Determinanti dell'inserimento professionale dei laureati. Analisi delle interazioni

Mariano Porcu, Giuseppe Puggioni, Nicola Tedesco[1]
Dipartimento di Ricerche Economiche e Sociali - Università degli Studi di Cagliari

Riassunto: Negli studi sull'inserimento professionale dei laureati risulta oggetto di interesse la definizione di un insieme di predittori dell'evento dicotomico lavorare/non lavorare. È del tutto evidente che i predittori esercitano la loro azione sulla variabile risposta non solo singolarmente ma interagendo fra essi. Con il presente lavoro ci si propone di studiare queste azioni congiunte attraverso l'applicazione di una tecnica di analisi di recente introduzione (Boolean logit) supportando la stessa con analisi esplorative basate su segmentazioni binarie.
Parole chiave: Inserimento professionale, determinanti, segmentazione, Boolean regression, logit.

1. Premessa

La ricerca delle determinanti che influiscono sul conseguimento di un'occupazione da parte dei laureati è uno dei temi più importanti affrontati in sede di valutazione dell'efficacia del sistema di formazione universitario; esso è stato studiato da diversi autori e secondo differenti approcci metodologici (Chiandotto, 2004; Civardi-Zavarrone, 2004). Fra questi, di un certo rilievo per la loro diffusione e la loro valenza esplicativa, sono quelli basati sulle relazioni di dipendenza causale di tipo logit. L'evento *lavorare/non-lavorare* può essere considerato, quindi, come una variabile risposta binaria il cui valore dipende da un insieme di variabili predittrici $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$.

I predittori influiscono sulla risposta singolarmente, in maniera congiunta e combinandosi fra loro e secondo i loro diversi livelli. Tale azione sulla risposta prospetta un quadro di analisi riconducibile a quelle che sono le categorie concettuali

[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 Luigi Fabbris, coordinatore del gruppo di Cagliari è Giuseppe Puggioni. Il lavoro è opera comune degli autori. In particolare si possono attribuire a M. Porcu i parr. 1, 2, 5 e 6, a G. Puggioni il par. 3 e a N. Tedesco il par. 4.

della *causazione complessa*. "*Concrete definitions of causal complexity are difficult to come by, perhaps because the concept is so slippery*". In sostanza, "*multiple causes interact with one other and the way in which they interact is described by the logical operators and and or*" (Braumoeller, 2003).

Come è noto sono diversi i concetti che possono essere compresi come esempi di causazione complessa:

- X_1 and X_2 and X_3 causano Y (causazione congiunta multipla);
- X_1 or X_2 or X_3 causano Y (sostituibilità);
- X_2 causa Y ma solo in presenza di X_1 (contestualità);
- X_1 and X_2 causano Y, X_1 or X_2 causano Y (condizioni necessarie e sufficienti);
- (X_1 and X_2) or (X_3 and X_4) causano Y (condizioni INUS[2]).

I meccanismi di causazione complessa sono problematici per la maggior parte delle tecniche statistiche standard. Essi, infatti, implicano delle forme di non additività che provengono dal processo cumulativo dell'influenza delle variabili indipendenti sulla variabile dipendente. Da un punto di vista applicativo sorge, quindi, il problema di come fare per *catturare* con i metodi statistici le implicazioni causali complesse o multiple. In questo campo le proposte metodologiche sono molteplici ed è costante l'attenzione che viene dedicata al problema (Frosini, 2004). Anche facendo riferimento all'evento dicotomico *lavorare/non-lavorare*, si può osservare come in numerosi studi si sia asserito che l'evento è l'esito di un rapporto di causazione complessa o di percorsi causali multipli (Granovetter, 1974; Reyneri, 2002).

2. Modellare l'interazione

Il noto modello di regressione logistica viene frequentemente impiegato per modellare la probabilità di un particolare evento come funzione di un insieme di variabili esplicative. L'influenza delle esplicative sulla variabile risposta viene considerata lineare su una scala logit

$$\log\left(\boldsymbol{\pi}/\left(1-\boldsymbol{\pi}\right)\right)=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{X}_1+\boldsymbol{\beta}_2\boldsymbol{X}_2$$

Per tenere conto dei possibili effetti congiunti esercitati dai predittori si inseriscono dei termini aggiuntivi riferiti al prodotto fra le covariate prese in esame (Hosmer e Lemeshow, 1989)

$$\log\left(\boldsymbol{\pi}/\left(1-\boldsymbol{\pi}\right)\right)=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{X}_1+\boldsymbol{\beta}_2\boldsymbol{X}_2+\boldsymbol{\beta}_3\left\{\boldsymbol{X}_1\times\boldsymbol{X}_2\right\}.$$

[2] L'acronimo INUS è stato creato da Mackie (Braumoeller, 2003) come definizione di un particolare tipo di relazione causale, riferendosi a "an insufficient but necessary part of a condition which is itself unnecessary but sufficient for the result".

