi condotte su più livelli possono apportare una maggiore precisione ai tentativi di comprendere la componente casuale comunque presente in un modello statistico, facendo uso intelligente dei dati nella fase di comprensione delle differenze tra le unità di secondo livello. Questa categoria di modelli non intende in alcun modo sostituire quelle che sono teorie ormai più che fondate, il fatto che introducano maggiori complessità, inoltre, può estendere sì le capacità interpretative, ma certamente non le rende più semplici.

Nelle pagine che seguono si preferisce prediligere la spiegazione di quelli che sono i modelli strutturalmente più semplici, ma la cui comprensione dettagliata non può che essere preliminare a qualsiasi ulteriore approfondimento.

La comprensione della categoria dei modelli *multilevel* lineari costituisce un presupposto fondamentale per qualsiasi tipo di analisi successiva compiuta su dati strutturati su più livelli e con una variabile di risposta che non si presenta più come funzione lineare dei parametri sia nella parte fissa che in quella casuale del modello.

L'ostacolo principale sta infatti nel chiarirsi il significato e le implicazioni che un procedimento di stima di un modello organizzato su più livelli può avere nei confronti del processo globale di spiegazione dei fenomeni socio-demografici.

L'individuazione di quelli che possono essere i vantaggi di un'analisi multilevel rispetto ad una regressione ordinaria o a tecniche di analisi della varianza o della covarianza avviene sostanzialmente riflettendo sulla categoria dei modelli lineari; la novità metodologica, ai fini di una migliore comprensione dei modi in cui la dimensione micro interagisce con il contesto macro, è apprezzata infatti più che altro illustrando quelle che sono le caratteristiche di un modello di base lineare a due livelli. Saranno poi i diversi problemi applicativi a richiedere il passaggio ad un tipo di modello non lineare piuttosto che ad un altro, esigenze di ricerca dettate queste dal tipo di fenomeno che si intende studiare.

Una volta compresa la struttura concettuale portante attraverso la categoria dei modelli lineari, i passi successivi consistono esclusivamente nell'estendere l'idea *multilevel* a categorie generali di modelli statistici necessari per spiegare variabili di risposta *one-category* o *multi-category* (Yang *et al.*, 1996), così come dati di sopravvivenza o di durata molto spesso associati allo studio di problemi di popolazione.

Nello studio della realtà sociale, demografica od economica, si presentano facilmente casi in cui si ha a che fare con un insieme di gruppi definiti da variabili che assumono modalità uguali per soggetti appartenenti al medesimo gruppo e diverse tra soggetti di gruppi diversi; allo stesso modo si

incontrano spesso situazioni in cui le osservazioni sono suddivise in gruppi, i quali a loro volta sono contenuti o *nested* in gruppi ulteriori, di posto gerarchicamente superiore o inferiore ai precedenti.

Di fronte ad un tale scenario l'interesse del ricercatore sociale può allora essere rivolto a capire in prima battuta se i gruppi sono differenti, per poi tentare di indagare il perché delle differenze osservate tra i gruppi o i livelli, realizzando una vera e propria *analisi contestuale* (cfr. Iversen, 1991).

Se il primo obiettivo può essere raggiunto facendo uso delle tecniche di analisi della varianza o della covarianza ad effetti fissi o casuali, il secondo quesito è risolto invece guardando alla categoria più generale dei modelli statistici ad effetti aleatori noti in letteratura come modelli *random coefficients*, modelli misti (Mason et al., 1983), modelli *multilevel* lineari (Goldstein et al. 1988) o modelli lineari gerarchici (Bryk e Raudenbush, *op. cit*).

Non è superfluo precisare che tutti i modelli statistici citati finora rientrano nella definizione generale di *modelli multilevel*, tra i quali si ritrovano quindi anche le tecniche di analisi della varianza e della covarianza sia ad effetti fissi che aleatori.

Con le pagine seguenti si intende allora presentare in modo più chiaro questa casistica fino ad ora sommariamente tratteggiata, utile per una scelta corretta e consapevole del modello statistico più adeguato allo studio del problema affrontato di volta in volta.

4.2 Un generico modello di regressione a due livelli

A tale scopo si ritiene utile partire da una versione generale del modello *multilevel* a due livelli per vedere poi quali tipi di sottomodelli rientrano nella definizione più ampia di modelli lineari gerarchici, semplicemente andando a modificare quelle che sono le parti costitutive del modello nella sua versione generale[6].

Si è scelto, inoltre, di presentare i vari casi contemplando un'unica variabile esplicativa, dal momento che la difficoltà maggiore non risiede nell'estendere il procedimento al caso di due o più covariate. In questa parte verranno presi in considerazione solo i modelli multilevel ad effetti casuali, tralasciando la trattazione dei modelli ad effetti fissi, ipotizzando il realizzarsi di tutte quelle condizioni favorevoli alla scelta di un approccio ad effetti casuali (non verranno quindi presentate le tecniche già note di analisi della varianza e della covarianza che non contemplino coefficienti casuali)[7].

[6] Oltre alla notazione adottata in questa sede è possibile ritrovare nel manuale di Byrk e Raudenbush (1992) l'intera trattazione della categoria dei modelli multilevel lineari in scarti dalla media.

[7] Si ricorda che mentre nei modelli ad effetti fissi l'interesse dello statistico è rivolto esclusivamente a spiegare ed interpretare i coefficienti che definiscono la parte fissa del modello, considerando gli effetti aleatori semplicemente come delle componenti di disturbo, nei modelli ad effetti aleatori lo sforzo metodologico è

L'insieme dei modelli derivante da un generico modello a due livelli raccoglie le seguenti tipologie :

1) modello *One-Way Anova* con effetti casuali/effetti fissi

2) modello *One-Way Ancova* con effeti casuali/effetti fissi

3) modello di regressione *with means-as-outcomes*

4) modello di regressione a coefficienti casuali (*random coefficients regression model*)

Tutte queste categorie rientrano nel termine unico di *multilevel models*, i quali non sono quindi da intendersi come sinonimi dei *random coefficients models*.

La distinzione fondamentale per tutti i modelli appartenenti alla categoria dei modelli multilevel contemplanti effetti casuali sta infatti nel distinguere tra i modelli per i quali solo l'intercetta varia casualmente tra le unità di secondo livello (*random intercept model* o *variance components*) e quelli per i quali entrambe i coefficienti di regressione sono spiegati da una particolare distribuzione casuale (*random slopes model* o *randomly varying slope models*). Quindi i modelli indicati nell'elenco precedente con 1, 2, 3, appartengono all'insieme dei modelli *random intercepts*; il modello 4 rientra invece nella categoria dei modelli *random coefficients*, i quali permettono una variabilità casuale sia per l'intercetta che per il coefficiente angolare. Con questi modelli è allora possibile andare a spiegare anche il perché della diversa "forma" osservata tra le equazioni di regressione.

Perchè la trattazione formale non risulti troppo avulsa dalla realtà, e soprattutto per rendere più chiara la logica sottostante ai modelli lineari gerarchici, si ipotizzi di lavorare con due livelli di osservazione, dei quali il primo è riferito all'individuo e il secondo al comune di residenza del soggetto considerato[8]. Supponendo di considerare solo il gruppo di donne di età pari a 30 anni[9] la

diretto a modellare anche la variabilità associata ai coefficienti di regressione, considerati nel primo caso come fissi. Oltre al diverso significato che viene dato alla componente casuale, i modelli ad effetti fissi vengono usati quando si possiedono dati su tutte le unità di secondo livello con le quali si sta lavorando, mentre quelli ad effetti casuali si adottano quando si hanno dati relativi solo ad un campione delle unità di secondo livello. Sebbene siano state campionate tute le unità di secondo livello la scelta può cadere ancora sui *random effects*, nel caso in cui il numero dei gruppi sia molto elevato. Questo indurrebbe a pensare che gli stessi gruppi siano stati campionati da una qualche «superpopolazione» definita concettualmente a priori.

