Utilizzo dell'ICC come indicatore dell'esistenza di una struttura gerarchica

Marta Nai Ruscone
Dipartimento di Statistica,
Università Cattolica del Sacro Cuore, Milano

Abstract Recent contributions concerning multilevel models may be found in Goldstein (2011), Hox (2002), Kreft, de Leeuw (1997), Raudenbush, Bryk, (2002) and Snijders, Bosker (1999). The present paper first illustrates the reason for adopting multilevel models: among these the presence of the so-called intraclass correlation, which shows the degree of dependence of observations within a group. Equivalent definitions of the intraclass correlation coefficient are then introduced, and a test of the hypothesis of null intraclass correlation, based on the score test, is presented. The statistic does not depend on a particular distribution and is related to the pairwise correlation coefficient. The test is also adjusted for models with explanatory variables.

1 Premessa

In ambito induttivo sperimentale la specificazione del modello di riferimento è l'aspetto fondamentale e più delicato, perchè dalla sua correttezza dipendono la validità e l'efficacia di tutte le fasi successive di analisi: occorre individuare le variabili in gioco e il loro ruolo, e esplicitare il tipo di legame funzionale tra queste ipotizzato, che può essere sinteticamente espresso con la notazione

$$y = f(x)$$

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

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

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

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

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

Il modello adottato stabilisce che la caratteristica Y (risposta) è influenzata da due tipi di fattori, il cui effetto si esprime in generale come contributo additivo e su cui si dovranno formulare opportune ipotesi. In alcuni contesti la specificazione della relazione funzionale $f(\cdot)$ deriva in modo immediato dalla natura del problema o dalla teoria che descrive il fenomeno. Circa i termini di errore associati al modello vengono in genere formulate ipotesi concernenti l'indipendenza (stocastica, in media o lineare) tra di essi e rispetto alle esplicative incluse nel modello, la legge di distribuzione e l'omoschedasticità.

La stima e la verifica del modello statistico vengono successivamente eseguite utilizzando i dati raccolti attraverso un opportuno campionamento casuale. Per una analisi più efficace è bene individuare la struttura dei dati, soprattutto quando questi presentano una configurazione di tipo gerarchico. Si osserva, infatti, che i caratteri delle unità elementari sono influenzati da variabili presenti, e quindi osservabili, a differenti livelli della gerarchia: ad esempio, uno studente può avere rendimenti diversi a seconda della scuola in cui è inserito, oppure l'opinione dei cittadini, in relazione a un determinato provvedimento legislativo, può dipendere dalla zona geografica di residenza. È importante notare che la struttura gerarchica esercita in genere il proprio effetto indipendentemente dalla sua genesi: infatti,

anche se gli studenti non hanno scelto di frequentare una data scuola, il fatto oggettivo di condividere uguali strutture didattiche, insegnanti e programmi scolastici rende quel gruppo di studenti più omogenei tra di loro e diversi da quelli di un'altra scuola. Talvolta, in fase di raccolta dei dati, anche lo stesso piano di campionamento, che in genere è basato esplicitamente sulla ipotizzata esistenza di una gerarchia, può procurare gli effetti sopra descritti; tuttavia, l'esistenza della gerarchia non è esclusivamente legata al piano di campionamento, per cui anche i dati raccolti con una procedura di campionamento casuale semplice possono richiedere l'utilizzo di modelli specifici che considerano l'esistenza di una struttura a più livelli.
In alcuni casi, come precedentemente accennato, il campione che si estrae dalla popolazione potrebbe essere, ad esempio, un campione a più stadi; si pensi a tal proposito all'estrazione di un campione di studenti da utilizzare per la stima della media di una loro caratteristica, quale l'altezza in cm. Si può partire dall'estrazione casuale di alcuni distretti scolastici, quindi, da ognuno di essi, di un campione di scuole, e così via (Snijders, Bosker, 1999). Come anche evidenziato in Kish (1995), questo tipo di campionamento produce particolari effetti sulla varianza campionaria, contrariamente ad altre procedure di campionamento. I modelli cosiddetti multilevel, permettono di tener conto di questi effetti e di quelli in generale prodotti dell'esistenza di una struttura di tipo gerarchico.

2 Caratteristiche della struttura gerarchica

La metodologia multilevel fornisce un insieme di strumenti adatti ad analizzare simultaneamente variabili classificate a livelli differenti di gerarchia, con riferimento a modelli statistici che specificano le varie possibili forme di dipendenza (le osservazioni all'interno di un gruppo sono infatti fra loro più simili rispetto a quelle di altri gruppi). I modelli multilivello, inoltre, considerano i vari livelli di osservazione: quello relativo alle unità statistiche e quelli cosiddetti contestuali, che possono derivare da aggregazioni delle unità a livelli via via crescenti della gerarchia. Storicamente, le analisi di dati gerarchicamente organizzati sono state inizialmente realizzate utilizzando le tecniche standard, come l'analisi della varianza o la regressione multipla, spostando tutte le variabili su un solo livello di interesse. Ciò avveniva mediante due distinte procedure: aggregazione, che è lo spostamento di variabili originariamente osservate su un livello basso della gerarchia verso un livello superiore; la disaggregazione, che è lo spostamento di variabili verso il livello più basso.
Ad esempio, con la cosiddetta regressione aggregata (pooled regression) si ignora la eventuale struttura gerarchica dei dati e si ipotizza che le differenze tra i gruppi siano spiegate solo dalle esplicative X (covariate) misurate a livello di gruppo, ignorando così i possibili effetti della struttura gerarchica nei dati. In tal modo, con la regressione su dati aggregati, tutta la variabilità viene attribuita alle differenze tra le medie dei gruppi, mentre all'interno di ciascun gruppo le unità sono implicitamente considerate perfettamente omogenee.

Analizzare variabili che appartengono a differenti livelli della gerarchia su un singolo e comune livello può risultare inadeguato e presentare degli inconvenienti, che diventano tanto più gravi quanto più la gerarchia è rilevante nella spiegazione del fenomeno analizzato. Da un lato, l'aggregazione comporta una sostanziale perdita di informazioni e, di conseguenza, l'analisi statistica perde precisione. Dall'altro, quando i dati vengono disaggregati, i test statistici ordinari considerano che i valori disaggregati siano, in genere, informazioni indipendenti provenienti dall'insieme della unità di basso livello. Invece, nelle situazioni in cui i dati sono gerarchicamente organizzati, tale ipotesi viene generalmente a cadere. Il comportamento degli individui è infatti influenzato dal contesto sociale nel quale sono inseriti e le caratteristiche di un gruppo sono influenzate dagli individui che formano il gruppo stesso: gli individui e il contesto sociale nel quale vivono possono essere visti come un sistema gerarchico di individui e gruppi, nel quale gli individui e i gruppi agiscono a livelli diversi. I test statistici tradizionali, basati sull'assunto di indipendenza, producono stime distorte degli errori standard e, di conseguenza, i risultati che si ottengono possono apparire "impropriamente" significativi.