Questo modo di precedere obbliga il ricercatore a mantenere le interazioni fra variabili ad un livello piuttosto elementare, al massimo si considerano interazioni del primo o del secondo ordine sia per ragioni tecniche (sparsità dei dati, potenza dei test) che teoriche (il principio di parsimonia). Come conseguenza, si inseririscono nel modello solo gli effetti principali nonostante siano gli effetti di interazione che dovrebbero essere più utili ai fini predittivi o per isolare gruppi di osservazioni, soprattutto in contesti applicativi quali quelli delle indagini in ambito sociale.

2.1 Il Boolean logit

Un metodo che tiene conto delle relazioni di complessità causale è il "Boolean logit" proposto da Braumoeller (2003). Tale metodo consente di stimare l'influenza sulla variabile *Y* esercitata dall'interazione fra le variabili indipendenti del modello. Viene postulato che la risposta binaria *Y* sia prodotta da una combinazione *Booleana* o *logica* di alcune *condizioni* $A_1, \ldots, A_k, \ldots$, del tipo, ad esempio:

$$A_1 \text{ and } (A_2 \text{ or } A_3) \rightarrow Pr(Y=1) = \pi = Pr(A_1) \times Pr(A_2 \cup A_3)$$

la probabilità che si verifichi ciascuna condizione

$$Pr(A_K) = p_k$$

viene espressa per mezzo di un modello logit o probit (Braumoeller 2003):

$$\boldsymbol{p_k} = \frac{\exp(\boldsymbol{\beta_k X})}{1 + \exp(\boldsymbol{\beta_k X})}$$

dove k sta ad indicare che ogni "condizione" dipende dalle sue variabili esplicative $X = \{X_j\}$ attraverso i parametri β_k ad esse associati. La stessa X_j può essere inserita in diversi p_k senza indurre multicollinearità nel modello (ovviamente, se la "condizione" è solo una il Boolean logit si riduce allo standard logit). Il Boolean logit trova impiego nella soluzione di problemi statistici di stima in presenza di situazioni di complessità causale. Il ricercatore deve postulare un modello di causazione per π, quindi, π viene espressa come funzione di un insieme di variabili esplicative e relativi parametri attraverso le diverse probabilità p_k. Ad esempio, se si è assunto che

$$\pi = Pr(A_1) \times Pr(A_2)$$
$$\text{logit}(p_1) = \mathbf{x}_1' \boldsymbol{\beta}_1 \quad e \quad \text{logit}(p_2) = \mathbf{x}_2' \boldsymbol{\beta}_2$$

il modello assumerà la forma

$$\pi_i = p_{1i} \times p_{2i}$$

e la verosimiglianza ad esso associata sarà:

$$\boldsymbol{Lik}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{i=1}^{n} (\boldsymbol{p}_{1i} \times \boldsymbol{p}_{2i})^{y_i} (1 - \boldsymbol{p}_{1i} \times \boldsymbol{p}_{2i})^{1-y_i}$$

Quindi, una volta che l'esito di un evento viene "spiegato" nel linguaggio della *causazione complessa*, le ipotesi conseguenti potranno essere espresse in termini di calcolo probabilistico[3].

3. I dati

I dati analizzati provengono da un'indagine CATI realizzata nel novembre del 2003 dall'Università degli Studi di Cagliari. Sono stati intervistati 1.112 laureati dell'Ateneo che hanno conseguito il loro titolo negli anni 1999 e 2000. Al termine dell'indagine gli intervistati sono stati classificati rispetto al loro *status* occupazionale come *occupati* (823), *disoccupati* (108), *ancora impegnati nella formazione* (137), *in cerca di prima occupazione* (42) e *inoccupati* (2). L'insieme degli occupati è stato poi distinto in due sotto-gruppi, quello di chi ha *iniziato a lavorare dopo la laurea* (756) e di chi aveva già un'*occupazione prima della laurea* (67). Il totale degli intervistati che non lavorano (disoccupati + in cerca di prima occupazione) è di 150.

Per le finalità del presente lavoro si è deciso di fissare i seguenti criteri di eleggibilità:

condizione professionale di:

- occupato;
- disoccupato o in cerca di prima occupazione;

per gli occupati:

- aver iniziato a lavorare dopo il conseguimento della laurea;
- non aver impiegato più di 36 mesi per trovare l'impiego.

Sulla base di questi criteri, per le successive analisi in cui verranno studiati i predittori dell'evento "*Y*" *lavorare/non-lavorare*, sono state prese in considerazione 837 osservazioni delle quali 687 riferite ad *occupati* (*Y*=1) e 150 a *non occupati* (*Y*=0).