[8] I livelli di analisi e le variabili esplicative proposte per calare in un contesto reale la formulazione dei modelli statistici multilevel lineari, sono stati scelti anche in relazione ad un obiettivo più ampio rivolto allo studio delle forme di dipendenza tra i comportamenti individuali procreativi e il contesto comunale di residenza. Considerando sempre l'individuo come unità di primo livello si potrebbe pensare ad una variabile dipendente continua data dal numero di visite specialistiche richieste in ciascuna delle unità socio-sanitarie locali a cui i cittadini di un grande comune afferiscono, caso questo in cui il primo livello di osservazione è rappresentato dall'individuo, mentre il secondo dalle varie U.S.S.L. Così come potrebbe essere oltremodo interessante

variabile dipendente Y_{ij} è definita invece dal numero di figli desiderati[10] ad un generico istante temporale[11] dalla donna *i*-esima residente in un comune *j*-esimo.

Si pensi inizialmente ad un unico comune all'interno del quale viene rilevato il numero di figli desiderati da ciascuna delle N donne (i = 1...N) residenti in tale comune; una prima relazione ipotizzabile è formalizzata dalla seguente equazione di regressione:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad [1]$$

in base alla quale la variabile dipendente Y_i osservata sull'individuo *i*-esimo nell'ambito di unico comune è funzione lineare di una o più variabili esplicative X_i relative alla singola donna, quali potrebbero essere il numero di anni di istruzione, la condizione lavorativa, il grado di partecipazione religiosa, la dimensione della rete familiare attiva quotidianalmente o settimanalmente[12], il tipo di strategia familiare adottata nella famiglia coniugale (come vengono suddivisi i compiti e le faccende relative al *menage* familiare), la vicinanza o meno del luogo di lavoro rispetto all'abitazione, , il grado di ambizione relativo alla propria carriera professionale, ecc..

Considerando ad esempio come variabile predittiva individuale il numero di anni di istruzione, l'intercetta β_0 definisce il valore atteso della variabile dipendente per una donna analfabeta. Il coefficiente angolare β_1 rappresenta invece la variazione attesa nella variabile dipendente Y_i associata ad un incremento unitario nella variabile anni di istruzione. La componente erratica e_i definisce invece l'unico effetto casuale associato all'individuo *i*-esimo; generalmente si assume che e_i si distribuisca normalmente con media nulla e varianza pari a σ^2, cioè $e_i \sim N(0; \sigma_e^2)$[13].

guardare ad una variabile dipendente quale può essere il numero di ore di lavoro quotidiano per una donna lavoratrice dipendente residente in uno dei principali capoluoghi italiani. In tal caso il primo livello sarebbe il soggetto femminile mentre il secondo è dato dal capoluogo.

[9] Non è particolarmente significativa la scelta di questa precisa età, dal momento che l'obiettivo ultimo è quello di dare dei nomi alle variabili, che altrimenti finirebbero per rendere eccessivamente teorica, la riflessione sugli aspetti più metodologici del problema di ricerca impostato con l'intero lavoro qui proposto.

[10] La scelta per la variabile dipendente continua è stata strategica: sebbene questo tipo di informazione sia meglio descritta da un modello per conteggi si è apportata una semplificazione per presentare i caratteri distintivi di un modello di regressione multilevel lineare.

[11] Questo generico istante temporale è definito generalmente dalla data in cui avviene l'indagine campionaria, il più delle volte di natura retrospettiva.

[12] Come dimensione della rete familiare è lecito intendere il numero di persone che un individuo ritiene più importante per sé e con il quale tesse rapporti regolari.

[13] Spesso perchè il significato dell'intercetta sia pienamente e immediatamente compreso si considera la variabile indipendente X espressa in scarti dal suo valore medio X_m: $(X_i - X_m)$. Questo è il considdetto

Estendendo ora l'analisi al caso ipotetico di due comuni, le equazioni di regressione associate rispettivamente al comune 1 e al comune 2 saranno le seguenti:

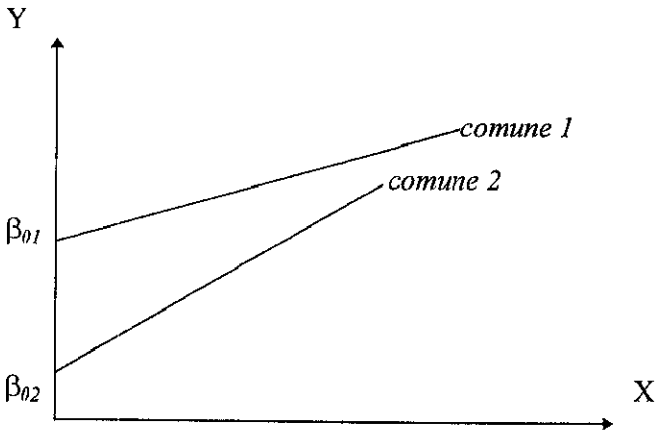
$$Y_{i1} = \beta_{01} + \beta_{11} X_{i1} + e_{i1} \qquad [2]$$

e

$$Y_{i2} = \beta_{02} + \beta_{12} X_{i2} + e_{i2} \qquad [3]$$

Si noti come questa volta le variabili sono accompagnate da un doppio indice, nel quale il numero 1 indica la retta di regressione relativa alla prima unità di analisi di secondo livello (il comune 1), mentre il numero 2 definisce l'equazione propria del secondo comune considerato; l'indice *i* rimane sempre riferito alle unità di analisi di primo livello e i suoi limiti generici di variazione sono come sopra $i = 1....N$.

La rappresentazione grafica seguente mostra come i due comuni generici presi in considerazione si differenziano per due elementi: mentre il comune 1 presenta un'intercetta maggiore del comune 2 ($\beta_{01} > \beta_{02}$) quindi un numero medio di figli desiderato maggiore, la variabile esplicativa prescelta risulta meno predittiva per il primo comune rispetto al secondo, il coefficiente angolare β_{11} è infatti minore di quello associato al comune 2, β_{12}.



Si ipotizzi ora di estendere la relazione studiata tra il numero di figli desiderati e il livello di istruzione nell'ambito di un'intera *popolazione* di comuni. Considerando allora un campione casuale

procedimento di *centering* in seguito al quale il valore dell'intercetta β_0 rappresenta proprio il numero medio di figli desiderati, o più genericamente, della variabile dipendente da spiegare (si veda Bryk e Raudenbusch, 1992).

di J comuni, con J pari ad un numero elevato, è possibile generalizzare l'equazione di regressione di partenza [1] per ogni j-esimo comune:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \qquad [4]$$

con $i = 1.... N_j$ e $j = 1.... J$

Sia l'intercetta che il coefficiente angolare sono ora indicizzate oltre che con i anche con la lettera j, il che permette ad ogni comune di essere definito da un'unico valore per il termine noto e per il coefficiente angolare; a differenza della regressione ordinaria ciascuna unità j-esima di secondo livello è caratterizzata quindi da differenti intercetta β_{0j} e coefficiente di regressione β_{1j} . In ogni comune si assume ancora che la componente erratica $e_{ij} \sim N(0, \sigma_e^2)$[14].

Come verrà esplicitato formalmente nelle pagine seguenti l'intercetta e il coefficiente angolare possono variare tra le unità di secondo livello; ed è per questo motivo che in letteratura si trova spesso l'espressione di *random coefficients*[15] per distinguere questa categoria di modelli da quella dei modelli di regressione multipla.

Soffermandoci ora sul significato dei parametri associati all'equazione [4], si considerino i valori generici delle seguenti delle statistiche intermedie:

$$E(\beta_{0j}) = \gamma_0 \qquad E(\beta_{1j}) = \gamma_1$$

$$\text{Var}(\beta_{0j}) = \sigma_{00} \qquad \text{Var}(\beta_{1j}) = \sigma_{11} \qquad \text{Cov}(\beta_{1j}; \beta_{0j}) = \sigma_{01}$$

dove

[14] La varianza associata alla j-esima scuola potrebbe essere indicata da σ_j, ma dal momento che la maggior parte dei modelli multilevel suppone un'omogeneità della varianza all'interno di ogni j-esima scuola, si specifica questa componente erratica genericamente con σ^2.

