Tecniche di segmentazione e modelli a componenti di varianza per l'analisi del tempo di ritardo nel conseguimento della laurea

Marisa Civardi, Emma Zavarrone[1]
Dipartimento Metodi Quantitativi, Università di Milano-Bicocca

Riassunto: Lo scopo del lavoro è quello di analizzare gli effetti, sulla permanenza degli studenti all'interno del sistema universitario oltre la durata legale del corso di studi, di caratteristiche, sia individuali sia legate al contesto da cui lo studente proviene ed in cui è inserito, attraverso la segmentazione e l'analisi di covarianza ad effetti random (RANCOVA).
Parole chiave: RANCOVA, RECPAM, COX REGRESSION

1. Introduzione

Il lavoro presenta un approccio esplorativo per l'analisi della permanenza degli studenti all'interno del sistema universitario oltre la durata legale del corso di laurea di iscrizione. Tale approccio deriva dalla fusione della tecnica di segmentazione con quella dei modelli a componenti di varianza.

Come è noto, nella valutazione delle determinanti del rischio di sopravvivenza, la presenza di livelli di stratificazione delle unità di analisi richiede il ricorso a modelli multi livello. Inoltre, in genere, in questo tipo di analisi risultano coinvolte contemporaneamente sia variabili individuali sia variabili, non direttamente osservabili, legate al contesto formativo da cui lo studente proviene. Il ricorso a tecniche di segmentazione, comunemente note come alberi di regressione o alberi di classificazione, può facilitare, nella fase esplorativa, l'individuazione di gruppi omogenei rispetto sia al tempo di permanenza nel sistema universitario sia a caratteristiche legate al percorso formativo dello studente. Sui gruppi così individuati viene successiva-

[1] Il presente lavoro è stato finanziato nell'ambito del progetto "La ricerca di determinanti del rischio mediante analisi di segmentazione di campioni", cofinanziato dal MIUR. Coordinatore nazionale è L. Fabbris. Pur essendo frutto del lavoro comune, il § 3 è opera di Emma Zavarrone, il § 4 di Marisa Civardi, mentre l'Introduzione, il § 2 e le conclusioni sono stati redatti da entrambi gli autori.

mente applicata l'analisi della covarianza ad effetti random (RANCOVA), al fine di verificare se la differenza tra i tempi di permanenza dei gruppi identificati attraverso la segmentazione siano imputabili solo alla variabilità non osservata o, invece, anche alle covariate introdotte nel modello.

2. L'ambito applicativo

L'applicazione riguarda il contesto universitario e la variabile oggetto d'analisi è il tempo di ritardo nel conseguimento della laurea. Si intendono identificare quali siano i fattori, individuali e di contesto, che influenzano, all'interno del sistema universitario, la sopravvivenza dello studente oltre il termine della durata legale del corso scelto (tempo t_0). Il periodo temporale oggetto della nostra analisi è quindi quello che intercorre tra t_0 e t, data corrispondente alla sua uscita dal sistema.

Tra le variabili che costituiscono la base dati, un primo gruppo appartiene in modo univoco alla dimensione micro, essendo costituito da variabili definite sulle unità finali (gli iscritti appartenenti ad una coorte) e riferite a caratteristiche strettamente individuali (Kreft and De Leew, 1998). Nella nostra applicazione, in particolare, in questo gruppo rientrano il genere ed il ritardo nella decisione di iscriversi all'università. Esistono poi variabili micro che invece caratterizzano aspetti della dimensione macro. A questo insieme appartengono il Corso di Laurea a cui lo studente è iscritto e l'anno di prima immatricolazione. La prima variabile, infatti, pur essendo rilevata a livello micro, identifica la tipologia della struttura didattica di appartenenza, la seconda, invece, identifica la coorte d'appartenenza e quindi, indirettamente, modalità tendenzialmente omogenee di fruizione della didattica[2].

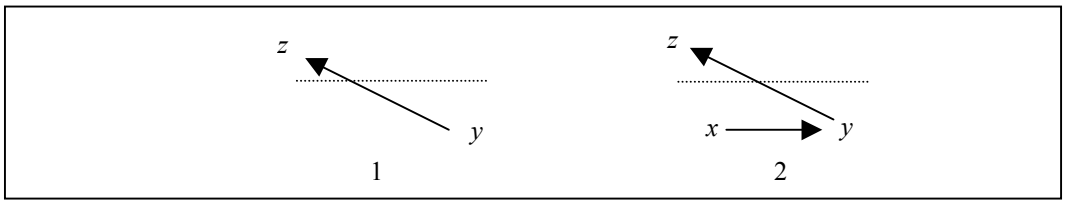
A questa seconda tipologia di variabili, a nostro avviso, appartengono anche quelle indicative del background formativo dello studente rappresentate, nella nostra applicazione, dal tipo di scuola da cui proviene, dalla regolarità della frequenza della stessa e dal voto di diploma conseguito. Si tratta, infatti, di variabili che possono essere considerate come la manifestazione osservabile di una variabile latente che esprime il tipo di formazione conseguito durante la scuola media superiore.