2.1 Dati ecologici e dati individuali

Sul finire degli anni '80 si assiste al tentativo di approdare ad un nuovo paradigma che, superando la dicotomia tra la dimensione macro (contestuale) e la dimensione micro (individuale), provi ad integrarle. Sempre negli stessi anni si sviluppano, dapprima in ambiti esclusivamente legati alla scienza dell'educazione (Goldstein, 1987; Raudenbush, Bryk, 2002; Aitkin, Longford, 1986), i modelli multilevel, finalizzati all'analisi dei livelli micro e macro, al fine di superare la prospettiva riduzionista, dal macro al micro, ed agevolare l'integrazione tra le due prospettive analitiche. Essi trovano ampia giustificazione nel risolvere le problematiche che si incontrano utilizzando dati a struttura complessa, tra cui, ad esempio, quella di analizzare i dati ad un certo livello e formulare le conclusioni ad un altro livello (fallacia del livello decisionale). Questo tipo di errore può assumere sostanzialmente due forme (Pintaldi, 2003):

1. Atomistic Fallacy: problema in cui si incorre quando si formulano inferenze ad un determinato livello della gerarchia basandosi su analisi che si riferiscono a un livello inferiore (Alker, 1969); si fanno ad esempio inferenze riguardanti associazioni a livello di gruppo mediante associazioni a livello individuale. In tal modo non si considera che i fattori che spiegano la variabilità tra individui all'interno dei gruppi non sono necessariamente gli stessi che spiegano la variabilità tra i gruppi (Hox, 1995), oppure non agiscono nel medesimo modo.

2. Ecological Fallacy: l'approccio ecologico consiste nell'interpretare dati macro come se fossero dati micro, facendo inferenze riguardanti il livello individuale sulla base dei dati inerenti il livello di gruppo, considerando cioè esclusivamente le aggregazioni a livello del gruppo cui gli individui appartengono (Robinson, 1950); in tal modo si utilizza la correlazione tra variabili a livello

di gruppo per fare affermazioni su relazioni di livello individuale (Snijders, Bosker, 1999).

Si è a lungo dibattuto se per dati con struttura di tipo gerarchico fosse da prediligere un approccio ecologico o un'analisi individuale: se da un lato non si può pensare che il singolo possieda in sè tutte le determinanti che lo conducono a certe scelte (e quindi appare limitativo procedere considerando il solo livello individuale), dall'altro il prediligere l'analisi ecologica, conferendo all'osservazione del comportamento medio dei gruppi una capacità esplicativa della variabilità dei componenti individuali, comporta il generarsi dell'*Ecological Fallacy*. Le relazioni tra gli aggregati si sono spesso rilevate inconsistenti, o addirittura opposte, una volta analizzati i comportamenti individuali. L'errore, essenzialmente di natura logica, è dovuto ad una imperfetta specificazione del modello di riferimento. Il modello deve opportunamente tener conto della non indipendenza delle osservazioni, ma consentire, allo stesso tempo, di analizzare simultaneamente la dipendenza da variabili classificate a diversi livelli della gerarchia.

2.2 Definizione del contesto di riferimento

Con i modelli multilevel si è giunti alla costruzione di algoritmi sempre più sofisticati, tali da prevedere strutture gerarchiche molto complesse (Goldstein, 2011); tuttavia rimangono ancora aperti quei problemi di ordine concettuale (a partire dalla definizione di contesto) connessi alla spiegazione del rapporto funzionale tra ambiente sociale e comportamento individuale, ovvero, tra legami relativi a unità situate a livelli gerarchici differenti.
A tal proposito è possibile individuare almeno tre definizioni di contesto (Zaccarin, Rivellini, 2002):

- Raggruppamento "naturale". Rappresenta il criterio di aggregazione più intuitivo e si può affermare che la modellistica multilevel nasce dalle riflessioni su questa modalità di raggruppamento: in questo caso la struttura gerarchica è intrinseca e i soggetti vengono naturalmente classificati come appartenenti ad un gruppo. E' il caso tipico degli alunni aggregati per classi, o di individui residenti nella stessa area geografica.

- Raggruppamento "ambientale". In questo caso la correlazione tra unità appartenenti allo stesso gruppo si ipotizza che possa derivare dall'esposizione allo stesso sistema relazionale (di lavoro, ad esempio) che può favorire una comunanza di valori atteggiamenti e comportamenti. Alcuni autori hanno sottolineato l'importanza dell'appartenenza di classe in relazione, ad esempio, alle scelte politiche e/o elettorali (Andersen, Heath, 2002; Charnock, 1996);

- Raggruppamento "teorico". Ci si riferisce ad aggregazioni formulate sulla base dei costrutti teorici fondati su fattori di tipo economico, sociale e culturale. Sicuramente tra le tre tipologie è quella più problematica da indagare, proprio per l'incertezza dei suoi confini.

Dal momento che il raggruppamento gioca un ruolo fondamentale nell'analisi dei comportamenti, non sembra banale evidenziare i limiti concettuali ed i problemi metodologici ed interpretativi che tali definizioni presentano. Innanzitutto vi è la questione dei confini. Alcuni gruppi presentano dei confini fissi e ben determinati. In questo caso l'individuo, o l'unità d'analisi gerarchicamente inferiore, può appartenervi oppure no. Non è prevista una situazione intermedia. E' il caso tipico dei raggruppamenti naturali: non si può appartenere a due comuni o a due province. Nell'ambito delle scienze sociali, tuttavia, quest'ultima condizione rappresenta l'eccezione piuttosto che la regola: nella maggior parte dei casi, infatti, il ricercatore si trova dinnanzi a strutture di gruppo fluide, dai contorni sfocati, dai confini incerti ed indeterminabili e diventa fondamentale stabilire, non tanto se un'unità appartiene ad un raggruppamento, bensì "in che misura" vi appartiene. Nel caso di esistenza di una struttura gerarchica le relazioni si possono presentare tra variabili osservate a differenti livelli. Gli individui interagiscono col contesto sociale cui appartengono, cioè i soggetti sono influenzati dalle caratteristiche dei gruppi di cui fanno parte e, a loro volta, le proprietà di questi gruppi risentono dell'influenza dei singoli individui. In simili circostanze, individui (unità) e gruppi (macro-unità) vanno presi in considerazione come un sistema gerarchico.

3 Modello multilevel generale a due livelli

Per individuare i principali effetti prodotti dall'esistenza di una struttura gerarchica delle osservazioni si può far riferimento, in modo conveniente e senza perdita di generalità, ad un modello a due soli livelli di raggruppamento. Più esattamente, si supponga che le singole unità di osservazione, nonchè elementari o di primo livello, siano aggregate in J gruppi di unità di secondo livello e si assuma che le unità elementari raggruppate entro il $j-esimo$ gruppo siano pari a n_j $(j = 1, ..., J)$. Della variabile oggetto di interesse Y, osservata sulle unità elementari, si desidera indagare in merito al legame con la variabile esplicativa X. Si suppone che tale legame sia di tipo lineare e, inoltre, che esso possa variare, da gruppo a gruppo, in relazione all'azione di una variabile esplicativa W, che interviene al secondo livello. Il modello multilivello si propone di collegare, con un'unica formulazione statistica, i modelli di regressione specificabili separatamente per i diversi gruppi. Formalmente, la relazione fra X e Y viene espressa tramite il seguente modello:

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

dove l'osservazione y_{ij} del fenomeno di interesse, effettuata sulla $i-esima$ unità elementare appartenente alla $j-esima$ unità di secondo livello, $(i = 1, ..., n_j; j = 1, ..., J)$ dipende dal valore x_{ij} assunto dalla variabile esplicativa X e dalla componente casuale d'errore e_{ij}.
Si assume, inoltre, che gli errori e_{ij}, nel seguito detti "di primo livello", siano generati da v.c. E_{ij} di valore atteso nullo, varianza costante pari a σ^2 e mutuamente incorrelate all'interno dello stesso gruppo, così come fra gruppi diversi. Il modello

definito dalla (1) è detto modello di livello 1. Come si nota i parametri β_{0j} (intercetta) e β_{1j} (coefficiente angolare) dipendono dall'indice j di gruppo: con ciò si vuole indicare che, al variare del gruppo, le rette di regressione possono essere caratterizzate da diversa intercetta e/o da diversa pendenza, e quindi che la variabile X può esercitare un'influenza, lineare, diversa da gruppo a gruppo.

