L'effetto degli studi universitari sull'occupazione: un'applicazione dell'approccio degli "strati principali" all'analisi causale

Leonardo Grilli, Fabrizia Mealli[1]
Dipartimento di Statistica "G. Parenti" - Università degli Studi di Firenze

Riassunto. Il lavoro mostra come valutare l'efficacia di due corsi di laurea rispetto allo status occupazionale usando l'approccio degli "strati principali" all'inferenza causale. L'applicazione riguarda la coorte 1992 degli iscritti ai corsi di laurea in Economia e Commercio e in Scienze Politiche presso l'Università di Firenze. L'articolo illustra un uso innovativo dei limiti non parametrici nell'ambito degli "strati principali", esaminando il ruolo di alcune assunzioni in ordine alla riduzione dell'incertezza. La seconda fase dell'analisi si basa su un modello parametrico adattato con la massima verosimiglianza. In quel contesto si discutono alcune rilevanti questioni relative alla modellizzazione, delineando una strategia generale per la specificazione del modello.
Parole chiave: effetti causali, efficacia, risultati potenziali, strati principali.

1. Introduzione

Le tradizionali analisi dell'effetto dei corsi di laurea sullo stato occupazionale (chiamate anche analisi di *efficacia esterna*) sono condotte soltanto sulla base degli studenti laureati, trascurando il fatto che l'insieme degli studenti che sono in grado di laurearsi in un dato corso di laurea è, in generale, diverso dall'insieme di studenti che sono in grado di laurearsi in un altro corso di laurea. In alter parole, due diversi corsi di laurea possono selezionare diverse tipologie di studenti, con differenti attitudini, capacità e prospettive lavorative. Un'analisi dello stato occupazionale basata soltanto sugli studenti laureati mescola l'effetto "diretto" del corso di laurea sull'occupazione con l'effetto "indiretto" che passa attraverso il raggiungimento della laurea.

[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 multidimensionale delle determinanti", cofinanziato dal MIUR (Coordinatore nazionale Luigi Fabbris, coordinatore del gruppo di Firenze Bruno Chiandotto). La nota è frutto del lavoro congiunto dei due autori ed è stata redatta da F. Mealli per quanto concerne i paragrafi 1, 2, 3 e 7 e da L. Grilli per i paragrafi 4, 5 e 6.

Da un punto di vista di politica dell'istruzione, scomporre i due effetti è molto importante. Per esempio, se esiste un effetto diretto sull'occupazione, allora il corso di laurea con minore efficacia dovrebbe modificare i propri contenuti al fine di soddisfare le richieste del mercato del lavoro. Se invece il successo sul mercato del lavoro di un corso di laurea è dovuto soltanto ai diversi criteri di selezione (ad es., un corso di studi risulta più difficile di un altro e quindi seleziona studenti migliori), allora il problema diventa un tema di politica dell'istruzione (nell'esempio posto, se sia o meno auspicabile per la società il permettere che si laureino studenti con minori capacità o permettere l'esistenza di corsi di laurea con diversi livelli di difficoltà).

Al fine di studiare l'effetto diretto dei corsi di laurea sull'occupazione, evitando distorsioni dovute ad un diverso processo di selezione/laurea, è necessario pianificare uno studio congiunto dei processi che conducono alla laurea e al lavoro. A questo proposito un approccio appropriato può essere quello della stratificazione principale (Frankgakis & Rubin, 2002), un importante sviluppo dell'approccio dei risultati potenziali all'inferenza causale (Rubin, 1974). L'approccio della stratificazione principale è stato recentemente utilizzato in Barnard *et al.* (2003) per l'analisi di un complesso esperimento casualizzato nell'ambito dell'istruzione. Nella presente applicazione, la *variabile di trattamento* è la variabile indicatrice di un corso di laurea (vs. un altro), mentre la *variabile intermedia* (*post-trattamento*) che definisce gli strati principali è la laurea (laureato/non laureato). Il punto cruciale è che, se uno studente non si laurea, la *variabile risultato*, che è lo stato occupazionale, non è definita se l'obiettivo è quello di studiare l'efficacia dei corsi di laurea rispetto al mercato del lavoro. Questo è un esempio della cosiddetta *censura per morte*, discussa in Zhang & Rubin (2004) nel caso ipotetico di un esperimento casualizzato riguardante due programmi educativi nella scuola superiore, dove la variabile intermedia è l'abbandono e la variabile risultato è il punteggio su un test finale.

In questo lavoro, l'approccio di Zhang & Rubin (2004) è applicato ad un caso di studio reale, che differisce dal loro esempio in molti aspetti: (i) il trattamento non è casualizzato; (ii) i due trattamenti sono sullo stesso piano, ovvero non vi è un trattamento *attivo* da confrontare con un trattamento di controllo; (iii) la variabile risultato è binaria e soggetta a non risposta; (iv) alcune covariate rilevanti sono disponibili.

La presente analisi è limitata al confronto di due soli corsi di laurea. L'estensione a più corsi comporta alcune difficoltà tecniche, ma l'approccio concettuale rimarrebbe essenzialmente inalterato.

I due corsi di laurea messi a confronto sono Economia e Commercio e Scienze Politiche, che sono piuttosto simili, almeno in Italia, rispetto ai contenuti dei corsi e alle opportunità lavorative. Alla luce di tali similarità, per un dato livello delle covariate osservate, la scelta di iscriversi ad un determinato corso è verosimilmente poco associata a caratteristiche osservabili che potenzialmente influiscono anche sulle chance di laurea e il successivo status occupazionale; quindi l'assunzione di ignora-

bilità dell'assegnazione al trattamento discussa più aventi sembra ragionevole. Si noti che la variabile binaria che identifica il corso di laurea a cui uno studente è iscritto sarà chiamata *indicatore di trattamento*, in modo da conformarsi all'usuale linguaggio dell'analisi causale, sebbene nella presente applicazione non vi sia alcuna randomizzazione e, inoltre, I due corsi di laurea sono sullo stesso piano, non essendoci un trattamento attivo vs. uno di controllo.

2. I dati

Un'analisi congiunta dei processi che conducono alla laurea e al successo sul mercato del lavoro richiede di unire due archivi di dati: un database amministrativo riguardante una (o più) coorte di studenti e i dati relative ad un indagine che rilevi lo stato occupazionale sui laureati di quella stessa coorte. In questo lavoro, riguardante corsi di laurea offerti dall'Università di Firenze le due fonti di dati sono:

- L'archivio amministrativo dalla coorte di matricole iscritte nel 1992 ad uno dei due corsi di laurea Economia e Commercio e Scienze Politiche;
- Tre indagini esaustive sullo stato occupazionale dei laureati negli anni 1998, 1999 e 2000.

I dataset sono stati uniti attraverso il numero di matricola. Le matricole della coorte esaminata sono 1941: 1068 iscritte a Economia e Commercio e 873 a Scienze Politiche. La scelta della coorte è stata motivata dalla disponibilità di dati da indagine per i laureati dal 1998 al 2000: la coorte 1992 è parsa la scelta migliore poiché solo 21 studenti di tale coorte si sono laureati prima del 1998, mentre la maggior parte degli studenti che non hanno abbandonato si sono laureati proprio nel triennio 1998-2000.

La carriera accademica degli studenti alla fine dell'anno 2000 è riassunta nella Tabella 1. Per gli studenti ancora iscritti, così come per gli abbandoni, non disponiamo di dati relativi al loro status occupazionale. Quindi, ai fini della presente analisi, la laurea è definita come "laurea entro nove anni dall'iscrizione. Questa restrizione nella definizione dello status di laureato non pare eccessivamente dannosa per l'analisi, considerato anche il fatto che il raggiungimento della laurea dopo nove anni coinvolge molti studenti che hanno già un lavoro regolare svolto durante gli studi.

Lo stato occupazionale al momento dell'intervista per il sottoinsieme di studenti laureati è riportato nella Tabella 2.

Tutti gli studenti intervistati hanno risposto alle domande sullo stato occupazionale. Fatta eccezione per i 21 studenti laureati prima del 1988, che erano fuori target dell'indagine, quasi tutti le interviste mancanti sono dovute a mancati contatti.