La presenza di variabili che caratterizzano il livello macro implica che le osservazioni effettuate sulle micro unità (l'outcome) debbano essere considerate dipendenti anche dalla loro collocazione all'interno delle macro unità che esprimono la struttura didattica, le modalità di fruizione della stessa, il tipo di formazione possedu-

[2] E' appena il caso di segnalare che la distinzione tra variabili che caratterizzano la dimensione micro e quelle che caratterizzano la dimensione macro non implica necessariamente l'assunzione di una gerarchia tra le due dimensioni ma, solamente, l'opportunità che il modello contempli la possibilità di quantificare le loro possibili interdipendenze.

to (Snijders and Bosker, 1999). Con riferimento in particolare a quest'ultimo aspetto, si deve cioè verificare se la formazione comune conseguita dagli studenti che hanno lo stesso tipo di percorso scolastico (liceo piuttosto che istituto tecnico), e che quindi hanno sperimentato un analogo tipo di formazione di base, possa influire sul tempo di ritardo nel conseguimento della laurea. Si tratta, in ultima analisi, di adottare un approccio di relazioni micro-a-macro, rendendo esplicito il legame che si instaura tra le variabili di livello micro e quelle relative al contesto (le variabili del livello macro).

Figura 1. Relazioni micro a macro



Nella figura 1 sono schematizzate le due situazioni possibili. La situazione (1) rappresenta il caso della proposizione micro-a-macro[3]. Essa è la formalizzazione della seguente possibile ipotesi: il tempo, *y*, impiegato per conseguire la laurea da uno studente iscritto ad un corso di laurea di una classe umanistica sarà inferiore se lo studente proviene dal liceo classico e quindi se ha una formazione, *z*, di tipo umanistico.

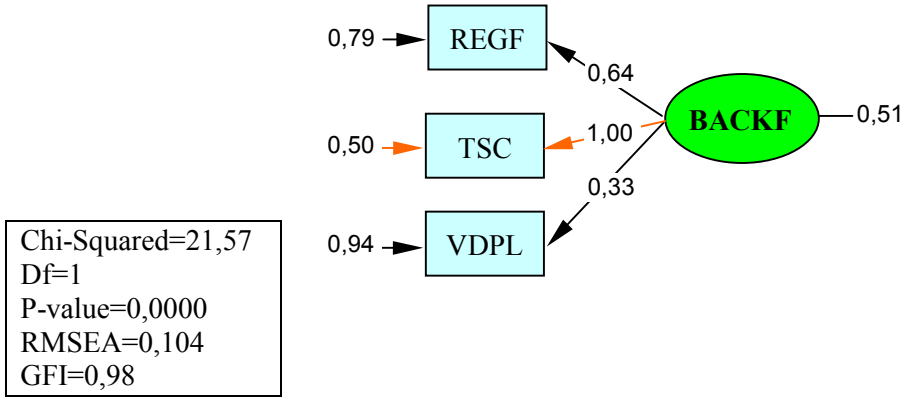
La situazione (2) è un caso speciale della (1). Essa indica che, pur esistendo sempre una relazione tra *y* e *z*, ora si tiene conto anche dell'effetto esercitato sulla *y* da un'altra variabile micro *x*. L'esempio precedente diventerebbe: dato il livello *x*, (espresso, ad esempio, dal voto conseguito al diploma) di preparazione nel tipo di formazione, il tempo impiegato per conseguire la laurea da uno studente iscritto ad un corso di laurea di una classe umanistica…..

Accettare che per le variabili relative al tipo di formazione esista una relazione micro-a-macro è dunque equivalente ad assumere che esse siano variabili osservate a livello individuale ma associate ad una variabile latente sottostante, riferita non all'individuo ma all'ambiente (contesto) in cui esso si è formato e che sintetizza il suo background formativo. Per verificare questa ipotesi, sul collettivo delle coorti di studenti iscritti al Corso di laurea in Economia dell'Università di Milano-Bicocca negli a.a. 92/93 – 96/97, è stata applicata l'analisi fattoriale confermativa (CFA). Essa ha evidenziato che le tre variabili utilizzate (tipo di scuola, voto di diploma e regolarità negli studi medi) possono essere considerate come buoni indicatori del

[3] È appena il caso di osservare che, speculari alle relazioni micro-macro, esistono ovviamente quelle macro-a-micro.

background formativo. La figura 2 riporta il *path diagram* ed i valori più significativi degli indici che consentono di valutare la bontà di adattamento[4].

Figura 2 Path diagram relativo al background formativo



3. Segmentazione: il metodo RECPAM

Per determinare le modalità in cui si articola il background formativo quando venga introdotto come variabile esplicativa in una funzione di sopravvivenza, sembra particolarmente utile la tecnica di segmentazione. A questo scopo si è scelto di utilizzare il metodo **RECPAM**, (**REC**ursive **P**artition **AM**algamation) introdotto da Ciampi et al. (1981, 1987, 1990) in ambito epidemiologico. **RECPAM** è una tecnica che combina la classe degli algoritmi usualmente utilizzati per la segmentazione con il modello lineare generalizzato al fine di individuare e tenere sotto controllo gli "effetti principali" delle interazioni che si instaurano tra i diversi predittori coinvolti nella segmentazione. Il risultato finale è la costituzione di gruppi i cui componenti presentino valori tendenzialmente simili di "sopravvivenza" nonché caratteristiche individuali il più possibile omogenee.