Nell'indagine, sono state raccolte numerose informazioni sulle caratteristiche demo-sociali degli intervistati e sono state registrate le loro valutazioni sui percorsi formativi e sulle eventuali esperienze lavorative e i relativi tempi di inserimento (Porcu-Tedesco, 2004; Porcu-Puggioni, 2004). Con riferimento a queste informazioni sono state condotte delle analisi esplorative che hanno portato ad isolare un insieme di variabili da noi ritenute particolarmente informative ai fini di questo lavoro:

- •sesso
- •tipo di diploma
- •voto di diploma
- •tipo di laurea[4]
- •età alla laurea
- •voto di laurea
- •frequenza corsi post-lauream
- •scolarità dei genitori.

[3] Un metodo alternativo per modellare l'interazione fra le variabili, anch'esso di recente proposta, è quello della *Logic Regression* (Ruczinski *et al.*, 2003).

Tabella 1. *Misure di associazione per coppie delle variabili considerate*

Variabili	Sesso	Voto di diploma	Tipo di diploma	Tipo di laurea	Voto di laurea	Età alla laurea	Corsi post-lauream
Voto di diploma[1]	0,547						
Tipo di diploma[2]	0,194	6,302					
Tipo di laurea[3]	17,443	28,901	7,979				
Voto di laurea[4]	148,988	40,538	1,338	6,629			
Età alla laurea[5]	10,197	14,652	6,197	69,207	14,660		
Corsi post-lauream[6]	11,956	0,534	3,163	4,059	18,758	3,408	
Anni scuola genitori[7]	16,079	0,675	27,888	0,013	21,263	9,617	0,013

[1] ≤ 90/100, > 90/100; [2] Liceo classico e scientifico, altro tipo di scuola secondaria;
[3] Corso di laurea scientifico, altro tipo di facoltà; [4] < 108/110 , > 108/110; [5] ≤ 26 anni, > 26 anni;
[6] Frequenza, non frequenza; [7] < 26 anni di scuola, ≥ 26 anni di scuola.
Numero di osservazioni valide 837 per tutti i caratteri e 815 per il carattere "Anni di scuola dei genitori"

Nelle successive analisi verrà esclusa la variabile "scolarità dei genitori" in quanto il suo impiego isolato (ad esempio, dalla professione) appare, allo stato delle nostre ricerche, non assumere adeguatamente il ruolo di proxy dell'estrazione socio-economica del laureato.

Nella Tabella 1, nella quale le variabili sono state dicotomizzate per motivi di coerenza con le applicazioni che verranno di seguito presentate, sono riportate alcune misure relative ai legami associativi fra le coppie di variabili considerate. Dall'esame dei valori ottenuti della statistica X^2 emerge, con tutta evidenza, che si è di fronte a relazioni funzionali complesse, per cui un simile approccio può fornire solo delle indicazioni di massima. In altri termini, pur prendendo atto della significatività statistica di alcune associazioni, da tali risultanze non è possibile cogliere le eventuali interrelazioni che possono esistere fra le diverse variabili in quanto ciascuna associazione così osservata non tiene conto dei valori assunti dalle restanti.

4. La scelta dei gruppi di variabili per lo studio delle interazioni

Il problema della scelta dei criteri con cui formare gruppi di predittori per costruire il modello di regressione Booleano rappresenta, verosimilmente, l'aspetto "debole" di questa metodologia. Evidentemente, una buona scelta di raggruppamento può essere realizzata sulla base di convinzioni od opinioni del ricercatore, basate sulla propria

[4] Le tipologie di laurea sono state classificate nel modo seguente: Gruppo Economico-Giuridico-Sociale (EGS): Economia, Giurisprudenza e Scienze Politiche. Gruppo Scientifico-Tecnico (SCT): Ingegneria, Fisica, Matematica, Chimica e Geologia. Gruppo Scienze della Vita-Salute (SVS): Medicina, Biologia, Scienze Naturali, Farmacia. Gruppo Umanistico-Educazione-Comportamento (UEC): Lettere, Lingue e Scienze della Formazione (Pedagogia e Psicologia).

esperienza nei riguardi dell'oggetto della ricerca. Tuttavia l'indeterminatezza o la soggettività di questo approccio rischia di indebolire il modello finale. Inoltre, i modelli booleani sono particolarmente avidi di risorse computazionali, per cui calcolare numerosi modelli per poi confrontarli potrebbe richiedere un tempo eccessivo.