La variabile di risultato per l'analisi è la variabile indicatrice del possesso di una lavoro permanente al momento dell'intervista, ovvero da uno a due anni dopo la

Tabella 1. *Carriera accademica degli studenti della coorte 1992 alla fine dell'anno 2000*

Status	*Economia e Commercio*		*Scienze Politiche*	
Abbandonato	545	51.0%	532	60.9%
Laureato	270	25.3%	176	20.2%
Ancora iscritto	253	23.7%	165	18.9%
Totale	1068	100.0%	873	100.0%

Tabella 2. *Status occupazionale al momento dell'intervista*

Status	*Economia e Commercio*		*Scienze Politiche*	
Laureati	270		176	
Intervistati	187	69.3%*	99	56.2%*
Lavoro permanente	97	51.9%**	36	36.4%**

* Intervistati/Laureati **Lavoro permanente/Intervistati

Tabella 3. *Medie campionarie delle covariate per corso di laurea*

Covariate	*Economia e Commercio (n=1068)*	*Scienze Politiche (n=873)*
Femmina	0.41	0.54
Residenza a Firenze	0.23	0.31
Liceo	0.34	0.45
Voto alto	0.37	0.25
Iscrizione con ritardo	0.06	0.22

laurea. La natura permanente del lavoro dipende dal tipo di contratto per I lavoratori dipendenti, mentre deriva da un'autovalutazione per i lavoratori autonomi. I lavori temporanei sono ignorati. I dati amministrativi includono informazioni aggiuntive sull'intera coorte, che sono state utilizzate per definire cinque covariate binarie, successivamente rappresentate per ogni studente dal vettore $\mathbf{x}_i$: genere, residenza (a Firenze vs. altro), diploma di maturità (Liceo vs. altro), Voto di maturità (voto alto 50-60 vs. voto basso 36-49), Iscrizione in ritardo. Nella Tabella 3 sono riportate le medie campionarie delle covariate.

Le covariate hanno distribuzione diversa nei due corsi di laurea, evidenziando come l'assegnazione al trattamento non sia completamente casuale. In particolare, il voto di maturità è più elevato tra gli studenti di Economia e Commercio; la differenza maggiore riguarda l'iscrizione in ritardo, che è piuttosto rara per gli studenti di Economia e Commercio, ma raggiunge il 22% tra gli studenti iscritti a Scienze Poli-

tiche, plausibilmente per la presenza di molti studenti lavoratori che decidono di iscriversi all'Università in un momento distante dall'ottenimento del diploma di maturità.

3. L'approccio degli Strati Principali

Sia n il numero totale di individui oggetto dello studio, ovvero la dimensione della coorte di immatricolati a Economia e Commercio e Scienze Politiche nel 1992. La *variabile di trattamento* Z_i è quindi definita come:

- $Z_i = 1$ se lo studente i è iscritto a Economia e Commercio;
- $Z_i = 0$ se lo studente i è iscritto a Scienze Politiche.

Sia adesso z_i il valore osservato di Z_i e sia $\boldsymbol{z}$ il vettore degli z_i per tutti gli n individui. Nell'approccio dei risultati potenziali ogni variabile post-trattamento dipende dal vettore dei trattamenti assegnati $\boldsymbol{z}$. Tuttavia, nella presente applicazione è ragionevole fare la seguente assunzione, che esclude la possibilità di interazioni tra individui:

Assunzione 1 (SUTVA - Stable Unit Treatment Value Assumption):
Per ogni individuo i ogni variabile post-trattameno dipende da $\boldsymbol{z}$ soltanto attraverso z_i.

Data l'ipotesi SUTVA, ogni variabile post-trattamento ha tante versioni "potenziali" quanti sono i possibili trattamenti (due nel nostro caso). Quindi, le variabili post-trattamento possono essere definite come segue.

La prima variabile post-trattamento è la *variabile intermedia* $S_i(z)$:

- $S_i(z) = 1$ se lo studente i si è laureato entro la fine del 2000 (ovvero entro 9 anni) se iscritto al corso z;
- $S_i(z) = 0$ se lo studente i non si è laureato entro la fine del 2000 (ovvero entro 9 anni) se iscritto al corso z.

Un'altra variabile post-trattamento è l'*indicatore di risposta* $R_i(z)$:

- $R_i(z) = 1$ se lo studente *i ha risposto* alla domanda sullo stato occupazionale se iscritto al corso z e si è laureato;
- $R_i(z) = 0$ se lo studente *i non ha risposto* alla domanda sullo stato occupazionale se iscritto al corso z e si è laureato.

L'ultima variabile post-trattamento è la *variabile risultato* $Y_i(z)$:

- $Y_i(z) = 1$ se lo studente *i,* se iscritto al corso z e si è laureato, *aveva un lavoro permanente* al momento dell'intervista;
- $Y_i(z) = 0$ se lo studente *i,* se iscritto al corso z e si è laureato, *non aveva un lavoro permanente* al momento dell'intervista.

Poiché per ogni individuo la variabile di trattamento assume un solo valore, per ogni variabile post-trattamento solo una delle due versioni potenziali può essere osservata. È quindi utile introdurre la seguente notazione:

$$S_i^{\text{obs}} = S_i(Z_i),\ R_i^{\text{obs}} = R_i(Z_i),\ Y_i^{\text{obs}} = Y_i(Z_i).$$

Essendo binarie sia la variabile di trattamento che la variabile intermedia, è possibile definire 4 strati principali identificati dai valori della variabile latente L_i:

- L_i = '*GG*' (Laureato, Laureato) se $S_i(1)=1$ e $S_i(0)=1$: studenti che sono in grado di laurearsi in entrambi i corsi di laurea;
- L_i = '*GN*' (Laureato, Non laureato) se $S_i(1)=1$ e $S_i(0)=0$: studenti che sono in grado di laurearsi se iscritti ad Economia e Commercio ma che non sono in grado di laurearsi se iscritti a Scienze Politiche;
- L_i = '*NG*' (Non laureato, Laureato) se $S_i(1)=0$ e $S_i(0)=1$: studenti che non sono in grado di laurearsi se iscritti ad Economia e Commercio ma che sono in grado di laurearsi se iscritti a Scienze Politiche;
- L_i = '*NN*' (Non laureato, Non laureato) se $S_i(1)=0$ e $S_i(0)=0$: studenti che non sono in grado di laurearsi in nessuno dei due corsi di laurea.

Si noti che ogni studente appartiene ad un singolo strato, sebbene i dati non siano in grado di rivelare in generale quale sia il suo strato di appartenenza. In altre parole, gli strati principali sono classi latenti e i dati permettono soltanto di stimare le probabilità che un dato studente appartenga ad una certa classe latente. Si noti che gli strati principali sono definiti da coppie di valori potenziali della variabile intermedia, quindi non sono influenzati dal trattamento e possono quindi essere considerati come covariate pre-trattamento non osservabili.

La relazione tra i gruppi osservati, definiti da Z_i e S_i^{obs}, e gli strati principali è descritta nella tabella 4, insieme ai corrispondenti supporti di R_i^{obs} e Y_i^{obs}.

Per le variabili post-trattamento S e Y le proporzioni campionarie nei due gruppi risultano:

- $P_{S,1}$ = 0.253: la proporzione campionaria di laureati tra gli studenti iscritti a Economia (Z_i=1);
- $P_{S,0}$ = 0.202: la proporzione campionaria di laureati tra gli studenti iscritti a Scienze Politiche (Z_i =0);

Tabella 4. *Gruppi osservati e strati principali*

Gruppo osservato $O(Z, S^{obs})$	Z_i	S_i^{obs}	R_i^{obs}	Y_i^{obs}	*Gruppo latente* L_i *(strato principale)*
$O(1,1)$	1	1	$\in \{0,1\}$	$\in \{0,1\}$	*GG* o *GN*
$O(1,0)$	1	0	non definito	non definito	*NG* o *NN*
$O(0,1)$	0	1	$\in \{0,1\}$	$\in \{0,1\}$	*GG* o *NG*
$O(0,0)$	0	0	non definito	non definito	*GN* o *NN*

- $P_{Y,1}$ = 0.516: la proporzione campionaria di individui con un'occupazione permanente tra gli studenti iscritti a Economia (Z_i=1) che si sono laureati (S_i^{obs} = 1) e hanno risposto all'intervista (R_i^{obs} = 1);
- $P_{Y,0}$ = 0.364: la proporzione campionaria di individui con un'occupazione permanente tra gli studenti iscritti a Scienze Politiche (Z_i =0) che si sono laureati (S_i^{obs} = 1) e hanno risposto all'intervista (R_i^{obs} = 1).

Tali proporzioni mostrano come ad Economia il tasso di laurea sia più elevato, così come il tasso di occupazione permanente tra i laureati. L'analisi dovrebbe permettere di valutare se la migliore *performance* di Economia sia da attribuirsi ad un effetto causale positivo.

Poiché l'obiettivo dello studio è quello di valutare l'efficacia dei corsi di laurea rispetto al mercato del lavoro, la variabile di risultato *Y* è definita solo per i laureati. Quindi l'effetto causale $Y_i(1)$-$Y_i(0)$ sull'occupazione è definito in modo appropriato solo per lo strato *GG*, ovvero per gli studenti che sono in grado di laurearsi in entrambi i corsi di studio. In generale, se i dati fossero disponibili, la variabile occupazione potrebbe essere definita per tutti gli studenti iscritti, anche se ai fini di valutare l'effetto del possesso di una laurea o di un'altra sull'occupazione ciò non sarebbe così rilevante.