Questa tecnica di segmentazione è stata sviluppata pensando a dati che presentano una struttura generale del tipo $(\boldsymbol{y},\boldsymbol{z})$ dove $\boldsymbol{z}$ è un vettore di predittori e $\boldsymbol{y}$ è un vettore di variabili dipendenti che non devono necessariamente essere considerate

[4] Gli indici considerati per la valutazione dell'adattamento si riferiscono a tre diversi tipi di verifiche: quella sul modello totale (indice GFI), quella basata sul numero di parametri da stimare (indice RMSEA) e quella basata sul confronto tra due distinti modelli (chi quadrato).

come "variabili risposta" [5]. Il vettore ***y*,** infatti, potrebbe essere costituito anche da un misto di variabili casuali e non casuali quali ad esempio la "risposta" e i trattamenti; la sua distribuzione deve essere almeno parzialmente specificata e l'oggetto della previsione è uno dei parametri di questa distribuzione, generalmente di tipo multidimensionale. Tale parametro è denotato con γ ed è definito "criterio". Il punto di partenza per la generazione dell'albero e dei suoi sotto alberi è l'applicazione, ad un dataset sufficientemente ampio D, della partizione binaria iterativa (funzione di split). Anche in RECPAM, la funzione di split è definita come in modo usuale con $\phi(s,g)$, dove *s* indica lo split e *g* il nodo. Essa esprime il contenuto di informazione (*information content, IC*) della partizione effettuata nel nodo sulla base del criterio γ. Tuttavia, a differenza degli altri metodi di segmentazione noti in letteratura, $\phi(s,g)$ è determinata come il rapporto di verosimiglianza condizionato (LRS) risultante dal confronto delle curve di sopravvivenza della coppia di nodi[6]. LRS assume la seguente forma (Cox, 1972):

$$L(\gamma) = \prod_{i=1}^{n} \left(\frac{exp(I_i(x)\gamma)}{\sum_{l \in R_i} exp(I_l(x)\gamma)} \right)^{\delta_i} \qquad [1]$$

ove $I_i(x)$ sono variabili dummy, funzioni dei predittori ***x***, indicanti l'appartenenza alla specifica classe (*class indicator variables*), δ_t è il parametro di *censoring*, R_i indica l'insieme a rischio al tempo *t*, ossia, nel nostro caso, tutti gli studenti che al tempo *t* non hanno conseguito ancora la laurea e che sono esposti al "rischio" di laurea.

La costruzione dell'albero di sopravvivenza con il metodo RECPAM rispetta tutte le fasi standard della costruzione degli alberi di classificazione. Nel disegnare l'albero si assume questa convezione: sul lato sinistro vanno le unità che appartengono alla classe I_k=1 (risposta si). Sul lato destro le altre. Poiché la risposta "si" implica γ_k>0, gli studenti del ramo sinistro avranno un tempo di permanenza minore di quelli del ramo destro. Pertanto, i gruppi che nelle indagini classiche di sopravvivenza presentano la prognosi peggiore, poiché sopravvivono meno, nella nostra applicazione saranno invece i migliori. Tradizionalmente, la classe di riferimento è costituita dai soggetti che presentano valore **no** (assenza) a tutte le domande: ciò significa che nel-

[5] E' possibile, infatti, applicare la tecnica RECPAM alle diverse tipologie di variabili dipendenti (discrete, continue, censored) ed associare a ciascuna tipologia di variabili appropriate regole di *split* e di *pruning* nonché di verifica della bontà di adattamento.

[6] Nella metodologia CART (1984), ad esempio, la funzione di split può essere stimata sia ricorrendo ai minimi quadrati (LS) che alla minima deviazione assoluta (LAD). Se si opta per i minimi quadrati, la funzione di split è data da $\phi(s,g) = SS(g) - SS(g_L) - SS(g_R)$ dove *SS* indica la somma dei quadrati all'interno del nodo. Lo split scelto è quello che massimizza $\phi(s,g)$.

la rappresentazione grafica la foglia più a destra corrisponderà alla prognosi migliore e quindi al tempo di sopravvivenza più lungo.

La verifica della bontà dell'adattamento viene effettuata sempre sulla base del rapporto LRS ricorrendo sia al criterio di Akaike (AIC), sia alla teoria di Gabriel (1969), basata sulla verifica di ipotesi simultanee. L'aspetto innovativo di RECPAM è la possibilità di individuare i gruppi più simili tra loro e di unirli assicurando la minima perdita di informazione (**AM**algamation). Il risultato di questo processo è la costruzione di una sequenza di partizioni *nested* che dovrebbe fornire l'albero più "onesto". Questa nuova struttura, quando sia individuata, nell'output viene indicata graficamente con un poligono esagonale.

La tecnica RECPAM consente di effettuare due distinti tipi di studi:

1. classificazione di tipo prognostico in presenza di una stratificazione *a priori*,
2. analisi per sottogruppi.

Nel primo caso, l'obiettivo è quello di trovare una classificazione prognostica sulla base di dati di sopravvivenza censurati. Gli strati sono in questo caso rappresentati dalle coorti e l'ipotesi nulla da sottoporre a verifica è l'assenza di interazioni fra l'appartenenza ad una coorte e i fattori "prognostici". Ciò equivale ad assumere le coorti come strati a priori.

Per ogni soggetto *i* il dataset assume la forma:

$$\left(t^{(i)}, \boldsymbol{\delta}^{(i)}, s^{(i)}, \boldsymbol{x}^{(i)}\right) \qquad i = 1,2,...,N \qquad [2]$$

dove:

$t^{(i)}$ = tempo di sopravvivenza → numero di giorni di permanenza nel sistema dopo t_0,

$\delta^{(i)}=1$ se lo studente è laureato; in questo caso t indica i giorni di ritardo laurea,

$\delta^{(i)}=0$ se lo studente presente nel sistema non è ancora laureato; in questo caso t è un tempo censurato,

$s^{(i)}$ = variabile discreta indicante lo strato (coorte),

$\boldsymbol{x}^{(i)}$ = vettore di predittori.

