Corso della Scuola della SIS
Metodi statistici per la valutazione e il monitoraggio della formazione universitaria
Firenze, 10-14 ottobre 2005

Introduzione ai modelli statistici per la valutazione e il monitoraggio

Leonardo Grilli
grilli@ds.unifi.it
www.ds.unifi.it/grilli

Dipartimento di Statistica "G. Parenti", Firenze

Indice

1. Modelli statistici
2. Introduzione ai modelli multilivello
3. Il modello lineare a due livelli
4. Modelli multilivello e valutazione
5. Inferenza
6. Software e libri

Modelli statistici

Modello

- Modello: schema teorico che descrive un fenomeno ipotizzando le *caratteristiche strutturali* più rilevanti
- Modello statistico: modello di tipo matematico con
  - una *componente deterministica*
  - una *componente stocastica*

Modello statistico

$$\boxed{y = f\left(\mathbf{x}, \mathbf{z}, \mathbf{w}\right)}$$

y	variabile di risposta
$\mathbf{x}$	variabili esplicative molto influenti su y e osservate
$\mathbf{z}$	variabili esplicative molto influenti su y e non osservate
$\mathbf{w}$	variabili esplicative poco influenti su y
$f(\cdot)$	funzione ignota

Modello statistico

Ipotesi di errori additivi:

$$\boxed{y = f\left(\mathbf{x}, \mathbf{z}\right) + e_{\mathbf{w}}}$$

$$\boxed{y = f\left(\mathbf{x}\right) + e_{\mathbf{z}} + e_{\mathbf{w}}}$$

y	variabile di risposta
$\mathbf{x}$	variabili esplicative molto influenti su y e osservate
$\mathbf{z}$	variabili esplicative molto influenti su y e non osservate
$\mathbf{w}$	variabili esplicative poco influenti su y
$f(\cdot)$	funzione ignota

Modello statistico

Ipotesi di linearità

-> modello di regressione lineare (multipla)

$$y = \mathbf{x\beta} + e_{\text{non linearità}} + e_z + e_w$$
$$= \mathbf{x\beta} + \quad e_{\text{totale}}$$

$\mathbf{x\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ Componente sistematica (segnale): quantità ignota e deterministica

e_{totale} Componente accidentale (rumore): quantità ignota e stocastica

Modello statistico

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

- Sia la componente sistematica che quella accidentale sono ignote perché includono dei parametri:
  - Componente sistematica: $k+1$ coefficienti di regressione
  - Componente accidentale: 1 o più parametri della distribuzione di e

Linearità nei parametri

$\Rightarrow$ sono ammesse trasformazioni delle variabili, es.

$$\log y = \beta_0 + \beta_1 \frac{1}{x} + e$$

I dati

Campione di n unità statistiche

$$i = 1, 2, \ldots, n$$

$$\begin{array}{cc} & \overbrace{\qquad\qquad\qquad\qquad}^{\text{variabili}} \\ \text{Unità statistiche} \left\{ \right. & \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \end{array}$$

Popolazione e campione

Il modello descrive la relazione fra la y e le x nella POPOLAZIONE e si assume valido per ogni unità del CAMPIONE

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + e_i$$

$$i = 1, 2, \ldots, n$$

Assunzioni sugli errori

Modello lineare classico: per la distribuzione degli errori (condizionatamente alle variabili esplicative) si assume

- errori a media nulla $\quad E(e_i) = 0 \quad \forall i$
- errori omoschedastici $\quad Var(e_i) = \sigma^2 \quad \forall i$
- errori incorrelati $\quad Cov(e_i, e_j) = 0 \quad \forall i \neq j$

Le assunzioni sugli errori determinano le proprietà degli stimatori (distorsione, varianza campionaria,...)

Regressione lineare semplice

Caso speciale con una sola variabile esplicativa (k=1)

$$y = \gamma_0 + \gamma_1 x + e_{y|x}$$

Parametri del modello:

$\gamma_0 =$ intercetta (nella popolazione)

$\gamma_1 =$ pendenza o coefficiente angolare (nella popolazione)

$\sigma^2_{y|x} = Var(e_{y|x}) =$ varianza residua

Regressione lineare semplice

Da $E\left(e_{y|x}\right)=0$ segue

$$\boxed{E(y\,|\,x)=\gamma_0+\gamma_1 x}$$

modello per la media condizionata di y (media di y dato x)

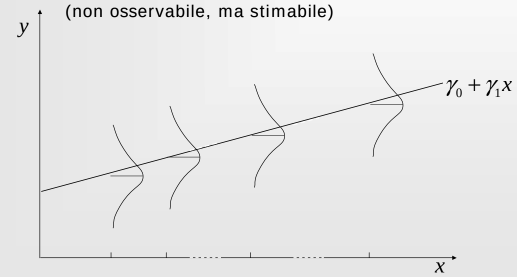
Interpretazione della pendenza:

$$\gamma_1 = E(y\,|\,x=x^*+1)-E(y\,|\,x=x^*)$$

Variazione della media condizionata di y corrispondente ad un *aumento unitario* di x