Solitamente, la natura variabile dei parametri β_{0j} e β_{1j} viene a sua volta descritta mediante modelli di regressione che prevedono la presenza di una variabile esplicativa W di secondo livello, che agisce con intensità differente da gruppo a gruppo, ma costante all'interno dello stesso gruppo $j = 1, ..., J$, secondo le relazioni:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j} \tag{2}$$

e

$$\beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}. \tag{3}$$

Tale modello è detto di livello 2 ed è caratterizzato dai parametri γ_{00}, γ_{10} e, se presenti, γ_{01} e γ_{11}, che non dipendono dalla struttura di gruppo. Le variabili U_{0j} e U_{1j} che generano le determinazioni u_{0j} e u_{1j} costituiscono, invece, la parte aleatoria dei modelli e vengono dette effetti casuali o errori di secondo livello.

Si assume che esse abbiano valore atteso nullo, varianze non necessariamente uguali e che possano essere correlate.

Sostituendo nell'equazione (1) di livello 1 le equazioni (2) e (3) di livello 2, si perviene alla definizione del cosiddetto modello combinato

$$y_{ij} = \gamma_{00} + \gamma_{01} w_j + \gamma_{10} x_{ij} + \gamma_{11} w_j x_{ij} + u_{0j} + u_{1j} x_{ij} + e_{ij} \tag{4}$$

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

- $\gamma_{00} + \gamma_{01} w_j + \gamma_{10} x_{ij} + \gamma_{11} w_j x_{ij}$ costituisce la parte deterministica del modello (γ_{00} è l'intercetta o la costante; γ_{01} rappresenta l'effetto della variabile esplicativa del livello 2; γ_{10} indica l'effetto dei predittori del livello 1; γ_{11} è l'effetto della interazione cross-level tra i predittori del livello 1 e quelli del livello 2);

- $u_{0j} + u_{1j} x_{ij} + e_{ij}$ costituisce la parte casuale del modello (le u_{0j}, determinazioni di variabili casuali U_{0j} IID, rappresentano la variabilità di livello 2, riguardo i valori dell'intercetta del livello 1, al netto della variabile esplicativa w_j; le u_{1j}, determinazioni di v.c. U_{1j} IID indicano la variabilità di livello 2 della pendenza del livello 1 al netto di x_{ij}; gli e_{ij}, errori di livello 1 al netto delle variabili esplicative del primo livello, rappresentano la variabilità tra le unità del livello 1).

La specificazione del modello (4) si completa esplicitando le seguenti assunzioni sulla componente casuale del modello

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

2. $Var(U_{0j}) = \tau_{00}$, $Var(U_{1j}) = \tau_{11}$, $Var(E_{ij}) = \sigma^2$; questo postula che gli errori hanno varianza costante[1] all'interno dei rispettivi livelli.

3. U_{0j}, U_{1j} e E_{ij}; sono normalmente distribuite.[2] Considerando unitamente le assunzioni (1) - (4) si ha che i residui di livello 2 sono descritti da una distribuzione normale bivariata con media nulla e matrice di varianza-covarianza:

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

mentre i residui del livello-1 si distribuiscono con una distribuzione normale con media nulla e varianza σ^2.

4. $Cov(U_{0j},\ E_{ij}) = Cov(U_{1j},\ E_{ij}) = 0$; questo implica che gli errori della pendenza e dell'intercetta sono incorrelati con gli errori delle unità del primo livello. Questa assunzione è necessaria per ottenere l'identificabilità del modello.

4 Conseguenze della struttura gerarchica

I residui del modello multilivello, considerati globalmente

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

costituiscono determinazioni di variabili casuali, che indicheremo con Δ_{ij}, che sono caratterizzate da media nulla e varianza non costante [3], dal momento che

$$\begin{aligned} Var(\Delta_{ij}) &= E[(U_{0j} + U_{1j}x_{ij} + E_j)^2] = \\ &= E[U_{0j}^2] + 2x_{ij}E[U_{0j}, U_{1j}] + x_{ij}^2 E[U_{1j}^2] + E[E_{ij}^2] = \\ &= \tau_{00} + 2x_{ij}\tau_{01} + x_{ij}^2\tau_{11} + \sigma^2. \end{aligned} \tag{5}$$

Abbiamo quindi che $Var(\Delta_{ij})$, e quindi $Var(Y_{ij})$, è in parte una funzione dei predittori di livello-1; quindi Δ_{ij} ha una varianza non costante (sebbene U_{0j}, U_{1j} e E_{ij} abbiano varianza costante per l'assunto (2)). Si avrà varianza costante solo nel caso in cui $U_{1j} = 0$, ovvero quando w_j definisce in maniera esatta (deterministica) le pendenze di livello-2.
Si fa inoltre osservare che i residui dei modelli multilivel sono anche correlati, se considerati tra le unità di livello-1 comprese nelle unità di livello-2. Indicati con

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

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

[3] Per la dimostrazione si considera l'assunzione addizionale che bisogna considerare un'assunzione addizionale, cioè che $Cov(E_{ij}, E_{kl}) = 0$ per $i \neq j$, $k \neq l$ (Goldstein, 1995).

δ_{ij} e δ_{kj} due generici residui del livello-2, abbiamo infatti che tra generici errori appartenenti allo stesso, ma generico gruppo j

$$\begin{aligned} Cov(\Delta_{ij}, \Delta_{kj}) &= E[(U_{0j} + U_{1j}x_{ij} + E_{ij})(U_{0j} + U_{1j}x_{kj} + E_{kj})] = \\ &= E[U_{0j}^2] + x_{ij}E[U_{0j}, U_{1j}] + x_{kj}E[U_{0j}, U_{1j}] + x_{ij}x_{kj}E[U_{1j}^2] = \\ &= \tau_{00} + x_{ij}\tau_{01} + x_{kj}\tau_{01} + x_{ij}x_{kj}\tau_{11}. \end{aligned} \tag{6}$$

Questa covarianza assumerà valore nullo nel caso in cui $U_{0j} = U_{1j} = 0$, ovvero quando le w_j definiscono in maniera esatta i valori delle unità di livello-2 intercetta e pendenza. Dalla covarianza (6) si ottiene:

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

che misura la correlazione tra due generiche osservazioni, tenuto conto anche della esistente suddivisione in k classi. Per tale motivo, come sarà approfondito nel seguito, fornisce anche una misura dell'omogeneità all'interno dei gruppi, rappresentando la proporzione di varianza (residua) attribuibile all'esistenza del raggruppamento (Kreft, De Leeuw, 1998).
Si fa osservare che il modello (4) è proposto nella sua formulazione più generale. Infatti, potrebbe non essere necessario introdurre nel modello tutte le componenti casuali, così come potrebbe non essere necessario spiegare la variazione dei parametri β_{0j} e/o β_{1j} mediante la variabile esplicativa W, e neppure inserire la variabile esplicativa X, come accade, ad esempio, nel modello di analisi della varianza ad effetti casuali. Ne discende che il modello può essere specificato nei modi più appropriati a seconda delle relazioni ipotizzate tra le variabili.