Un modello con un buon adattamento ai dati nel processo di classificazione è basato sulla seguente funzione di rischio:

$$h(t; s, \boldsymbol{x}) = \exp\{\gamma_1 I_1(\boldsymbol{x}) + \gamma_2 I_2(\boldsymbol{x}) + \ldots + \gamma_p I_p(\boldsymbol{x})\} h_s(t) \qquad [3]$$

dove:

$h_s(t)$ è la funzione hazard *base-line* per lo studente della coorte s;

I_k sono variabili dummy, funzioni dei predittori $\boldsymbol{x}$, che indicano l'appartenenza a una delle p classi diverse da quella di riferimento ($I_k(\boldsymbol{x})$ =1 se lo studente rientra nella classe k, $I_k(\boldsymbol{x})$ =0 negli altri casi).

γ_k è il LOG-RELATIVE hazard della classe k; esso quindi non è altro che il logaritmo del rapporto dell'hazard della classe k rispetto a quello della classe di riferimento.

Le classi sono p+1, cioè la classe di riferimento e le p classi aggiuntive. La generica *k-esima* classe è caratterizzata da un rapporto di hazard, rispetto alla classe di riferimento, costante nel tempo e dato da $\exp(\gamma_k)$. Si assume inoltre che tale rapporto sia indipendente dallo strato (coorte). Si tratta di un modello di regressione di Cox stratificato i cui regressori sono le variabili che indicano le classi di appartenenza.

L'analisi per sottogruppi (secondo caso), è equivalente ad identificare i gruppi di studenti per i quali la relazione fra tempo di permanenza (tempo di sopravvivenza) ed i fattori che lo determinano sia significativamente diversa. Ad esempio, se come fattore determinante si assume una variabile dummy indicante il tipo di scuola di provenienza, l'obiettivo della segmentazione sarà quello di classificare gli studenti in accordo con l'effettiva efficacia di un tipo di scuola rispetto agli altri. L'ipotesi nulla da sottoporre a verifica potrebbe essere allora formulata nel seguente modo: il tempo di ritardo della laurea degli studenti provenienti dai licei è uguale a quello degli studenti provenienti dagli altri tipi di scuola. Il parametro da stimare sarà il coefficiente della variabile dummy in un modello semplice di Cox, e l'albero fornirà predittori distinti di tale coefficiente per ogni ramo.

In questo caso il dataset è del tipo:

$$\left(t^{(i)}, \delta^{(i)}, \boldsymbol{z}^{(i)}, s^{(i)}, \boldsymbol{x}^{(i)}\right) \qquad i = 1,2,...,N \qquad [4]$$

dove $\boldsymbol{z}^{(i)}$ denota un vettore di varabili di speciale interesse (le **determinanti**). Il vettore di predittori $\boldsymbol{x}^{(i)}$ ha un duplice ruolo:

1) determina lo strato prognostico $j(\boldsymbol{x})$;
2) determina p classi per ciascuna delle quali una specifica equazione di regressione descrive gli effetti indotti sulla sopravvivenza dalle variabili determinanti.

Nella nostra applicazione, il vettore multivariato $\boldsymbol{z}$ è costituito da: tipo di scuola, voto di diploma e regolarità degli studi medi.

Il modello diventa:

$$h(t; \boldsymbol{z}, \boldsymbol{x})=\exp[(\gamma_1 \boldsymbol{z})I_1(\boldsymbol{x})+ (\gamma_2 \boldsymbol{z})I_2(\boldsymbol{x})+\ldots+ (\gamma_p \boldsymbol{z})I_p(\boldsymbol{x})]h_{sj}(t) \qquad [5]$$

h_{sj}= hazard base-line per i soggetti appartenenti allo strato s (coorte), definito a priori, e allo strato prognostico $j = j(\boldsymbol{x}) = 1,2,...,L$, determinato dai valori del vettore di predittori $\boldsymbol{x}$. Si assume assenza di informazione sugli hazard base-line.

3.1 RECPAM: risultati

Per costruire l'albero di sopravvivenza si è fatto ricorso ad una macro che utilizza la procedura PHREG[7]. Come variabile obiettivo si è adottata la variabile binaria indi-

[7] Tale macro, scritta in linguaggio IML attiva la procedura PHREG di SAS sviluppata per l'analisi di sopravvivenza (Carinci, 2001).

cante il conseguimento della laurea (*Dumlau*), mentre le variabili determinanti (*tree covariates*) sono il tipo di scuola di provenienza (*Tscuola*, codificato da 1 a 8), il voto di diploma (*VDP,* codificato con valori da 1 a 6[8]) e la regolarità di frequenza della scuola superiore (*Regfress*, variabile binaria, 0= regolare, 1= non regolare). L'obiettivo di questo modello è quello di studiare una partizione in gruppi tenendo sotto controllo le seguenti variabili micro: Genere (*Sex*) e Ritardo in giorni nell'iscrizione all'università (*Iscrug*). La variabile che scandisce il trascorrere del tempo è la Durata in giorni di permanenza nel sistema universitario (*Duratag*) a partire da t_0 (data prima sessione di laurea possibile). Essa compare all'interno dell'analisi di sopravvivenza come tempo principale in base al quale si manifesta l'evento laurea[9]. Poiché il collettivo era costituito dagli iscritti appartenenti a cinque coorti, la prima delle quali relativa all'anno di immatricolazione 1992/1993 e l'ultima all'anno di immatricolazione 1996/1997, per assicurare a tutte le unità di analisi lo stesso tempo disponibile per il conseguimento della laurea si è scelto di porre per ogni iscritto il momento finale di osservazione a 749 giorni. Tale periodo corrisponde ad un intervallo, a partire da t_0, pari a due anni accademici (ricordiamo che la sessione di laurea di ogni anno accademico termina ad aprile dell'anno successivo). Questa scelta ha anche comportato l'esclusione dall'analisi della coorte 1996/1997. Per i laureati delle prime tre coorti con tempo di permanenza superiore a 749, la variabile *Dumlau* è stata posta uguale a zero mentre alla variabile *Duratag* è stato assegnato il valore 749.