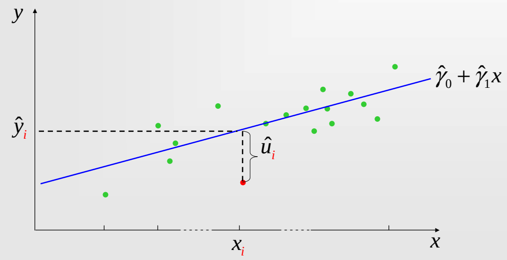
Regressione lineare semplice

Modello: relazione nella popolazione
(non osservabile, ma stimabile)



Regressione lineare semplice

Dati e relazione stimata



Regressione lineare multipla

k variabili esplicative

$$\boxed{y=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k+e_{y|x_1\ldots x_k}}$$

Parametri del modello:

β_0 = intercetta (nella popolazione)

$\beta_1,\ldots,\beta_k$ = pendenze o coeff. angolari (nella popolazione)

$\sigma^2_{y|x_1\ldots x_k}=Var(e_{y|x_1\ldots x_k})=$ varianza residua

Regressione lineare multipla

Da $E\left(e_{y|x_1\ldots x_k}\right)=0$ segue

$$\boxed{E(y\,|\,x_1,\ldots,x_k)=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}$$

modello per la media condizionata di y (media di y dato $x_1,\ldots,x_k$)

Interpretazione della pendenza β_1:

$$\beta_1 = E(y\,|\,x_1^*+1,x_2^*,\ldots,x_k^*)-E(y\,|\,x_1^*,x_2^*,\ldots,x_k^*)$$

Variazione della media condizionata di y corrispondente ad un *aumento unitario* di x_1
a parità di $x_2,\ldots,x_k$

Regressione lineare multipla

β_1 = effetto di x_1 su y "a parità di" $x_2,\ldots,x_k$

"al netto di"
"controllando per"

Il modello di regressione consente di fare esperimenti virtuali per valutare come cambia la variabile di risposta "muovendo" una variabile esplicativa alla volta (cioè, "tenendo ferme" tutte le altre)

Effetti lordi e netti

$y = \gamma_0 + \gamma_1 x_1 + e_{y|x_1}$

effetto lordo di x_1

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e_{y|x_1 \ldots x_k}$

effetto netto di x_1

In generale $\beta_1 \neq \gamma_1$

L. Grilli - Scuola SIS 2005 19

Modelli lineari generalizzati

La specificazione lineare per la media condizionata

$$E(y \mid x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

è molto conveniente, ma talvolta non è adeguata, in particolare quando per costruzione la media condizionata assume valori in un intervallo limitato

Esempio: y binaria $\Rightarrow E(y \mid x_1, \ldots, x_k) \in (0,1)$
mentre $(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k) \in \Re$

L. Grilli - Scuola SIS 2005 20

Modelli lineari generalizzati

In un modello lineare **generalizzato** la specificazione lineare non viene applicata direttamente alla media condizionata ma a una sua trasformazione:

$$g(E(y \mid x_1, \ldots, x_k)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$g(\cdot)$ funzione di link (invertibile con codominio $\Re$)

Esempio: y binaria & $g(\cdot)$ logit $\Rightarrow$ modello logit

L. Grilli - Scuola SIS 2005 21

Qual è il modello giusto?

- La realtà è troppo complessa per poter essere rappresentata in modo esaustivo da un modello
- Pertanto: un modello è “buono” quando
  - Coglie gli aspetti salienti del fenomeno (ruolo descrittivo)
  - Aiuta a rispondere ai quesiti della ricerca (ruolo strumentale)

“Tutti i modelli sono sbagliati, ma alcuni sono utili” (G.E.P. Box)

L. Grilli - Scuola SIS 2005 22

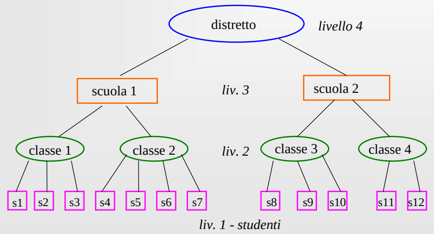
Qual è il modello giusto?

- La specificazione del modello (in particolare la scelta delle variabili esplicative) è guidata da:
  - Conoscenza del fenomeno (teoria)
  - Dati (evidenza empirica)
- Lo statistico è chiamato a stabilire, caso per caso, un ragionevole compromesso tra *parsimonia* e *complessità*

L. Grilli - Scuola SIS 2005 23

Introduzione ai modelli multilivello

Un esempio di struttura gerarchica



Modello multilivello

- Modello multilivello: modello di regressione con una struttura di errore complessa che rispecchia una struttura gerarchica
- Sinonimi:
  - M. gerarchico
  - M. a coefficienti casuali
  - M. a effetti casuali
  - M. a effetti misti

Esempi di strutture gerarchiche rilevanti nella valutazione dei sistemi universitari

- **Valutazione della didattica (1):** studenti, insegnamenti, corsi di laurea, facoltà, atenei
- **Valutazione della didattica (2):** studenti, insegnamenti, settori scientifico-disciplinari, atenei
- **Valutazione degli sbocchi occupazionali**: laureati, corsi di laurea, atenei

Quale struttura adottare dipende dal contenuto informativo dei dati e dalle finalità dell'analisi

L'esistenza della struttura gerarchica *non* dipende dal piano di campionamento (che può essere semplice, a più stadi, ecc.)