Nella presente analisi l'effetto causale di principale interesse è l'effetto causale medio per lo strato *GG*. Quando l'interesse è rivolto solo alla popolazione effettivamente osservata, questo effetto è semplicemente la differenza tra le medie dei due risultati potenziali *Y(1)* e *Y(0)* per gli individui che appartengono allo strato *GG*: $\overline{Y}_{GG}(1) - \overline{Y}_{GG}(0)$. Tuttavia, in quanto segue, l'interesse è rivolto al più generale processo di generazione dei dati, e quindi i risultati sono implicitamente riferiti ad una superpopolazione ed espressi in termini di probabilità: la differenza tra le probabilità di avere un lavoro permanente sempre per lo strato *GG*: $E(Y_{GG}(1)) - E(Y_{GG}(0)) = P(Y_{GG}(1) = 1) - P(Y_{GG}(0) = 1)$.

Poiché *Z* non è casualizzato ci potrebbero essere delle variabili di confondimento che influiscono contemporaneamente su *Z* e *S* o su *Z* e *Y*: in questo caso l'effetto di *Z* su *Y* non potrebbe essere interpretato come un effetto causale. Le covariate disponibili $\mathbf{x}_i$, descritte in Tabella 3, possono aiutare ad alleviare questo problema, nel modo contenuto nella seguente assunzione:

Assunzione 2 (Assenza di confondimento dell'assegnazione del trattamento):

$$Z_i \perp S_i(0), S_i(1), Y_i(0), Y_i(1) \mid \mathbf{x}_i .$$

Nella presente applicazione, questa assunzione sarebbe violata se studenti a parità di caratteristiche osservate basassero la loro decisione di iscriversi ad un corso di laurea, piuttosto che ad un altro, su valutazioni circa le proprie *chances* di laurea e di lavoro, confrontando le *chances* di laurea e lavoro di studenti simili anche relati-

vamente a caratteristiche non osservate ma influenti su entrambi i risultati. Tuttavia tale comportamento appare poco plausibile.

I dati sui risultati dei laureati soffrono anche del problema delle mancate risposte: infatti, la variabile *Y* è disponibile solo per coloro che hanno risposto all'intervista. Nel seguito assumiamo che l'informazione su *Y* sia mancante a caso:

Assunzione 3 (Missing at random): $R_i(z) \perp Y_i(z) | \mathbf{x}_i$, $S_i(z)=1$ per ogni z=0,1.

Sotto l'assunzione 3, il meccanismo di risposta è ignorabile, quindi l'analisi si può basare sui dati disponibili (condizionatamente alle variabili osservate). Poiché i dati mancanti sono dovuti principalmente a mancati contatti ed è verosimile che la difficoltà di contatto sia maggiore per le persone che lavorano, una possibile conseguenza potrebbe essere la sottostima del tasso di occupazione. Tuttavia le ripercussioni sull'effetto di interesse, che è una differenza tra probabilità, dovrebbero essere trascurabili. Assunzioni alternative sul meccanismo di mancata risposta sono discusse in Mealli *et al.* (2004).

4. La struttura probabilistica

Sotto le assunzioni 1-3 il processo generatore dei dati può essere definito in base ai due seguenti gruppi di probabilità:

A. Probabilità degli strati principali:

- $\pi_{GG:i} = \Pr(L_i = 'GG' | \mathbf{x}_i)$
- $\pi_{GN:i} = \Pr(L_i = 'GN' | \mathbf{x}_i)$
- $\pi_{NG:i} = \Pr(L_i = 'NG' | \mathbf{x}_i)$
- $\pi_{NN:i} = \Pr(L_i = 'NN' | \mathbf{x}_i)$.

Per esempio, $\pi_{GN:i}$ è la probabilità che lo studente *i* appartenga allo strato principale *GN*, ovvero lo studente è in grado di laurearsi entro nove anni a Economia ma non a Scienze Politiche.

B. Probabilità della variabile riposta, condizionata allo strato principale:

- $\gamma_{1,GG:i} = \Pr(Y_i^{obs} = 1 | Z_i = 1, L_i = 'GG', \mathbf{x}_i) = \Pr(Y_i(1) = 1 | L_i = 'GG', \mathbf{x}_i)$
- $\gamma_{0,GG:i} = \Pr(Y_i^{obs} = 1 | Z_i = 0, L_i = 'GG', \mathbf{x}_i) = \Pr(Y_i(0) = 1 | L_i = 'GG', \mathbf{x}_i)$
- $\gamma_{1,GN:i} = \Pr(Y_i^{obs} = 1 | Z_i = 1, L_i = 'GN', \mathbf{x}_i) = \Pr(Y_i(1) = 1 | L_i = 'GN', \mathbf{x}_i)$
- $\gamma_{0,NG:i} = \Pr(Y_i^{obs} = 1 | Z_i = 0, L_i = 'NG', \mathbf{x}_i) = \Pr(Y_i(0) = 1 | L_i = 'NG', \mathbf{x}_i)$.

Per esempio, $\gamma_{1,GG:i}$ è la probabilità che lo studente *i* abbia un lavoro permanente se appartiene allo strato principale *GG* , si è iscritto e laureato in Economia e Commercio (Z_i=1). Si noti che le probabilità che corrispondono a combinazioni tra corsi di laurea e strati principali diverse dalle quale presentate non sono definite nella presente applicazione.

La struttura probabilistica è analoga a quella dei modelli a classi latenti, fatta eccezione per il fatto che in questo caso l'appartenenza ad una certa classe latente non soltanto influisce sulla distribuzione di probabilità di *Y*, ma anche sulla sua esistenza, ovvero se *Y* sia o meno definita.

Le quantità oggetto di stima sono le differenze (o una loro sintesi) tra le probabilità di *Y* relative agli individui appartenenti allo strato *GG*, $\gamma_{1,GG:i} - \gamma_{0,GG:i}$, una per ogni combinazione delle covariate. Inoltre, anche probabilità degli strati principali $\left(\pi_{GG:i}, \pi_{NG:i}, \pi_{GN:i}, \pi_{NN:i}\right)$ sono interessanti poiché esse fanno luce sulle dinamiche del processo che conduce alla laurea nei due corsi di laurea. Infatti, l'effetto causale sulla probabilità di laurea è dato da

$$\Pr(S_i(1)=1) - \Pr(S_i(0)=1) = (\pi_{GG:i} + \pi_{GN:i}) - (\pi_{GG:i} + \pi_{NG:i}) = \pi_{GN:i} - \pi_{NG:i}. \qquad (1)$$

Quindi la probabilità $\pi_{GG:i}$ di appartenenza allo strato *GG* è irrilevante per l'effetto causale sulla laurea, nonostante il suo valore possa descrivere scenari anche molto diversi. In particolare, quando $\pi_{GG:i}$ diminuisce, i laureati dei due corsi di laurea tendono ad essere più eterogenei tra loro e quindi ci saranno maggiori opportunità di incrementare i tassi di laurea attraverso appropriate politiche di orientamento.

Anche nel caso di una popolazione omogenea, le probabilità π e γ non sono direttamente stimabili dai dati senza assunzioni aggiuntive. Infatti, risultano tre π non ridondanti e quattro γ, a fronte di sole quattro proporzioni campionarie ($P_{S,1}$, $P_{S,0}$, $P_{Y,1}$, $P_{Y,0}$). In particolare, nel paragrafo successivo si mostra come i $P_{S,1}$ e $P_{S,0}$ permettono di ottenere una stima puntuale dei π soltanto dopo avere fissato uno di loro, assumendo che i π siano gli stessi in entrambi i gruppi di trattamento. Inoltre, i γ non possono essere direttamente stimati, poiché sono definiti condizionatamente agli strati principali. I dati tuttavia permettono di stimare (attraverso $P_{Y,1}$ e $P_{Y,0}$) le seguenti probabilità:

- $\gamma_{1:i} = \Pr(Y_i^{obs}=1 \mid Z_i=1, S_i^{obs}=1, \mathbf{x}_i) = \Pr(Y_i(1)=1 \mid S_i(1)=1, \mathbf{x}_i)$
- $\gamma_{0:i} = \Pr(Y_i^{obs}=1 \mid Z_i=0, S_i^{obs}=1, \mathbf{x}_i) = \Pr(Y_i(0)=1 \mid S_i(0)=1, \mathbf{x}_i)$.

Queste probabilità sono infatti misture di probabilità condizionate allo strato principale:

$$\gamma_{1:i} = \gamma_{1,GG:i} \frac{\pi_{GG:i}}{\pi_{GG:i} + \pi_{GN:i}} + \gamma_{1,GN:i} \frac{\pi_{GN:i}}{\pi_{GG:i} + \pi_{GN:i}} \qquad (2)$$

$$\gamma_{0:i} = \gamma_{0,GG:i} \frac{\pi_{GG:i}}{\pi_{GG:i} + \pi_{NG:i}} + \gamma_{0,NG:i} \frac{\pi_{NG:i}}{\pi_{GG:i} + \pi_{NG:i}}, \tag{3}$$

e quindi la stima richiede qualche scomposizione della mistura.