L'albero di sopravvivenza ottenuto utilizzando questa base dati (Figura 3), a causa del numero esiguo delle foglie finali, non consente di procedere alla fase di *amalgamation*.

In aggiunta, la fase di pruning[10], che dovrebbe portare all'individuazione dell'albero più "onesto", ha segnalato l'esistenza di due soli gruppi. Si è pertanto scelto di prendere in considerazione la segmentazione ottenuta allo step precedente che evidenzia l'esistenza di 4 gruppi con tempi di sopravvivenza significativamente distinti e crescenti cosicché il Gruppo 1 è quello degli studenti migliori, a cui è associata la più alta probabilità di tempi di permanenza nel sistema (ritardo alla laurea) più ridotti, mentre il Gruppo 4 è quello con la più alta probabilità di rimanere nel sistema più a lungo.

---

[8] Il voto di diploma, espresso in sessantesimi, è stato suddiviso nelle seguenti classi:

Classe	Voto (VDP)	Classe	Voto (VDP)
1	36	5	53≤VDP≤55
2	37≤VDP≤ 41	6	56≤VDP≤59
3	42≤VDP≤48	7	60
4	49≤VDP≤52		

[9] La procedura PHREG prevede la possibilità di inserire più variabili temporali.

[10] Questa fase si arresta quando viene raggiunto il minimo valore dell'indice AIC.

Tabella 1. Modalità delle variabili che caratterizzano il fattore formazione

Gruppo	Tipo scuola	VDP	Gruppo	Tipo scuola	VDP
1=FORM1	Liceo scientifico Liceo classico Titolo straniero	VDP>48/60	3=FORM3	Liceo Linguistico Istituto tecnico Istituti profess. Istituti magistr. Altre scuole	48/60<VDP<55/60
2=FORM2	Liceo linguistico Istituto: tecnici Istituto profess. Istituto magistr. Altre scuole	VDP>55/60	4=FORM4	Tutte le scuole	VDP≤48/60

Nell'ordine, i quattro gruppi individuati sono schematizzati nella tabella 1.

Più in dettaglio, l'albero di figura 1 evidenzia che la variabile in grado di discriminare meglio gli studenti in relazione al rischio di ritardo nel conseguimento della laurea è il voto conseguito al diploma. Indipendentemente dal tipo di scuola scelto, gli studenti che si sono diplomati con al più 48/60 presentano la minore probabilità di laurearsi in tempi brevi (Gruppo 4). E precisamente, il loro rischio di ritardo alla laurea è 2,734 volte quello del gruppo 1 (studenti, provenienti da licei classico e scientifico o con titolo straniero, con voto di diploma >48). Il rischio associato agli studenti del gruppo 3 (titoli diversi da liceo classico scientifico e straniero, ma con voto di diploma compreso tra 48 e 55) sempre rispetto al gruppo 1, indica un hazard ratio (HR) pari a 2,715. Tale rapporto scende invece a 1,328 quando la comparazione sia fatta con gli studenti del gruppo 2 (titoli diversi da liceo classico scientifico e straniero, ma con voto di diploma maggiore di 55). La tabella 2 evidenzia la sintesi delle variabili coinvolte nel processo di segmentazione.