Livelli gerarchici

- Per semplicità consideriamo solo due livelli gerarchici:
  - Livello 1 (es. laureato)
  - Livello 2 (es. corso di laurea)

Sinonimi per "unità di liv. 1":	Sinonimi per "unità di liv. 2":
unità micro	unità macro
unità within	unità between
individuo	gruppo (cluster)

Variabili e livelli gerarchici

Le variabili sono riferite ad un certo livello della gerarchia:

Es. Valutazione degli sbocchi occupazionali:

laureato (liv. 1):	genere, voto di laurea
corso di laurea (liv. 2):	classe di appartenenza, voto medio di laurea

Nota: le variabili di livello 2 (o superiore) si distinguono in

GLOBALI: caratteristiche intrinseche delle unità macro (gruppi) che vengono rilevate separatamente e per le quali non esiste la corrispondente misura individuale: es. classe di appartenenza del CdL, rapporto studenti/docenti

CONTESTUALI: indicatori macro ottenuti per aggregazione delle corrispondenti misure individuali; esprimono la misura collettiva delle caratteristiche del singolo: es. voto medio di laurea, proporzione di femmine

Il dilemma dell'unita' di analisi

$j = 1, \ldots, J$ gruppi

$i = 1, \ldots, n_j$ individui nel gruppo j

dimensione totale $N = \sum_{j=1}^{J} n_j$

- Si può scegliere di analizzare i dati
  - a livello individuale (es. laureato) -> analisi disaggregata (archivio con N record)
  - a livello di gruppo (es. CdL) -> analisi aggregata (archivio con J record, ottenuti calcolando le medie di gruppo)
- Entrambe queste scelte danno luogo a dei problemi

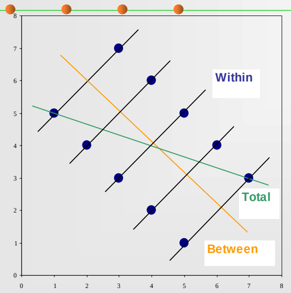
Problemi dell'analisi disaggregata

- **Inferenza sui gruppi**: impossibile fare inferenza sui gruppi, cioè trattare i gruppi osservati come un campione casuale da una popolazione di gruppi
- **Errata dimensione campionaria delle variabili di livello 2** (che è J e non N)
- **Dipendenza**: Le osservazioni all'interno di un gruppo sono fra loro più simili rispetto a quelle di altri gruppi, per cui si ha una correlazione positiva all'interno dei gruppi $\Rightarrow$ viene violata l'ipotesi di indipendenza tipica dei metodi tradizionali (es. GLM), che quindi forniscono un'**errata stima degli errori standard** (spesso si ha una sottostima degli errori standard -> errori del I tipo più alti del livello nominale α)

Problemi dell'analisi aggregata

- **Shift of meaning**: le variabili aggregate si riferiscono al gruppo e non all'individuo, per cui non possono nemmeno concettualmente essere usate per indagare le relazioni a livello di individuo
- **Ecological fallacy (distorsione da aggregazione)**: Le relazioni a livello di gruppo (cioè tra le medie di gruppo) sono diverse dalle corrispondenti relazioni a livello individuale
- **Interazione tra livelli**: l'analisi aggregata non consente di studiare le relazioni tra livelli gerarchici

Relazioni entro e tra gruppi



Esempio da Snijders & Bosker, p. 27

i	X_ij	X_.j	Y_ij	Y_.j
1	1	2	5	6
2	3	2	7	6
1	2	3	4	5
2	4	3	6	5
1	3	4	3	4
2	5	4	5	4
1	4	5	2	3
2	6	5	4	3
1	5	6	1	2
2	7	6	3	2

Differenza Between-Within: "Ecological fallacy"

Relazioni entro e tra gruppi

$\widehat{Y}_{ij} = 5.33 - 0.33 x_{ij}$ Regressione totale

$\widehat{\overline{Y}}_{.j} = 8.00 - 1.00 \overline{x}_{.j}$ Regressione tra le medie di gruppo

$\widehat{Y}_{ij} = \overline{Y}_{.j} + 1.00\left(x_{ij} - \overline{x}_{.j}\right)$ Regressione entro i gruppi

$\widehat{Y}_{ij} = 8.00 - 1.00 \overline{x}_{.j} + 1.00\left(x_{ij} - \overline{x}_{.j}\right)$ Regressione multilivello