5. Limiti non parametrici asintotici

Come primo passo dell'analisi è utile determinare l'insieme dei valori ammissibili delle probabilità degli strati principali alla luce dei dati disponibili e determinare i corrispondenti limiti dell'effetto causale di interesse, che è una sintesi di $\gamma_{1,GG:i}$ - $\gamma_{0,GG:i}$ (in particolare, una media marginale o condizionata). I calcoli sono effettuati sotto l'assunzione che il trattamento sia assegnato a caso e che la popolazione sia omogenea, per cui il deponente *i* viene omesso.

Nella presente applicazione ci sono quattro strati principali, la cui distribuzione è definita da tre probabilità non ridondanti. Quando il trattamento è assegnato a caso la distribuzione degli strati principali è la stessa per ogni livello del trattamento; pertanto, con l'aggiunta di un vincolo, le probabilità degli strati principali possono essere stimate a partire dalle due proporzioni osservate di laureati nei due corsi di laurea, $P_{S,1}$ e $P_{S,0}$. Quando il campione è sufficientemente grande gli errori campionari possono essere trascurati, per cui si ottengono le seguenti equazioni:

$$P_{S,1}=\pi_{GG} + \pi_{GN}; \quad 1 - P_{S,1}=\pi_{NG} + \pi_{NN}; P_{S,0}=\pi_{GG} + \pi_{NG}; \quad 1 - P_{S,0}=\pi_{GN} + \pi_{NN}.$$

Da queste equazioni segue che π_{GG} è compreso nell'intervallo

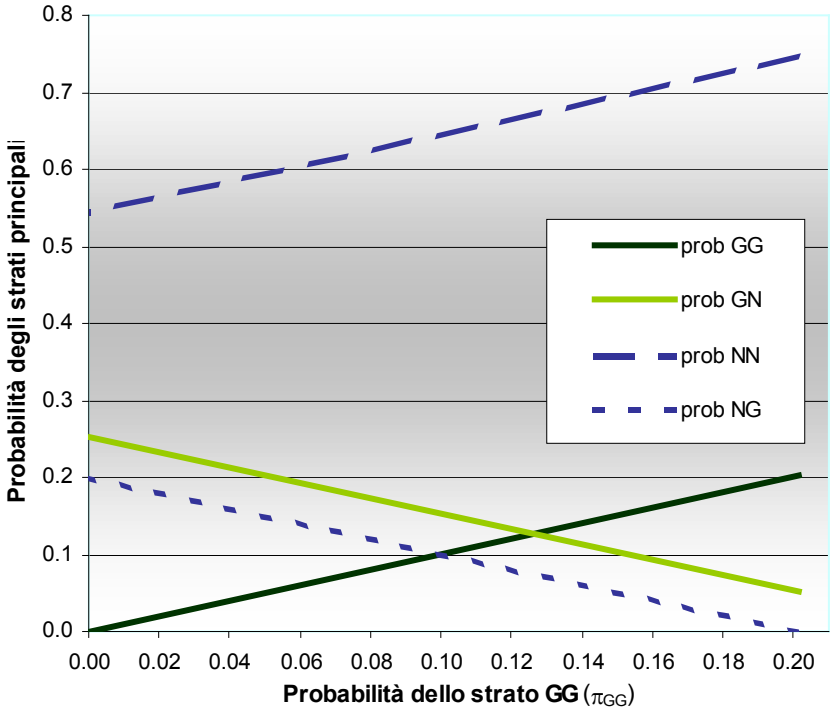
$$\max(P_{S,0} + P_{S,1} - 1, 0) \le \pi_{GG} \le \min(P_{S,0}, P_{S,1}). \tag{4}$$

Fissando π_{GG} a uno dei suoi valori ammissibili le probabilità degli altri strati principali sono

$$\pi_{GN} = P_{S,1} - \pi_{GG}; \qquad \pi_{NG} = P_{S,0} - \pi_{GG}; \qquad \pi_{NN} = 1 - P_{S,1} - P_{S,0} + \pi_{GG}. \tag{5}$$

La Figura 1 mostra le quattro probabilità degli strati principali come funzioni di π_{GG} per i dati a disposizione, dove π_{GG} può variare tra 0 e 0.202. Si noti che la differenza tra le due rette parallele discendenti, $\pi_{GN} - \pi_{NG}$, è l'effetto causale sulla laurea definito dall'equazione (1) e stimato da $P_{S,1}$- $P_{S,0}$. Pertanto la Figura 1 può essere vista come la rappresentazione di diversi scenari caratterizzati dallo stesso effetto causale stimato sulla laurea. In particolare, il massimo valore ammissibile di π_{GG} corrisponde allo scenario in cui gli strati *GN* e *NG* sono al loro minimo ammissibile, cioè $\pi_{GN} = P_{S,1}$- $P_{S,0}$ e π_{NG}=0.

Figura 1. *Valori ammissibili delle probabilità degli strati principali*



I limiti dell'effetto causale medio nello strato *GG*, $\gamma_{1,GG} - \gamma_{0,GG}$, sono calcolati per ogni valore fissato di π_{GG} considerando gli scenari migliori e peggiori. Dall'equazione (2) segue che

$$\gamma_{1,GG} = \frac{\gamma_1 - \gamma_{1,GN}(1-\varphi_{1,GG})}{\varphi_{1,GG}}, \tag{6}$$

dove $\varphi_{1,GG} = \pi_{GG} / (\pi_{GG} + \pi_{GN})$. Allora $\gamma_{1,GG}$ raggiunge il suo minimo quando $\gamma_{1,GN} = 1$ e il suo massimo quando $\gamma_{1,GN} = 0$, dando luogo ai seguenti limiti:

$$\max\left(1 - \frac{1-\gamma_1}{\varphi_{1,GG}}, 0\right) \le \gamma_{1,GG} \le \min\left(\frac{\gamma_1}{\varphi_{1,GG}}, 1\right). \tag{7}$$

Questi limiti dipendono da due quantità ignote che devono essere stimate: $\varphi_{1,GG}$ è stimato tramite le equazioni (5), per cui dipende dal valore ipotizzato di π_{GG} e dalla proporzione campionaria della variabile intermedia per *Z*=1, $P_{S,1}$; d'altra parte, γ_1 è stimato dalla proporzione campionaria della variabile risposta per *Z*=1, $P_{Y,1}$.

In modo analogo, l'equazione (3) implica che i limiti per $\gamma_{0,GG}$ siano dati da

$$\max\left(1-\frac{1-\gamma_0}{\varphi_{0,GG}},0\right)\le\gamma_{0,GG}\le\min\left(\frac{\gamma_0}{\varphi_{0,GG}},1\right), \tag{8}$$

dove $\varphi_{0,GG}=\pi_{GG}/\left(\pi_{GG}+\pi_{NG}\right)$ dipende dal valore ipotizzato di π_{GG} e dalla proporzione campionaria della variabile intermedia per Z=0, $P_{S,0}$, mentre γ_0 è stimato dalla proporzione campionaria della variabile risposta per Z=0, $P_{Y,0}$.

Infine, i limiti dell'effetto causale nello strato GG, $\gamma_{1,GG}-\gamma_{0,GG}$, derivano dai limiti (7) e (8):

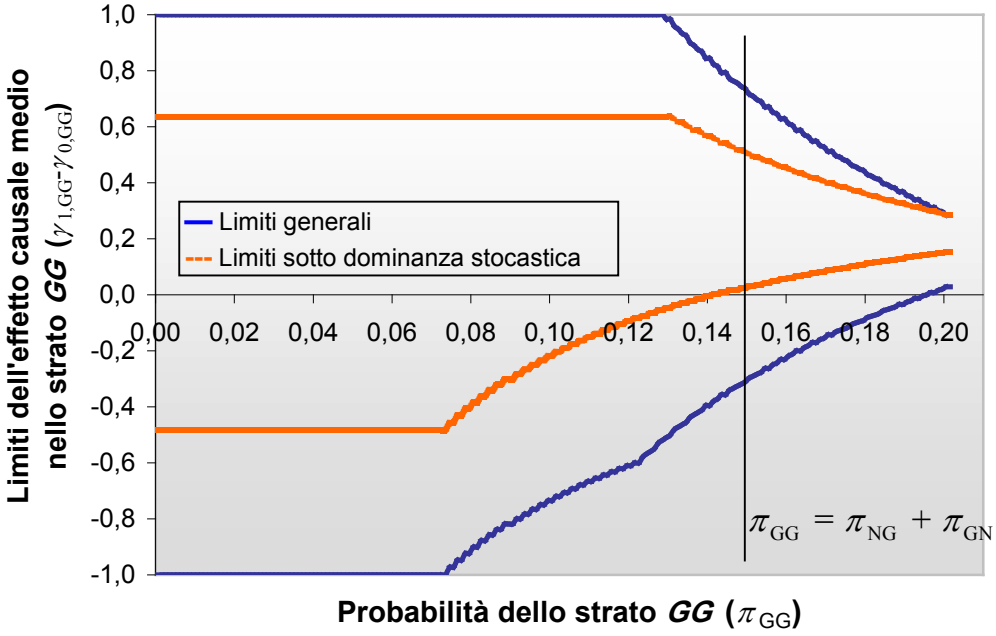
$$\begin{aligned}\max\left(1-\frac{1-\gamma_1}{\varphi_{1,GG}},0\right)-\min\left(\frac{\gamma_0}{\varphi_{0,GG}},1\right)&\le\gamma_{1,GG}-\gamma_{0,GG}\\ &\le\min\left(\frac{\gamma_1}{\varphi_{1,GG}},1\right)-\max\left(1-\frac{1-\gamma_0}{\varphi_{0,GG}},0\right).\end{aligned} \tag{9}$$

