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

Anna Gottard, Leonardo Grilli, Carla Rampichini[1]
Dipartimento di Statistica "Giuseppe Parenti", Università degli studi di Firenze

Riassunto. Obiettivo principale del presente lavoro è l'analisi dell'inserimento lavorativo dei laureati sfruttando le potenzialità dei modelli grafici a catena, estesi al caso di dati correlati. Dopo una breve introduzione ai modelli multilivello, vengono descritte le indipendenze condizionali derivanti dal modello e definiti i grafi a catena per modelli multilivello. La metodologia dei modelli grafici multilivello viene utilizzata per l'analisi della probabilità di occupazione dei laureati e diplomati dell'anno 2000 presso l'ateneo di Firenze. Il lavoro si chiude con alcune indicazioni di ricerca futura.

Parole chiave: Modelli grafici a catena, modelli multilivello, regressione logistica.

1. Introduzione

La valutazione nel settore dell'istruzione universitaria richiede la predisposizione di adeguati metodi e modelli statistici che siano in grado di coglierne appieno la complessità. La complessità della valutazione è imputabile a vari aspetti tra i quali:

(a) la presenza di una struttura gerarchica delle osservazioni rilevante ai fini dell'analisi: studenti in classi, classi in corsi di studio, ecc.. La struttura gerarchica dei dati comporta problemi di correlazione tra le osservazioni e la considerazione di effetti ai vari livelli della gerarchia e delle loro interazioni. L'ignorare la gerarchia può comportare problemi, principalmente una sottovalutazione dell'importanza degli effetti di gruppo e il fatto che i modelli di regressione usualmente impiegati per lo studio delle relazioni tra variabili

[1] Il presente lavoro è stato finanziato nell'ambito del progetto "Transizioni Università-Lavoro e valorizzazione delle competenze professionali dei laureati: modelli e metodi di analisi multidimensionali delle determinanti", cofinanziato dal MIUR. Coordinatore nazionale è Luigi Fabbris, coordinatore del Gruppo di Firenze è Bruno Chiandotto.

(p.e. GLM) non forniscono risultati corretti, soprattutto in termini di stima degli errori standard dei coefficienti di regressione.

(b) Un altro aspetto di complessità del fenomeno è la presenza di variabili esplicative riferite a diversi momenti temporali del processo formativo (p.e. titolo di studio dei genitori, voto di maturità, voto agli esami, voto alla laurea). Ciò comporta la necessità di tenere in considerazione l'ordinamento logico-temporale delle variabili al fine di rendere esplicito il meccanismo di formazione del risultato finale (p.e. ottenere un'occupazione dopo il conseguimento del titolo), distinguendo tra effetti diretti e indiretti delle variabili coinvolte.

I modelli multilivello (Snijders e Bosker, 1999; Goldstein, 2003) consentono di affrontare in maniera adeguata i problemi della correlazione tra le osservazioni e dell'analisi degli effetti di gruppo evidenziati in (a) e per tale motivo sono ampiamente utilizzati nell'ambito della valutazione di efficacia relativa; d'altra parte, i modelli grafici a catena (Cox e Wermuth, 1996) sono un utile strumento per la rappresentazione del processo descritto in (b).

Nel presente lavoro si propone un'integrazione tra queste due metodologie idonea a modellare la relazione tra l'inserimento nel mercato del lavoro, il percorso di studi e le caratteristiche individuali, mettendo in luce il contributo del corso di studi seguito e distinguendo, al contempo, tra effetti diretti e indiretti delle variabili di *background* e di percorso di ogni studente.

Nel Par. 2 sarà brevemente introdotto il modello lineare a due livelli e l'estensione di tale modello al caso di variabile di risposta dicotomica e nel Par. 3 si illustreranno i modelli grafici multilivello, derivanti dall'integrazione tra modelli multilivello e modelli grafici a catena. Nel Par. 4 si presentano i dati e i risultati dell'analisi. Infine, nel Par. 5 si forniscono alcune indicazioni di ricerca futura.

2. Il modello multilivello a intercetta casuale

Si consideri una struttura gerarchica a due livelli, indicando con Y_{ij} la variabile di risposta per l'i-mo individuo (unità di primo livello) del j-mo gruppo (unità di secondo livello), $i=1,2,\ldots,n_j$, $j=1,2,\ldots,J$. Per ciascun individuo si disponga inoltre di un vettore $\mathbf{X}_{ij}$ di caratteristiche individuali (p.e. sesso, voto di maturità) e di gruppo (p.e. numero di iscritti allo stesso corso di laurea).

Assumendo che Y_{ij} sia una variabile quantitativa e che vi sia un legame lineare tra la risposta Y_{ij} e le covariate $\mathbf{X}_{ij}$, possiamo specificare il seguente modello a intercetta casuale a due livelli:

$$\begin{aligned} Y_{ij} &= \alpha_j + \boldsymbol{\beta}' \mathbf{X}_{ij} + \varepsilon_{ij} \\ \alpha_j &= \alpha + U_{0j} \end{aligned} \tag{1}$$

dove ε_{ij} denota gli errori (residui) a livello di individuo e U_{0j} gli errori a livello di gruppo. Si assume che, a entrambi i livelli, gli errori seguano una distribuzione normale con media nulla e varianza Var(ε_{ij})=σ^2 a livello individuale e Var(U_{0j})= τ^2 a livello di gruppo. Si assume inoltre che errori a livelli differenti siano indipendenti e che le covariate siano incorrelate agli errori. Le relazioni di indipendenza tra le osservazioni che derivano da tale modello sono:

$$\begin{aligned} &Y_{ij} \not\perp\!\!\!\perp Y_{i'j} \mid \mathbf{X}, && \forall i \neq i', \forall j \\ &Y_{ij} \perp\!\!\!\perp Y_{i'j'} \mid \mathbf{X}, && \forall i, i', \forall j \neq j', \end{aligned} \tag{2}$$

dove $\mathbf{X} = \left\{ \mathbf{X}_{ij} : i = 1, 2, \text{K}\ , n_j,\ j = 1, 2, \text{K}\ J \right\}$.