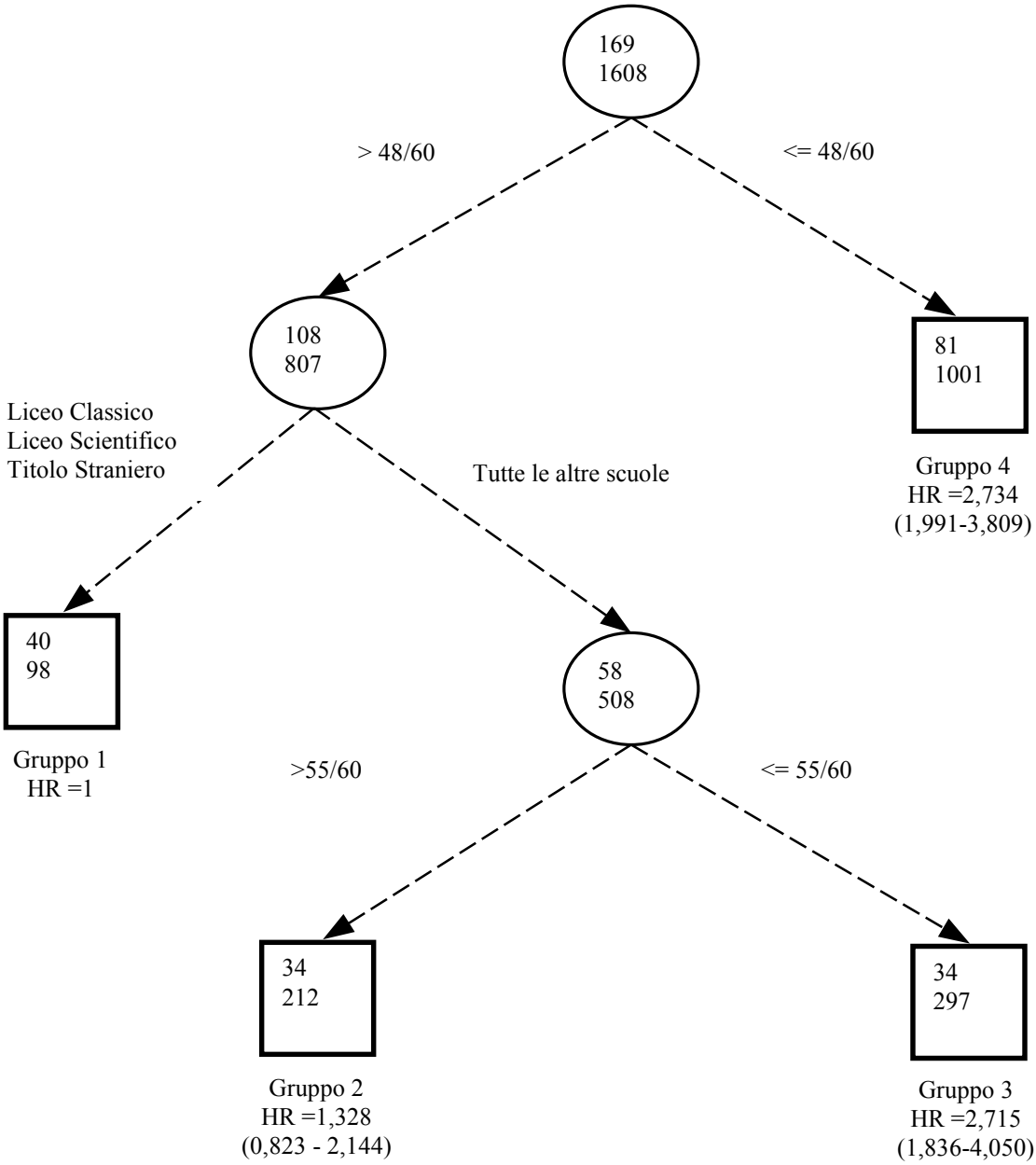
Tabella 2. Valori medi delle variabili utilizzate nel processo di segmentazione

Caratteristiche	FORM1	FORM2	FORM3	FORM4
Valori di HR	1	1/1,328	1/2,715	1/2,734
DURATAGN	624,54	655,51	664,78	673,48
% REGFRESS [(a)]	91,0	88,1	88,9	72,4
% ISCRUG [(b)]	95,7	90,8	91,3	89,2
% di studenti maschi	42,0	36,5	44,7	64,8
% di studenti laureati	29,0	13,2	9,1	5,7

[(a)] Percentuale di studenti con regolarità di frequenza della scuola superiore

[(b)] Percentuale di studenti iscrittisi subito dopo il conseguimento del diploma

Figura 1. Albero di sopravvivenza[(c)]



[(c)] In ogni nodo, il valore superiore indica il numero di laureati, quello sottostante il numero di studenti. Tra parentesi è riportato l'intervallo di variazione dell'HR.

4. Metodologia multilivello: modelli a componenti di varianza

I risultati della segmentazione sono preliminari alla successiva analisi, riconducibile alla classe dei modelli multilivello. Tali modelli, come noto, sono utilizzati nello studio di dati con strutture gerarchiche. Sostanzialmente, si tratta di descrivere come il valore atteso della variabile obiettivo possa essere spiegato mediante funzioni specifiche per ogni livello, così da tener conto delle componenti di variabilità imputabili non soltanto ad attributi relativi alle unità elementari (microunità) ma anche alle caratteristiche dei gruppi (Goldstein, 1991).

Questo lavoro si propone l'obiettivo di determinare, per ogni coorte e per ogni gruppo di formazione individuato, la relazione funzionale tra il tempo di permanenza nel sistema universitario e un insieme di covariate, verificando se tra i gruppi sia possibile identificare un effetto comune. In particolare, è stata condotta un'analisi della covarianza ad effetti random (RANCOVA), basata su un disegno fattoriale con una struttura a blocchi incompleta e non bilanciata (Rao, 1997). Tale scelta è giustificata dalla tipologia dei dati disponibili caratterizzati dalla presenza di blocchi distinti (le 4 coorti), di più trattamenti in ogni blocco (i 4 gruppi di background formativo) e di numerosità differenti in ciascun trattamento (Littel et al., 1996). Ciò che maggiormente differenzia la RANCOVA dall'ANCOVA é il vincolo di uguaglianza dell'effetto per tutti i gruppi (*common effect*) (Bryk and Raudenbush, 1992). Questo implica che mentre nell'ANCOVA gli effetti associati ad ogni gruppo sono considerati costanti non note (*fixed*), nella RANCOVA sono, invece, assunti casuali (*random*) e varia solo l'intercetta (Longford, 1993).
Il modello presenta pertanto la seguente forma:

$$y_{ijl} = \alpha_i + \beta_i x_{ijl} + b_j + e_{ijl} \qquad (i,j) \in \mathbf{B} \qquad \begin{matrix} i = 1,2,..,4 \\ j = 1,2,..,4 \\ l = 1,2,...,n_j \end{matrix} \qquad [6]$$

dove:

- y_{ijl} indica i giorni di permanenza nel sistema universitario, a partire da t_0, dell'l-esimo studente appartenente al blocco j –mo (coorte) e con trattamento i-esimo (tipo di formazione),
- α_i indica l'intercetta riferita al trattamento i-esimo,
- β_i indica la pendenza riferita al trattamento i-esimo,
- x_{ijl} indica il predittore per il j-mo blocco, l'i-mo trattamento e l'l-mo studente,
- b_j $\approx N(0, \sigma_b^2)$ indica l'effetto random per il j-esimo blocco (coorte),
- e_{ijl} $\approx N(0, \sigma_e^2)$ indica gli errori casuali associati alle unità d'analisi,
- **B** è l'insieme delle combinazioni blocchi (j)- trattamenti (i); b_j ed e_{ijl} sono variabili casuali indipendenti.