La proposta che si fa in questo lavoro è di operare la scelta dei gruppi di predittori seguendo i risultati di una procedura di analisi esplorativa dei dati basata sulla segmentazione binaria che, come è noto, è in grado di fornire informazioni sull'importanza dell'influenza sulla variabile risposta esercitata dai diversi predittori e sull'esistenza di eventuali interazioni tra essi. Anche in questo caso, tuttavia, i risultati che si possono ottenere varieranno in funzione del tipo di segmentazione scelto (in particolare il tipo di funzione criterio adottata), ma si ritiene che la scelta di un metodo di segmentazione binario basato sulla funzione criterio del rapporto di verosimiglianza, possa rappresentare una scelta di buon senso in quanto libera il ricercatore dalla necessità di scegliere una misura di distanza (Tedesco, 2002). In aggiunta, la segmentazione è pur sempre una rappresentazione della complessità causale delle variabili osservate in riferimento al campione utilizzato e, quindi, si ritiene opportuno in sede di costruzione dei gruppi Booleani, non perdere questa importante informazione. Sostanzialmente non si vuole far prevalere l'idea astratta del ricercatore rispetto alle informazioni che il campione può fornire. Il software impiegato per la segmentazione è RECPAM[5], il criterio scelto è, come accennato, quello della massimizzazione del rapporto di verosimiglianza del logit *lavorare/non-lavorare* rispetto a tutte le combinazioni, a due a due, tra le diverse modalità delle covariate, mentre si è scelto di avere per ogni nodo/foglia almeno 40 soggetti di cui almeno 10 occupati.

La suddivisione è avvenuta ad un livello α del 5%, piuttosto restrittivo, al fine di avere un albero sintetico e non troppo articolato. L'obiettivo, infatti, è quello di esplorare i dati per la costruzione dei gruppi booleani.

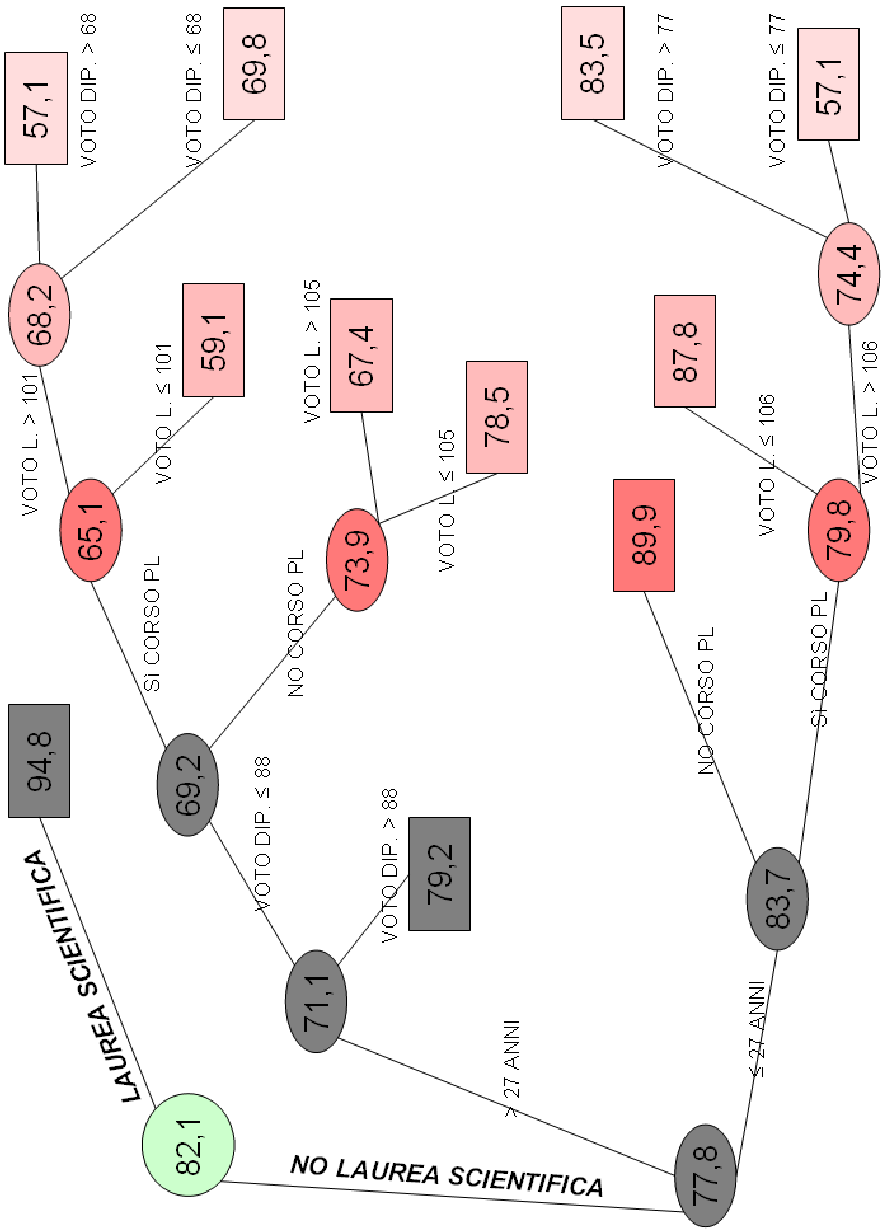
Le variabili inserite nella segmentazione, oltre alla dicotomica lavora/non lavora, sono: Frequenza corsi *post-lauream* (Sì/No), *sesso* (M/F), *tipo di diploma* (Liceo/Non Liceo), *tipo di laurea* (SCT, EGS, SVS, UEC), *voto di diploma* (in centesimi), *voto di laurea* (110-mi), *età alla laurea* (in anni compiuti). Si è deciso di lasciare le ultime tre variabili nella loro scala di misura continua al fine di ottenere soglie di suddivisione determinate direttamente dai dati campionari.