Dalle relazioni (2) discende che, condizionatamente alle covariate **X**, osservazioni appartenenti a gruppi diversi sono indipendenti, mentre osservazioni appartenenti allo stesso gruppo sono dipendenti. Il coefficiente di correlazione intraclasse

$$\rho = Corr(Y_{ij}, Y_{i'j'}) = \begin{cases} 0 & \text{se } j \neq j' \\ \dfrac{\tau^2}{\tau^2 + \sigma^2} & \text{se } j = j' \end{cases}$$

fornisce una misura della dipendenza tra le osservazioni. Condizionatamente a **X** e a U_{0j} anche le osservazioni dello stesso gruppo sono indipendenti:

$$Y_{ij} \perp\!\!\!\perp Y_{i'j} \mid \mathbf{X}, U_{0j}, \qquad \forall i \neq i', \forall j. \tag{3}$$

Per ogni gruppo j la distribuzione di probabilità congiunta fattorizza nel seguente modo[2]:

$$\begin{aligned} & f(\mathbf{y}_j, u_{0j}, \mathbf{x}) = f(\mathbf{y}_j \mid u_{0j}, \mathbf{x}) f(u_{0j} \mid \mathbf{x}) f(\mathbf{x}) \\ &= f(\mathbf{y}_j \mid u_{0j}, \mathbf{x}) f(u_{0j}) f(\mathbf{x}) \\ &= \left[\prod_{i=1}^{n_j} f(y_{ij} \mid u_{0j}, \mathbf{x}) \right] f(u_{0j}) f(\mathbf{x}) \end{aligned} \tag{4}$$

[2] Il considerare le sole covariate relative agli individui del gruppo j è equivalente ai fini della fattorizzazione (4) a considerare le covariate di tutti gli individui **X**. Per semplicità di notazione, si è optato per la seconda soluzione.

dove $\mathbf{y}_j' = \left\{ y_{1j}, y_{2j}, \mathrm{K}, y_{n_j j} \right\}$. Infatti, $f(u_{0j} \mid \mathbf{x}) = f(u_{0j})$ per l'indipendenza tra u_{0j} e **X**, mentre $f(\mathbf{y}_j \mid u_{0j}, \mathbf{x})$ corrisponde alla produttoria in n_j delle densità individuali per l'indipendenza condizionale riportata nella (3).

In generale, l'effetto di una covariata di primo livello (cioè, misurata a livello di individuo) sulla variabile di risposta può essere scomposto nelle parti *tra* ed *entro* gruppi, concordemente alla scomposizione della variabilità della covariata (Snijders e Bosker, 1999). Ad esempio, nel modello lineare a due livelli con una sola covariata *X*, considerando le stime di minimi quadrati, il coefficiente totale $\hat{\beta}_{\text{tot}}$ è una combinazione lineare del coefficiente di regressione tra le medie di gruppo $\hat{\beta}_{\text{tra}}$ e del coefficiente di regressione all'interno dei gruppi $\hat{\beta}_{\text{entro}}$:

$$\hat{\beta}_{\text{tot}} = \hat{\eta}_X^2 \cdot \hat{\beta}_{\text{tra}} + \left(1 - \hat{\eta}_X^2\right) \cdot \hat{\beta}_{\text{entro}} \tag{5}$$

dove $\hat{\eta}_X^2$ è il rapporto di correlazione di *X*. Di conseguenza $\hat{\beta}_{\text{tot}}$ assume un valore intermedio tra $\hat{\beta}_{\text{tra}}$ e $\hat{\beta}_{\text{entro}}$. I due coefficienti *tra* ed *entro* hanno un'interpretazione diversa e possono assumere valori anche di segno opposto, per cui è possibile ottenere un coefficiente totale nullo a fronte di coefficienti *tra* ed *entro* significativi ma di segno contrario. Pertanto, è opportuno specificare il modello in modo da stimare entrambi questi effetti.

Un modo per ottenere la stima dei due coefficienti di interesse consiste nell'inserire nel modello tanto la covariata X_{ij} quanto la sua media di gruppo $\bar{X}_{.j}$:

$$Y_{ij} = \mathrm{K} + \beta_{\text{entro}} X_{ij} + (\beta_{\text{tra}} - \beta_{\text{entro}}) \bar{X}_{.j} + \mathrm{K} \tag{6}$$

Nella specificazione (6) il coefficiente associato alla media di gruppo è la differenza fra i coefficienti *tra* ed *entro*, per cui il classico test per la significatività del coefficiente di $\bar{X}_{.j}$ va interpretato come un test per la significatività della differenza fra i coefficienti *tra* ed *entro*. Se il coefficiente di $\bar{X}_{.j}$ non è statisticamente significativo la distinzione *tra* ed *entro* può essere ignorata, lasciando fra i regressori solo la variabile non centrata X_{ij}.

Si noti che l'inserimento nel modello della media di gruppo di una covariata non solo permette di scomporre l'effetto nelle componenti *tra* ed *entro*, ma permette anche di eliminare l'eventuale correlazione tra la covariata e l'effetto casuale u_{0j} (Snijders e Bosker, 1999).