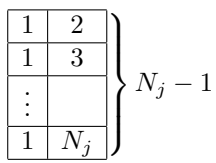
In un modello di regressione ordinario la varianza del termine di errore, indicato come varianza residua, rappresenta la quota di variabilità non spiegata dai regressori. In genere l'inserimento di un nuova variabile comporta una riduzione della varianza residua, la cui entità dipende dal suo potere esplicativo. La situazione è più complessa nel modello qui trattato, nel quale la varianza non spiegata dai regressori viene scomposta in due parti: la componente *between* σ_u^2, ovvero la varianza non spiegata dai regressori e che è attribuibile agli effetti casuali, ovvero alla struttura gerarchica; la componente within σ_e^2, ovvero la varianza residua in senso stretto, che non è spiegata nè dai regressori, nè dall'appartenenza ai gruppi, ma che è legata alla variabilità individuale. L'effetto dell'inserimento di nuove variabili sulle componenti di varianza dipende dalla loro caratterizzazione in (Longford, 1993, pp. 29-30):

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

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

5 Il coefficiente di correlazione intraclasse (ICC)

La correlazione intraclasse è una particolare misura del grado di dipendenza degli individui: più gli individui condividono le esperienze comuni, dovute alla vicinanza nel tempo e nello spazio, più sono simili. Il più alto livello di dipendenza può presentarsi, ad esempio, tra osservazioni di gemelli monozigoti, oppure tra bambini nati e cresciuti nella stessa famiglia. Un altro esempio riguarda le cosidette "misure ripetute", eseguite sulla stessa persona. La caratteristica principale dell'analisi multilevel è costituita dal fatto che in genere, trattandosi di dati gerarchicamente organizzati, le osservazioni individuali non sono indipendenti. La correlazione esistente tra individui appartenenti allo stesso gruppo viene detta *intra-class correlation*, generalmente indicata con il simbolo ρ, e può essere spiegata in diversi modi: ad esempio, può anche essere definita come misura di omogeneità all'interno di un gruppo. A partire da Fisher (1954), la correlazione intraclasse è stata considerata nel quadro dell'analisi della varianza (ANOVA): data la suddivisione in J gruppi, determinata, ad esempio, dall'esistenza delle cosiddette unità di livello 2, il coefficiente di correlazione intraclasse proposto dall'autore consiste nel rapporto tra la media dei prodotti degli scarti da μ (media generale) di tutte le $N_j(N_j - 1)$ coppie distinte che si possono formare con le N_j osservazioni, che indicheremo ancora con y_{ij}, contenute all'interno di ognuno dei J gruppi (vedi figura seguente), e



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

$$\vdots$$

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

il prodotto degli scarti quadratici (e quindi la varianza) delle osservazioni che formano le coppie, calcolata attraverso tutte le $N = \sum_{j=1}^{J} N_j$ osservazioni. Esso è dato cioè dalla seguente espressione

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

dove

$$\mu = \sum_{j=1}^{J} \sum_{i=1}^{N_j} \frac{y_{ij}}{N}$$

mentre

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

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

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

Si considerino infatti N coppie di dati $(y_{n,1}, y_{n,2})$ per $n = 1, ..., N$. Il coefficiente di correlazione intraclasse proposto da Fisher (1954) è

$$\frac{1}{Ns^2}\sum_{n=1}^{N}(y_{n,1}-\bar{y})(y_{n,2}-\bar{y})$$

dove

$$\begin{aligned}\bar{y} &= \frac{1}{2N}\sum_{n=1}^{N}(y_{n,1}+y_{n,2})\\ s^2 &= \frac{1}{2N}\{\sum_{n=1}^{N}(y_{n,1}-\bar{y})^2+\sum_{n=1}^{N}(y_{n,2}-\bar{y})^2\}\end{aligned}$$

generalizzabile anche per gruppi con più di due valori. Ad esempio, per gruppi di $k = 3$ soggetti:

$$\begin{aligned}\bar{y} &= \frac{1}{3N}\sum_{n=1}^{N}(y_{n,1}+y_{n,2}+y_{n,3})\\ s^2 &= \frac{1}{3N}\{\sum_{n=1}^{N}(y_{n,1}-\bar{y})^2+\sum_{n=1}^{N}(y_{n,2}-\bar{y})^2+\sum_{n=1}^{N}(y_{n,3}-\bar{y})^2\}\\ \rho &= \frac{1}{3Ns^2}\sum_{n=1}^{N}(y_{n,1}-\bar{y})(y_{n,2}-\bar{y})+(y_{n,1}-\bar{y})(y_{n,3}-\bar{y})+(y_{n,2}-\bar{y})(y_{n,3}-\bar{y}).\end{aligned}$$

L'ultima espressione può porsi nella forma equivalente

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

dove $\bar{y}_n$ è la media dell'$n-esimo$ gruppo (Harris, 1913). Per K elevato, questo coefficiente di correlazione intraclasse risulta poi

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

che può essere allora interpretato come la frazione della varianza totale imputabile alla varianza tra i gruppi.
Questo indice coincide con il coefficiente di correlazione intraclasse di Pearson nel caso in cui il numero dei gruppi tenda all'infinito e la numerosità all'interno dei gruppi diverga. Nel caso di dati organizzati in una struttura gerarchica a due livelli, l'*intra-class correlation* è definita quindi come proporzione di variabilità

attribuibile ai gruppi o, equivalentemente, come correlazione fra due generiche unità dello stesso generico gruppo. Se si è in presenza di correlazione intraclasse, come potrebbe succedere con questo tipo di dati, il presupposto della indipendenza delle osservazioni non è rispettato. Un effetto di tale violazione è l'incremento non controllabile della probabilità di commettere l'errore di prima specie (livello α), che in letteratura è associato alla presenza della correlazione intraclasse. I test statistici tradizionali sono basati sull'assunto di indipendenza tra le osservazioni e quando questa ipotesi risulta violata le stime degli errori standard prodotte dalle procedure convenzionali risultano distorte per difetto e, di conseguenza, i risultati che si ottengono potrebbero essere "impropriamente" significativi.