L'albero ottenuto mostra risultati interessanti. Innanzitutto vi è una forte asimmetria, dovuta al fatto che i laureati in discipline del gruppo scientifico presentano un elevato tasso di occupazione (94,8%) e non si suddividono più. Ciò significa che per questi laureati il tipo di laurea è l'unico e fondamentale fattore di occupazione, senza distinzioni particolari tra i due sessi, per voto o tipo di diploma e per voto o età alla laurea. Al contrario, per tutti gli altri laureati, il tasso di occupazione appare

[5] RECPAM è una macro utilizzabile in ambiente SAS, realizzata da F. Carinci (2001) su idea di A. Ciampi (1991). Cfr., ad es., Tedesco (2002).

decisamente inferiore (77,8%), ma raggiunge valori piuttosto alti per particolari profili di soggetti. Nella fattispecie, tra i laureati giovani (< 27 anni), il non frequentare corsi *post-lauream* (89,9%), mentre tra i laureati meno giovani (≥ 27 anni) appare piuttosto penalizzante la bassa votazione al diploma, la frequenza di corsi *post-lauream* e la bassa votazione alla laurea.

Figura 1. *Albero di segmentazione*



Appare evidente, quindi, che tra i laureati in materie non del gruppo scientifico, conta in via prioritaria la giovane età alla laurea, requisito sempre molto apprezzato da chi offre lavoro, piuttosto che la formazione aggiuntiva, evidentemente perché questa è fornita direttamente dalle aziende, almeno quelle di medie–grandi dimensioni[6]. Interessante appare, poi, l'importanza del voto di diploma, che interviene più volte nella segmentazione, rispetto alla totale assenza del tipo di diploma e del sesso.

Ciò sembra suggerire che la "qualità" di un laureato non è solo il prodotto del processo degli studi universitari, ma affonda le radici nella formazione secondaria che, se fatta bene, forma un individuo in maniera efficace. Sempre tra i laureati in discipline non scientifiche, giovani e che hanno un titolo *post-lauream*, colpisce la non utilità del voto di laurea (l'87,8% è occupato con un voto laurea ≤ 106), mentre tra i migliori (voto laurea > 106) sembra influire la votazione al diploma.

In conclusione, tenendo conto anche dei valori dell'indice GPI (Tabella 2)[7], appare evidente che il voto di diploma e l'età alla laurea hanno un effetto congiunto sulla variabile risposta, ma solo tra i laureati in discipline non scientifiche.

Tabella 2. *Valori del GPI*

Covariate	GPI
Voto diploma	100
Tipo laurea	96
Età laurea	90
Voto laurea	60
Corso PL	55
Tipo diploma	28
Sesso	27

[6] A tal proposito occorre ricordare come è apparso evidente (Porcu-Tedesco, 2004) che sovente la formazione PL sia più una forma di prolungamento del "parcheggio" in attesa di un'occupazione, piuttosto che la reale esigenza di incrementare le proprie competenze.

[7] Si ricorda che il GPI (Global Predictive Index) è un indice che misura il grado di predittività di una covariata sulla base della somma degli incrementi nel valore della LRS per ogni covariata ad ogni nodo, rispetto al valore della LRS senza quel predittore; in buona sostanza è una misura del guadagno di informazione dovuto all'*i-esimo* predittore. Determinate tutte le *i* somme (una per ciascuna covariata), si pone uguale a 100 quella maggiore e, quindi, le altre sono calcolate in rapporto a questa. Per tale motivo la covariata che ha il potere predittivo più grande, ha un valore del GPI pari a 100 (Ciampi, 1991).

5. Modellare l'evento *Y* "lavorare/non-lavorare"

Per modellare l'evento lavorare (Y=1) vs non-lavorare (Y=0), sono state prese in esame, anche in considerazione dei risultati dell'analisi di segmentazione, le seguenti variabili dicotomiche 1/0 (1=Sì):

- •sesso maschile (SEXM);
- •diploma di liceo classico o scientifico (LICCS);
- •voto di diploma ≥90/100 (DIP90);
- •laurea del Gruppo Scientifico-Tecnico (SCIEN);
- •laurea entro i 26 anni (LAU26);
- •voto di laurea ≥108 (VOTOHIGH);
- •ha fatto formazione post-lauream (CORPOST).