Si può supporre che il modello lineare a intercetta casuale (1) valga per la variabile latente continua Y_{ij}, che genera la variabile osservata dicotomica Y_{ij}^{obs} nel modo seguente:

$$Y_{ij}^{obs} = \begin{cases} 0 & \text{se } Y_{ij} \leq 0 \\ 1 & \text{se } Y_{ij} > 0 \end{cases}$$

Assumendo una distribuzione logistica standard per l'errore ε_{ij} che compare nella (1), si ottiene un modello logistico a intercetta casuale per la probabilità di risposta:

$$P(Y_{ij}^{obs} = 1 \mid u_{0j}, \mathbf{x}_{ij}) = \frac{1}{1 + \exp(-(\alpha + \boldsymbol{\beta}'\mathbf{x}_{ij} + u_{0j}))}.$$

Le proprietà di tale modello e i metodi di stima sono descritti in dettaglio nei testi di analisi multilivello (tra gli altri, Snijders e Bosker, 1999).

3. Modelli grafici a catena per dati gerarchici

I modelli grafici costituiscono una classe di modelli probabilistici per lo studio della distribuzione congiunta di un insieme di variabili aleatorie, la cui struttura di indipendenza condizionale può essere descritta mediante un grafo.

Un grafo è costituito da un insieme di nodi e da un insieme di archi (non orientati) o frecce tra i nodi: ciascun nodo rappresenta una variabile aleatoria (per convenzione un nodo "vuoto" ○ rappresenta una variabile discreta, un nodo "pieno" ● rappresenta una variabile continua), mentre la connessione tra due nodi mediante un arco rappresenta una qualche associazione tra le rispettive variabili. Più precisamente, l'assenza di connessione tra due nodi sta ad indicare un'indipendenza condizionale sulla base delle proprietà Markoviane del grafo.

I modelli grafici a catena costituiscono una classe di modelli probabilistici che ammette relazioni di tipo simmetrico o asimmetrico, sfruttando l'ordinamento logico-temporale assunto tra le variabili. Parallelamente, un grafo a catena è un grafo che ammette sia archi non orientati, sia frecce: le frecce indicano le relazioni di tipo asimmetrico, mentre gli archi non orientati le relazioni simmetriche.

Nei grafi a catena non sono ammessi cicli (parzialmente) orientati, ovvero non è possibile, partendo da un determinato nodo, tornare ad esso dopo aver compiuto un percorso seguendo archi e frecce del grafo.

Le variabili del modello grafico a catena possono essere suddivise in una catena ordinata di blocchi in modo che le variabili di un blocco possano essere considerate esplicative delle variabili nel blocco successivo. I nodi di uno stesso blocco possono essere connessi solo da archi non orientati, mentre due nodi appartenenti a blocchi diversi solo da frecce. Ne consegue che si assume di poter costruire un ordinamento parziale tra le variabili, distinguendo l'insieme delle variabili puramente esplicative (generalmente poste nell'ultimo blocco a destra), quello delle variabili puramente di risposta (ultimo blocco a sinistra) e quello delle variabili intermedie, che sono al tempo stesso risposta ed esplicative, rappresentate nei blocchi intermedi.

La topologia del grafo consente di rendere esplicita la struttura di indipendenza condizionale tra le variabili sulla base delle proprietà di Markov specifiche per i grafi a catena. Le proprietà a cui si fa riferimento in questo lavoro sono state introdotte da Lauritzen e Wermuth (1989) e da Fridenberg (1990) e sono comunemente indicate con la sigla LWF. La più importante delle proprietà markoviane è la proprietà globale la cui deduzione si basa sul concetto di *grafo moralizzato*. Un grafo moralizzato si ottiene unendo mediante archi i nodi da cui partono frecce che puntano su uno stesso nodo e quindi trasformando le frecce in archi non orientati.

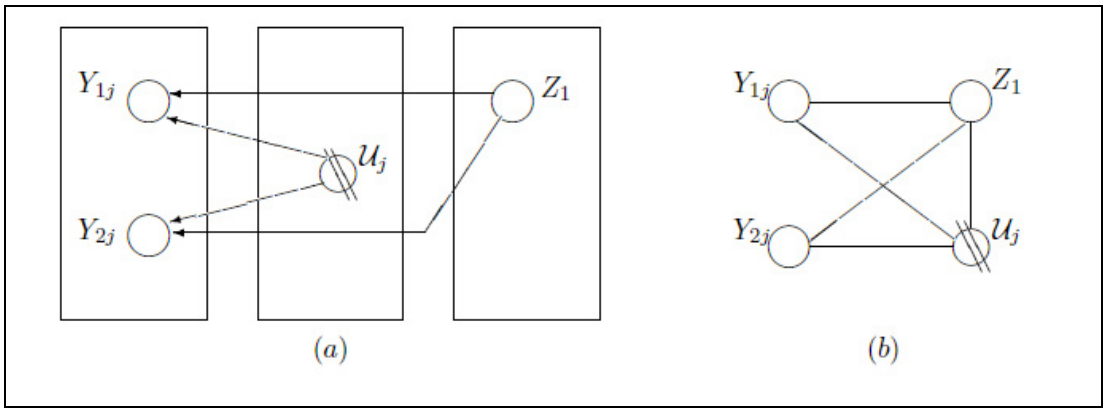
La proprietà di Markov globale associa il concetto di indipendenza a quello di separazione tra i nodi del grafo moralizzato: se l'insieme dei nodi A è separato dall'insieme B da i nodi in S, ovvero se tutti i percorsi da un nodo di A ad un nodo di B passano, nel grafo moralizzato, necessariamente per almeno un nodo di S, ciò implica che $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_S$, dove $\mathbf{X}_k$ rappresenta il vettore di variabili aleatorie rappresentate dai nodi in k, $k = A, B, C$. Le proprietà di Markov inducono una fattorizzazione della distribuzione congiunta delle variabili in esame. Si veda Lauritzen (1996) per una trattazione esaustiva dell'argomento.