Questi limiti sono simili a quelli derivati da Zhang & Rubin (2004): la differenza è che questi Autori utilizzano una variabile Y continua e calcolano i limiti per mezzo di una procedura basata sui valori ordinati di Y; tuttavia, quando Y è binaria, come nel caso presente, la loro procedura fornisce gli stessi risultati della nostra, a meno di approssimazioni dovute alla natura discreta dei dati. Si noti inoltre che Zhang & Rubin (2004) analizzano i dati di un esperimento in cui sono presenti veri trattamenti e controlli, per cui studiano i limiti come funzioni di π_{NG}, mentre nella nostra applicazione, dato che i due trattamenti sono sullo stesso piano, è più naturale studiare i limiti come funzioni di π_{GG}.

I limiti (9), stimati dalle proporzioni campionarie, sono disegnati come funzioni di π_{GG} in Figura 2 con la dizione "Limiti generali". Si noti che i limiti si allargano mano a mano che π_{GG} diventa più piccolo: per valori elevati di π_{GG} (tra 0.196 e il massimo 0.202) gli estremi sono entrambi positivi, per cui il segno dell'effetto causale è determinato; poi i limiti si allargano fino a raggiungere l'intervallo [-1,1], diventando inutili.

I limiti appena calcolati sono asintotici, nel senso che in grandi campioni stimano i veri limiti quasi senza errore e non c'è bisogno di considerare esplicitamente bande di confidenza rappresentanti l'incertezza dovuta alla stima. In generale, sia i limiti superiori che quelli inferiori dovrebbero essere inclusi in bande di confidenza: questo permetterebbe di tenere in considerazione la possibilità che un dato modello fornisca una stima dell'effetto causale medio che cade al di fuori dei limiti calcolati. Nella presente applicazione le bande di confidenza non sono mostrate, poiché l'uso principale dei limiti è quello di esplorare i dati e di giudicare qualitativamente la plausibilità dei risultati prodotti dal modello.

Figura 2. *Limiti dell'effetto causale medio nello strato GG*



I limiti sintetizzano l'incertezza che caratterizza la stima dell'effetto causale medio nello strato *GG* indipendentemente dalla dimensione campionaria: il messaggio è che persino in un grande campione c'è un intero intervallo di valori ammissibili per la quantità oggetto di stima, la cui ampiezza dipende dalla struttura della popolazione, in particolare dalla dimensione dello strato *GG*.

I limiti possono essere ristretti facendo opportune assunzioni sulle probabilità degli strati principali o sulle probabilità della variabile risultato.

Per quanto riguarda le probabilità degli strati principali, un'assunzione standard è quella di *monotonicità*, ovvero la non esistenza del gruppo *NG*, cioè $\pi_{NG} = 0$. Questa assunzione viene fatta spesso in studi in cui si confrontano un trattamento attivo con un placebo poiché, rispetto alla variabile intermedia *S*, il gruppo *NG* ha una performance negativa ($S_i = 0$) sotto il trattamento attivo ($Z_i = 1$) e una performance positiva ($S_i = 1$) sotto il controllo ($Z_i = 0$). Tuttavia nell'applicazione presente i due gruppi di trattamento sono sullo stesso piano, per cui è verosimile che entrambi i gruppi *NG* e *GN* siano presenti. L'assunzione di monotonicità è dunque poco plausibile.

Un vincolo sulle probabilità degli strati principali che sembra ragionevole nel presente contesto è che gli studenti in grado di laurearsi in entrambi i corsi, π_{GG}, siano una maggioranza nel gruppo degli studenti in grado di laurearsi in almeno uno dei

corsi, cioè nel gruppo con probabilità $\pi_{GG} + \pi_{NG} + \pi_{GN}$. Questo porta a formulare la seguente assunzione:

Assunzione 4 (Maggioranza relativa dello strato GG):
per ogni i, $\pi_{GG:i} \geq \pi_{NG:i} + \pi_{GN:i}$.

Assumendo omogeneità della popolazione e casualizzazione del trattamento, dalle equazioni (5) segue che l'Assunzione 4 equivale a $3\pi_{GG} - (P_{S,1} + P_{S,0}) \geq 0$. Poiché i limiti si ampliano mano a mano che π_{GG} diminuisce, i limiti più ampi che soddisfano l'Assunzione 4 corrispondono a quell'unico valore di π_{GG} per il quale la disuguaglianza diviene un'uguaglianza, cioè $\pi_{GG} = (P_{S,1} + P_{S,0})/3$, purché tale valore di π_{GG} sia ammissibile. Questo caso è rappresentato in Figura 2 dalla linea verticale passante attraverso π_{GG} = 0.152. I corrispondenti limiti sono [-0.290, 0.708], ovviamente molto più informativi dell'intervallo [-1,1].

Per quanto riguarda le probabilità della variabile risposta, è ragionevole assumere che gli studenti in grado di laurearsi in entrambi i corsi (strato *GG*) abbiano più probabilità di ottenere un lavoro permanente rispetto agli studenti in grado di laurearsi in un corso ma non nell'altro (strati *NG* e *GN*). Questa considerazione porta alla seguente assunzione:

Assunzione 5 (Dominanza stocastica): per ogni i, e per ogni numero reale t,
$\Pr(Y_{GG:i}(1) \leq t) \leq \Pr(Y_{GN:i}(1) \leq t)$; $\Pr(Y_{GG:i}(0) \leq t) \leq \Pr(Y_{NG:i}(0) \leq t)$.

Questa assunzione viene utilizzata da Zhang & Rubin (2004) nel caso di una variabile risultato continua. Nel contesto attuale la variabile risultato Y è binaria, per cui in termini di probabilità di occupazione, la dominanza stocastica è equivalente a $\gamma_{1,GG:i} \geq \gamma_{1,GN:i}$ e $\gamma_{0,GG:i} \geq \gamma_{0,NG:i}$.

Assumendo la dominanza stocastica i limiti sono più stretti che nel caso generale, poiché l'espressione (9) diviene

$$\gamma_1 - \min\left(\frac{\gamma_0}{\varphi_{0,GG}}, 1\right) \leq \gamma_{1,GG} - \gamma_{0,GG} \leq \min\left(\frac{\gamma_1}{\varphi_{1,GG}}, 1\right) - \gamma_0 \qquad (10)$$

Si noti che, quando $\pi_{GG} = \pi_{NG} + \pi_{GN}$, i limiti stimati (10) sono [0.030, 0.494], per cui l'effetto causale medio è necessariamente positivo. Questo è un risultato interessante, poiché mostra che due assunzioni deboli, come la 4 e la 5, possono essere sufficienti a determinare il segno dell'effetto senza bisogno di affidarsi ad un modello parametrico.

I limiti qui calcolati sono validi solo se l'assunzione di non confondimento (Assunzione 2) vale marginalmente (cioè, non condizionatamente alle covariate). Nel caso presente non vi è casualizzazione e il trattamento è stato liberamente scelto dagli individui, per cui marginalmente il non confondimento potrebbe non valere. Una possibile miglioria, basata sull'assunzione meno restrittiva che il non confondimento valga condizionatamente alle covariate, è di derivare i limiti per ogni cella definita dalle covariate e poi ricostruire i limiti non condizionati attraverso una media pesata con le frequenze di cella. L'applicazione di tale tecnica ai nostri dati non porta a cambiamenti di rilievo.

6. Analisi basata su modello

Un modo efficiente di sfruttare l'informazione insita nelle covariate, al costo di aggiungere altre assunzioni, è quello di costruire un modello parametrico, che può essere adattato ai dati sia con metodi frequentisti che con metodi Bayesiani. La specificazione del modello e la stima sono compiti impegnativi, poiché nell'approccio degli strati principali i gruppi latenti portano a misture di distribuzioni difficili da scomporre. Le covariate sono estremamente utili per identificare il modello: l'identificazione può essere raggiunta attraverso diversi tipi di vincoli la cui plausibilità deve essere valutata caso per caso, come illustrato da Jo (2002) nel caso particolare di *noncompliance* con due gruppi latenti. Tuttavia, la funzione di verosimiglianza è solitamente piuttosto piatta, per cui la sua massimizzazione non è agevole. L'approccio Bayesiano (Imbens & Rubin, 1997) può aiutare a superare queste difficoltà, ma, a parte la complessità computazionale, la scelta di appropriate distribuzioni a priori è tutt'altro che facile. Nella presente applicazione effettuiamo un'analisi di massima verosimiglianza, che risulta efficace per il problema allo studio.