[15] Naturalmente non si assume che tali coefficienti siano completamente casuali; la speranza sta infatti nell'essere in grado di spiegare parte di questa variabilità casuale, introducendo nel modello variabili definite al livello gerarchicamente superiore a quello individuale. Comunque nella maggior parte dei casi non sarà possibile spiegare interamente questa componente di variabilità; di conseguenza, anche dopo aver introdotto variabili esplicative di secondo livello parte di questa variabilità casuale potrebbe rimanere non spiegata, da qui il senso di una delle espressioni usate per definire questi modelli «random coefficient model», il nome «random component model» si riferisce proprio al problema statistico della stima di parte di questa variabilità casuale (cfr. Hox, 1995).

γ_0 = il valore atteso per il numero medio di figli desiderati relativamente all'intera popolazione dei comuni

σ_{00} = la variabilità osservata tra i valori medi del numero di figli sull'intera popolazione

γ_1 = il valore atteso per il coefficiente angolare relativamente all'intera popolazione dei comuni

σ_{11} = la variabilità osservata tra i coefficienti angolari nell'intera popolazione dei comuni.

σ_{01} = la covarianza osservata tra i coefficienti angolari e le intercette nell'intera popolazione.

Conseguentemente un valore positivo di σ_{01} ad esempio implica che comuni con un numero medio di figli maggiore tendono anche ad avere un coefficiente di regressione positivo[16].

Supponendo di simulare diverse situazioni che mostrano come le regressioni associate agli *N* comuni variano in termini di intercetta e di coefficiente angolare potrebbe risultare significativo, ai fini di una migliore comprensione del fenomeno studiato, sviluppare un modello per la previsione dei due coefficienti β_{0j} e β_{1j}. In particolare si potrebbero inserire variabili osservate al livello comunale, quali potrebbero essere il tasso di attività femminile, il numero medio di figli per comune, l'età media, il reddito medio pro-capite od altri indicatori della realtà sociale, economica, demografica o culturale con l'obiettivo di caratterizzare in maniera più precisa il contesto in cui la donna è inserita e realizza le proprie scelte di vita.

Ipotizzando allora che tra gli *N* comuni i coefficienti di regressione abbiano una loro distribuzione casuale con una certa media e una data varianza, il passo successivo per i modelli di regressione lineari gerarchici consiste nel formulare un'equazione per ciascuno dei coefficienti β_j introducendo variabili esplicative di secondo livello.

In tal caso all'ultima equazione formulata [4] si aggiungono le seguenti due relazioni:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \qquad [5]$$

e

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j} \qquad [6]$$

[16]Considerando che la maggior parte delle volte si lavora con un campione di osservazioni rilevate sia a livello individuale che comunale e che quindi raramente si conoscono i valori effettivi dei parametri di popolazione prima elencati , cosic come i valori di β_{0j} e β_{1j} , è immediato dedurre che si otterranno stime di questi parametri dai dati a disposizione. Per ora si tralascia la trattazione delle procedure di stima, per permettere al lettore di concentrarsi esclusivamente sul significato associato ai vari parametri precedentemente elencati, in modo da vedere come può essere arricchito il modello per spiegare quanta più variabilità osservabile sia possibile.

La prima di queste due equazioni stabilisce che l'intercetta può essere prevista mediante la variabile Z osservata nel comune j-esimo: in corrispondenza di un valore positivo del coefficiente γ_{01} si concluderebbe allora che il numero medio di figli desiderati da una donna rilevato sull'intera popolazione dei comuni (intercetta β_{0j}) è correlato positivamente alla variabile di secondo livello Z, quale potrebbe essere la percentuale di soggetti praticanti la religione cattolica. Viceversa, se il coefficiente γ_{01} risultasse negativo il valore atteso del numero dei figli avuti sarebbe minore laddove la percentuale di soggetti che mostrano un elevato grado di partecipazione religiosa è più alta.

L'interpretazione dell'equazione [6] non è invece così immediata: essa asserisce che la relazione (formalizzata dal coefficiente β_{1j}) tra la variabile risultato Y_{ij} e la variabile esplicativa X_{ij} relativa al soggetto i-esimo residente nell'unità j-esima di secondo livello varia al variare delle modalità assunte dalla variabile Z osservata al livello gerarchicamente superiore a quello individuale. Così, comuni con un tasso di attività femminile più elevato (ad esempio $Z_1 > Z_2$) mostrano un atteggiamento più prolifico tra le donne maggiormente istruite se il coefficiente γ_{11} è positivo, contrariamente se lo stesso è negativo; l'essere inseriti in un contesto ad elevata occupazione femminile fa si che ad un più elevato grado di istruzione raggiunto da una donna corrispondano scelte di vita da lei adottate meno prolifiche.

Guardando invece ai termini u_{oj} e u_{1j} essi rappresentano le componenti erratiche casuali di secondo livello: si assume che tali componenti u_j abbiano media nulla e siano indipendenti dalle componenti erratiche di primo livello, quelle che nell'equazione 3 erano indicate con e_{ij}. La varianza associata al termine residuale u_{0j} è indicata con σ_{00}, quella relativa alla componente u_{1j} con σ_{11}, la covarianza tra gli u_j termini residuali è data invece da σ_{01} e generalmente è assunta essere diversa da zero. Si noti che queste componenti di varianza e covarianza sono ora *conditional* o *residual* (Bryk, Raudenbush, 1993), dal momento che esse rappresentano la variabilità residua associata ai coefficienti β_{0j} e β_{1j} dopo avre controllato per le variabili di secondo livello Z_j .

Si noti, inoltre, come nelle equazioni 4 e 5 i coefficienti di regressione γ non compaiono più con l'indice j, segno questo del fatto che sono assunti fissi e quindi non più variabili tra le unità j-esime di secondo livello.

Sostituendo le due ultime espressioni nell'equazione [4] si perviene alla forma in singola equazione del **modello generale lineare gerarchico a due livelli**:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij} + u_{1j}X_{ij} + u_{0j} + e_{ij} \qquad [7]$$

I primi quattro termini del membro di destra ($\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij}$) formano quella che è anche nota come parte fissa del modello dal momento che contiene solo i coefficienti fissi; i termini rimanenti ($u_{1j}X_{ij} + u_{0j} + r_{ij}$) definiscono invece la parte *random* o stocastica del modelo lineare

gerarchico. Il termine $Z_j X_{ij}$ è una componente di interazione che compare nel modello a seguito dell'inserimento della variabile Z nel processo di spiegazione del coefficiente di regressione β_{1j} ; si dice dunque che l'effetto moderatore di Z sulla relazione tra la variabile dipendente Y e quella indipendente X è espresso in termini di *cross-level interaction.*[17]

Si osservi come l'equazione [7] non rappresenta il tipico modello lineare assunto nel procedimento standard dei minimi quadrati ordinari, per il quale l'accuratezza del test di ipotesi e l'efficacia delle stime richiedono che le componenti erratiche casuali siano indipendenti, normalmente distribuite ed abbiano una varianza costante. Al contrario il modello statistico a cui ci si riferisce presenta una parte *random* molto più complessa ($u_{1j} X_{ij} + u_{0j} + e_{ij}$). Tali componenti erratiche sono infatti dipendenti all'interno di ciascuna unità *j*-esima di secondo livello, considerato che i termini u_{1j} e u_{0j} accomunano ogni soggetto appartenente alla medesima unità *j*. Gli errori hanno inoltre una varianza diversa, dal momento che la componente ($u_{1j}X_{ij} + u_{0j} + e_{ij}$) dipende oltre che da u_{1j} e u_{0j} che variano tra le unità di secondo livello, anche dal valore di X_{ij}, diverso da individuo a individuo[18]. Ovviamente la nullità dei termini u_{1j} e u_{0j} renderebbe il modello perfettamente equivalente ad un modello di regressione OLS.

Sintetizzando quanto visto fino ad ora e generalizzando la terminologia adottata, l'equazione [4] può essere anche definita *modello level-1*, le equazioni [5] e [6] rappresentano entrambe il *modello level-2*, mentre l'ultima equazione considerata definisce invece il *combined model* .