Il coefficiente di correlazione intraclasse è generalmente definito facendo riferimento ad un modello lineare ad effetti casuali (Donner, Koval, 1980a, 1980b, Donner, Wells, 1986). In tale situazione (Snijders, Bosker, 1999), si parte dalla usuale scomposizione della varianza totale in *within* (infra-gruppi) e *between* (inter-gruppi), ovvero:

$$\tau^2 + \sigma^2.$$

Se ne consideri ora la corrispondente formulazione campionaria relativa ad un campione di numerosità N; si indica con J il numero di macro unità osservate (gruppi) e con n_j il numero delle micro unità nella $j - esima$ macro unità, quindi $N = \Sigma_j n_j$. Si considerino, ad esempio, le ricerche effettuate nel campo dell'istruzione, attraverso le quali ci si propone di rilevare l'esistenza di differenze tra classi (gruppi) di studenti (unità statistiche) sulla base di una certa misura individuale di risultato, tenendo conto del fatto che sia le caratteristiche degli studenti che quelle delle classi possono essere rilevanti nel determinare tale risultato (Aitkin, Longford, 1986; Goldstein, Spiegelhalter, 1996). Vengono definite:

- la media della macro unità j

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

- la media generale

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

- la varianza della macro unità j

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

La varianza *within* si può interpretare come sintesi delle variabilità delle singole osservazioni all'interno delle macro unità; sarà quindi definita come media ponderata delle varianze entro le macro unità, ovvero

$$\begin{aligned} s^2_{within} &= \frac{1}{N-J}\sum_{i=1}^{n_j}\sum_{j=1}^{J}(y_{ij}-\bar{y}_j)^2 = \\ &= \frac{1}{N-J}\sum_{i=1}^{n_j}(n_j-1)s^2_j. \end{aligned}$$

Per quanto riguarda la varianza *between*, occorre distinguere due differenti situazioni. Per gruppi di uguale numerosità essa è definita come

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

mentre, per gruppi di diversa numerosità, il contributo dei vari gruppi viene pesato nel modo seguente

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

dove $\widetilde{n}$ è definito come

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

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

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

la varianza delle dimensioni delle macro unità.
La varianza totale s^2 può, allora, essere scritta come combinazione lineare delle varianze *within* e *between* sopra definite:

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

Il valore atteso della variabile casuale corrispondente alla varianza *within* è

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

mentre il valore atteso della variabile casuale corrispondente alla varianza *between* è

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

Le stime $\hat{\sigma}^2$ e $\hat{\tau}^2$ delle variaze degli errori di primo e secondo livello sono calcolate come:

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

e

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

Il coefficiente di correlazione intraclasse ρ indica quindi sia la correlazione tra due individui dello stesso gruppo che la quota di variabilità totale a livello di gruppo. Nel caso in cui il coefficiente di correlazione è significativamente diverso da zero si può affermare che parte della variabilità è attribuibile ai gruppi, e che quindi il macro livello influeza il micro.

Tale indice è stato proposto anche da Donner (1986), nella forma

$$\hat{\rho}(ICC) = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2}$$

dove la classe è identificata dal proprio livello medio, ovvero dallo scarto del livello medio rispetto alla media generale.

La correlazione intraclasse ha la caratteristica che, per un numero sufficientemente elevato di gruppi, fornisce la proporzione della varianza attribuibile alla differenza tra le medie dei gruppi.

Tale misura viene anche usata per la valutazione della coerenza o della riproducibilità delle misurazioni fatte da osservatori differenti sulle stesse quantità (Aitkin, Longford, 1986; Goldstein, Spiegelhalter, 1996).

6 ICC e particolari modelli multilevel

Volendo introdurre l'inferenza sui modelli multilevel è opportuno richiamare brevemente la struttura di alcuni modelli che risultano essere casi particolari di quelli generali a due livelli con variabili esplicative X e Z operanti, rispettivamente, al primo e al secondo livello. Il confronto con questi modelli costituisce, infatti, l'oggetto delle procedure inferenziali di validazione.

6.1 Empty model

Questo modello è caratterizzato da una struttura estremamente semplice: è infatti un particolare modello ANOVA ad effetti casuali che non prevede variabili esplicative. Tale modello considera cioè la presenza di gruppi estratti casualmente e la sola variabilità delle singole osservazioni: esso è denominato *Empty Model* e definisce il livello della variabile dipendente come somma della media generale γ_{00}

e degli effetti casuali u_{0j} a livello di gruppo e e_{ij} a livello individuale. Si assume che gli effetti casuali sono determinazioni di variabili casuali di media nulla e mutuamente indipendenti. Come nel modello più generale la varianza di Y può essere scomposta come somma delle varianze a livello 1 e a livello 2 nel seguente modo:

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

La covarianza tra due individui i e i' appartenenti allo stesso gruppo j corrisponde alla varianza di U_{0j}

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

e la loro correlazione

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

corrisponde al coefficiente di correlazione intraclasse.

6.2 Il modello ad intercetta casuale

Anche questo modello rappresenta un caso particolare del modello gerarchico lineare a due livelli, ed è conosciuto col nome di *Random Intercept Model* (Snijders, Bosker, 1999). Come nel classico modello di regressione lineare, si è in presenza di una variabile dipendente Y e di un predittore X, misurato al livello degli individui. In particolare, il modello assume la forma seguente:

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

dove y_{ij} rappresenta la variabile risposta dell'$i - esimo$ individuo appartenente alla $j - esima$ unità di secondo livello[4]. L'obiettivo è quello di stimare il valore atteso di y_{ij}, considerando l'effetto del predittore X, e ipotizzando che la variabile esplicativa sia caratterizzata da livelli medi differenti in ogni gruppo. Tale modello descrive l'effetto gruppo del predittore attraverso le variazioni dell'intercetta β_{0j}, mentre il coefficiente di regressione è costante nei gruppi (parallelismo). Gli e_{ij} sono gli errori a livello degli individui. L'intercetta variabile a livello di gruppo viene poi modellata come

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

dove γ_{00} rappresenta l'intercetta media di tutti i gruppi, mentre u_{0j} rappresenta la componente d'errore. Sostituendo quest'ultima equazione nella precedente si ottiene il modello completo

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

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

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

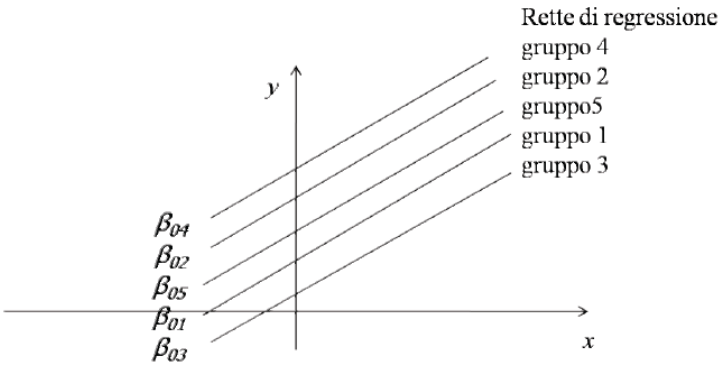


Figura 1: *Random intercept model*

Si assume che gli errori u_{0j} e e_{ij} sono determinazioni di variabili casuali U_0 e E_{ij} mutuamente indipendenti, con medie nulle e varianze rispettivamente τ_0^2 e σ^2. La variabile casuale U_0 può essere vista come descrittiva degli errori a livello di gruppo (effetti di gruppo) non spiegati da X. Dal momento che gli errori casuali contengono quella parte di variabilità della variabile dipendente che non è considerata come funzione di variabili esplicative, si può affermare che questo modello contiene variabilità non spiegata a due livelli annidati. La partizione della variabilità non spiegata sui vari livelli è l'essenza dei modelli gerarchici ad effetti casuali. All'interno del modello, γ_{00} è sempre l'intercetta media dei gruppi e β_1 può essere interpretato, nel modo usuale, come aumento teorico di Y derivante da un aumento unitario del livello di X. La varianza residua condizionata al valore generico di X è

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

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

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

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

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

e per il livello 2 questa frazione è

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

Della correlazione tra due individui dello stesso gruppo, una parte può essere spiegata dai rispettivi valori di X, dando luogo al cofficiente di correlazione intraclasse residuo:

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

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

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

che include anche, rispetto all'*Empty Model* gli effetti della variabile esplicativa.

