Efficacia esterna della formazione universitaria:
il progetto OUTCOMES

Un'analisi multilivello della probabilità di occupazione dei laureati dell'ateneo fiorentino con modelli grafici a catena

Carla Rampichini, Leonardo Grilli, Anna Gottard
Dipartimento di "Statistica G. Parenti"
Università di Firenze

Padova, 3 - 5 Febbraio 2005

Obiettivo principale

Analisi dell'inserimento lavorativo dei laureati sfruttando le potenzialità dei modelli grafici a catena, estesi al caso di dati correlati

- Indice della presentazione
  - Introduzione ai modelli multilivello
  - Descrizione delle indipendenze derivanti dal modello
  - Definizione dei grafi a catena per modelli multilivello
  - Risultati
  - Conclusioni e lavoro futuro

Modelli multilivello

- Molti dati presentano una struttura gerarchica
  - Studenti in classi, classi in scuole, ecc.
  - Misure ripetute nel tempo su uno stesso soggetto
  - Pazienti negli ospedali, ospedali nei distretti, ...

struttura gerarchica non trascurabile

- Trascurare la struttura gerarchica comporta
  - Sottovalutare importanza effetto di gruppo
  - Invalidità di molte tecniche statistiche utilizzate nell'analisi della relazione tra variabili

Struttura gerarchica a due livelli

$$\mathbf{Y} = (\underbrace{Y_{11},\dots Y_{n_1 1}}, \underbrace{Y_{21},\dots Y_{n_2 2}}, \dots, Y_{n_J J}) \qquad \sum_{j=1}^{J}\sum_{i=1}^{n_j} n_{ij} = n$$

Dato un insieme di covariate $\mathbf{X}$

Osservazioni stesso gruppo: $Y_{ij} \not\perp Y_{i'j} \mid \mathbf{X}, \forall i \neq i', \quad i,i' = 1,2,\dots n_j$

gruppi diversi: $Y_{ij} \perp Y_{i'j'} \mid \mathbf{X}, \forall j \neq j', \quad j,j' = 1,2,\dots J$

Osservazioni appartenenti allo stesso gruppo sono correlate!

Modello a intercetta casuale

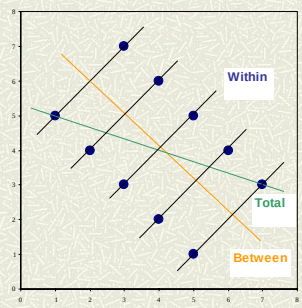
$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{0j} + \varepsilon_{ij}$$

u_{0j} v.c. latente: rappresenta la dipendenza tra le Y dello stesso gruppo

Hp:

$$u_{0j} \overset{iid}{\sim} N(0, \tau^2)$$
$$Y_{ij} \perp Y_{i'j} \mid \mathbf{X}, u_{0j} \quad \forall i \neq i', \quad i,i' = 1,2,\dots n_j$$
$$u_{0j} \perp \varepsilon_{ij}, \quad u_{0j} \perp \mathbf{X}$$

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 (2)

Nel modello a intercetta casuale con una covariata

$$\hat{\gamma}_{\text{tot}} = \hat{\eta}^2_X \cdot \hat{\gamma}_{\text{tra}} + \left(1 - \hat{\eta}^2_X\right) \cdot \hat{\gamma}_{\text{entro}}$$

$\hat{\eta}^2_X = \dfrac{SSB_X}{SST_X}$ (rapporto di correlazione di X)

$\hat{\gamma}_{\text{tot}}$ assume un valore intermedio tra $\hat{\gamma}_{\text{tra}}$ e $\hat{\gamma}_{\text{entro}}$

Regressioni entro e tra gruppi

1) $Y_{ij} = \ldots + \gamma_{tot} x_{ij} + \ldots$
2) $Y_{ij} = \ldots + \gamma_{entro} x_{ij} + (\gamma_{tra} - \gamma_{entro})\bar{x}_j + \ldots$
3) $Y_{ij} = \ldots + \gamma_{entro}(x_{ij} - \bar{x}_j) + \ldots$
4) $Y_{ij} = \ldots + \gamma_{entro}(x_{ij} - \bar{x}_j) + \gamma_{tra}\bar{x}_j + \ldots$