Le componenti erratiche e_{ij} rappresentano gli effetti casuali di livello 1, mentre gli errori u_{1j} e u_{0j} definiscono i *random effects* di livello 2. La componente di varianza di primo livello è data da Var(e_{ij}) mentre Var(u_{0j}), Var (u_{1j}) e Cov (u_{0j} ; u_{1j}) costituiscono le componenti di varianza e covarianza associate al secondo livello di osservazione. Infine la variabile esplicativa di primo livello è data da X_{ij}, mentre quella di secondo da Z_j.

Ma provare a dimostrare l'applicabilità di una categoria di modelli statistici in tutta la loro estensione vuol dire anche saper riconoscere in alcuni tipi di modelli - il cui uso può rivelarsi più comune - casi particolari del modello generale al quale sono state dedicate le pagine precedenti.

[17] L'interpretazione dei termini di interazione nell'analisi di regressione multipla può risultare alquato complessa. In generale l'interpretazione sostanziale dei coefficienti in un modello che contiene interazioni è molto più semplice se le variabili che formano l'interazione sono espresse in scarti dalla rispettiva media. A tale proposito sia la media generale che la media calcolata all'interno dei singoli gruppi sono considerate valide (cfr, Hox, 1995).

[18] Sebbene la regressione ordinaria non sia adeguata ad analizzare strutture di dati caratterizzati da queste forme di dipendenza, i parametri di questi modelli possono essere stimati mediante procedimenti iterativi di massima verosimiglianza.

I sottomodelli che seguono vanno dal più semplice al più complesso ed includono categorie più o meno comuni all'insieme degli strumenti metodologici utilizzati dal ricercatore sociale.

4.3 Modello Anova One-Way con effetti casuali

Nell'analisi di dati strutturati in modo gerarchico l'interesse del ricercatore potrebbe essere rivolto inizialmente ad indagare semplicemente se sono presenti delle differenze tra i valori medi assunti dalla variabile risultato in corrispondenza degli N gruppi in cui risultano suddivise le informazioni, riflettendo sui valori della varianza osservata tra i gruppi (*between groups variance*) ed entro i gruppi (*within groups variance*). Questo vorrebbe dire trascurare qualsiasi tipo di variabile predittiva sia al primo che al secondo livello, effettuando una comune analisi della varianza. Infatti il modello lineare gerarchico più semplice risulta proprio essere equivalente ad un *modello Anova con effetti casuali*, nel quale l'equazione di primo livello è data da:

$$Y_{ij} = \beta_{0j} + e_{ij} \qquad [8]$$

mentre quella relativa all'unico coefficiente di regressione è definita da:

$$\beta_{0i} = \gamma_{00} + u_{0j} \qquad [9]$$

dove γ_{00} rappresenta il valore medio rilevato nella popolazione per la variabile risultato Y_{ij}, mentre u_{0j} è l'effetto casuale associato alla j-esima unità di secondo livello con media nulla e varianza pari a σ_{00}. Il modello Anova ad effetti casuali può essere visto anche nella forma di *single equation* sostituendo l'espressione [9] nella [8] in modo da avere:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \qquad [10]$$

Il modello è definito ad effetti casuali dal momento che la componente u_{0j} è supposta variare casualmente tra le J unità di secondo livello[19]; il termine e_{ij} rappresenta invece l'effetto associato al singolo individuo i-esimo.

La variabilità di Y_{ij} è scomponibile allora in due elementi:

[19] Il modello analogo di analisi della varianza ad effetti fissi è usato quando si hanno dati su tutti i gruppi in cui risulta suddivisa la popolazione oggetto di indagine. Quando invece si hanno osservazioni solo su un campione di tali gruppi il modello richiesto è quello ad effetti casuali.

$$Var\ (Y_{ij}) = Var\ (u_{0j} + e_{ij}) = \sigma_{00} + \sigma^2 \qquad [11]$$

delle quali il parametro σ^2 cattura la varianza *within groups*, mentre σ_{00} quella *between groups*. Un altro parametro che potrebbe rivelarsi interessante nella fase interpretativa di questo tipo di modello è dato dal *coefficiente di correlazione intraclasse*, la cui formula è data da:

$$\rho = \frac{\sigma_{00}}{\sigma_{00} + \sigma^2} \qquad [12]$$

Esso misura la parte di variabilità dovuta all'effetto di raggruppamento e quindi di dipendenza tra le osservazioni nested in unità dello stesso livello; aspetto che, come si è visto, conduce ad optare per la categoria di modelli *multilevel*.

Ma il più grande svantaggio che presenta l'analisi della varianza *one way* consiste nel fatto che l'unica variabile esplicativa considerata è l'appartenenza ad un gruppo o ad un livello gerarchico di osservazione: nessun altro tipo di proprietà nè dell'individuo, nè delle unità di secondo livello sono inserite nella formulazione del modello. Conseguentemente se il ricercatore arriva a scoprire che i gruppi sono significativamente diversi l'uno dall'altro, non è ancora in grado con questo tipo di strumento metodologico di indagare le possibili ragioni di queste differenze, le quali potrebbero essere causate da variabili osservate al livello individuale, così come al livello gerarchicamente superiore.

4.4 Modello Ancova One-Way con effetti casuali

Il passo successivo conduce allora il ricercatore a studiare le diversità osservate tra i comportamenti medi rilevati nelle diverse unità di livello 2 controllando per variabili *individual-level*.

Così, riferendosi al modello lineare gerarchico generale formalizzato nell'equazione [7], costringendo i coefficienti γ_{01} e γ_{11} insieme alla componente erratica u_{1j} (per ogni valore di j) al valore nullo, il modello risultante di primo livello è dato da:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \qquad [13]$$

mentre quello di secondo livello è definito dalle seguenti due equazioni:

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad [14]$$

$$\beta_{1j} = \gamma_{10} \qquad [15]$$

Il modello combinato detto modello *Ancova con effetti casuali* si presenta allora nella foma seguente:

$$Y_{ij} = \gamma_{00} + \gamma_{10} Xij + u_{0j} + e_{ij} \qquad [16]$$

L'unica differenza rispetto al modello standard di analisi della covarianza consiste nell'effetto di secondo livello u_{0j} qui concepito come casuale, mentre nel modello convenzionale è considerato un effetto fisso. Diversamente dal modello esposto precedentemente compare in quest'ultimo caso la variabile esplicativa X_{ij} osservata al livello individuale. La parte casuale del modello rimane invece immutata: la componente casuale di primo livello è ancora espressa dal termine e_{ij} mentre quella di secondo livello da u_{0j}.

Si comprende facilmente allora come il principale vantaggio dell'analisi della covarianza nei confronti dell'analisi della varianza consista nel poter inserire nel processo di spiegazione anche variabili *individual-level*; rispetto alla regressione ordinaria vi è invece la possibilità di considerare l'appartenenza ad un gruppo o livello (Iversen, 1991).

Rimane comunque ancora un passo da fare: le proprietà dei gruppi e quindi le variabili osservate al secondo livello non sono ancora entrate in gioco, dal momento che non sono incluse nella formalizzazione statistica. Tutto ciò vorrebbe dire che, sebbene il ricercatore sia in grado di avviare il processo di comprensione delle differenze osservate tra i gruppi, ancora non può capire a fondo come mai tra le j-esime unità di secondo livello cambi la forma del legame tra la variabile dipendente Y e quella indipendente X.

4.4.1 L'inserimento di variabili esplicative di secondo livello

In virtù di quanto è stato appena messo in evidenza vale la pena esplicitare formalmente come si presenta un rmodello *random intercept* con variabili esplicative osservate anche al secondo livello.