Di seguito verranno presentati, dapprima i risultati dell'adattamento di uno standard logit, successivamente quelli dell'applicazione del Boolean logit.

5.1 Adattamento di un logit standard

I risultati dell'applicazione sono riportati nella Tabella 3. Dal suo esame (valori negativi di $\hat{\beta}$ indicano una minore probabilità per l'evento) si può rilevare che le sole variabili che paiono esercitare un effetto significativo (α = 0,05) sulla risposta siano DIP90, SCIEN, LAU26 e CORPOST.

Tabella 3. *Stime puntuali ($\hat{\beta}$) e corrispondenti z*-score *($z = \hat{\beta} / SE(\hat{\beta})$) per il modello logit standard di base*

Covariate	$\hat{\beta}$	**|z|-score**
SEXM	0,1967	0,916
LICCS	-0,2801	1,417
DIP90	0,5453	**2,068**
SCIEN	1,4855	**4,315**
LAU26	0,6134	**2,875**
VOTOHIGH	-0,2575	1,301
CORPOST	-0,4493	**2,309**
Log*Lik*	-362,937	

Adattando un modello che teneva conto delle interazioni del primo ordine fra le variabili, è stato osservato come nessuna di esse influenzi significativamente la risposta.

Tabella 4. *Stime puntuali ($\hat{\boldsymbol{\beta}}$) e corrispondenti z-*score *($z = \hat{\boldsymbol{\beta}} / \boldsymbol{SE}(\hat{\boldsymbol{\beta}})$) per alcuni modelli logit standard.*

Covariate	Mod. Base		Senza SCIEN		Con SEXM × SCIEN	
	$\hat{\boldsymbol{\beta}}$	|z|-score	$\hat{\boldsymbol{\beta}}$	|z|-score	$\hat{\boldsymbol{\beta}}$	|z|-score
SEXM	0,1967	0,916	0,5464	0,916	0,1780	0,787
LICCS	-0,2801	1,417	-0,2030	1,417	-0,2770	1,398
DIP90	0,5453	2,068	0,8068	2,068	0,5454	2,069
SCIEN	1,4855	4,315	–	–	1,3758	2,544
LAU26	0,6134	2,875	0,5251	2,470	0,6128	2,873
VOTOHIGH	-0,2575	1,301	-0,3212	1,633	-0,2579	1,303
CORPOST	-0,4493	2,309	-0,5115	2,668	-0,4445	2,274
SEXM × SCIEN	–	–	–	–	0,1778	0,256
log*Lik*	-362,937		-374,985		-362,904	

Come è noto, relazioni di tipo causale, come quella appena descritta, pongono al ricercatore dei problemi di interpretazione degli effetti esercitati dal complesso dei predittori sulla risposta *Y*. Ad esempio, se dal modello base della Tabella 3 viene escluso il predittore SCIEN si osserva come la variabile SEXM acquisti un significativo potere predittivo (anche se, in termini di log*Lik*, il modello è meno soddisfacente). Tuttavia, adattando un altro modello che comprende il termine di interazione fra SCIEN e SEXM si riscontra la non significatività statistica dello stesso.

5.2 Adattamento di un Boolean logit

Per l'adattamento del modello Boolean logit[8] sono state considerate le stesse variabili prese in esame per lo standard logit (SEXM, LICCS, DIP90, SCIEN, LAU26, VOTOHIGH, CORPOST). Come detto nel § 2.1, per procedere all'adattamento di un modello Boolean logit è necessario ipotizzare preliminarmente alcune *condizioni*; anche sulla base delle indicazioni date dai risultati della segmentazione binaria (§ 4) tali condizioni sono state definite come:

- A_1 = "Possesso di requisiti *vincenti* per il mondo del lavoro"
- A_2 = "Possesso di alcuni fattori caratterizzanti la formazione"

A_1 è definita da un insieme di covariate riferite a ciò che caratterizza in maniera più incisiva chi si candida ad entrare nel mondo del lavoro e, cioè, l'*età* e le *competenze* possedute: LAU26 e SCIEN.

[8] Per il calcolo dei parametri è stata utilizzata la libreria "Boolean" in ambiente R (http://www.R-project.org).

A_2 è definita da un insieme di covariate riferite al profilo formativo del laureato alle quali si aggiunge la variabile "sesso": SEXM, DIP90, LICCS, VOTOHIGH e CORPOST.