DISTRIBUZIONE CONGIUNTA

- I gruppi sono indipendenti
- La distribuzione di probabilità congiunta per ogni gruppo *j* fattorizza:

$$f(\mathbf{Y}_j, u_{0j}, \mathbf{X}) = f(\mathbf{Y}_j \mid u_{0j}, \mathbf{X}) f(u_{0j} \mid \mathbf{X}) f(\mathbf{X})$$
$$= f(\mathbf{Y}_j \mid u_{0j}, \mathbf{X}) f(u_{0j}) f(\mathbf{X}) \quad [u_{0j} \perp \mathbf{X}]$$
$$= \left[\prod_{i=1}^{n_j} f(Y_{ij} \mid u_{0j}, \mathbf{X})\right] f(u_{0j}) f(\mathbf{X}) \quad [Y_{ij} \perp Y_{i'j} \mid \mathbf{X}, u_{0j}]$$

con $\mathbf{Y}_j = \left\{Y_{1j}, Y_{2j}, \ldots, Y_{n_j j}\right\}'$

Modelli grafici a catena per dati gerarchici

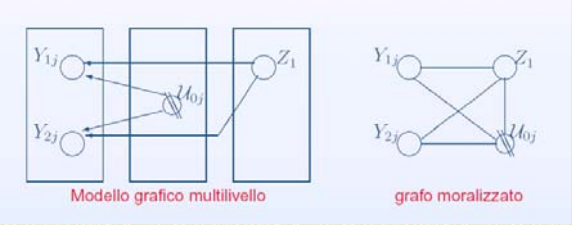
I modelli grafici rappresentano le RELAZIONI DI INDIPENDENZA CONDIZIONALE TRA VARIABILI, considerando **le unità statistiche come indipendenti**

questo non è vero per strutture di dati gerarchici

La soluzione proposta sfrutta le proprietà note dei **grafi a catena**, richiedendo solo alcune **definizioni supplementari**

(Gottard e Rampichini, 2004)

Esempio: modello grafico a intercetta casuale



Il vantaggio principale di questa formulazione è che valgono **le usuali proprietà di Markov per modelli grafici a catena e il criterio di fattorizzazione (Lauritzen, 1996)**

Inserimento lavorativo dei laureati

- Università di Firenze: laureati anno 2000 (vecchio ordinamento) che all'intervista lavorano o sono in cerca di occupazione (Alma Laurea+integrazione Valmon)
- Indagine a circa 2 anni dalla laurea
- Y: condizione occupazionale all'intervista:

Y lavoro	Freq	%
0 no+temporaneo	1595	53.79
1 stabile	1370	46.21
TOTALE	2965	100.00

- 56 corsi di laurea
- N. laureati per corso variabile da 4 (Chimica) a 504 (Architettura), mediana 22

Definizione delle variabili e ordinamento in blocchi

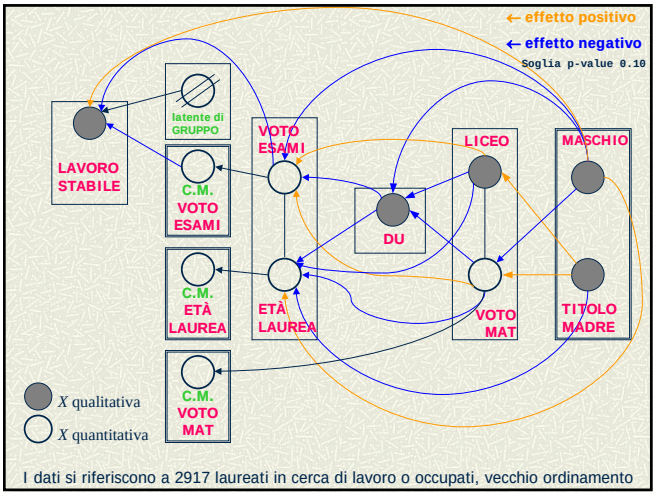
blocco	variabile	modalità
1 esogene	maschio	1=maschio, 0=femmina
	titolo madre	obbligo (rif), diploma, laurea
2 intermedie	liceo	1=liceo, 0=altra maturità
	voto mat	36-60 (media=48)
3 intermedie	du	1=corso di diploma, 0=corso laurea
4 intermedie	età alla laurea	21-50 (media=27.6)
	voto medio esami	18-30 (media=26.8)
5.1 medie di gruppo	C.M. voto esami	
	C.M. età laurea	
	C.M. voto mat	
5.2 nodo latente di gruppo	u_oj	
6 risposta	lavoro stabile	1=lavoro stabile, 0=altrimenti

$$f(\mathbf{Y}_j, u_{0j}, \mathbf{X}) = f(\mathbf{Y}_j \mid u_{0j}, \mathbf{X}) f(u_{0j}) f(\mathbf{X})$$
$$f(\mathbf{X}) = f(\mathbf{X}_4 \mid \mathbf{X}_3, \mathbf{X}_2, \mathbf{X}_1) f(\mathbf{X}_3 \mid \mathbf{X}_2, \mathbf{X}_1) f(\mathbf{X}_2 \mid \mathbf{X}_1)$$
$$f(\overline{x}_j \mid x_{1j}, \ldots, x_{n_j j}) = 1$$

fattorizzazione distribuzione congiunta

Strategia di stima

- La stima di un modello grafico a catena richiede più passi
- Modelli di regressione univariati appropriati alla natura della variabile dipendente in base alla natura recursiva del modello grafico a catena (Cox and Wermuth, 1996)
- Modello multilivello per Y dicotomica stimato con procedura **`gllamm`** di Stata