Come notato nella Sezione 4, il processo generatore dei dati può essere definito in termini di due insiemi di probabilità: i π, che danno origine al sottomodello relativo agli strati principali, e i γ, che danno origine al sottomodello relativo alla variabile risultato. Le variabili disponibili per ogni individuo sono Z_i, S_i^{obs}, R_i^{obs}, Y_i^{obs} (se R_i^{obs} =1) e il vettore di covariate $\mathbf{x}_i$. Nella presente applicazione i 19 individui con valori mancanti delle covariate sono semplicemente eliminati, per cui le covariate sono trattate come completamente osservate. Estensioni per gestire valori mancanti delle covariate sono state sviluppate da Barnard *et al.* (2003).

Raccogliamo ora i parametri nel vettore $\boldsymbol{\theta}$ e le variabili per gli n individui nei vettori $\mathbf{Z}$, $\mathbf{S}^{obs}$, $\mathbf{R}^{obs}$ and $\mathbf{Y}^{obs}$ e nella matrice $\mathbf{X}$. La verosimiglianza può essere scritta come prodotto sui quattro gruppi osservabili definiti da Z_i e S_i^{obs}, dove $i \in O(k,h)$ sta per $Z_i = k$ e $S_i^{obs} = h$:

$$
\begin{aligned}
&L\left(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{S}^{obs}, \mathbf{R}^{obs}, \mathbf{Y}^{obs}, \mathbf{X}\right) = \\
&\prod_{i \in O(1,1)} \left\{ \pi_{GG:i} \left[\left(\gamma_{1,GG:i}\right)^{Y_i^{obs}} \left(1-\gamma_{1,GG:i}\right)^{1-Y_i^{obs}} \right]^{R_i^{obs}} + \pi_{GN:i} \left[\left(\gamma_{1,GN:i}\right)^{Y_i^{obs}} \left(1-\gamma_{1,GN:i}\right)^{1-Y_i^{obs}} \right]^{R_i^{obs}} \right\} \\
&\times \prod_{i \in O(1,0)} \left\{ \pi_{NG:i} + \pi_{NN:i} \right\} \\
&\times \prod_{i \in O(0,1)} \left\{ \pi_{GG:i} \left[\left(\gamma_{0,GG:i}\right)^{Y_i^{obs}} \left(1-\gamma_{0,GG:i}\right)^{1-Y_i^{obs}} \right]^{R_i^{obs}} + \pi_{NG:i} \left[\left(\gamma_{0,NG:i}\right)^{Y_i^{obs}} \left(1-\gamma_{0,NG:i}\right)^{1-Y_i^{obs}} \right]^{R_i^{obs}} \right\} \\
&\times \prod_{i \in O(0,0)} \left\{ \pi_{GN:i} + \pi_{NN:i} \right\}
\end{aligned}
\tag{11}
$$

Il modello è basato sulle Assunzioni da 1 a 3 (SUTVA, non confondimento del trattamento e *missing at random*).

Nella verosimiglianza (11) gli individui che non hanno risposto all'intervista (R_i^{obs} =0) non contribuiscono alla stima dei γ, ma contribuiscono comunque alla stima dei π. In generale, i π sono stimati a partire da tutti gli individui del campione, mentre l'informazione sui γ è fornita solo dagli individui che si sono laureati e che sono stati intervistati (15% del campione), per cui l'informazione per la stima dei γ è limitata.

Come nella maggior parte delle attuali applicazioni dell'approccio degli strati principali, le variabili trattamento e intermedia sono entrambe binarie, originando quattro strati principali. Tuttavia, mentre in molti contesti è ragionevole assumere che certi strati siano vuoti (ad es. l'assunzione di assenza di *defiers* in un esperimento con *noncompliance*), nel contesto attuale tali assunzioni non sono plausibili alla luce della simmetria dei due trattamenti, per cui tutti gli strati in principio sono non vuoti. Questo livello di generalità comporta un notevole incremento della complessità del modello poiché, come risulta chiaro dalla verosimiglianza (11), *ogni* gruppo osservato $O(k,h)$ è generato da una mistura di due distribuzioni che deve essere scomposta.

Le probabilità degli strati principali π sono soggette ad alcuni vincoli poiché devono stare nell'intervallo [0,1] e la loro somma deve essere uguale a uno. Pertanto per modellare la dipendenza di tali probabilità dalle covariate è utile operare una trasformazione in un insieme di parametri non vincolati, usando la specificazione logistica multinomiale (dove *NN* è la categoria di riferimento):

$$\pi_{GG:i} = \frac{\exp(\eta^{\pi}_{GG:i})}{1+\exp(\eta^{\pi}_{GG:i})+\exp(\eta^{\pi}_{GN:i})+\exp(\eta^{\pi}_{NG:i})}$$
$$\pi_{GN:i} = \frac{\exp(\eta^{\pi}_{GN:i})}{1+\exp(\eta^{\pi}_{GG:i})+\exp(\eta^{\pi}_{GN:i})+\exp(\eta^{\pi}_{NG:i})}$$
$$\pi_{NG:i} = \frac{\exp(\eta^{\pi}_{NG:i})}{1+\exp(\eta^{\pi}_{GG:i})+\exp(\eta^{\pi}_{GN:i})+\exp(\eta^{\pi}_{NG:i})}$$
$$\pi_{NN:i} = \frac{1}{1+\exp(\eta^{\pi}_{GG:i})+\exp(\eta^{\pi}_{GN:i})+\exp(\eta^{\pi}_{NG:i})} \quad .$$

Per le probabilità della variabile risultato γ la trasformazione in un insieme di parametri non vincolati si può ottenere con specificazioni logistiche separate:

$$\gamma_{1,GG:i} = \frac{1}{1+\exp(-\eta^{\gamma}_{1,GG:i})} \qquad \gamma_{0,GG:i} = \frac{1}{1+\exp(-\eta^{\gamma}_{0,GG:i})}$$
$$\gamma_{1,GN:i} = \frac{1}{1+\exp(-\eta^{\gamma}_{1,GN:i})} \qquad \gamma_{0,NG:i} = \frac{1}{1+\exp(-\eta^{\gamma}_{0,NG:i})} \quad .$$

Si assume poi che i parametri η^{π} e η^{γ} dipendano linearmente dalle covariate. Nella versione più generale del modello ognuno di questi parametri ha il suo insieme distinto di coefficienti di regressione.

Nell'applicazione corrente la versione più generale del modello è caratterizzata da una specificazione lineare non vincolata degli η^{π},

$$\begin{aligned} \eta^{\pi}_{GG:i} &= \alpha^{\pi}_{GG} + \boldsymbol{\beta}^{\pi}_{GG}{}' \mathbf{x}_i \\ \eta^{\pi}_{GN:i} &= \alpha^{\pi}_{GN} + \boldsymbol{\beta}^{\pi}_{GN}{}' \mathbf{x}_i \\ \eta^{\pi}_{NG:i} &= \alpha^{\pi}_{NG} + \boldsymbol{\beta}^{\pi}_{NG}{}' \mathbf{x}_i \end{aligned} \qquad (12)$$

e da una particolare specificazione lineare degli η^{γ},

$$\begin{aligned} \eta^{\gamma}_{1,GG:i} &= \alpha^{\gamma}_{1,GG} + \boldsymbol{\beta}^{\gamma}{}' \mathbf{x}_i \\ \eta^{\gamma}_{0,GG:i} &= \alpha^{\gamma}_{0,GG} + \boldsymbol{\beta}^{\gamma}{}' \mathbf{x}_i \\ \eta^{\gamma}_{1,GN:i} &= \alpha^{\gamma}_{1,GN} + \boldsymbol{\beta}^{\gamma}{}' \mathbf{x}_i \\ \eta^{\gamma}_{0,NG:i} &= \alpha^{\gamma}_{0,NG} + \boldsymbol{\beta}^{\gamma}{}' \mathbf{x}_i \, . \end{aligned} \qquad (13)$$

La specificazione degli η^{γ} è particolare perché assume che ogni covariata abbia lo stesso effetto in ogni strato principale e che l'effetto causale sulla scala logistica, $\alpha^{\gamma}_{1,GG} - \alpha^{\gamma}_{0,GG}$, sia additivo, cioè lo stesso per tutti i valori delle covariate. Nella nostra applicazione questa specificazione sembra ragionevole. Altre specificazioni potrebbero essere adottate (Jo, 2002), ma nel caso presente la loro adozione è ostacolata dalla scarsità dell'informazione campionaria.