Nell'ambito di un modello di regressione lineare a due livelli, l'equazione di secondo livello relativa all'intercetta può presentare anche una variabile Z_j specifica per ogni singola unità di livello gerarchicamente superiore all'individuo. In al caso le equazioni di primo e secondo livello si presentano nella forma seguente:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \qquad [17]$$

con

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \quad [18]$$

e

$$\beta_{1j} = \gamma_{10} \quad [19]$$

Quindi considerando il modello in forma combinata si avrà:

$$Y_{ij} = \gamma_{00} + \gamma_{10} Xij + \gamma_{01} Zj + u_{0j} + e_{ij} \quad [20]$$

Gli ultimi tre casi presentati si riferiscono ad una situazione di *variabilità semplice* tra i comportamenti medi osservati per gli individui raggruppati in unità di livello superiore. Tale concetto può apparire più chiaro riflettendo sulla fig. 1:

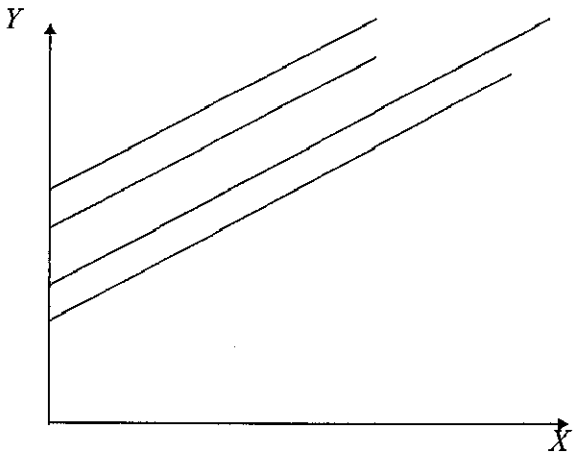


Fig. 1 Una situazione di variabilità semplice

Ad ogni comune è associata una linea di regressione; le diverse linee non mostrano coefficienti angolari diversi. Si suppone quindi che il parametro β_{1j} sia assunto fisso. Una situazione di questo tipo risponde ad un'esigenza preliminare che spinge il ricercatore a valutare se i gruppi si mostrano tra loro differenti

4.5 Means as Outcomes Regression

Un altro problema statistico abbastanza comune consiste nel considerare i valori medi di ciascuno dei gruppi come un risultato o *outcome* da essere previsto tramite le caratteristiche di gruppo. Questo tipo di sottomodello è definito allora dall'equazione di primo livello seguente:

$$Y_{ij} = \beta_{0j} + e_{ij} \quad [21]$$

mentre per il secondo livello si ha:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \qquad [22]$$

e

$$\beta_{10} = \gamma_{10} \qquad [23]$$

Quindi sostituendo la seconda equazione nella prima si ottiene:

$$Y_{ij} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} + r_{ij} \qquad [24]$$

Se nel modello di analisi della varianza ad effetti casuali la componente casuale u_{0j} veniva considerata come la deviazione di ogni *j*-esimo valore medio di gruppo dalla media generale, essa rappresenta ora la parte residuale dopo aver controllato per la variabile di secondo livello Z_j:

$$u_{0j} = \beta_{0j} - \gamma_{00} - \gamma_{01} Z_j \qquad [25]$$

Similmente la varianza della componente casuale rappresenta la varianza residuale o condizionale in β_{0j} dopo aver controllato per Z_j.

4.6. Modello di regressione con coefficienti casuali

I modelli visti finora costituiscono tutti esempi di *random intercepts models* o anche modelli *a componenti di varianza*, caratterizzati dall'avere *random* solo l'intercetta β_{0j} : nelle equazioni definenti il modello di secondo livello, infatti, solo all'intercetta era associata la componente casuale u_{0j}.

Ma una più ampia classe di applicazioni dei modelli lineari multilevel richiede studi in cui anche il coefficiente angolare β_{1j} sia concepito come variabile casualmente tra le unità di popolazione di secondo livello.

In tal caso l'apparato metodologico va arricchito con i *random coefficients regression models* o modelli a coefficienti di regressione casuali, tra cui il più semplice è quello in cui sia l'intercetta che il coefficiente angolare sono assunti casuali, ma senza che venga inserita nel processo di spiegazione e quindi di previsione del fenomeno alcuna variabile di secondo livello. Le equazioni che definiscono il modello sono allora le seguenti, rispettivamente per il primo e il secondo livello:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \qquad [26]$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad [27]$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \qquad [28]$$

Ad entrambe i coefficienti sono associate questa volta le componenti casuali u_j.

Il passo logicamente successivo consiste allora nel tentare di modellare la variabilità di questi coefficienti di regressione (sia l'intercetta che il coefficiente angolare e cioè σ^2_{u0} e σ^2_{u1}) osservata tra le unità di secondo livello, prendendo in considerazione tra le variabili esplicative anche quelle definite al secondo livello di analisi.

La fig. 2 sottostante, da confrontare con quella proposta poco sopra, permette di tradurre in maniera grafica, quanto emerso dall'introduzione dei coefficienti casuali. Le linee di regressione mostrano infatti coefficienti angolari variabili da comune a comune. Questo spiega il perchè della formulazione di equazioni apposite per i coefficienti.

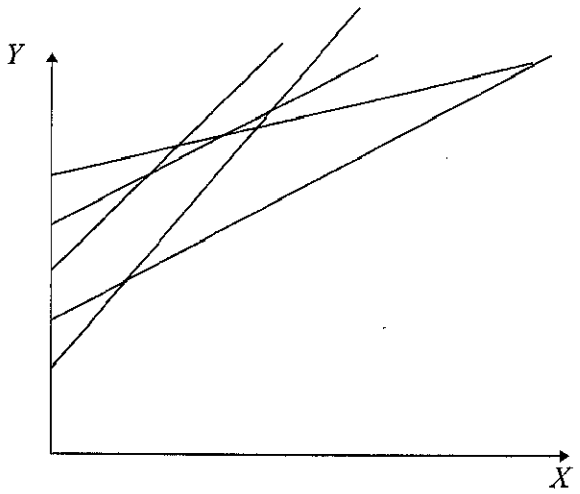


Fig. 1 Una situazione di variabilità complessa

Ecco allora che il cerchio si chiude ritornando con quest'ultimo caso al modello generale (si veda l'equazione [7]) presentato nelle prime pagine e che ha dato avvio alla specificazione delle diverse categorie di modelli *multilevel* lineari.

5. Modelli *multilevel* non lineari: per un maggiore adattamento alle caratteristiche delle informazioni e delle variabili di analisi

I modelli che sono stati presentati finora sono definiti lineari, nel senso che la variabile di risposta è una funzione lineare dei parametri nella parte fissa del modello e che gli elementi della matrice di

covarianze V sono funzioni lineari dei parametri anche in quella che è la parte casuale del modello (cfr. Goldstein, 1995).

Ma spesso per alcune applicazioni si rivela più appropriato considerare tipologie di modelli in cui la parte fissa o la parte casuale del modello o entrambe, contemplino funzioni non lineari. Sembra inoltre limitativo pensare che nella ricerca sociale si possano incontrare solo informazioni e variabili di analisi studiabili con modelli lineari. Questo è uno dei motivi che hanno indotto i metodologici ad estendere recentemente la categoria dei modelli *multilevel* anche al caso non lineare, arricchendo ulteriormente questa strumentazione statistica, la quale sembra ora adattabile a qualsiasi tipo di problematica applicativa.

Sono state formulate anche specifiche procedure di stima, per la cui trattazione si rimanda ad una serie di lavori ormai noti in letteratura (Longford, 1993; Goldstein 1991; Zeger et. al. 1988).

Non si può tralasciare di presentare anche se in maniera non esaustiva alcuni dei modelli appartenenti alla categoria dei modelli multilevel non lineari. Un caso tipico è quello relativo allo studio della crescita (Jenss and Bayley, 1937) dove veniva proposta la seguente equazione per decrivere la crescita nel peso dei bambini piccoli:

$$Y_{ij} = \alpha_0 + \alpha_1 t_{ij} + u_{\alpha 0j} + u_{\alpha 1j} + e_{\alpha ij} - \exp(\beta_0 + \beta_1 t_{ij} + u_{\beta 0j} + u_{\beta 1j} t_{ij} + e_{\beta ij}) \qquad [29]$$

Si nota chiaramente come la variabile di risposta Y_{ij} definente il peso del bambino j-esimo all'i-esima occasione di misura, se nella prima parte dell'equazione è legata alla variabile esplicativa età (t_{ij}) da una relazione lineare, nella seconda parte il legame è stabilito, invece, da un'esponenziale, funzione non lineare.