La regressione multilivello consente di studiare contemporaneamente le relazioni between e within

ANOVA ad effetti casuali ...ovvero il modello multilivello più semplice

$j = 1, \ldots, J$ gruppi $\quad i = 1, \ldots, n_j$ individui nel gruppo j

$$\boxed{Y_{ij} = \mu + u_j + \varepsilon_{ij}}$$

$\varepsilon_{ij} \sim \text{iid}, \quad \mathrm{E}(\varepsilon_{ij}) = 0, \quad Var(\varepsilon_{ij}) = \sigma^2$

$u_j \sim \text{iid}, \quad \mathrm{E}(u_j) = 0, \quad Var(u_j) = \tau$

ε_{ij} indipendente da $u_j \quad \forall i, j$

Attenzione: τ è la varianza e non la deviazione std.

Questo modello ha 3 parametri

media generale: μ
varianza di livello 1: σ^2
varianza di livello 2: τ

ANOVA ad effetti casuali ...ovvero il modello multilivello più semplice

Esempio: analisi dei tempi di laurea (liv. 1 laureato, liv. 2 CdL)

$J =$ numero di CdL nell'archivio
$n_j =$ numero di laureati del CdL j
$Y_{ij} =$ tempo impiegato dal laureato i del CdL j
$\mu =$ tempo medio generale (tutto l'Ateneo)
$u_j =$ scostamento del tempo medio del CdL j da quello generale
$\tau = Var(u_j) =$ varianza dei tempi attribuibile ai CdL
$\varepsilon_{ij} =$ scostamento del tempo del laureato i rispetto al tempo medio del CdL j
$\sigma^2 = Var(\varepsilon_{ij}) =$ varianza residua dei tempi (cioè non attribuibile ai CdL)

ANOVA ad effetti casuali: varianze e covarianze

$$\boxed{Y_{ij} = \mu + u_j + \varepsilon_{ij}}$$

$$Var(Y_{ij}) = Var(u_j + \varepsilon_{ij}) = \tau + \sigma^2 \qquad Cov(Y_{ij}, Y_{i'j'}) = \begin{cases} 0 & \text{se } j \neq j' \\ \tau & \text{se } j = j' \text{ e } i \neq i' \end{cases}$$

- La variabilità di Y_{ij} viene scomposta in una quota legata alla variabilità tra gruppi (τ) ed una alla variabilità individuale (σ^2) (componenti di varianza)
- Le osservazioni appartenenti allo stesso gruppo sono correlate positivamente

OSSERVAZIONE: la correlazione è necessariamente positiva perché è generata da una variabile latente condivisa u_j (è la stessa idea fondamentale dell'analisi fattoriale, in cui u_j è chiamata *fattore*)

ANOVA ad effetti casuali: matrice di covarianza

$$\boxed{\text{Esempio } J = 2, \quad n_1 = 2, \quad n_2 = 3}$$

$$Var(\mathbf{Y}) = \begin{bmatrix} \tau+\sigma^2 & \tau & & & \\ \tau & \tau+\sigma^2 & & & \\ & & \tau+\sigma^2 & \tau & \tau \\ & & \tau & \tau+\sigma^2 & \tau \\ & & \tau & \tau & \tau+\sigma^2 \end{bmatrix}$$

ANOVA ad effetti casuali: coefficiente di correlazione intraclasse

ρ denota l'ICC (intraclass correlation coefficient)

$$\rho = Corr(Y_{ij}, Y_{i'j}) = \frac{\tau}{\tau+\sigma^2} = \frac{\text{varianza dovuta ai gruppi}}{\text{varianza totale}} \qquad \rho \in [0,1]$$

- ρ fornisce una misura del grado di omogeneità tra osservazioni appartenenti allo stesso gruppo.
- Maggiore è il valore di ρ e tanto più importante è utilizzare una procedura di stima adeguata che tenga conto della dipendenza

Il modello lineare a due livelli

Esempio: valutazione delle scuole

- **Livelli di analisi**: 1° livello, studenti; 2° livello, scuole
- **Variabile risposta Y**: punteggio test finale
- **Variabile esplicativa di 1° livello X** : punteggio test d'ingresso

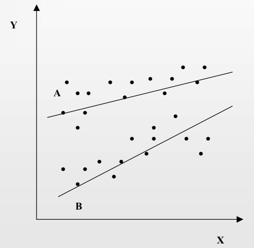
Approccio delle regressioni separate: analizzare ogni scuola separatamente, ad es. con il modello di regressione lineare semplice:

$$\boxed{Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}} \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

L'indice j è presente anche nei parametri -> ogni scuola ha una diversa intercetta e pendenza!