I modelli grafici rappresentano le relazioni di indipendenza condizionale tra variabili, considerando le unità statistiche come indipendenti. Questa assunzione non è valida per strutture di dati gerarchici. L'idea proposta da Gottard e Rampichini (2004) per risolvere tale problema consiste nel rappresentare nel grafo la distribuzione congiunta delle variabili aleatorie relative a un gruppo: ad esempio, in un modello a due livelli, si rappresenta nel grafo il vettore $(\mathbf{Y}_j, U_{0j}, X_{1j}, \ldots, X_{Kj})$ relativo al gruppo j-esimo, perché i J gruppi sono indipendenti e identicamente distribuiti. Dato che i gruppi possono avere numerosità diversa (n_j), è stato proposto di rappresentare nel grafo solo due osservazioni elementari per gruppo, ossia il sottografo più piccolo che consente la deduzione di tutte le relazioni condizionali implicate: l'inserimento di 3 o più nodi individuali, invece di 2, non modifica le relazioni tra variabili deducibili dal grafo. Questa soluzione sfrutta le proprietà note dei grafi a catena, richiedendo solo poche definizioni supplementari.

Si denomina *nodo individuale* un nodo del grafo che rappresenta una variabile aleatoria relativa ad una specifica unità statistica e *nodo latente di gruppo*, un nodo associato a una variabile aleatoria latente U_{0j} che rappresenta i fattori non osservati a livello di gruppo e che ⦸ costituisce il separatore tra i nodi individuali, per cui vale: $Y_{ij} \perp\!\!\!\perp Y_{i'j} | U_{0j}$. Tale nodo è rappresentato nel grafo dal simbolo. Infine, si denomina *deterministico* un nodo del grafo che rappresenta una variabile aleatoria la cui distribuzione condizionata è degenere. Tale nodo è rappresentato nel grafo in un blocco a doppia linea.

La struttura di indipendenza condizionale di un modello a due livelli con intercetta casuale può essere rappresentata da un grafo a catena, dove nell'ultimo blocco, quello delle variabili risposta, sono posti due nodi individuali, nel blocco immediatamente precedente viene posto il nodo latente di gruppo, mentre i restanti blocchi includono le altre variabili, esplicative ed intermedie. Un esempio di modello a due livelli con intercetta casuale ed una sola variabile esplicativa è riportato nella Figura 1. Il vantaggio principale di questa formulazione è che valgono le usuali proprietà di Markov per modelli grafici a catena e il criterio di fattorizzazione della distribuzione congiunta.

Figura 1 *Esempio di modello grafico multilivello a intercetta casuale (a) e relativo grafo moralizzato (b).*



Nel caso sub a) della Figura 1, per la proprietà markoviana a coppie (*pairwise*), la variabile latente U_{0j} è marginalmente indipendente dalla variabile esplicativa Z_1. Analizzando il grafo moralizzato sub (b), per la proprietà markoviana globale, la stessa variabile latente U_{0j} non è indipendente da Z_1 condizionatamente alla variabile risposta $\mathbf{Y}_j$.

4. Inserimento lavorativo dei laureati

Il modello grafico multilivello descritto nel Par. 3 sarà ora utilizzato per analizzare i dati relativi alla condizione occupazionale dei laureati o diplomati dell'anno solare 2000 dell'Università Firenze intervistati telefonicamente a circa due anni dal conseguimento del titolo.

L'obiettivo principale è quello di determinare i fattori che influenzano l'inserimento nel mondo del lavoro, con riferimento sia alle caratteristiche individuali che al titolo conseguito. I dati riguardano 2917 laureati e diplomati (occupati o in cerca di lavoro), dei quali circa il 46% aveva un'occupazione stabile alla data dell'intervista, mentre il restante 54% risultava disoccupato o con occupazione temporanea. I laureati e diplomati considerati provengono da 56 corsi di laurea/diploma, con un numero di unità per corso variabile da 4 (Chimica) a 504 (Architettura) e un valore mediano di 22.

La variabile di risposta è dicotomica: assume la modalità 1 se il laureato/diplomato lavora stabilmente alla data dell'intervista, e 0 altrimenti.

Il ricorso ad un modello grafico consente di mettere in luce relazioni dirette e indirette tra le variabili esplicative e la variabile di risposta. In questo senso il presente contributo si differenzia da analisi simili che, pur ricorrendo ad un modello multilivello per la variabile di risposta, non analizzano esplicitamente le relazioni tra le variabili esplicative (tra gli altri, Chiandotto e Bacci, 2004).

La specificazione del modello grafico a catena richiede, prima di tutto, di ordinare le covariate secondo un ordine logico-temporale. Le variabili utilizzate per la presente analisi[3] e il loro ordinamento in blocchi sono riportati nella Tabella 1.

Le variabili che compaiono nel blocco 5 sono medie di gruppo delle corrispondenti variabili individuali. L'inserimento della media di gruppo consente di scomporre l'effetto totale della variabile individuale nelle due parti 'tra' e 'entro' gruppi, come mostrato nella (6). Dato che la distribuzione condizionata della media di gruppo è degenere, cioè $f(\bar{x}_j \mid x_{1j}, \mathrm{K}, x_{n_j j}) = 1$, la media di gruppo viene rappresentata nel grafo da un nodo deterministico, posizionato in un blocco successivo a quello che contiene la corrispondente variabile individuale e antecedente il blocco contenente la variabile latente di gruppo.

Il blocco 6 contiene un'unica variabile inosservabile, l'effetto casuale U_{0j}, che entra in gioco contribuendo alla varianza della risposta. Tale variabile è rappresentata nel grafo da un nodo latente di gruppo.