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

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

6.3 Il modello a coefficienti casuali: random slopes

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

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij} \tag{7}$$

dove β_{0j} è la classica intercetta, β_{1j} è l'usuale coefficiente di regressione per la variabile esplicativa X, misurata sul livello degli individui, mentre e_{ij} rappresenta il termine d'errore. Come nel *random intercept model*, anche in questo caso la differenza rispetto al modello di regressione non gerarchico consiste nel fatto che ogni gruppo possiede una diversa intercetta, β_{0j}, ma ora anche un differente coefficiente di regressione, β_{1j}. Inoltre, si assume che, all'interno di ciascun gruppo, gli errori al livello individuale siano indipendenti e normalmente distribuiti con media nulla e varianza comune σ^2, $E_{ij} \sim N(0; \sigma^2)$. A causa della variazione tra le unità di livello superiore, i coefficienti in esame prendono il nome di coefficienti casuali. Le macro-unità sono ancora viste come un campione proveniente da una più vasta popolazione di gruppi: i coefficienti β_{0j} e β_{1j} del modello di regressione gerarchico vengono esplicitati dalle seguenti relazioni

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

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

per le quali si assume che i termini di errore u_{0j} e u_{1j}, spesso denominati macro-errori, siano normalmente distribuiti con media nulla e varianze τ_0^2 e τ_1^2, rispettivamente. Inoltre, si assume che i macro-errori siano indipendenti tra i gruppi e dagli errori di livello individuale e_{ij}; con σ_{u01}^2 viene indicata la covarianza tra i macro-errori u_{0j} e u_{1j}. In sintesi

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

Sostituendo le due precedenti espressioni nella (7) il modello di regressione multilevel può essere scritto in un'unica forma

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} z_j + \gamma_{11} x_{ij} z_j + u_{1j} x_{ij} + u_{0j} + e_{ij}. \tag{8}$$

Il termine $x_{ij} z_j$ è denominato cross-level interaction, poichè risente dell'effetto moderante delle variabili esplicative misurate su differenti livelli della gerarchia come mostrato in figura 2.
La parte

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

viene denominata parte sistematica del modello, mentre la parte

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

che contiene i termini casuali di errore, viene denominata parte aleatoria del modello. Essa costituisce una struttura complessa di errore e, come si può notare, gli errori all'interno delle macro unità sono correlati poichè u_{0j} e u_{1j} risultano comuni per le osservazioni che appartengono al medesimo gruppo. Il modello implica non solo che gli individui all'interno dello stesso gruppo abbiano valori di Y correlati, ma anche che questa correlazione, così come la varianza di Y, dipende dal valore di X (il termine d'errore u_{1j} è connesso con x_{ij}). Da ciò deriva che l'errore totale

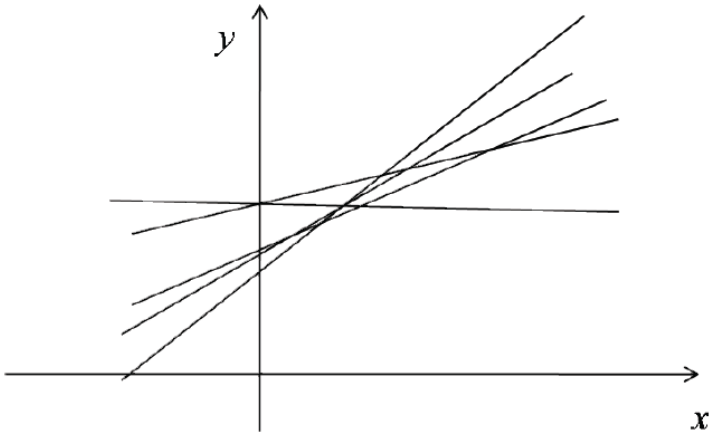


Figura 2: *Random slope model*

sarà differente per differenti valori di X, situazione questa, che nei modelli di regressione ordinari, prende il nome di eteroschedasticità. Risultano pertanto violate le assunzioni di indipendenza e di omoschedasticità degli errori, su cui si basano i modelli di regressione ordinari. Attraverso la (8) è, dunque, possibile stimare i coefficienti degli effetti fissi, degli effetti indipendenti delle variabili di secondo livello, di quelle di primo livello e la loro interazione. Il modello multilevel, inoltre, permette di quantificare la variabilità nei diversi livelli della gerarchia:

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

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

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

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

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

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

Quindi la quantità

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

risulta analoga all'usuale cofficiente di correlazione intraclasse, ma ora i parametri τ_0^2 e σ^2 sono riferiti alle varianze del modello (8) che include rispetto all'*Empty Model* valori differenti per l'intercetta e per il coefficiente di regressione.

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

7 Test di nullità di ICC

Da quanto fin qui esposto risulta giustificabile la proposta di verificare, in via preliminare, l'eventuale esistenza della struttura gerarchica/multilevel, attraverso un test sul coefficiente di correlazione intraclasse. Il coefficiente ICC può infatti assumere valore nullo, in assenza di raggruppamenti, oppure positivo, nel caso di presenza di raggruppamenti. Il sistema d'ipotesi più opportuno appare allora essere il seguente

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{cases} \tag{9}$$

nel quale l'ipotesi H_1 corrisponde all'esistenza di struttura gerarchica. Ci si potrebbe chiedere se il sistema di ipotesi sopra introdotto sia il più adeguato, ovvero se anche piccoli valori di ρ, che inducano una debole violazione dell'assunzione di osservazioni indipendenti, comportano effetti di rilievo sulle procedure inferenziali. In Barcikovski (1981) viene presentata una simulazione riguardante la modifica

	ρ			
n_j	0.00	0.01	0.05	0.20
10	0.05	0.06	0.11	0.28
25	0.05	0.08	0.19	0.46
50	0.05	0.11	0.30	0.59
100	0.05	0.17	0.43	0.70

Tabella 1: One Way ANOVA, valore effettivo di α (Barcikovski, 1981)

prodotta sul livello effettivo dell'errore di primo tipo $\alpha = 0.05$, anche in presenza di piccoli valori di ρ. L'autore indica infatti che, a parità delle numerosità di gruppo n_j, il valore di α già si modifica anche per piccoli valori di ρ e cresce all'aumentare di ρ come dettagliato in Tabella 1.

Come si può notare, con solo 10 soggetti per gruppo e un ICC di 0.01, il vero valore di α risulta essere 0.06, poco superiore a quello nominale. Se si considerano però più soggetti o un ICC leggermente superiore a 0.01 il vero valore di α risulta molto superiore a 0.05. Ne deriva che, se non si considera in maniera adeguata la presenza di ICC nei dati da analizzare, gli standard error dei parametri stimati risulteranno sottostimati.

7.1 ICC: la statistica test

Si consideri un campione casuale di N soggetti raggruppati in J gruppi di numerosità n_j $(j = 1, ..., J)$, che si suppongono anch'essi estratti casualmente da una popolazione di gruppi, e si adotti come modello di riferimento, quello ad effetti casuali in cui si trascurerà, senza perdita di generalità, la presenza di variabili esplicative e denominato, nei precedenti paragrafi, modello ad intercetta casuale.