Esempio: valutazione delle scuole

- Sia l'intercetta che la pendenza sono importanti per la valutazione:
  - La scuola A è più **efficace** (valori previsti di Y più elevati per tutto il range di X)
  - La scuola A è anche più **equa** (pendenza inferiore)



Modello gerarchico lineare a due livelli

Ad ogni scuola corrisponde una coppia (β_{0j}, β_{1j})

Approccio delle regressioni separate : per ogni scuola (β_{0j}, β_{1j}) sono dei parametri e non vi è relazione tra i parametri di scuole diverse

Approccio gerarchico (multilivello): le coppie (β_{0j}, β_{1j}) delle varie scuole sono realizzazioni indipendenti da una distribuzione di probabilità, tipicamente normale bivariata

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ & \tau_{11} \end{bmatrix} \right)$$

Altre distribuzioni sono possibili ma la normale è usualmente preferibile

(β_{0j}, β_{1j}) indipendenti da ε_{ij}

Modello gerarchico lineare a due livelli

Assumere che

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ & \tau_{11} \end{bmatrix} \right)$$

equivale ad assumere che le scuole presenti nell'archivio siano un *campione casuale semplice* da una popolazione di scuole in cui l'intercetta media è γ_{00} e la pendenza media è γ_{10}

Modello gerarchico lineare a due livelli

Parametri del modello

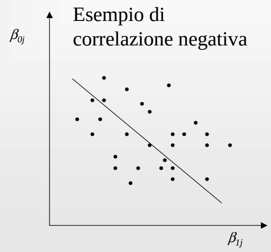
γ_{00}	intercetta media	Parametri fissi
γ_{10}	pendenza media	(= degli effetti fissi)
τ_{00}	varianza intercetta	Parametri casuali
τ_{11}	varianza pendenza	(= di varianza-covarianza)
τ_{01}	covarianza intercetta-pendenza	
σ^2	varianza residua (livello 1)	

NOTA: il modello è molto parsimonioso: ha 6 parametri indipendentemente dal numero di scuole nel campione!

Modello gerarchico lineare a due livelli

Correlazione tra intercette e pendenze:

$$\rho(\beta_{0j}, \beta_{1j}) = \frac{\tau_{01}}{\sqrt{\tau_{00}\tau_{11}}}$$



Modello gerarchico lineare a due livelli

Modello di 1° livello: $Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$

Modello di 2° livello: $\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \end{cases}$

Modello combinato:

$$Y_{ij} = (\gamma_{00} + u_{0j}) + (\gamma_{10} + u_{1j}) x_{ij} + \varepsilon_{ij}$$

$$= \underbrace{\gamma_{00} + \gamma_{10} x_{ij}}_{\text{Parte fissa}} + \underbrace{u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij}}_{\text{Parte aleatoria}}$$

Modello gerarchico lineare a due livelli

Errori di livello 2 (**effetti casuali**): $\begin{cases} u_{0j} = \beta_{0j} - \gamma_{00} & \mathrm{Var}(u_{0j}) = \tau_{00} \\ u_{1j} = \beta_{1j} - \gamma_{10} & \mathrm{Var}(u_{1j}) = \tau_{11} \end{cases}$

scarti non spiegati tra il valore del parametro per il gruppo j e il valore medio del parametro nella popolazione

Le variabili esplicative possono spiegare in parte questi scarti, cioè possono ridurre le loro varianze

Di solito le ipotesi distribuzionali vengono espresse con riferimento agli effetti casuali anziché ai beta:

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ & \tau_{11} \end{bmatrix} \right) \qquad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \text{indip da } \varepsilon_{ij}$$

Modello gerarchico lineare a due livelli: alcuni casi

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

pendenza casuale

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

intercetta casuale

$\tau_{00} = 0, \tau_{11} = 0 \quad \Rightarrow \quad Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \varepsilon_{ij}$

regressione ordinaria

Nota: $\tau_{11} = 0 \rightarrow \tau_{01} = 0$

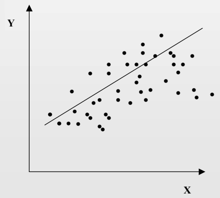
Nota: nel modello a pendenza casuale anche l'intercetta è casuale

Modello gerarchico lineare a due livelli: caso speciale "regressione ordinaria"

$\tau_{00} = 0, \tau_{11} = 0 \quad \Rightarrow \quad Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \varepsilon_{ij}$

regressione ordinaria

La variabilità tra i gruppi è nulla e quindi i coefficienti sono fissi

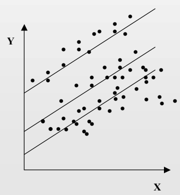


Modello gerarchico lineare a due livelli: caso speciale "intercetta casuale"

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

intercetta casuale

- La varianza del coefficiente di regressione è nulla (e quindi anche la covarianza tra i coefficienti)
- La varianza dell'intercetta non dipende da X (la centratura di X è irrilevante)
- Le rette di regressione relative ai gruppi sono fra loro parallele
- E' possibile ordinare i gruppi

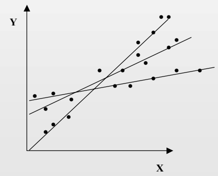


Modello gerarchico lineare a due livelli: caso generale "pendenza casuale"