---

[3] Le variabili utilizzate sono state scelte sulla base della conoscenza del fenomeno e dei risultati di numerose altre analisi condotte in precedenza. L'obiettivo è quello di giungere ad un modello relativamente semplice che colga gli aspetti essenziali del processo oggetto di studio.

Tabella 1 *Ordinamento in blocchi delle variabili*

Blocco	Variabile	Descrizione e modalità
1 esogene	MASCHIO	Genere: 1=maschio, 0=femmina
	TITOLO MADRE	Titolo di studio della madre: obbligo (rif), diploma, laurea
2 intermedie	LICEO	Tipo scuola superiore: 1=liceo, 0=altra maturità
	VOTOMAT	Voto di maturità: 36-60 (media=48)
3 intermedie	DU	Tipo di corso: 1=corso di diploma, 0=corso laurea
4 intermedie	ETALAU	Età alla laurea: 21-50 (media=27.6)
	VOTO ESAMI	Voto medio agli esami di profitto:18-30 (media=26.8)
5 medie di gruppo (cluster mean, c.m.)	c.m. VOTO ESAMI	Media di corso di voto esami
	c.m. ETALAU	Media di corso età alla laurea
	c.m. VOTOMAT.	Media di corso voto di maturità
6 nodo latente di gruppo	$U_{0j} \sim N(0,\tau^2)$	Variabile latente di corso
7 risposta	LAVORO STABILE	Posizione lavorativa: 1=lavoro stabile, 0=altrimenti

La fattorizzazione della distribuzione congiunta per il generico corso *j*, derivante dall'ordinamento riportato nella Tabella 1 e dalle assunzioni di indipendenza del modello multilivello, è la seguente:

$$\begin{aligned} f(\mathbf{y}_j, u_{0j}, \mathbf{x}) &= f(\mathbf{y}_j \mid u_{0j}, \mathbf{x})\, f(u_{0j})\, f(\mathbf{x}) \\ f(\mathbf{x}) &= f(\mathbf{x}_{[4]} \mid \mathbf{x}_{[3]}, \mathbf{x}_{[2]}, \mathbf{x}_{[1]})\, f(\mathbf{x}_{[3]} \mid \mathbf{x}_{[2]}, \mathbf{x}_{[1]})\, f(\mathbf{x}_{[2]} \mid \mathbf{x}_{[1]}) \end{aligned} \tag{7}$$

dove con $\mathbf{X}_{[k]}$ abbiamo indicato le variabili del *k*-mo blocco, *k*=1,2,3,4, ad esempio $\mathbf{X}_{[2]}$={LICEO, VOTOMAT}.

L'adattamento del modello grafico corrispondente alla fattorizzazione (7) comporta la stima di 4 modelli di regressione, alcuni dei quali multivariati. Data l'alternanza tra variabili qualitative e continue in blocchi successivi, la procedura di stima suggerita da Cox e Wermuth (1996), che si basa sull'adozione delle proprietà di Markov LWF, consiste nell'adattare, per ogni variabile endogena, un modello di regressione univariato appropriato alla natura della variabile, inserendo come

variabili esplicative tutte e sole le variabili che stanno nei blocchi precedenti e nello stesso blocco della variabile considerata. Quando la variabile dipendente è quantitativa si stima un modello di regressione lineare, mentre nel caso di variabile dipendente dicotomica si adatta un modello logistico. Il modello multilivello (intercetta casuale) è assunto solo per la variabile LAVORO STABILE.

Per la stima dei modelli di regressione si utilizza il metodo della massima verosimiglianza. Il modello logistico a due livelli per la variabile di risposta è stimato approssimando la verosimiglianza con quadratura numerica gaussiana adattiva, tramite la procedura `gllamm` di Stata (Rabe-Hesketh *et al.*, 2001).

Il modello grafico selezionato è riportato nella Figura 2. Le frecce sono presenti quando il *p-value* del corrispondente coefficiente di regressione è minore di 0.10.

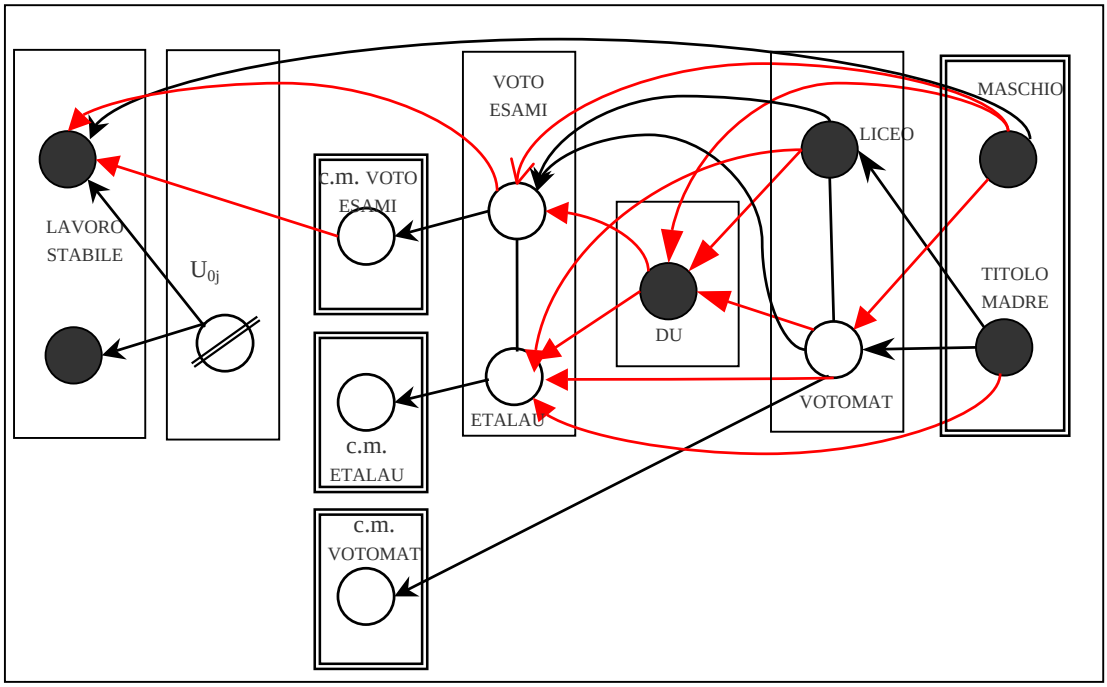
Per facilitare la lettura del grafo gli effetti positivi sono rappresentati da frecce nere, mentre quelli negativi da frecce grigie. Si osservi che le frecce che arrivano sulle medie di gruppo rappresentano un legame di tipo deterministico (come evidenziato anche dal blocco a doppia linea), mentre il segno del legame tra la variabile di risposta e il nodo latente U_{0j} non è identificabile. Inoltre, i due nodi individuali sono da considerarsi identicamente distribuiti, cioè la struttura di dipendenza dei due nodi è la stessa anche se, per semplificare la lettura del grafo, le frecce sono state tracciate solo per il primo dei due nodi individuali.