I dati si riferiscono a 2917 laureati in cerca di lavoro o occupati, vecchio ordinamento

Stime modello per lavoro stabile

Parametro	Stima	Std.err.	p-value
Intercetta	4.925	6.024	0.414
MASCHIO	0.372	0.087	0.000
TITOLO MADRE (obbligo)	-0.020	0.090	0.820
TITOLO MADRE (laurea)	-0.115	0.135	0.397
LICEO	-0.089	0.087	0.308
VOTO MAT	-0.003	0.007	0.595
C.M. VOTO MAT	0.070	0.050	0.164
DU	0.612	0.396	0.122
VOTO ESAMI	-0.055	0.030	0.069
C.M. VOTO ESAMI	-0.221	0.119	0.063
ETA' LAUREA	-0.005	0.016	0.737
C.M. ETA' LAUREA	-0.034	0.106	0.747
var. latente di gruppo	0.510	-0.148	

Test su varianza tra gruppi LRT=108.8 (1 df)

ICC=0.134

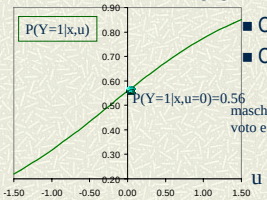
voto esami:
entro= -0.055
tra= -0.276
(-0.055)+(-0.221)

Stime modello per lavoro stabile

- LAVORO STABILE dipende direttamente solo da MASCHIO e da VOTO ESAMI
- VOTO ESAMI media l'effetto di altre variabili (es. DU, LICEO) che quindi hanno solo un effetto indiretto
- VOTO ESAMI ha un diverso effetto *entro* (`-0.055`) e *tra gruppi* (`-0.276`): entrambi sono negativi ma quello *tra* è molto più importante ➔ l'effetto negativo del voto è per lo più un effetto di Corso di Laurea
- Stimando senza scomporre VOTO ESAMI si ottiene effetto totale -0.068 (s.e. 0.029), simile a *entro* ma non interpretabile

Stime modello per lavoro stabile

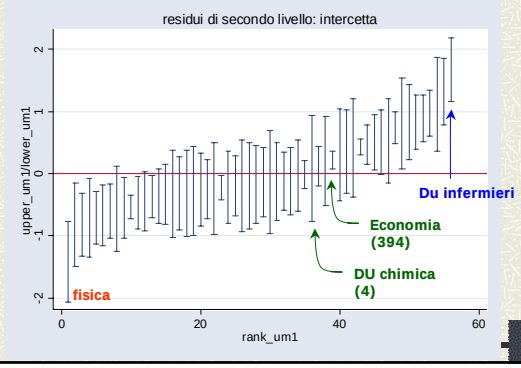
- Le variabili non osservate a livello di Corso di Laurea contano molto:
  - Stima ICC = 0.134
  - Considerando un corso medio (U=0) e un individuo con P(Y=1|U,X)=0.5



- Corso 'migliore' (+2τ) ➔ P=0.81
- Corso 'peggiore' (-2τ) ➔ P=0.19

maschio, madre diplomata, altra maturità, voto mat=48, voto esami=26 eta lau=27

Residui a livello di CdL con intervalli per confronti a coppie



19

Vantaggi dell'approccio

- Esplicitazione delle ipotesi a priori:
  - Ordinamento delle variabili in blocchi
  - indipendenze condizionali sottostanti il modello multilivello
- Visualizzazione degli effetti diretti e indiretti
- Modellazione della distribuzione congiunta
- Lettura delle indipendenze condizionate (proprietà markoviane del grafo)

20

Vantaggi della rappresentazione proposta

- Utilizzo della teoria dei grafi esistente
- I nodi di risposta individuali evidenziano che l'unità di analisi è il gruppo, mentre la risposta è multivariata
- rappresentazioni alternative:
  - Buntine (1994) (WinBugs)
  - Johnson and Hoeting (http://www.stat.colostate.edu/ jah/papers/graph.pdf)

21

Sviluppi futuri

- Più regressioni multilivello nello stesso grafo
- Modellazione del processo di formazione dei gruppi
- Regressione di variabili di gruppo su variabili individuali
- Sfruttamento delle proprietà markoviane del grafo

22