L'identificazione del modello è possibile solo con un adeguato numero di covariate. Denotando con *k* il numero di covariate, il modello definito dalle espressioni (12) e (13) ha 3(*k*+1) parametri per i π e (4+*k*) parametri per i γ, per un totale di (7+4*k*) parametri. D'altra parte, se il trattamento e le *k* covariate sono tutte variabili binarie, allora ci sono $2^{(k+1)}$ celle con al più due proporzioni campionarie, una per la variabile intermedia *S* e una per la variabile risultato *Y*, per cui il massimo numero di proporzioni campionarie è $2^{(k+2)}$. È importante notare che alcune celle potrebbero essere completamente o parzialmente vuote, per cui il numero effettivo di proporzioni campionarie, e di conseguenza il numero di gradi di libertà, deve essere controllato caso per caso. Comunque sono necessarie almeno due covariate per rendere possibile l'identificazione.

Nella nostra applicazione (*k*=5) il modello ha 27 parametri, mentre il trattamento e le cinque covariate danno luogo a 64 celle e 128 proporzioni campionarie teoriche. Poiché 3 celle sono completamente vuote e altre 23 celle hanno la risposta mancante solo per la variabile risultato, le proporzioni campionarie disponibili sono 99, ben oltre il numero di parametri. Tuttavia uno sguardo ai valori delle proporzioni campionarie fa prevedere dei problemi di stima legati all'alto numero di proporzioni campionarie uguali a zero oppure ad uno: infatti, su 61 proporzioni campionarie disponibili per la variabile intermedia, 19 sono zero e 1 è uno, mentre su 38 proporzioni campionarie disponibili per la variabile risultato, 5 sono zero e 3 sono uno.

La stima di massima verosimiglianza è stata ottenuta per mezzo della procedura NLMIXED del SAS (SAS Institute, 1999). Come suggerito dal nome, tale procedura è designata alla stima di modelli non lineari misti e, in effetti, una delle componenti essenziali è l'algoritmo per l'integrazione numerica. Tuttavia la NLMIXED è anche una procedura generale di massimizzazione della verosimiglianza, poiché può gestire funzioni di verosimiglianza arbitrarie scritte dall'utente. Per la presente applicazione è sufficiente scrivere la verosimiglianza usando il linguaggio SAS e lanciare la procedura senza integrazione numerica. La procedura ha diversi algoritmi di massimizzazione, fra cui quello di default è di tipo quasi-Newton con aggiornamento BFGS (Broyden, Fletcher, Goldfarb e Shanno) del fattore di Cholesky della matrice hessiana approssimata.

La verosimiglianza del modello più generale, cioè del modello definito dalle equazioni (12) e (13) senza ulteriori vincoli, è piuttosto piatta. Per affrontare un compito così difficile alcuni dei valori iniziali (quelli di $\alpha^{\pi}_{GG}, \alpha^{\pi}_{GN}, \alpha^{\pi}_{NG}$) sono stati

scelti attraverso una ricerca su griglia. Inoltre sono stati provati diversi algoritmi di stima: nonostante si ottenesse sempre la convergenza, gli algoritmi fornivano risultati sensibilmente diversi per un sottoinsieme di parametri legati ai π e caratterizzati da valori stimati molto negativi ed errori standard elevati. Questo significa che per certi valori delle covariate alcuni strati principali sono vuoti. In particolare, per l'individuo base, che è quello con la configurazione di covariate più frequente nel campione e caratterizzato dall'avere il valore zero in tutte le covariate, lo strato *NG* sembra vuoto, poiché il corrispondente valore sulla scala logistica multinomiale è -7.826 (errore standard 14.763). Pertanto, al fine di seguire una strategia di selezione del modello semplice e chiara, abbiamo ridefinito la codifica della covariata in modo da ottenere un nuova definizione dell'individuo base con probabilità sensibilmente diverse da zero in tutti gli strati. Questo obiettivo è stato conseguito semplicemente invertendo la codifica della covariata *Iscritto con ritardo*, che d'ora in avanti chiameremo *Iscritto senza ritardo*.

I risultati della stima ottenuti con l'algoritmo di default e basati sulla nuova codifica sono riportati in Tabella 5 nella colonna denominata "Modello iniziale". Sei dei β^{π} stimati sono inferiori a –5, con errori standard enormi o non disponibili: ciò significa che quando la covariata passa da zero a uno il corrispondente strato principale scompare. In particolare, con l'eccezione di alcuni studenti iscritti con ritardo, lo strato *NG* risulta vuoto. Questo non è sorprendente, poiché la proporzione complessiva di laureati è modesta ed è minore per Z_i =0, per cui lo strato *NG* ("Non laureato" se Z_i =1 e "Laureato" se Z_i =0) è necessariamente molto limitata. Anche lo strato *GN* contrapposto sembra essere vuoto in certi casi.

La selezione del modello prosegue fissando a $-\infty$ i suddetti β^{π} e porta ai risultati mostrati in Tabella 5 nella colonna denominata "Modello finale". La riduzione da 27 a 21 parametri comporta una riduzione irrilevante della devianza, mentre gli altri parametri ed errori standard sono sostanzialmente invariati. Alcuni dei β^{π} sono non significativi ai livelli convenzionali, per cui il sottomodello degli strati principali potrebbe essere ulteriormente semplificato. Tuttavia la selezione del modello è stata arrestata a questo punto, poiché avere un sottomodello degli strati principali con pochi parametri non è un obiettivo di interesse sostanziale ed ha uno scarso effetto sulla precisione delle stime del sottomodello della variabile risultato.

Nel sottomodello della variabile risultato i β^{γ} non sono significativi ai livelli convenzionali, sebbene due di loro (*Liceo* e *Iscrizione senza ritardo*) abbiano valori elevati: servirebbero più dati per stabilire l'influenza delle covariate sulla variabile risultato. Nonostante ciò, l'oggetto principale dell'inferenza, cioè l'effetto causale su scala logistica, $\alpha^{\gamma}_{1,GG} - \alpha^{\gamma}_{0,GG}$, ha una stima pari a 0.666 con errore standard 0.301, per cui è significativamente diverso da zero al livello 5%.

Tabella 5. *Stime dei parametri (ed errori standard) dell'analisi basata su modello*

	Modello iniziale		*Modello finale*	
Numero di parametri	27		21	
Devianza (-2*logL*)	2231.8		2231.8	
Sottomodello strati principali (π's)				
α^{π}_{GG}	-4.403	(0.449)	-4.402	(0.448)
α^{π}_{GN}	-2.644	(0.749)	-2.647	(0.752)
α^{π}_{NG}	-3.206	(0.836)	-3.207	(0.835)
$\beta^{\pi}_{GG,liceo}$	1.275	(0.157)	1.275	(0.157)
$\beta^{\pi}_{GN,liceo}$	-5.757	(n.d.)	- ∞	
$\beta^{\pi}_{NG,liceo}$	-15.041	(n.d.)	- ∞	
$\beta^{\pi}_{GG,voto\,alto}$	1.204	(0.146)	1.205	(0.146)
$\beta^{\pi}_{GN,voto\,alto}$	1.113	(0.653)	1.113	(0.652)
$\beta^{\pi}_{NG,voto\,alto}$	-8.092	(114.022)	- ∞	
$\beta^{\pi}_{GG,iscrizione\,senza\,ritardo}$	2.024	(0.425)	2.023	(0.425)
$\beta^{\pi}_{GN,iscrizione\,senza\,ritardo}$	-0.012	(0.788)	-0.009	(0.792)
$\beta^{\pi}_{NG,iscrizione\,senza\,ritardo}$	-8.140	(64.473)	- ∞	
$\beta^{\pi}_{GG,femmina}$	0.117	(0.137)	0.117	(0.137)
$\beta^{\pi}_{GN,femmina}$	-0.617	(0.753)	-0.622	(0.755)
$\beta^{\pi}_{NG,femmina}$	0.988	(1.112)	0.991	(1.111)
$\beta^{\pi}_{GG,Firenze}$	0.280	(0.144)	0.280	(0.144)
$\beta^{\pi}_{GN,Firenze}$	-13.499	(559.599)	- ∞	
$\beta^{\pi}_{NG,Firenze}$	-10.353	(533.855)	- ∞	
Sottomodello risultato (γ's)				
$\alpha^{\gamma}_{1,GG}$	1.257	(1.240)	1.262	(1.241)
$\alpha^{\gamma}_{0,NG}$	-1.357	(1.561)	-1.365	(1.568)
$\alpha^{\gamma}_{0,GG}$	0.593	(1.185)	0.596	(1.185)
$\alpha^{\gamma}_{1,GN}$	0.498	(1.057)	0.484	(1.058)
β^{γ}_{liceo}	-0.405	(0.374)	-0.410	(0.374)
$\beta^{\gamma}_{voto\,alto}$	-0.035	(0.262)	-0.036	(0.263)
$\beta^{\gamma}_{iscrizione\,senza\,ritardo}$	-0.933	(0.979)	-0.932	(0.979)
$\beta^{\gamma}_{femmina}$	0.072	(0.272)	0.070	(0.272)
$\beta^{\gamma}_{Firenze}$	0.106	(0.333)	0.104	(0.333)
Effetto causale $\alpha^{\gamma}_{1,GG} - \alpha^{\gamma}_{0,GG}$	0.664	(0.301)	0.666	(0.301)