Un generico modello di regressione non lineare strutturato su due livelli può essere rappresentato dalla seguente equazione (Goldstein, 1995):

$$Y_{ij} = X_{1ij}\beta_1 + Z_{1ij}^{(2)} u_{1j} + Z_{1ij}^{(1)} e_{1ij} + f(X_{2ij}\beta_2 + Z_{2ij}^{(2)} u_{2j} + Z_{2ij}^{(1)} e_{2ij} + \ldots) \qquad [30]$$

dove la funzione f è ovviamente non lineare e dove il segno di addizione seguito dai puntini sta ad indicare che funzioni non lineari addizionali possono essere inserite, contenenti ulteriori variabili esplicative appartenenti alla parte fissa (X) o ulteriori componenti casuali associate al livello 1 o al livello 2, rispettivamente rappresentate nel caso dell'equazione precedente da $Z^{(1)}$ e $Z^{(2)}$.[20]

[20]Questo generico modello non lineare può essere linearizzato mediante l'espansione in serie di Taylor, procedura che conduce a trasformare il legame con le variabili esplicative presenti nella funzione f in un legame di tipo lineare, usando la derivata della funzione non-lineare di primo e di secondo ordine. Comprendere in tutti i suoi passaggi questo procedimento di linearizzazione è interessante più che altro per riflettere sul tipo di procedura da utilizzarsi

Le due assunzioni principali che erano alla base dei modelli finora trattati, e cioè che la variabile risposta abbia una distribuzione continua insieme al fatto che la componente erratica di primo livello sia spiegata da una distribuzione normale con media nulla e varianza pari a σ^2_e, può succedere che siano violate. Ci sono situazioni in cui l'assunzione di normalità per la componente erratica è chiaramente messa in discussione. Per esempio quando la variabile dipendente è rappresentata da una variabile dicotomica sia l'assunzione di normalità che quella di continuità non appaiono per nulla calzanti. O ancora se la variabile di risposta è data da una proporzione i problemi sono meno garvi, ma in ogni caso le due assunzioni continuano a non essere rispettate.

L'approccio piu' recente e piu' diffuso per la soluzione del problema di non normalità è quello che vede l'inserimento esplicito del modello statistico di un'opportuna trasformazione e la scelta di un'appropriata distribuzione dell'errore. Questa classe di modelli statistici è meglio nota con il nome di modelli lineari generalizzati, i quali sono definiti dalla concorrenza di tre elementi:

- un'equazione di regressione lineare
- una specifica distribuzione dell'errore
- una funzione cosiddetta di link che rappresenta la trasformazione che collega il valore previsto per la variabile dipendente con i valori osservati.

Si comprende quindi come nel caso piu' semplice rappresentato dalla funzione di *link* data dall'identità ($f(x)=x$) e da una distribuzione dell'errore normale, il modello lineare generalizzato si esplicita in un'ordinaria analisi di regressione multipla. Per altre funzioni di link e distribuzioni dell'errore il modello lineare generalizzato è stimato mediante complesse procedure di massima verosimiglianza (McCullagh&Nelder,1989), ma i risultati possono essere interpretati per la maggior parte come se provenissero da un ordinario modello di regressione lineare. Modelli multilevel generalizzati sono stati trattati recentemete da Wong e Mason (1985), Longford (1988, 1990) e Goldstein (1991)[21].

Tra i modelli multilvel lineari generalizzati si ritrovano:

per la stima dei parametri. Si fa riferimento in tal caso alla differenza esistente tra il «*marginal* (*quasi likelihood model*)» e il «*penalized* or *predictive* (*quasi likelihood model*). Il secondo tipo di modello si presenta nel caso in cui le stime correnti residuali al secondo livello vengono aggiunte alla componente lineare della funzione f al fine di ottenere una migliore approssimazione, mentre nel primo tipo di modello il procedimento di stima conduce a stime sicuramente piu' distorte soprattutto nel caso in cui sono presenti poche unità di primo livello per ogni unità di secondo livello.

[21] Il software più recente per l'analisi dei modelli mulitlevel (MLn) messo a punto dal Multilevel Models Project, permette di impostare la stima di una ampia gamma di modelli lineari generalizzati mediante la definizione di opportune macro che contengono gli elementi essenziali per la definizione delle caratteristiche distintive del modello prescelto

1) modelli per dati a risposta discreta (il caso di proporzioni come risposta);

2) modelli per dati a risposta multipla (multi-category response model) dove per variabile di risposta si ha un vettore di proporzioni;

3) modelli per conteggi:

4) modelli logistici *hazard* a tempo discreto

Nella trattazione seguente si tralasceranno i modelli a risposta multipla, in quanto risultano essere quelli meno vicini alle applicazioni proposte per studiare il comportamento procreativo in alcune delle sue possibili sfaccettature.

5.1 L'analisi di dati a risposta binaria strutturati su piu' livelli nella ricerca socio-demografica

Una tecnica statistica molto diffusa per trattare variabili di risposta date da proporzioni o da variabili dicotomiche è la regressione logistica. Ma perché questa metodologia permetta di modellare una struttura multilevel è necessario allora introdurre una versione a due livelli del modello logistico.

Sia Y_{ij} il valore della variabile dicotomica dato da sole due possibili modalità ($Y_{ij} = 0$ o $Y_{ij} = 1$). Se definiamo π_{ij} come $Pr(Y_{ij} = 1)$, introducendo la funzione di trasformazione logit ed indicando genericamente con f la parte fissa del modello, si avrà:

$$\log it\pi_{ij} = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = f_{ij} + r_j \qquad [31]$$

che può anche essere scritta come

$$\pi_{ij} = \frac{\exp(f_{ij} + r_j)}{1 + \exp(f_{ij} + r_j)} \qquad [32]$$

Il modello completo si presenta allora come:

$$Y_{ij} = \pi_{ij} + e_{ij} = \frac{\exp(f_{ij} + r_j)}{1 + \exp(f_{ij} + r_j)} + e_{ij} \qquad [33]$$

dove Y_{ij} rappresenta il valore (0,1) per la variabile di risposta osservato sul soggetto i-esimo appartenente al livello j-esimo.

Nella formulazione standard del modello si assume che la componente erratica e_{ij} abbia una distribuzione binomiale. In generale, si può ammettere una distribuzione extra-binomiale e cioè che la varianza sia data da:

$$\sigma_1^2 \frac{\pi_{ij}(1-\pi_{ij})}{N_{ij}} \qquad [34]$$

dove N_{ij}= 1 se si modella la risposta di ciascun individuo (cfr. Woodhouse, 1995).

La prevalenza nell'uso di metodi contraccettivi potrebbe essere una variabile di potenziale interesse per gli studi del demografo, così come la probabilità di avere il primo o il secondo figlio. In tal caso se si volesse stimare un modello logistico *random coefficients*, ipotizzando di trovarsi nel caso di una variabilità complessa dei comportamenti osservati tra le J unità di secondo livello, il modello sarebbe opportunamente specificato dalle seguenti equazioni rispettivamente di primo e secondo livello:

$$\log it(\pi_{ij}) = \beta_{0j} + \beta_{1j} X_{ij} \qquad [35]$$

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

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j} \qquad [36]$$

e in forma di singola equazione

$$\log it(P_{ij}) = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij} \qquad [37]$$

In tal caso si assume che i coefficienti di regressione varino casualmente tra le unità di livello gerarchicamente superiore (in questo caso il livello è il secondo) e questa variazione è modellata inserendo anche una variabile definita al secondo livello. Si noti infatti come questo modello contempli anche un fattore di interazione tra la variabile osservata al primo livello e quella al secondo livello[22].

[22]E' importante ricordare che l'interpretazione dei coefficienti non deve essere immediatamente condotta in termini della variabile di risposta data dalla proporzione che intendiamo analizzare, quanto piuttosto considerando la funzione di *link* utilizzata per la linearizzazione del modello. La funzione *logit* trasforma infatti le proporzioni osservate, che per definizione assumono valori compresi tra 0 e 1, in valori che variano tra $-\infty$ e $+\infty$. Essendo inoltre la funzione *logit* non lineare in corrispondenza degli estremi 0 e 1 diventa molto più difficile generare un cambiamento nella variabile dipendente e cioè nella proporzione. Per comprendere totalmente le implicazioni dei

E' immediato allora derivare tutte le altre possibili formulazioni riprendendo le pagine precedenti sui vari casi di modelli lineari.