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

pendenza casuale

- La varianza dell'intercetta (τ_{00}) e la covarianza intercetta-pendenza (τ_{01}) si riferiscono a X=0 e dipendono da X
- Poiché spesso l'origine di X è arbitraria è bene non vincolare a zero la covarianza
- Non esiste un ordinamento univoco dei gruppi: l'ordinamento varia al variare del valore X considerato



Modello gerarchico lineare a due livelli

(una covariata di livello 1 + una covariata di livello 2)

Introduzione di variabili esplicative di livello 2:

- Le variabili di livello 2 rappresentano caratteristiche dei gruppi che servono a
  - Definire un modello per i parametri del modello di livello 1 (β_{0j}, β_{1j}) ovvero
  - a ridurre le varianze di livello 2
- Ad esempio: W variabile binaria,

  1=scuola pubblica; 0=scuola privata

Modello gerarchico lineare a due livelli

(una covariata di livello 1 + una covariata di livello 2)

Modello di 1° livello: $Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$

Modello di 2° livello: $\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} \mathrm{w}_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} \mathrm{w}_j + u_{1j} \end{cases}$

Modello combinato:

Interazione cross-level

$$Y_{ij} = \gamma_{00} + \gamma_{01} \mathrm{w}_j + \gamma_{10} x_{ij} + \gamma_{11} \mathrm{w}_j x_{ij} \qquad \text{Parte fissa}$$

$$+ u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij} \qquad \text{Parte aleatoria}$$

Modello gerarchico lineare a due livelli

(una covariata di livello 1 + una covariata di livello 2)

- Modello di 2° livello :

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j} & \mathrm{Var}(u_{0j}) = \tau_{00} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j} & \mathrm{Var}(u_{1j}) = \tau_{11} \end{cases}$$

γ_{01} differenza nell'intercetta media tra scuola privata e pubblica
γ_{11} differenza nella pendenza media tra scuola privata e pubblica
u_{0j} effetto unico della scuola j sull'intercetta media
u_{1j} effetto unico della scuola j sulla pendenza media

Attenzione: le ipotesi sugli errori del modello sono le stesse, ma l'interpretazione delle varianze cambia poiché le varianze sono di tipo residuale rispetto alle covariate

Centratura delle covariate e regressioni entro e tra gruppi

Una covariata di livello 1 x_{ij} può essere centrata rispetto a:

A) una costante, ad es. la media generale $\bar{x}_{..}$
$\Rightarrow$ si sostituisce x_{ij} con $(x_{ij} - \bar{x}_{..})$
$\Rightarrow$ i coefficienti di x_{ij} e $(x_{ij} - \bar{x}_{..})$ sono identici
$\Rightarrow$ cambia solo l'intercetta γ_{00}

B) la media di gruppo $\bar{x}_{.j}$
$\Rightarrow$ si sostituisce x_{ij} con $(x_{ij} - \bar{x}_{.j})$
$\Rightarrow$ i coefficienti di x_{ij} e $(x_{ij} - \bar{x}_{.j})$ sono diversi!

Centratura delle covariate e regressioni entro e tra gruppi

1) $Y_{ij} = \ldots + \gamma_{total} x_{ij} + \ldots$

2) $Y_{ij} = \ldots + \gamma_{within} x_{ij} + (\gamma_{between} - \gamma_{within}) \bar{x}_{.j} + \ldots$

3) $Y_{ij} = \ldots + \gamma_{within} (x_{ij} - \bar{x}_{.j})$

4) $Y_{ij} = \ldots + \gamma_{within} (x_{ij} - \bar{x}_{.j}) + \gamma_{between} \bar{x}_{.j} + \ldots$

Generalizzazioni del modello

- Più variabili esplicative di livello 1 e di livello 2: X_1, X_2, …,W_1, W_2, … es. modello a intercetta casuale in notazione matriciale

$$Y_{ij} = \gamma_{00} + \boldsymbol{\gamma}_{10}{}^T \mathbf{x}_{ij} + \boldsymbol{\gamma}_{01}{}^T \mathbf{w}_j + u_j + \varepsilon_{ij}$$

- Struttura degli errori più complessa:
  - A livello 1: eteroschedasticità, es. $Var(\varepsilon_{ij}) = \sigma^2 x_{ij}$
  - A livello 2: più coefficienti casuali
- Più di due livelli gerarchici: es. modello a intercetta casuale a 3 livelli

$$Y_{ijk} = [parte\ fissa] + v_k + u_{jk} + \varepsilon_{ijk}$$

Modelli multilivello e valutazione

Indicatori di performance

- Le *valutazioni comparative di efficacia* (c.d. effiacia relativa) di un insieme di Istituzioni sono basate su **indicatori di performance** (misure statistiche di sintesi che rispecchiano certi aspetti del funzionamento di una Istituzione)
  - Indicatori di input, di output, di outcome