Il test si basa sulla scomposizione della varianza della variabile risposta generica Y_{ij}, la cui distribuzione viene indicata con $f(\cdot; u_j)$, essendo le u_j (errori di secondo livello) determinazioni di v.c. indipendenti con funzione di ripartizione che indicheremo con $H(\cdot)$. La varianza di Y_{ij} si può scomporre in funzione dei suoi momenti condizionali

$$Var(Y_{ij}) = E[Var(Y_{ij}|u_j)] + Var[E(Y_{ij}|u_j)].$$

Il primo termine risulta essere la varianza entro i gruppi, mentre il secondo termine è la varianza tra i gruppi, ovvero l'eterogeneità tra i gruppi delle medie condizionate.

In Commenges, Jacqmin (1994) si mostra che il modello ad intercetta casuale, del tipo $y_{ij} = \gamma_{00} + u_j + e_{ij}$, appartiene alla cosiddetta classe ICRE *Intraclass Correlation Random Effect*, e che per il quale si può porre

$$u_j = \bar{u} + \theta^{1/2} \upsilon_j$$

con $\bar{u} = E(U_j)$ e $\theta = Var(U_j)$, e dove i υ_j hanno distribuzione di media nulla e varianza unitaria e funzione di ripartizione che indicheremo con $\tilde{H}$.

La distribuzione delle y_{ij} può quindi scriversi come $f(\cdot; \upsilon_j, \bar{u}, \theta)$, dove le costanti $\bar{u}$ e θ assumono il ruolo di parametri *nuisance* (non di interesse), mentre i υ_j quello di parametri naturali (di interesse).

Utilizzando le precedenti notazioni il test (9) può essere allora riformulato come $H_0 : \theta = 0$, che implica sia $Var(U_j) = 0$, che $\rho = 0$, oppure come

$$H_0 : (u_1, u_2, ..., u_J) = (u_0, u_0, ..., u_0)$$

che propone il vettore $(u_1, u_2, ..., u_j)$ essere quello dei parametri di interesse in un problema inferenziale equivalente al (9).

Come indicato in Liang (1987) la funzione di log-verosimiglianza per il gruppo $j - esimo$ risulta

$$l_j = log \int \prod_{i=1}^{n_j} f(y_{ij}; \bar{u} + \theta^{\frac{1}{2}} \upsilon_j) d\widetilde{H}(\upsilon_j)$$

mentre la *score statistic*, per H_0, è data da

$$S = \frac{1}{2} \sum_{j=1}^{J} \{ [\sum_{i=1}^{n_j} \frac{\partial}{\partial \upsilon_j} \log f(y_{ij}; \upsilon_j, \bar{u}, \theta)]^2 + [\sum_{i=1}^{n_j} \frac{\partial^2}{\partial \upsilon_j^2} \log f(y_{ij}; \upsilon_j, \bar{u}, \theta)] \}$$

dove $\bar{u}$ e θ, parametri di disturbo, vanno sostituiti con le corrispondenti stime, consistenti, calcolate sotto H_0.

Se si considera il caso in cui f appartiene alla famiglia esponenziale (McCullagh, Nelder, 1983), con parametro canonico (valore atteso di Y_{ij}) funzione di un effetto casuale u_j, mentre il parametro di dispersione non dipende da u_j, ovvero

$$f(y_{ij}; \upsilon_j, \bar{u}, \theta) = \exp\{[y_{ij}\upsilon_j - b(\upsilon_j)]\theta^{-1} + c(y_{ij}; \bar{u}, \theta)\},$$

si ha che lo score è proporzionale allo scarto dalla media, cioè

$$\frac{\partial}{\partial \upsilon_j} \log f(y_{ij}; \upsilon_j, \bar{u}, \theta) = (y_{ij} - \gamma_{00})\theta^{-1}$$

dove $\gamma_{00} = E(Y_{ij}) = b'(\upsilon_j)$; la statistica S diventa allora

$$S = \frac{1}{2} \sum_{j=1}^{J} \{ [\sum_{i=1}^{n_j} \frac{y_{ij} - \gamma_{00}}{\theta}]^2 \frac{b''(\upsilon_j) n_j}{\theta} \}$$

dove $Var(Y_{ij}) = \theta b''(\upsilon_j) = Var(U) = \sigma^2$. Un test statistico equivalente risulta essere quello basato su:

$$S' = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{i'=1}^{i-1} (y_{ij} - \gamma_{00})(y_{i'j} - \gamma_{00}) + \frac{1}{2} \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \gamma_{00})^2 - \frac{1}{2} N\sigma^2.$$

Se γ_{00} e σ^2 non sono note, si ottiene una versione asintotica equivalente considerando gli stimatori di massima verosimiglianza

$$\hat{\gamma_{00}} = \frac{\sum_j \sum_i y_{ij}}{N}$$

$$\hat{\sigma}^2 = \frac{\sum_j \sum_i (y_{ij} - \hat{\gamma}_{00})^2}{N},$$

per cui la statistica si riduce alla seguente espressione

$$S' = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{i'=1}^{i-1} (y_{ij} - \hat{\gamma}_{00})(y_{i'j} - \hat{\gamma}_{00}).$$

Il test può quindi essere eseguito attraverso la statistica

$$C = \frac{S'(\hat{\gamma}_{00})}{Var(S'(\hat{\gamma}_{00}))^{\frac{1}{2}}}$$

avente, per N elevato, distribuzione normale.

Quando la numerosità dei gruppi è costante ($n_j = n$), la statistica C è equivalente alla statistica F, tipica dei modelli lineari, del tipo

$$F = \frac{SSB}{(k-1)} / \frac{(SS - SSB)}{(N-k)}$$

dove $SSB = nS'$ è la devianza tra i gruppi e $SS = \sum_{j=1}^{J} \sum_{i=1}^{n} (y_{ij} - (\hat{\gamma}_{00}))^2$ è la devianza totale. Si dimostra infatti che F è una funzione monotona crescente della statistica C (Commenges, Jacqmin , 1994).

Nel caso bilanciato, $n_j = n$, se la distribuzione degli errori è normale, F è un test UMPU *Uniformly Most Powerful Unbiased* (Kendall, Stuart, 1977, Cap. 37). Nel caso non bilanciato, con distribuzione normale, F e C non sono più coincidenti, e il test F non gode più di proprietà ottimali.

Nella maggior parte dei casi pratici i gruppi sono non bilanciati, per cui la distribuzione della statistica C sarà diversa dalla distribuzione normale. Considerando il modello con variabili esplicative ed errori con distribuzione appartenente alla famiglia esponenziale, si può utilizzare una versione nonparametrica del test C (Commenges, Jacqmin, 1994).

8 Considerazioni sulla robustezza di ICC

Da ultimo, si è ritenuto opportuno completare l'esame delle caratteristiche dell'ICC presentando i risultati di una simulazione, eseguita allo scopo di saggiare la robustezza dell'indicatore in assenza di normalità. Si sono generati due campioni di 100 gruppi di 50 soggetti ciascuno, secondo il modello multilivello a due livelli, dapprima utilizzando errori distribuiti secondo la v.c. normale, successivamente

secondo la v.c. *skew-normal* (SN) (Azzalini, 2005). Quindi, seguendo la procedura proposta in Bliese (2009, p. 43), si sono confrontate, graficamente, le medie ordinate dei gruppi precedentemente creati, con quelle stimate sulla base di un ricampionamento di tutte le osservazioni per formare dei gruppi di natura puramente casuale. Per il confronto tra i valori iniziali, generati secondo una struttura gerarchica, e le medie dei gruppi estratti casualmente dalle precedenti osservazioni, si sono costruiti anche gli intervalli di confidenza delle stime bootstrap delle medie, ordinate, dei gruppi casuali.