La probabilità di essere occupato, $Pr(Y=1) = \pi$ viene modellata come interazione fra A_1 e A_2, cioè:

$$\pi = Pr(A_1) \times Pr(A_2)$$

Le condizioni A_1 e A_2 vengono espresse come funzioni additive delle esplicative in esame:

- A_1 = LAU26 + SCIEN
- A_2 = SEXM + DIP90 + LICCS + VOTOHIGH + CORPOST

Come si può ricavare dalla Tabella 5 i risultati ottenuti sono simili a quello dello standard logit in termini di log verosimiglianza e di stima dei parametri. Tuttavia, i modelli che li hanno prodotti sono sostanzialmente differenti. Infatti, nel modello logit standard, nessuno dei termini di interazione ha mostrato di esercitare effetti significativi sulla variabile risposta, risultato questo che implica, da un punto di vista sostanziale, che ciascuna variabile influenza la probabilità di conseguire un'occupazione indipendentemente dalle altre variabili. Nel modello Boolean logit, invece, la risposta Y è prodotta dall'interazione fra i vettori di covariate. Ciò implica che l'essere o meno occupato dipende *congiuntamente* da A_1 e A_2: i parametri assumono, quindi, un "tacito" significato di interazione. Come si può vedere i parametri DIP90 e CORPOST mostrano ora di non influenzare in maniera significativa la risposta Y e ciò sta a significare che *interagendo* con le altre queste covariate perdono il loro potere predittivo.

Tabella 5. *Stime puntuali ($\hat{\boldsymbol{\beta}}$) e corrispondenti z-*score *($z = \hat{\boldsymbol{\beta}} / \boldsymbol{SE}(\hat{\boldsymbol{\beta}})$) per il modello logit standard e per due modelli Boolean*

Covariate	Standard		Boolean 1		Boolean 2	
	$\hat{\boldsymbol{\beta}}$	|z|-score	$\hat{\boldsymbol{\beta}}$	|z|-score	$\hat{\boldsymbol{\beta}}$	|z|-score
LAU26	0,6134	2,875	1,0330	2,996	1,0363	1,904
SCIEN	1,4855	4,315	2,2442	2,813	2,2573	1,192
LICCS	-0,2801	1,417	-1,4221	1,094	0,0070	0,008
SEXM	0,1967	0,916	0,5035	0,796	0,4973	0,501
DIP90	0,5453	2,068	1,3693	1,715	1,3640	1,326
VOTOHIGH	-0,2575	1,301	-1,2291	1,251	-1,2127	0,527
CORPOST	-0,4493	2,309	-1,6930	1,066	-1,6632	0,408
LICCS	–	–	–	–	-1,4294	0,904
log*Lik*	-362,937		-360,640		-360,639	

Come detto, il Boolean logit permette di adattare modelli in cui la stessa covariata viene inserita in più di una "condizione". Ad esempio, la variabile *provenienza dal liceo classico o scientifico* potrebbe essere inserita tanto nella condizione A_1 che nella condizione A_2. I risultati dell'adattamento di questi modello sono riportati nella Tabella 5 nelle colonne intestate "Boolean 2"; nonostante la non significatività statistica della variabile LICCS è interessante osservare come essa agisca in direzioni opposte in A_1 e in A_2.

6. Conclusioni

L'impiego dello standard logit per modellare la probabilità di un evento dicotomico come effetto di un rapporto causale di dipendenza rispetto a un insieme di esplicative offre al ricercatore notevoli vantaggi. Essi risiedono principalmente nell'interpretazione sostantiva dei parametri stimati; la loro lettura in termini di log-*odds ratio* permette di valutare direttamente l'influenza di ogni parametro sulla variabile risposta "controllando" il livello delle altre covariate prese in esame. In un contesto come quello della modellazione della probabilità di conseguire o meno un'occupazione per i laureati considerato in questo studio, lo standard logit consente di evidenziare l'esistenza di alcuni fattori frenanti che intervengono abbassando la probabilità dell'evento occupazione. Fra essi, appaiono di un certo interesse quelli relativi al possesso di un voto alto alla laurea e all'aver frequentato dei corsi di specializzazione dopo il conseguimento del titolo; verosimilmente, essi possono essere visti come fattori che influiscono sull'età con cui ci si presenta sul mercato del lavoro innalzandola e rendendo in questo modo meno competitivo lo stesso laureato. Altri fattori, si è visto, agiscono in direzione contraria (contribuiscono ad aumentare la probabilità dell'evento) e fra essi si distinguono quelli riferiti al possesso di una laurea di tipo scientifico-tecnico e, non inaspettatamente, quelli relativi alla giovane età del laureato.

Sempre in termini sostantivi, però, non va dimenticato che un modello logit standard quale quello adattato, non tenendo in considerazione le relazioni esistenti fra le covariate prese in esame, implica una forma di dipendenza causale additiva che non permette di "catturare" appieno la complessità del fenomeno.

Il Boolean logit, non deve essere inteso come alternativo (e, tantomeno, superiore) al modello logit standard. Il vantaggio che offre rispetto a quest'ultimo risiede nel fatto che esso permette al ricercatore di adattare dei modelli in cui viene preso in considerazione un rapporto di causazione *complessa*. I meccanismi di causazione complessa permettono (Braumoeller, 2003) di migliorare il potere predittivo dei modelli adattati per spiegare un determinato evento risposta.