Consideriamo gli aspetti statistici legati all'utilizzo di **indicatori di outcome** per valutazioni di efficacia relativa

Indicatori di performance

- Principali questioni statistiche:
  - I) Produrre indicatori *netti*, cioè aggiustati per le condizioni che caratterizzano l'Istituzione e i suoi Utenti -> necessario per effettuare confronti equi, alla pari (ceteris paribus)
  - II) Quantificare l'incertezza -> necessario per evitare di trarre conclusioni influenzate dalla variabilità campionaria e da altre fonti di errore

Le graduatorie grezze di istituzioni (cosiddette 'League Tables') ignorano entrambi gli aspetti (Goldstein & Spiegelhalter, 1996)

Metodologia statistica

I) Produzione di indicatori *netti* &
II) Quantificazione dell'incertezza

Modelli di regressione

Ma i modelli standard non sono adeguati perché si basano sull' assunzione irrealistica di indipendenza tra gli outcome degli Utenti (mentre gli outcome degli Utenti di una stessa Istituzione sono tipicamente correlati)
-> errata quantificazione dell'incertezza

Soluzione: modelli di regressione multilivello

Metodologia statistica

- I modelli multilivello
  - rappresentano adeguatamente la struttura di correlazione -> corrretta quantificazione dell'incertezza
  - rappresentano esplicitamente la nozione di efficacia tramite gli effetti casuali u_j

Istituz. 1 .... Istituz. J

Utente 1 ... Utente n_1 Utente 1 ... Utente n_J

Modello multilivello

i = Utente (livello 1)
j = Istituzione (livello 2)

Covariate di livello 1: Caratteristiche dell'Utente

Covariate di livello 2: Caratteristiche dell'Istituzione e caratteristiche ambientali

$$Y_{ij} = \gamma_{00} + \gamma_{10}^T \mathbf{x}_{ij} + \gamma_{01}^T \mathbf{w}_j + u_j + \varepsilon_{ij}$$

Variabile di risposta: Outcome dell'Utente

Effetto casuale sull'intercetta: Efficacia dell'Istituzione

Efficacia di Tipo A o B a seconda delle covariate

Inferenza sull'efficacia basata sul modello multilivello

- Stima dell'efficacia dell'Istituzione j $\Rightarrow$ residuo di livello 2 $\hat{u}_j$
- Con i residui si possono costruire graduatorie *nette* delle istituzioni (che però ignorano l'incertezza!)
- Per tener conto dell'incertezza si può usare lo standard error dei residui per costruire
  - Intervalli di confidenza univariati
  - Intervalli di confidenza per confronti a coppie

Inferenza

Stima e previsione

$$Y_{ij} = \gamma_{00} + \gamma_{01} w_j + \gamma_{10} x_{ij} + \gamma_{11} w_j x_{ij}$$ **Parte fissa**

$$+ u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij}$$ **Parte aleatoria**

Approccio di massima verosimiglianza:

step 1: stima dei **parametri fissi** (γ_{00}, γ_{01}, γ_{10}, γ_{11})

e dei **parametri casuali** (σ^2, τ_{00}, τ_{01}, τ_{11})

step 2: previsione degli **effetti casuali** (u_{0j}, u_{1j} : j=1,…,J)

(detta anche calcolo dei residui di livello 2)

Stima dei parametri

- La stima di massima verosimiglianza (MV) si basa su algoritmi iterativi (IGLS, Fisher scoring, EM)
- Sotto deboli condizioni di regolarità gli stimatori di MV hanno buone proprietà asintotiche:
  - Consistenza
  - Normalità
  - Efficienza

Osservazione: la teoria asintotica vale all'aumentare del numero di gruppi (l'aumento della dimensione dei gruppi non è sufficiente)

Previsione degli effetti casuali (calcolo dei residui di livello 2)

Nel caso dell'ANOVA ad effetti casuali $Y_{ij} = \mu + u_j + \varepsilon_{ij}$ esistono semplici formule per i residui:

$\hat{u}_j^{OLS} = \bar{Y}_{.j} - \hat{\mu}$ Residuo OLS grezzo

$\hat{u}_j^{EB} = \lambda_j \hat{u}_j^{OLS}$ Residuo Empirical Bayes o *shrinkage*

$\lambda_j = \dfrac{\tau}{\tau + \sigma^2 / n_j}$ Affidabilità del gruppo j (cresce con la dimensione del gruppo ed è circa uguale a 1 per i gruppi grandi)

$\hat{u}_j^{EB}$ è preferibile a $\hat{u}_j^{OLS}$ perché minimizza l'errore quadratico medio di previsione di u_j

Previsione degli effetti casuali (calcolo dei residui di livello 2)

per i gruppi grandi $\hat{u}_j^{EB}$ è simile a $\hat{u}_j^{OLS}$,

ma per i gruppi piccoli $\hat{u}_j^{EB}$ è notevolmente "ristretto" rispetto a $\hat{u}_j^{OLS}$

per tener conto della sua scarsa affidabilità

$\Rightarrow$ $\hat{u}_j^{EB}$ e $\hat{u}_j^{OLS}$ producono due graduatorie diverse!