I grafici di figura 3, il primo a sinistra relativo agli errori con distribuzione normale, il secondo a destra con errori SN, vengono proposti come metodo grafico di verifica empirica della presenza di una gerarchia: quante più medie bootstrap finiscono fuori dalle bande, tra le 100 in totale, tanto più evidente è l'effetto del raggruppamento. Si fa osservare che le classi non centrali presentano bande di confidenza più ampie, essendo caratterizzate da più elevata variabilità.

Come può notarsi dal confronto dei due grafici, la presenza di errori non normali (grafico di destra) riduce la capacità dell'indice ICC di evidenziare l'effettiva presenza di gruppi. Anche il valore di ICC, si osservi, passa da 0.0282 a 0.0245, con una riduzione di circa il 15% per il solo effetto della non normalità.

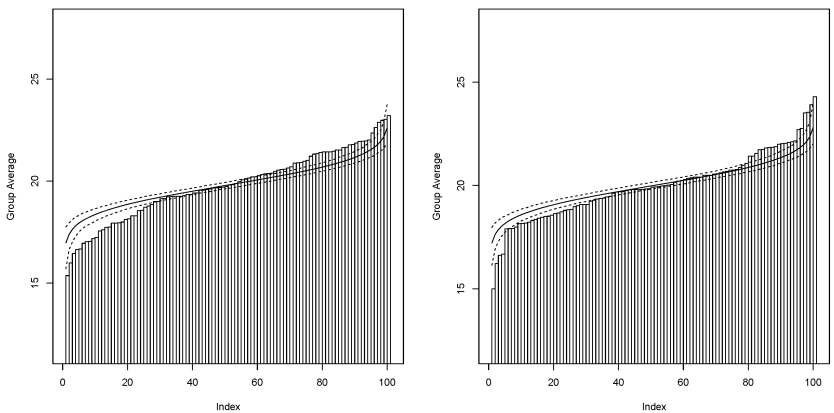


Figura 3: Simulazione ICC per modello multilivello con errori distribuiti normalmente (grafico a sinistra) e con errori distribuiti secondo una v.c. SN (grafico a destra)

9 Ringraziamenti

L'autore ringrazia il Prof. Boari per la collaborazione alla stesura del lavoro, alla revisione e alla correzione dello stesso, nonchè per i preziosi e continui suggerimenti.

Riferimenti bibliografici

[1] Aitkin M., Longford N. T. (1986) *Statistical modeling issues in school effectiveness studies (con discussione)*, J.R. Statist. Soc. A, 149, 1-43.

[2] Alker H. R., (1969) *A typology of fallacies*, in M. Dongan e S. Rokkan, Eds. Quantitative ecological analysis, The Social Science, Cambridge Ma. M.I.T. Press.

[3] Andersen R., Heath A. (2002) *Class matters. The persisting effects of contextual social class on individual voting in Britain*, European Sociological Review, 18, 1964-97.

[4] Azzalini A., (2005) *The skew-normal distribution and related multivariate families*, Scand. J. Statist., 32(2), 159-188.

[5] Barcikowski R. S., (1981) *Statistical power with group mean as the unit of analysis*, Journal of the Educational Statistics, 6 (3), 267-285.

[6] Bliese P., (2009) *Multilevel Modeling in R (2.3)*, $http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf$.

[7] Browne, Bryk W. J., Draper D., Goldstein H., Rasbash J. (2000) *Bayesian and Likehood Methods for Fitting Multilevel Modeling* , Computational Statistics and Data Analysis, 39 (2), 203-225.

[8] Charnock D., (1996) *Class and voting in the 1996 Australian Federal Election*, Electoral Studies, 16 (3), 281-300.

[9] Commenges D., Jacqmin H., (1994) *The Intraclass Correlation Coefficient: Distribution-Free Definition and Test*, Biometrics, 50, 517-526.

[10] Donner A., Koval J.J., (1980a) *The estimation of intraclass correlation in the analysis of family data* , Biometrics, 36, 19-25.

[11] Donner A., Koval J.J., (1980b) *The large sample variance of an intraclass correlation*, Biometrika, 67, 719-722.

[12] Donner A., Wells G., (1986) *A comparison of confidence interval methods for the intraclass correlation coefficient*, Biometrics, 36, 401-412.

[13] Donner A., (1986) *A review of inference procedures for the intraclass correlation coefficient in the one-way random effect model*, International Statistical Review, 54, 67-82.

[14] Fisher R. A., (1954) *Statistical Methods for Research Workers (Twelfth ed.)*,Oliver and Boyd, http://psychclassics.yorku.ca/Fisher/Methods/.

[15] Goldstein H., (1987) *Multilevel Covariance Component Models*, Biometrika, 74, 4300-431.

[16] Goldstein H., (1995) *Hierarchical Data Modeling in the Social Sciences*, Journal of Educational and Behavioral Statistics, 20, 201-204.

[17] Goldstein H., (2011) *Multilevel Statistical Models*, 4nd ed., John Wiley and Sons, Chichester, UK.

[18] Goldstein H., Spiegelhalter D. J., (1996) *League Tables and their Limitation: Statistical Issues in Comparisons of Istitutional Performance*, Journal of the Royal Statistical Society Serie A (Statistics in Society), 153 (3), 385-443.

[19] Goldstein H., Rabash J., Plewis I., Draper D., Browne W., Yang M., Woodhouse G. e Healy M.J.R., (1998) *A User's Guide to MLwiN*, Institute of Education, Londra.

[20] Harris A., (1913) *On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large*, Biometrika, 9, 446-472.

[21] Hox J.J., (1995) *Applied Multilevel Analysis*, TT-Publikaties, Amsterdam.

[22] Hox J.J., (2002) *Multilevel Analysis: Techniques and Applications*, Erlbaum, New Jersey.

[23] Kendall N., Stuart A. (1977) *The Advanced Theory of Statistics*, Vol. 3, London Griffin.

[24] Kreft Ita G. G., De Leeuw J. (1998) *Introducing Multilevel Modeling*, Sage, London.

[25] Kish L., (1995) *Survey Sampling*, New York, Wiley e Sons.

[26] Liang K. Y., (1987) *A Locally Most Powerful Test for Homogeneity with Many Strata*, Biometrika, 74, 259-264.

[27] Longford N., (1993) *Random Coefficient Models*, Oxford, Clarendon Press.

[28] McCullagh P., Nelder J. A., (1989) *Generalized linear models*, 2nd ed., Chapman and Hall, Londra.

[29] Pintaldi F., (2003) *I dati ecologici nella ricerca sociale*, Carocci, Roma.

[30] Raudenbush S. W., Bryk A. S., (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed., SAGE publications, Newbury Park CA.

[31] Robinson W. S., (1950) *Ecological Correlations and the Behavior of Individuals*, American Sociological Review, 15, 351-357.

[32] Snijders T.A.B., Bosker R.J., (1999) *Multilivel analysis: An introduction to basic and advanced multilevel modeling*, SAGE, Londra.

[33] Thum Y. M., (1997) *Hierarchical Linear Models for Multivariate Outcomes*, Journal of Educational and Behavioral Statistics, 22, 77-88.

[34] Zaccarin S., Rivellini G., (2002) *Multilevel analysis in social research: an application of a cross-classified model*, Statistical Methods and Applications, 11(1), 95-108.