5.2 Modelli multilevel per conteggi

Invece di guardare ad una variabile dicotomica o ad un set di proporzioni come variabile risposta, può capitare nella ricerca di sociale di avere a che fare con il conteggio di certi eventi. Un esempio immediato potrebbe essere dato dal set di informazioni elettorali. Se si suppone di aver classificato gli elettori in tre diverse classi sociali e in tre generazioni di appartenenza, in ciascuna delle nove celle ottenute mettendo insieme queste informazioni si conterà il numero di voti registrati per i vari partiti candidati.

Calandosi nell'ambito degli studi sui comportamenti riproduttivi un modello per conteggi ben si adatta a rappresentare la variabile Y_{ij} data dal numero di figli nati vivi fino al momento di realizzazione di un'indagine da una donna *i*-esima residente nel comune *j*-esimo.

Più specificatamente, il valore atteso del conteggio è dato da:

$$m_{ij} = \exp(\beta_{0j} + \beta_{1j} X_{ij} + u_{oj}) \qquad [38]$$

mentre, assumendo una distribuzione poissoniana dell'errore al primo livello si avrà:

$$\text{var}(Y_{ij}|m_{ij}) = m_{ij} \qquad [39]$$

da cui il modello Mln è espresso da:

$$Y_{ij} = m_{ij} + e_{ij} z_{ij} \quad z_{ij} = \sqrt{m_{ij}}, \ \sigma_e^2 = 1^{23}. \qquad [40]$$

coefficienti di regressione sulle proporzioni che si stanno modellando è necessario allora operare la trasformazione dei valori stimati nella loro scala di misurazione originaria, tramite la funzione inversa. Sempre con l'obiettivo di chiarire il piu' possibile il processo di interpretazione dei risultati delle stime, si può optare nella fase applicativa per la predisposizione di una cosiddetta tabella MCA (Retherford e Choe).

[23] Si può permettere una variabilità extra rimuovendo il vincolo posto su σ^2_e.Inoltre è possibile stimare dei modelli specifici, più complessi, ipotizzando che la varianza di primo livello sia una funzione più generale del valore previsto (Goldstein, 1995)

Riprendendo l'equazione [31] il termine $u_{0j} \sim N(\mu, \sigma^2{}_{u_0}$ rappresenta l'unica componente di variabilità osservata al secondo livello: si assume infatti come fisso il coefficiente associato alla variabile esplicativa x_{1ij}.[24].

Prevedendo l'inserimento di variabili osservate al livello comunale (Z_j) l'equazione [38] si modificherà in:

$$m_{ij} = \exp(\gamma_{00} + \gamma_{01} Z_j + \beta_1 x_{1ij} + u_{0j}) \qquad [41]$$

Ancora una volta ritroviamo i connotati fondamentali di un modello che è stato definito precedentemente come *random intercept.*

5.3 Modelli multilevel per dati di durata

Ma le caratteristiche dell'informazione spesso utilizzata dal demografo possono essere tali da richiedere il soddisfacimento di ulteriori esigenze applicative.

Quando la riflessione si sposta sui tempi della prolificazione, come nel caso dell'età della donna al primo figlio o alla nascita del secondo, gli strumenti metodologici che risultano più adeguati ad impostare correttamente le analisi sono rappresentati dai *modelli hazard*, appartenenti alla categoria dei modelli per l'analisi di dati durata (si veda Allison 1982; Yamaguchi, 1991). Tra i modelli multilevel non lineari rientra infatti anche la cetegoria dei modelli *event history*, nota anche con il nome di *survival time models* o *event duration models* (Goldstein, 1995)[25]. Gli individui, per esempio, possono essere protagonisti di ripetuti periodi trascorsi in varie condizioni occupazionali, tra cui la disoccupazione rappresenta un caso specifico. Questa particolare situazione definisce allora un modello a 2 livelli con l'individuo al livello 2 e le osservazioni effettuate nei diversi istanti temporali al livello 1. Così come si possono avere individui raggruppati in unità di secondo livello e su questi vengono misurate singole variabili tempo dipendenti.

Tra i modelli per dati di durata la riflessione si snoda esclusivamente sui modelli logistici *hazard* a tempo discreto, considerato anche il tipo di applicazione che verrà sviluppata successivamente.

Uno dei vantaggi principali associati a questo tipo di approccio consiste, infatti, nella possibilità di includere nello studio anche i *casi censurati*. La realizzazione di un'indagine retrospettiva in un dato momento temporale può portare a non conoscere il valore della variabile dipendente per soggetti che

[24]L'ipotesi semplificatrice sottostante è che non vari la «forma» del legame osservato tra la variabile di risposta y_{ij} e la generica covariata x_{ij}

[25] A tale proposito c'è una letteratura sia teorica che applicativa abbastanza consistente soprattutto nel campo della biostatistica; un sommario direi utile e completo è proposto da Clayton (1988).

non hanno ancora sperimentato l'evento osservato (come ad esempio la nascita del primo figlio) alla data dell'intervista, oppure che non lo sperimenteranno: l'esclusione delle informazioni relative a questi soggetti potrebbe generare problemi di forti distorsioni, soprattutto quando il numero dei casi censurati è elevato.

Un *modello hazard a tempo discreto multilevel* si presenta essenzialmente come un modello logistico multilevel (Hox, 1995): anch'esso appartiene quindi alla categoria dei modelli *multilevel* non lineari (Goldstein, 1991).

Considerando allora, un ipotetico caso di studio, y_{tij} rappresenta la variabile di risposta binaria osservata nell'intervallo di durata t per la i-esima donna vivente nel j-esimo comune, con $y_{tij} = 0$ se la donna non ha ancora vissuto la maternità nel generico intervallo di età t e $y_{ti\,j} = 1$ nel caso opposto in cui la donna abbia sperimentato l'evento di interesse.

Dato $\pi_{tij} = \Pr(y_{tij}=1)$, il modello hazard multilevel a tempo discreto si presenta allora nella seguente forma(Diamond et al., 1996);

$$y_{tij} = \pi_{tij} + e_{tij} \qquad [42]$$

dove

$$\log\left(\frac{\pi_{tij}}{1-\pi_{tij}}\right) = \mathbf{x}'_{tij}\beta + u_j \qquad [43]$$

$\mathbf{x}_{tij}$ è un vettore di covariate che, per modellare l'effetto dell'età della donna sullo stato di maternità, include una variabile categoriale rappresentante gli intervalli di durata. Associato al vettore delle covariate $\mathbf{x}_{tij}$ vi è un vettore delle stime dei parametri considerati costanti al variare dei comuni. Le componenti erratiche osservate al livello delle singole unità temporali e al livello comune di residenza sono rispettivamente e_{tij} (Bernoulli) e u_j ($N(0,\sigma^2_u)$).

Si tratta quindi di un modello *random intercept*, o modello a componenti di varianza, dal momento che l'intercetta è assunta come variabile casualmente tra i comuni considerati nell'analisi.

6. Un accenno ai procedimenti di stima dei parametri

Se queste sono le principali categorie di modelli *multilevel* quali sono invece i procedimenti di stima utilizzabili ? Solo pochi accenni sono stati fatti al funzionamento dei procedimenti di stima sia per la categoria dei modelli lineari che per quella dei non lineari: si è preferito infatti illustrare i

vantaggi associati a questa diversa metodologia di analisi, facendo anche il punto su alcuni dei limiti e degli ostacoli che il ricercatore applicato può incontrare nel loro utilizzo.

Guardando un pò più da vicino alla stima dei parametri di un generico modello lineare, l'algoritmo utilizzato è di tipo iterativo e conduce a delle stime finali sia per i coefficienti del modello che per la matrice delle varianze. A tal fine sono stati messi a punto tre diversi approcci di stima (cfr. Borra e Racioppi, *op. cit*):

1. un metodo basato sull'algoritmo iterativo dei minimi quadrati generalizzati (IGLS) proposto da Goldstein e una sua versione vincolata (Goldstein, 1989 e Hox, 1995) che sotto l'ipotesi di normalità delle distribuzioni dei parametri porta alle stime di massima verosimiglianza vincolate. L 'applicazione di tale algoritmo è possibile mediante il pacchetto *MLn*, versione successiva a *ML3*, che stima modelli linerai gerarchici a *n* livelli

2. un metodo proposto da Mason *et al* (1983)basato su una generalizzazione bayesiana dell'algoritmo EM (Dempster et al, 1981). Le stime che si ottengono sono di massima verosimiglianza vincolata (REML). L'algoritmo EM per tali stime è utilizzabile con il pacchetto HLM. Un altro prodotto software che permette di realizzare simili approcci di stima è GENMOD.