Le stime relative ai modelli per le variabili intermedie non sono riportate nel presente lavoro, in quanto l'informazione essenziale ai fini dell'analisi è contenuta nella Figura 2. I parametri relativi al modello logistico a intercetta casuale per la probabilità di occupazione stabile sono riportati nella Tabella 2.

Dalla Figura 2 si nota che la variabile di risposta, LAVORO STABILE, dipende direttamente solo da MASCHIO e VOTO ESAMI. Le variabili MASCHIO e VOTO ESAMI costituiscono quindi un insieme separatore tra la variabile di risposta e le altre covariate, nel senso che LAVORO STABILE è indipendente da ETALAU, DU, LICEO, VOTOMAT e TITOLO MADRE condizionatamente a MASCHIO e VOTO ESAMI.

Se ci si fosse limitati alla stima del modello relativo alla variabile di risposta, si sarebbe giunti alla conclusione che gli unici fattori rilevanti sull'inserimento lavorativo sono le variabili MASCHIO e VOTO ESAMI, mentre in realtà anche le altre covariate influenzano il risultato, sebbene non direttamente. Queste considerazioni mettono in luce le potenzialità del modello grafico a catena che, dato un ordinamento delle variabili in blocchi, consente di individuare relazioni dirette e indirette delle covariate sulla variabile di risposta.

Figura 2 *Modello grafico selezionato: laureati/diplomati (vecchio ordinamento) in cerca di lavoro o occupati all'intervista. Ateneo di Firenze, anno 2000.*



La presenza della freccia tra la media di gruppo di VOTO ESAMI e la variabile di risposta sta ad indicare che VOTO ESAMI ha un diverso effetto *entro* e *tra* gruppi, come discusso nel Par. 2. Dall'espressione (6) discende che l'effetto *entro* è il coefficiente della variabile individuale (-0.055), mentre l'effetto *tra* va calcolato come somma tra il coefficiente della variabile individuale e quello della media di gruppo[4] (-0.221) e quindi è -0.276. Entrambi gli effetti sono negativi ma quello tra gruppi è molto più forte, per cui l'effetto negativo del voto è per lo più un effetto del corso di studi: ad un voto più alto corrisponde una minore probabilità di occupazione stabile perché i voti alti sono più frequenti in corsi di studio che forniscono scarse opportunità occupazionali (tra gli altri, quelli della Facoltà di Lettere). L'effetto individuale, benché significativo e negativo, è di modesta entità. Una possibile spiegazione del fatto che l'effetto individuale sia negativo è che gli studenti con voto più alto hanno maggiori ambizioni e quindi sono più selettivi nell'accettare le offerte di lavoro.

[4] Si tenga presente che il modello considerato è di tipo logistico e pertanto il coefficiente di regressione β di una data covariata rappresenta l'effetto sul logaritmo dell'*odds ratio*. L'effetto marginale sulla probabilità di occupazione è invece $\beta^{*}P(Y=1|\mathbf{X},U)^{*}[1- P(Y=1|\mathbf{X},U)]$.

Tabella 2 *Modello logit a intercetta casuale per la stima della probabilità di occupazione 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
VOTOMAT	-0.003	0.007	0.595
c.m. VOTOMAT	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
ETALAU	-0.005	0.016	0.737
c.m. ETALAU	-0.034	0.106	0.747
τ^2	0.510	-0.148	

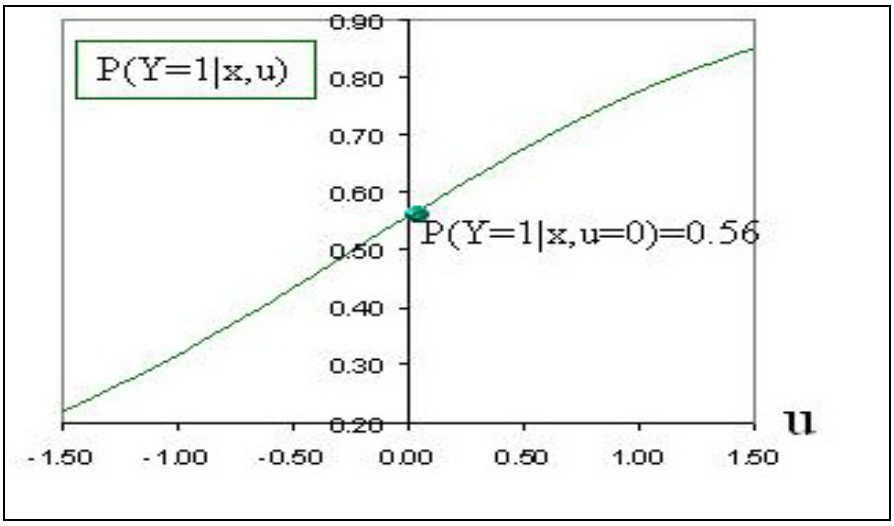
Se si ignora la distinzione tra effetti *entro* e *tra* gruppi, adottando un modello in cui sia presente la variabile VOTO ESAMI senza la corrispondente media di gruppo, si ottiene una stima del coefficiente pari a -0.068 (s.e. 0.029). Tale coefficiente è l'effetto totale della variabile VOTO ESAMI e quindi è difficilmente interpretabile. Se si interpretasse questo effetto come interamente individuale si giungerebbe a conclusioni errate.