Tabella 3. Voto di laurea e ritardo alla laurea (in giorni) per coorte e tipo di formazione

TIPO DI FORMAZIONE		COORTE: 1992-93		COORTE 2: 1993-94		COORTE 3:1994-95		COORTE 4: 1995-96	
		Voto di Laurea	Giorni di ritardo	Voto di Laurea	Giorni di ritardo	Voto di Laurea	Giorni di ritardo	Voto di Laurea	Giorni di ritardo
FORM 1	Media	106,40	553,75	102,00	604,00	104,00	597,28	103,40	546,60
	C.V.%	1,2	11,9	1,2	7,9	2,5	5,5	2,0	9,2
	Min-Max	100-110	267-739	96-105	280-732	93-110	505-744	92-110	380-749
FORM 2	Media	102,60	554,20	104,40	572,70	103,00	535,26	102,60	671,00
	C.V.%	2,0	15,8	1,7	9,4	1,3	9,3	1,2	3,0
	Min-Max	99-110	267-739	97-110	280-729	97-110	147-745	95-110	537-749
FORM 3	Media	100,00	659,00	97,70	592,00	101,20	527,20	103,20	660,83
	C.V.%	3,5	8,1	2,7	18,0	2	14,1	2,1	2,9
	Min-Max	94-107	513-739	91-102	280-729	94-105	364-745	95-108	628-748
FORM 4	Media	99,00	651,00	99,60	604,86	100,40	624,75	96,20	644,94
	C.V.%	1,6	5,4	1,2	4,9	1,4	4,1	1,4	3,5
	Min-Max	89-106	375-739	93-110	374-729	93-110	510-745	85-108	380-749

Il modello [6] è noto in letteratura anche come *equicovariance regression model* (Longford,1993).

Nella nostra applicazione la variabile esplicativa *x* è rappresentata dal voto di laurea (*Voto*); l'obiettivo è quello di determinare gli effetti sul ritardo di laurea connessi con la presenza di blocchi (le coorti) e di trattamenti (il tipo di formazione) al netto dell'influenza di questa variabile individuale che indubbiamente può condizionare la decisione di procrastinare il momento della laurea. Quest'ipotesi, d'altra parte, sembra supportata dall'analisi descrittiva preliminare in cui, per ciascuna coorte e per ciascuno dei quattro tipi di formazione, si sono calcolati sia il voto di laurea sia la durata media di permanenza nel sistema. Come emerge dalla tabella 3, ad esclusione dei laureati appartenenti alla coorte 2 per i quali si registrano differenze molte contenute, tra le due variabili sembra esistere una relazione di tipo discorde. Infatti, passando dal primo al quarto gruppo, aumenta la durata media di permanenza e, contemporaneamente, diminuisce il voto di laurea.

4.1 RANCOVA: i risultati

Nel ricercare il modello più idoneo per descrivere la relazione fra il ritardo nel conseguimento della laurea (variabile risposta *y*), la variabile di classificazione del background formativo, la covariata *x* (voto di laurea) e le unità d'analisi (gli immatricolati all'Ateneo) è opportuno prendere le mosse dal modello di regressione più semplice, almeno per la parte che riguarda le covariate, e verificarne l'adattabilità. In quest'ottica, si dovrà, come primo passo, verificare se un modello senza covariate sia in grado di descrivere i dati. Ciò significa, in altri termini, verificare l'ipotesi che le pendenze siano tutte uguali a zero e quindi:

H_0: $\beta_1=\beta_2=\beta_3=\beta_4=0$ vs. H_a: (no H_0)

Se l'evidenza empirica porta al rifiuto di H_0, il passo successivo consiste nel determinare se, per descrivere i dati, sia invece idoneo un modello con una pendenza unica per tutti i trattamenti.
Si dovrà quindi verificare:

H_0: $\beta_1=\beta_2=\beta_3=\beta_4=\beta$, in cui β non è specificato, vs. H_a: (no H_0) .

Se è possibile accettare questa nuova H_0, il processo di confronto tra i modelli di regressione relativi ai singoli trattamenti risulta decisamente semplificato. Infatti, se esiste una pendenza comune, per ogni dato valore della covariata *x*, la stima della media di *y*, per il trattamento *i*, è determinata come:

$$\hat{\mu}(y|x)_i = \hat{\alpha}_i + \hat{\beta}_i x \qquad i=1,2,..,4$$

La differenza tra le medie stimate di due trattamenti *i* e *j*, per un dato valore della covariata $x=x^*$, è indipendente dalla covariata poiché risulta uguale a:

$\hat{\alpha}_i - \hat{\alpha}_j$.

Le quantità $\hat{\mu}\left(y\middle|x^*\right)_i$ sono i valori stimati, sulla base delle rette di regressione per $x=x^*$ e sono chiamate medie aggiustate. I valori di queste medie nel punto $x = \bar{x}$ nel sistema SAS sono indicati con LSMEANS.