Nei modelli più complessi dell'ANOVA le formule sono complicate, ma valgono gli stessi principi di *shrinkage* e *borrowing strength*

Confronto fra residui

- I residui di livello 2 sono le previsioni dei corrispondenti effetti casuali u_j
- Spesso l'interesse non è sul valore di u_j per un singolo cluster, ma sul confronto fra gli u_j di due diversi cluster (ad es. per capire se l'istituzione A è più efficace dell'istituzione B)
- Problema statistico: *gli effetti casuali di due cluster arbitrari sono o non sono significativamente diversi ad un certo livello?*

Confronto fra residui

- Errore comune: pensare che due quantità i cui intervalli al 95% sono disgiunti siano significativamente diverse al 5%

$X \sim N(\mu_X, \sigma^2)$ $\Rightarrow$ $X \pm 1.96 \cdot \sigma$

$Y \sim N(\mu_Y, \sigma^2)$ $\Rightarrow$ $Y \pm 1.96 \cdot \sigma$

Se X e Y sono indipendenti

$X - Y \sim N(\mu_X - \mu_Y, 2\sigma^2)$ $\Rightarrow$ $(X - Y) \pm 1.96\sqrt{2} \cdot \sigma$

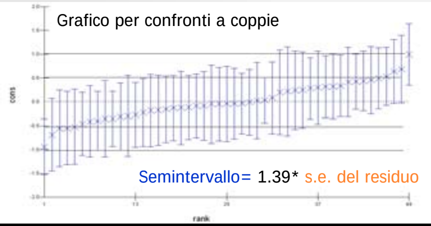
μ_X è significativamente diverso da μ_Y al livello 95% se e solo se:

• la distanza (in unità di σ) fra X e Y è maggiore di 1.96√2 = 2.77

• gli intervalli *univariati* di raggio 2.77/2 = 1.39 sono disgiunti

Confronto fra residui

- Il valore 1.39 si basa su alcune ipotesi (distribuzione normale, indipendenza, identica varianza) che tipicamente non sono plausibili -> il valore 1.39 va inteso come approssimazione e il livello di significatività 5% come livello medio dei confronti



Approfondimenti in Goldstein & Healy (1995)

73

Software e libri

Software per la stima di modelli multilivello

- Software specializzato (es. MLwiN, HLM)
- Procedure in pacchetti di uso generale (es. Proc Mixed e Proc Nlmixed in SAS, xtmixed e gllamm in Stata)

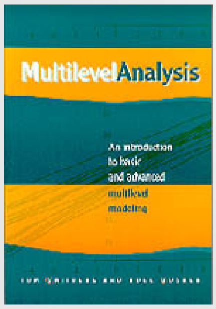
Rassegna critica: **http://multilevel.ioe.ac.uk/softrev/index.html**

Sito con dati ed esercitazioni svolte: Multilevel Modeling Resources at UCLA **http://www.ats.ucla.edu/stat/mlm/**

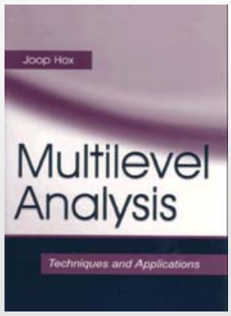
L. Grilli - Scuola SIS 2005 75

Libri sui modelli multilivello

Snijders & Bosker



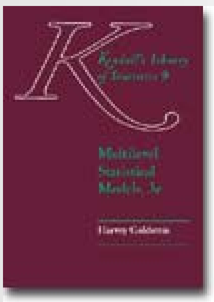
Hox



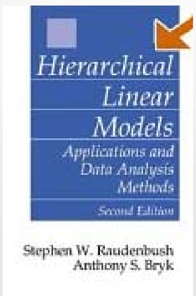
L. Grilli - Scuola SIS 2005 76

Libri sui modelli multilivello

Goldstein



Raudenbush & Bryk



L. Grilli - Scuola SIS 2005 77

Modelli multilivello e valutazione

- Raudenbush S.W. & Willms J.D. (1995) The Estimation of School Effects. *Journal of Educational and Behavioral Statistics*, Vol. 20, No. 4, pp. 307-335.
- Goldstein, H. & Spiegelhalter, D.J. (1996). Leage tables and their limitations: statistical issues in comparisons of institutional performances (con discussione). *Journal of the Royal Statistical Society*, series A, Vol. 159, pp. 385-443.
- Goldstein H. & Lewis T. (eds) (1996) Assessment: problems, developments and statistical issues: a volume of expert contributions. Wiley, Chichester.
- Gori E. & Vittadini G. (a cura di) (1999) Qualità e valutazione nei servizi di pubblica utilità. ETAS, Milano.

L. Grilli - Scuola SIS 2005 78