Le variabili non osservate a livello di corso di studi sono rilevanti. Il test del rapporto di verosimiglianza che confronta il modello con e senza componente di varianza è significativo (statistica test 108.8, con 1 gdl) e il coefficiente di correlazione intraclasse ρ assume il valore 0.134, a significare che il 13% circa della varianza non spiegata è da attribuirsi al corso di laurea/diploma. Tale valore è piuttosto elevato considerato sia l'ambito applicativo che il tipo di modello utilizzato.

Nella Fig. 3 si riporta l'andamento della probabilità di occupazione stabile $P(Y_{ij}=1|u_{0j},\mathbf{x})$ al variare di u_{0j} per un individuo tipo (cioè con covariate **x** fissate ai seguenti valori: maschio, madre diplomata, altra maturità, voto maturità 48, voto medio esami 26, età alla laurea 27 anni). Per questo individuo la probabilità di occupazione stabile è 0.56 se laureato/diplomato in un corso medio ($u_{0j}=0$), sale a 0.72 se laureato/diplomato in un corso che fornisce ottime opportunità di lavoro

stabile ($u_{0j} = +2\hat{\tau}$) e scende a 0.38 se laureato/diplomato in un corso con pessime opportunità di lavoro stabile ($u_{0j} = -2\hat{\tau}$).

Figura 3 *Probabilità di occupazione stabile per un individuo tipo al variare dell'effetto casuale di corso di studi*



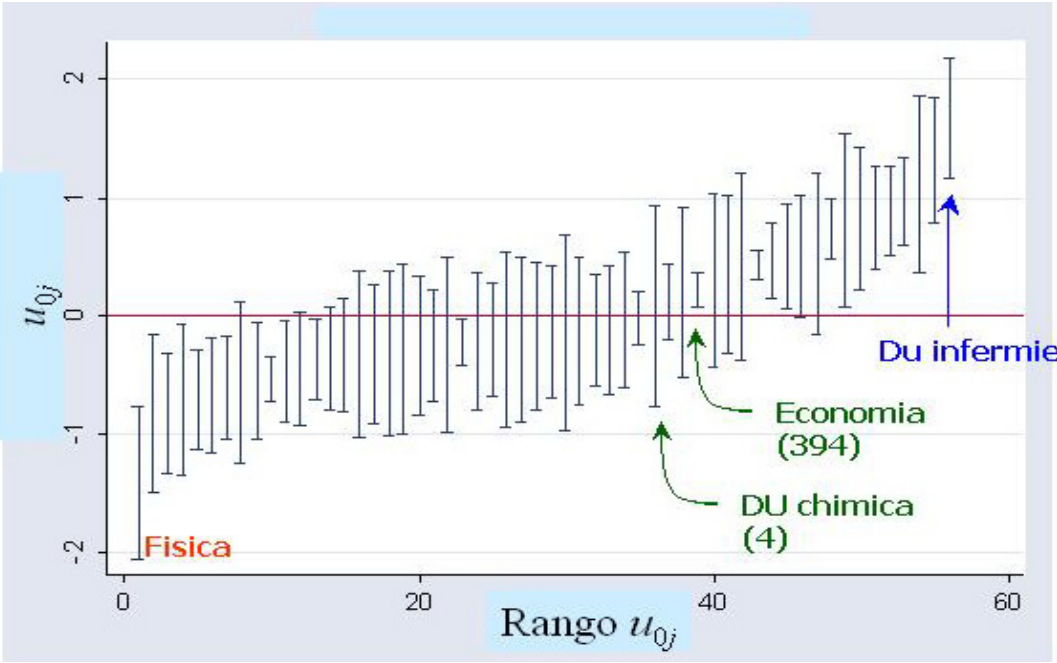
Una volta stimati i parametri del modello, i residui $\hat{u}_{0j}$ possono essere calcolati con il metodo bayesiano empirico (Snijders e Bosker, 1999). I corsi di studio con un residuo positivo (negativo) hanno laureati/diplomati con una probabilità di occupazione stabile superiore (inferiore) a quanto previsto dalle covariate disponibili.

Per confrontare i residui a coppie si può utilizzare la Figura 4: due residui sono significativamente diversi (ad un livello di fiducia di circa 95%) se e solo se i corrispondenti intervalli non si sovrappongono. Per ogni corso di studi l'ampiezza del corrispondente intervallo è funzione decrescente della numerosità campionaria: si considerino, ad esempio, le ampiezze degli intervalli relativi al DU Chimica (4 diplomati) e al CdL Economia (394 laureati). Si noti che per i corsi di studio con pochi laureati/diplomati, data la notevole ampiezza del corrispondente intervallo, il confronto con altri corsi è improponibile da un punto di vista statistico, in quanto le differenze non sono quasi mai significative. Per quanto riguarda gli estremi della graduatoria dei residui, al primo posto si trova il DU Infermieri, mentre all'ultimo posto si trova il CdL Fisica.