In base ai risultati numerici, si rifiuta la prima ipotesi nulla (p=0,0042), e quindi si accetta che il tempo di permanenza nel sistema universitario vari al variare del voto di laurea. La verifica della seconda ipotesi H_0, relativa all'esistenza di pendenze uguali per ciascun tipo di formazione, conduce alla sua accettazione, anche se il livello di significatività (p=0,06) è vicino al limite di rifiuto. Si conclude quindi che i gruppi di formazione presentano la medesima pendenza ma con intercette diverse. Al valore β, costante per tutti i tipi di formazione grazie all'accettazione dell'ipotesi che le pendenze sono uguali tra loro, si perviene attraverso la costruzione del modello RANCOVA. La procedura SAS fornisce, infatti, i valori delle medie aggiustate di *y* per il modello con la pendenza comune, nel punto $x = \bar{x}$, in corrispondenza cioè del valor medio del voto di laurea (pari a 101,3).

I valori di tali medie sono i seguenti:

- FORM1 Permanenza in giorni pari a 558,52
- FORM2 Permanenza in giorni pari a 603,47
- FORM3 Permanenza in giorni pari a 607,63
- FORM4 Permanenza in giorni pari a 614,10

Essi confermano che effettivamente, passando dal gruppo di studenti che la segmentazione aveva indicato come quello migliore (FORM1) ai successivi, la durata di permanenza nel sistema aumenta. La componente di varianza stimata σ_b^2 (varianza dell'effetto coorte) risulta uguale a 161,48 mentre la σ_e^2 (varianza dell'errore) ammonta a 17559,14. Poiché la stima della componente di varianza legata ai singoli soggetti è circa 100 volte quella della componente di varianza dovuta alle coorti, l'effetto di queste ultime sul ritardo alla laurea può essere considerato praticamente trascurabile. Infine, le differenze tra le LSMEANS dei quattro tipi di formazione risultano abbastanza simili e non significative. Il confronto a coppie evidenzia comunque differenze rispetto alla media stimata del primo gruppo più accentuate di quanto non siano quelle esistenti fra i restanti tre gruppi

CONCLUSIONI

Il tentativo di utilizzazione congiunta di tecniche di segmentazione e modelli a componenti di varianza qui sperimentato non può ovviamente essere considerato generalizzabile, dal momento che i risultati a cui si perviene sono fortemente dipendenti dalla natura dei dati utilizzati. In particolare, il contesto che ha riguardato quest'applicazione, più che la ricerca di livelli gerarchici naturali, implicava l'individuazione di modalità significative di classificazione di caratteristiche, rilevate a livello individuale, ma esprimenti macro-dimensioni. L'obiettivo principale era quello di determinare, per ogni coorte e per ogni gruppo di formazione individuato, la relazione funzionale tra il tempo di permanenza nel sistema universitario e un insieme di covariate, al fine di verificare se tra i gruppi fosse possibile identificare un effetto comune.

I risultati ottenuti confermano che la durata di permanenza nel sistema cresce passando dal gruppo di studenti, indicato dalla segmentazione come quello con la più elevata dotazione di capitale formativo, a quelli successivi. Sembra quindi lecito ipotizzare che, soprattutto nella presente fase di riorganizzazione del sistema didattico universitario, il ricorso ad un approccio di questo tipo possa contribuire ad individuare, ai vari livelli decisionali, politiche più efficaci per la riduzione dei tempi di permanenza degli studenti nell'Ateneo.

Riferimenti bibliografici

BRYK e RAUDENBASH (1992), *Hierarchical Linear Models*, Newbury Park.

CARINCI, F., NICOLUCCI, A. PELLEGRINI, F. (2001), "Regression Trees in health services and outcome research: an application of RECPAM approach using quality of care as a criterion", *Technical Report* disponibile su:
http://med.monash.edu.au\publichealth.

CIAMPI A. (1991), "Generalized Regression Tree", *Comput. Stat. Data Analyisis*, 12.

CIAMPI A. e NEGASSA, A.et al. (1995), "Tree-structured prediction for censored survival data and the Cox Model", *J. Clin. Epidemiol*. Vol. 48, n. 5.

COX, D. R. (1972) "Regression Models and Life Tables (with discussion)," *Journal of the Royal Statistical Society*, B, vol. 34.

GOLDSTEIN, H. (1986), "Multilevel mixed linear model analysis using iterative generalized least squares", *Biometrika*, 73.

GOLDSTEIN, H. (1995), *Multilevel Statistical Models*, Arnold, London.

KREFT, I. e DE LEEUW, J. (1998), *Introducing multilevel modelling*, Sage, London.
LEYLAND L. H. e GOLDSTEIN, H. (2001), *Multilevel Modelling of Health Statistics*, Wiley, London.
LITTELL R. C., et al. (1996), *Sas System for Mixed Models*, Sas Institute, Cary, Nc.
LONGFORD N. (1993), *Random Coefficient Models*, Clarendon Press, Oxford.
RAO, P. S. (1997), *Variance Components Estimation*, Chapman and Hall, London.
SNIJDERS T. e BOSKER R. (1999), *Multilevel Analysis*, Sage, London.

Segmentation techniques and variance components models for the analysis of the time of delay in the achievement of the degree

Summary. *This paper highlights the possibility to link different methodologies as segmentation and multilevel analysis (RANCOVA) in the field of the university education. In particular, this analysis regards the conjoint effects of the social-economic background and university context on the permanence of the students in the university system beyond the legal duration of the course of studies.*
Keywords*. RANCOVA, RECPAM, COX REGRESSION .*