3. un metodo basato sull'algoritmo Fisher scoring sviluppato da Longford (1987). Anche questo porta alle stime di massima verosimiglianza ed è applicabile con il package VARCL.

Nel caso di modelli non lineari il dibattito è ancora aperto : non è stato raggiunto un esplicito accordo sui temi che da poco hanno catturato l'attenzione degli studiosi.

Con il presente lavoro lo sforzo principale è stato quello di mostrare alcuni dei versanti applicativi del *multilevel modelling*, con particolare riguardo a quelli confinanti con gli interessi del demografo, rivolto ad approfondire lo studio delle dinamiche dei comportamenti riproduttivi. Queste pagine non esauriscono quindi la casistica, sebbene ritengo che possano offrire gli strumenti necessari per ritrovare tra la letteratura prodotta quanto consono alle proprie ipotesi di ricerca.

RIFERIMENTI BIBLIOGRAFICI

- AITKIN M., LONGFORD N. (1986), Statistical modelling issues in school effectivness studies, *Journal of Royal Statistical Society A*, 149, part 1, pp. 1-43.
- AMIN S., DIAMOND I., STEELE F., (1996) Contraception and Religious Practice in Bangladesh, The Population Council,Working Papers n. 83.
- ALLISON P. D. (1982), Discrete-time methods for the analysis of event histories, in *Sociological Methodology* (ed. S. Leinhardt), pp. 61-98, San Francisco: Jossey-Bass.
- ANGELI A., RAMPICHINI C. e SALVINI S. (1996), La contraccezione in Brasile: un'analisi attraverso un modello a componenti di varianza, Comunicazione presentata al II Convegno dei Giovani Studiosi dei Problemi della Popolazione, Roma, Dipartimento di Scienze Demografiche.
- BILLARI F., RIVELLINI G. (1996) Alla ricerca di un effetto contesto. Riflessioni sulla realizzazione di un'indagine sugli anziani di Milano, Comunicazione presentata al II Convegno dei Giovani Studiosi dei Problemi della Popolazione, Roma, 25-27 giugno 1996.
- BLALOCK , H. M., jr. (1979) Measurment and conceptualizazion problems: the major obstacle to integrating theory and research, *American Sociological Review*, 44: 881-894.
- BLIEN U., WIEDENBECK M., ARMINGER G. (1994), Reconciling Macro and Micro Perspectives by Multilevel Models: An Application to Regional Wage Differences, in Borg J., Mahler P. (1994), *Trends and Perspectives in empirical social research*, New York.
- BORRA S., RACIOPPI F. (1995), Modelli di analisi per dati complessi: l'integrazione tra micro e macro nella ricerca *multilevel*, (Modeling complex data structures: a bridge between micro and macro analysis in multilevel research) Convegno Sis 20-21 Aprile, Arcavacata di Rende.
- BRYK A. S. & RAUDENBUSH S. W. (1992) *Hierarchical Linear Models*, Newbury Park, CA Sage.
- CASTERLINE J. B., (1985) Community Effects on Fertility, in Casterline J. B., ed (1985), *The Collection and Analysis of Community Data*: 243-248, Voorburg, Netherlands, International Statistical Institute.
- CISLAGHI C., BRAGA M., DAL CASON M., TASCO C., (1996), Uso di modelli gerarchici per l'analisi di livelli multipli nello studio delle struttura di correlazione dei comportamenti sanitari (The Use of Multilevel Models for the Analysis of Health care Behaviour at Aggregate Levels), Atti del Convegno "Salute e Famiglia", Cleup Padova.
- CLAYTON, D. G., 1988 The analysis of event history data: a review of progress and outstanding problems, *Statistics in Medicine* 7, 819-41.

- DAVIES R. B., MARTIN A. M., PENN R., (1988), Linear modelling with clustered observations: an illustrative example of earnings in the engineering industry, *Environment and Planning*, Vol. 20, pp. 1069-1084.
- DE ROSE A. (1995), Uniformità di modelli individuali e divergenze di modelli collettivi nello studio dei comportamenti familiari, Atti del Convegno SIS, Arcavacata di Rende, 20-21 Aprile 1995.
- GALLETTI A., PINNELLI A. (1996), Determinanti dello stato di salute della popolazione italiana considerando due livelli di analisi (Determinants of health status of Italian population considering two levels of analysis), Atti del Convegno "Salute e Famiglia", Cleup Padova.
- GOLDSTEIN H. (1989) Restricted unbiased iterative generalized least-squares estimation, *Biometrika*, 76, 3,pp.622-3.
- GOLDSTEIN H., Mc DONALD P. R., (1988), A general model for the analysis of multlevel data, *Psycometrika*, n. 4, 455-467.
- GOLDSTEIN, H. (1995), *Multilevel Statistical Models*, London, Edward Arnold, New York: Halsted.
- HOX J.J., (1995), *Applied Multilevel Analysis*, TT-Publikaties, Amsterdam.
- JENSS, R. M. and BAYLEY, N. (1937), A mathematical method for studying the growth of a child, *Human Biology* 9, 556-63.
- KREFT I. G.G. (1996), Are Multilevel Techniques Necessary ? An Overview, including Simulation Studies, California State University, Los Angeles.
- IVERSEN G. R., (1991) *Contextual Analysis*, Newbury Park, Sage.
- LONGFORD,N. T., (1993), *Random Coefficient Models*, Oxford University Press.
- MASON W.M., WONG G. Y., ENTWISLE B., (1983) Contextual Analysis through the Multitilevel Linear Model, in S. Leinhardt (ed), *Sociological Methodology*, San Francisco, Jossey-Bass 1983-1984 pp. 72-103.
- MICHELI, G. A., (1995), Prima di dipingere predisporre la cornice: preliminari ad una analisi e una ricerca multilevel (Before painting prepare your *frame:* preliminaries to a multilevel analysis and research) Convegno SIS, 20-21 Aprile, Arcavacata di Rende.
- NAMBOODIRI K., (1994) The Human Ecological Approach to the Study of Population Dynamics, *Population Index* 60(4): 517-39.
- PINNELLI A. (1995), Dimensioni Micro e Macro dei Comportamenti Demografici. Quadri Concettuali e modelli di Analisi, Atti del Convegno SIS, Arcavacata di Rende, 20-21 Aprile 1995.

- RACIOPPI F. (1994), I modelli multilevel per le relazioni micro-macro nell'analisi di strutture gerarchiche di dati, in Ciucci L., Racioppi F. (a cura di), Studi di popolazione. Nuovi approcci per la descrizione e l'interpretazione, Dip. Scienze Demografiche, Roma.
- RETHERFORD, R. D. - CHOE, M. K. *Statistical Models for Causal Analysis*, John Wiley, New York.
- WERNER L. H., (1985), Creating Community-Level data: Experiences in Kenya, in Casterline J. B., ed (1985), The Collection and Analysis of Community data: 243-248, Voorburg, Netherlands, International Statistical Institute.
- WOODHOUSE, G. (1995) (Ed.) A Guide to Mln for New Users. London, Multilevel Models Project, University of London.
- YANG M., GOLDSTEIN H., RASBASH J., Mln Macros for Advanced Multilevel Modelling, Multilevel Models Project, Institute of Education, University of London.
- YAMAGUCHI K, (1991) *Event History Analysis*, Applied Social Research Methods Series, Volume 28, Sage.
- ZACCARIN, S., (1995), Women's condition and reproductive behaviour: a multilevel approach to the analysis of Italian data, European Population Conference, Milan 4-8 September 1995, Irp, Roma.
- ZEGER S. L., LIANG K. Y., ALBERT P. S., (1988) Models for longitudinal data: a generalized estimating equation approach, *Biometrics*, vol. 44, 1049:1060.