Si osservi che i residui di secondo livello $\hat{u}_{0j}$ incorporano tutti i fattori non osservati a livello di corso di laurea/diploma e quindi possono essere interpretati

come una misura dell'efficacia esterna del corso al lordo della situazione del mercato del lavoro. Per esempio, il fatto che gli infermieri trovino lavoro più facilmente di altri laureati, dipende non solo dalla qualità del corso di diploma, ma anche dall'elevata richiesta del mercato rispetto a questa figura professionale (si vedano a questo proposito anche Chiandotto e Grilli, 2003).

Figura 4 *Intervalli per confronti a coppie tra i residui a livello di corso di studi (livello di confidenza medio 95%)*



5. Conclusioni

Nel presente lavoro è stata proposta e applicata una metodologia di analisi basata sull'integrazione tra modelli grafici a catena e modelli multilivello. Tale metodologia offre il vantaggio di rendere esplicite le ipotesi a priori relativamente all'ordinamento logico-temporale delle variabili e le indipendenze condizionali sottostanti il modello multilivello. Il modello grafico consente di visualizzare gli effetti diretti e indiretti sulla variabile di risposta e di leggere in maniera semplice e diretta le indipendenze condizionate tra le covariate coinvolte, grazie alle proprietà markoviane del grafo.

Nell'applicazione le variabili coinvolte non hanno distribuzione congiunta normale multivariata, né condizionatamente normale (*Conditional Gaussian*), per cui

le stime ottenute dipendono dall'ordinamento assunto tra le variabili. Modificando l'ordine tra le variabili le stime e quindi le relative indipendenze condizionali potrebbero risultare diverse. Tuttavia, nel caso in esame, l'ordinamento utilizzato è molto plausibile, perché dettato dall'ordinamento logico-temporale tra le variabili.

Le potenzialità di questa classe di modelli devono essere ancora esplorate. In particolare può essere utile estendere la metodologia nelle seguenti direzioni: più regressioni multilivello nello stesso grafo, modellazione del processo di formazione dei gruppi, regressione di variabili di gruppo su variabili individuali.

Nell'applicazione si sono utilizzati i soli dati relativi ai laureati/diplomati che hanno cercato un lavoro dopo la conclusione degli studi, quindi la distribuzione congiunta considerata è condizionata a questo sottoinsieme di laureati/diplomati. Per esempio, la relazione tra scelta del corso di studi (DU: corso di diploma verso corso di laurea) e le caratteristiche dello studente (LICEO, VOTOMAT, MASCHIO, TITOLO MADRE) sono analizzate solo per i laureati/diplomati in cerca di lavoro e non per l'insieme degli studenti.

La scelta di condizionarsi sempre al sottoinsieme dei laureati/diplomati in cerca di lavoro deriva dalla necessità di avere un unico campione cui riferire la distribuzione congiunta, com'è usuale nei modelli grafici. La considerazione di un campione più ampio, ad esempio quello di una coorte di immatricolati, richiede la soluzione del problema della rappresentazione dei dati mancanti (non MAR) nel modello grafico: per esempio, per gli immatricolati che non concludono gli studi entro l'anno considerato, tutte le variabili relative alla laurea/diploma e quelle relative alla successiva ricerca di lavoro non sono osservabili. La definizione di modelli grafici multilivello che consentano di rappresentare anche il processo di selezione delle unità statistiche sarà oggetto della ricerca futura.

Riferimenti bibliografici

CHIANDOTTO B., BACCI S. (2004) Un modello multilivello per l'analisi della condizione occupazionale dei laureati dell'Ateneo fiorentino. In: CROCETTA C. (a cura di) *Modelli di statistici per l'analisi della transizione Università–lavoro*, CLEUP, Padova: 1-22

CHIANDOTTO B., GRILLI L. (2003) *La domanda di lavoro nella provincia di Firenze: Analisi integrative sui dati dell'indagine Excelsior 2003*, Università di Firenze (http://www.unifi.it/aut_dida/indexval.html)

COX D., WERMUTH N. (1996) *Multivariate Dependencies. Models, Analysis and Interpretation*, Chapman and Hall, London

FRYDENBERG M. (1990) The chain graph Markov property, *Scandinavian Journal of Statistics*, **17**: 333-353

GOLDSTEIN H. (2003) *Multilevel Statistical Models, 3rd edition*, Arnold, London

GOTTARD A., RAMPICHINI C. (2004) Chain Graphs for Multilevel Models, *Working Paper* Dipartimento di Statistica 'Giuseppe Parenti', n.8/2004, Firenze

LAURITZEN S. (1996) *Graphical Models*, Clarendon Press, Oxford

LAURITZEN S., WERMUTH N. (1989) Graphical models for the association between variables, some of which are qualitative and some quantitative, *Annals of Statistics*, **17**: 31–57

RABE-HESKETH S., PICKLES A., SKRONDAL A. (2001) GLLAMM manual, Technical Report 2001/01, Department of Biostatistics and Computing, Institute of Psychiatry, King's College, London

SNIJDERS T., BOSKER R. (1999) *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling*, Sage, London

A Multilevel Analysis of the Employment Probability of Graduates through Chain Graph Models

Summary. *The main goal of this paper is the analysis of the working position of graduates using the potentialities of multilevel models and chain graph models, extended to the case of correlated data. After a brief introduction of multilevel models and a description of the conditional independencies derived from the model, the paper defines chain graphs for multilevel models. The proposed model is used to analyse the factors influencing the graduates job position, using data on the graduates of the year 2000 of the University of Florence. Our work ends with some indications for future research.*

Keywords*. Chain graph models, Multilevel models, Logistic regression, Graduates, University of Florence.*