Tabella 6. *Probabilità stimate(%) per alcune configurazioni delle covariate*

Probabilità	**00000**	**00100**	**00110**	**00101**	**01100**	**10100**	**11100**	**11111**
$\pi_{GG:i}$	1.1	8.0	9.1	10.9	20.3	24.9	52.5	62.2
$\pi_{GN:i}$	6.3	6.0	3.3	0.0	14.0	0.0	0.0	0.0
$\pi_{NG:i}$	3.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\pi_{NN:i}$	89.0	86.0	87.6	89.1	65.7	75.1	47.5	37.8
$\gamma_{1,GG:i}$	77.9	58.2	59.9	60.7	57.3	48.0	47.1	51.5
$\gamma_{0,GG:i}$	64.5	41.7	43.4	44.2	40.8	32.2	31.4	35.3
$\gamma_{1,GN:i}$	61.9	39.0	40.7	41.5	38.1	29.8	29.0	32.8
$\gamma_{0,NG:i}$	20.3	9.1	9.7	10.0	8.9	6.3	6.1	7.1
Effetto causale $\gamma_{1,GG:i} - \gamma_{0,GG:i}$	13.5	16.5	16.5	16.4	16.5	15.8	15.7	16.2

Nota: la configurazione $(x_1, x_2, x_3, x_4, x_5)$ sta per: $Liceo = x_1, Voto\ alto = x_2, Iscrizione\ senza\ ritardo = x_3,$ $Femmina = x_4, Firenze = x_5$.

Per aiutare l'interpretazione dei risultati, la Tabella 6 riporta le probabilità stimate dal modello finale per alcune configurazioni delle covariate, con le configurazioni in ordine crescente di $\pi_{GG:i}$. Le proporzioni stimate di studenti appartenenti al gruppo *GG* variano molto con le covariate, da un minimo dell'1.1% a un massimo del 62.2%. Inoltre, le proporzioni stimate di studenti appartenenti ai gruppi *GN* e *NG* (cioè studenti in grado di laurearsi in un solo corso di laurea) tendono a diminuire mano a mano che lo strato *GG* cresce, nonostante che lo strato *NN* diminuisca. Ad un estremo, l'individuo con tutte le covariate uguali a uno (una femmina residente a Firenze, proveniente da un liceo, con un voto elevato e iscrizione senza ritardo) ha un'alta probabilità di laurearsi (62.2%), interamente attribuita al gruppo *GG*; all'altro estremo, l'individuo di base (un maschio residente fuori Firenze, proveniente da un liceo, con voto basso e iscrizione con ritardo) ha una bassa probabilità di laurearsi in almeno uno dei due corsi di laurea (11.0%), attribuita principalmente ai gruppi *GN* e *NG*.

Poiché la differenza tra i due corsi di laurea in termini di tassi di laurea è originata dai gruppi *GN* e *NG*, essendo $\pi_{GN:i} - \pi_{NG:i}$ come spiegato nella Sezione 4, segue che i due corsi di laurea hanno un diverso effetto sulla probabilità di laurea solo per gli studenti che hanno un background debole. Le politiche di orientamento dovrebbero quindi essere indirizzate in modo particolare a questo tipo di studenti.

Dall'analisi basata su modello sembra che l'assunzione di maggioranza relativa dello strato *GG* (Assunzione 4: $\pi_{GG:i} \geq \pi_{NG:i} + \pi_{GN:i}$), usata nella costruzione dei limiti, valga in generale, con l'eccezione degli individui che si sono iscritti con ritardo.

Guardando ora all'effetto sull'occupazione, ci sono alcuni risultati da sottolineare. Innanzitutto, l'assunzione di dominanza stocastica (Assunzione 5), usata per derivare alcuni dei limiti, sembra essere soddisfatta: infatti, condizionatamente alle covariate, gli studenti appartenenti ai gruppi *GN* e *NG* hanno una probabilità di essere occupati sempre inferiore a quella degli studenti del gruppo *GG*. Il livello della probabilità di occupazione varia molto con le covariate, oscillando tra 47.1% e 77.9% per i laureati in Economia, e tra 31.4% e 64.5% per i laureati in Scienze Politiche. L'effetto causale sull'occupazione per il gruppo *GG*, che si è ipotizzato costante sulla scala logistica per evitare problemi di identificazione, genera un differenziale pari a circa 15% nelle probabilità di occupazione. Naturalmente l'affidabilità e anche l'importanza sostanziale di tale differenziale dipende dall'ampiezza dello strato *GG*: ad esempio, l'effetto causale nello strato *GG* ha poca rilevanza per l'individuo base, che ha una probabilità di appena 1.1% di essere *GG*.

7. Conclusioni

In questo lavoro abbiamo confrontato due corsi di laurea dell'Università di Firenze al fine di valutare la loro efficacia rispetto allo status occupazionale dopo la laurea. L'approccio degli strati principali all'inferenza causale è stato usato per definire un quadro concettuale per l'analisi di questo fenomeno, con una definizione precisa delle quantità di interesse. In questo quadro sono stati derivati dei limiti non parametrici per l'effetto causale di interesse: i limiti non parametrici permettono di restringere lo spettro delle possibili inferenze sulla base di un insieme minimo di assunzioni, la cui validità deve essere giudicata caso per caso.

La successiva analisi basata su modello, condotta in un contesto frequentista, ha consentito di sfruttare in modo efficiente l'informazione insita nelle covariate, al costo di aggiungere alcune ulteriori assunzioni. La strategia di selezione del modello ha richiesto alcune accortezze per includere la possibilità che alcuni strati principali siano vuoti. Naturalmente, i risultati sono più informativi di quelli ottenuti per mezzo dei limiti non parametrici. In particolare, l'effetto causale per lo strato *GG* (ossia gli studenti in grado di laurearsi in entrambi i corsi di laurea) è positivo (ovvero in favore di Economia) e statisticamente significativo, rinforzando le impressioni ottenute con l'analisi non parametrica. Inoltre, il modello consente di approfondire l'analisi, poiché mostra come la struttura degli strati principali cambi con le covariate: questa informazione è cruciale per capire il processo di laurea ed anche per interpretare in modo consapevole l'effetto causale stimato (in quanto riferito ad uno specifico strato).

Purtroppo, a causa della limitatezza dell'informazione campionaria sulla condizione occupazionale, molti parametri del sottomodello per la variabile risultato so-

no risultati non significativi; per questo motivo è risultato impossibile costruire un sottomodello più sofisticato per la variabile risultato.

In alternativa, l'analisi basata su modello può essere sviluppata con metodi Bayesiani, che comportano diverse difficoltà (specificazione delle distribuzioni a priori, problemi computazionali), ma offrono alcuni vantaggi che divengono cruciali nel caso di modelli molto complessi, come in Barnard *et al.* (2003).

Riferimenti bibliografici

BARNARD J., FRANGAKIS C.E., HILL J.L. & RUBIN D.B. (2003) Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City, *Journal American Statistical Association*, **98**: 299-323.

FRANKGAKIS C.E. & RUBIN D.B. (2002) Principal stratification in causal inference, *Biometrics*, **58**: 21-29.

IMBENS G.W. & RUBIN D.B. (1997) Bayesian inference for causal effects in randomized experiments with noncompliance, *Annals of Statistics*, **25**: 305-327.

JO B. (2002) Estimation of intervention effects with noncompliance: alternative model specifications, *Journal of Educational and Behavioral Statistics*, **27**: 385-409.

MEALLI F., IMBENS G.W., FERRO S. & BIGGERI A. (2004). Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes, *Biostatistics*, **5**: 207-222.

RUBIN D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, **66**: 668-701.

SAS INSTITUTE (1999) *SAS/STAT User's Guide Version 8*. SAS Institute Inc, Cary.

ZHANG J. & RUBIN D.B. (2004) Estimation of causal effects when some outcomes are censored by death, In corso di stampa su *Journal of Educational and Behavioral Statistics.*

The effect of university studies on job opportunities: an application of the principal strata approach to causal inference

Summary. *The paper shows how to evaluate the effectiveness of two degree programmes with respect to the employment status using the principal strata approach to causal inference. The application concerns the 1992's cohort of freshmen of the University of Florence enrolled in the degree programmes of Economics and Political Science. The paper shows an innovative use of non parametric bounds in the principal strata framework, examining the role of some assumptions in reducing the uncertainty. The second phase of the analysis relies on a parametric model fitted by maximum likelihood. In that context we discuss some relevant modelling issues, sketching a general strategy for model building.*
Keywords: *causal effects, effectiveness, potential outcomes, principal strata.*