Il principale limite di un modello Boolean risiede nelle scelte soggettive che si operano per la definizione degli *statements* (condizioni) Booleani, anche se la possibilità di ricorrere a criteri basati sulla verosimiglianza mitiga questa soggettività. In questo senso, ricorrere a metodi di segmentazione binaria del tipo di quelli adottati in questo lavoro, può realmente aiutare il ricercatore ad operare scelte meno soggettive e più coerenti con le informazioni che il campione fornisce. Altro notevole limite è quello che deriva dalla non interpretabilità dei parametri stimati in termini di log-*odds ratio* rispetto alla risposta modellata ed, infine, non va sottovalutato che l'algoritmo di stima è "avido" sia di dati (data consumpting) sia di tempo computazionale.

Tuttavia, tenendo in considerazione i risultati ottenuti in questa e in altre applicazioni (Muggeo-Porcu, 2004), si può concludere che il Boolean logit si candida per essere un utile strumento per implementare analisi di sensibilità di altri modelli per risposte causali e quindi impiegabile per rafforzare le evidenze emerse sul significato sostantivo delle esplicative prese in esame.

Riferimenti bibliografici

AGRESTI A. (2002) *Categorical Data Analysis*, Wiley-Interscience, Hoboken NJ.

AKAIKE H. (1973), Information theory and an extension of the maximum likelihood principle, in *Proceedings of the Second International Symposium on Information Theory*, B.N. Petrov & Csaki, eds. Akademiai Kiado, Budapest: 267-281.

BRAUMOELLER B.F. (2003), Causal Complexity and the study of politics, *Political Analysis*, 11: 209-233.

CARINCI F., PELLEGRINI F. (2001), *RECPAM/SAS (Recursive Partitioning and Amalgamation): a statistical tool for criterion-driven data-mining*, Technical Report, in http://med.monash.edu.au\publichealt.

CHIANDOTTO B. (2004), "La situazione occupazionale dei laureati: dall'indagine alla pianificazione degli interventi sui percorsi formativi", in M. CIVARDI (a cura di), *Transizione Università-Lavoro: la definizione delle competenze*, vol. 4, CLEUP, Padova: 1-18.

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

CIVARDI M., ZAVARRONE E. (2004), "Proposta di un modello generatore delle competenze acquisite attraverso la formazione universitaria", in: E. AURELI CUTILLO (a cura di), *Strategie metodologiche per lo studio della transizione Università-Lavoro*, vol. 5, CLEUP, Padova: 141-152.

FROSINI B.V. (2004), Causality and Causal Models, in *Atti della XLII Riunione della Società Italiana di Statistica*, v. 1, Bari: 3-32.

GRANOVETTER M. (1974), *Getting a Job: a Study of Contacts and Careers*, Harvard University Press, Cambridge MA.

HOSMER D.W., LEMESHOW S. (1989) *Applied Logistic Regression*, John Wiley & Sons, New York.

MUGGEO V, PORCU M.. (2004), Factors that Cause University Students to Drop Out. An Alternative Modelling of Interaction Terms in Logistic Regression Models, in *Atti della XLII Riunione della Società Italiana di Statistica*, v. 2, Bari: 511-514.

PORCU M., PUGGIONI G. (2004), "L'esportazione del capitale umano: prima valutazione del fenomeno per i laureati dell'Ateneo di Cagliari" (in corso di stampa).

PORCU M., TEDESCO N. (2004), "Dall'Università al Lavoro: analisi dei tempi di passaggio dei laureati dell'Ateneo di Cagliari", in: E. AURELI CUTILLO (a cura di), *Strategie metodologiche per lo studio della transizione Università-Lavoro*, vol. 5, CLEUP, Padova: 281-295.

REYNERI E. (2002), Sociologia del Mercato del Lavoro, il Mulino, Bologna.

RUCZINSKI I., KOOPERBERG C., LEBLANC M. (2003), Logic Regression, *Journal of Computational and Graphical Statistics*, 12:. 475-511.

R DEVELOPMENT CORE TEAM (2003), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna; http://R-project.org.

TEDESCO N. (2002), "Analisi di segmentazione di una coorte di immatricolati dell'Università di Cagliari", in: G. PUGGIONI (a cura di), *Modelli e metodi per l'analisi dei rischi sociali e sanitari*, vol. 2, CLEUP, Padova: 141-160.

Determinants of occupational placement of graduates. An analysis of interactions

Summary. *In the analysis of occupational placement of graduates it is interesting to define the role paid by some covariates assembled to predict the dichotomous event occupation/not-occupation. It is well known that these covariates influence the response not only singularly but also jointly. This work propose an evaluation of this joint effect by means of a recently introduced technique named Boolean logit. An exploratory binary segmentation is also presented to support the analysis.*

Keywords*. Occupational placement, determinants, segmentation, Boolean regression, logit.*