UNIVERSITÀ DEGLI STUDI DI FIRENZE
Dipartimento di Statistica "G. Parenti"
Dottorato in Statistica Applicata XII ciclo

Sbocchi occupazionali e scelte formative dei diplomati: un'analisi multilivello

Leonardo Grilli

Relatore:
Prof. Luigi Biggeri

Correlatore:
Prof.ssa Carla Rampichini

Coordinatore:
Prof. Giovanni M. Marchetti

Questa tesi deve la sua forma definitiva al contributo di numerose persone. Desidero innanzitutto ringraziare il Prof. Luigi Biggeri per la guida sicura e le osservazioni puntuali, e la Prof.ssa Carla Rampichini, che mi ha seguito con dedizione e competenza. Desidero inoltre ringraziare tutti coloro che hanno letto parti della tesi e fornito utili suggerimenti, fra cui il Prof. Bruno Chiandotto, la Prof.ssa Fabrizia Mealli e la Dott.ssa Matilde Bini. A quest'ultima va un ringraziamento speciale, per l'entusiasmo con cui mi ha aiutato a risolvere tanti piccoli problemi.

Un contributo essenziale a questa tesi è stato fornito dal Servizio Istruzione-Formazione-Lavoro dell'Istat, che ha reso disponibili, prima della pubblicazione, numerose informazioni relative all'indagine sui Percorsi di Studio e Lavoro dei Diplomati. Un ringraziamento particolare va poi al Dott. Marco Centra dell'Isfol, che ha costruito, su mia richiesta, la serie storica di ingressi al lavoro dipendente impiegata nell'analisi presentata nel quinto capitolo.

La tesi di Dottorato conclude un intenso periodo di formazione che mi ha introdotto al mondo della ricerca universitaria. Questa esperienza è stata resa proficua dalla guida sicura dei coordinatori Prof. Bruno Chiandotto e Prof. Giovanni M. Marchetti, dal contributo di conoscenze ed esperienza apportato da tutti coloro che hanno tenuto i corsi di Dottorato e dalla collaborazione continua, sempre in un clima sereno ed amichevole, con i miei compagni Leonardo Fabbroni, Andrea Mercatanti, Andrea Neri, Alessandro Rossi e Simona Toti. Un sincero grazie va infine al personale del Dipartimento di Statistica "G. Parenti", che si è sempre prodigato con competenza e cortesia, contribuendo a rendere piacevole questa esperienza di studio e ricerca.

Leonardo Grilli

Firenze, 30 dicembre 1999

Indice

Premessa

Le scelte formative e gli esiti occupazionali dei giovani diplomati, per le loro implicazioni di carattere sociale ed economico, costituiscono da sempre un importante argomento di studio e discussione. Nell'ultimo decennio l'interesse nei confronti di questi fenomeni è addirittura aumentato, sia per le dimensioni preoccupanti assunte dalla disoccupazione giovanile, che per la crescente attenzione nei confronti della valutazione del sistema formativo in un'ottica di efficienza ed efficacia. Tale valutazione risulta, infatti, di estremo interesse al fine di garantire gli investimenti pubblici e privati, i diritti dell'utenza diretta, cioè degli studenti e delle loro famiglie, e quelli dell'utenza indiretta, cioè dei datori di lavoro che sono interessati alle capacità che i giovani hanno acquisito nel periodo formativo.

Nel presente lavoro ci proponiamo di individuare e valutare i fattori che determinano le scelte formative e gli esiti occupazionali dei diplomati, con particolare riferimento ai fattori di contesto associati alla scuola frequentata e all'area di residenza, anche nell'ottica di valutare l'efficacia delle scuole relativamente al successo nel mondo del lavoro dei rispettivi diplomati.

Il perseguimento di tali obiettivi richiede innanzitutto l'utilizzo di informazioni statistiche adeguate, sotto forma di dati individuali. L'analisi non sarebbe pertanto stata possibile senza la disponibilità del file standard di dati individuali dell'indagine campionaria sui Percorsi di Studio e Lavoro dei Diplomati (PSLD) svolta dall'Istat nel 1998.

Le finalità della nostra analisi richiedono inoltre l'impiego di un'appropriata metodologia statistica, atta a stimare gli effetti netti esercitati dai vari fattori sulle scelte dei giovani e ad effettuare confronti validi fra le scuole. Nel presente contesto una metodologia adatta allo scopo è quella dell'analisi multilivello (Goldstein, 1995), che considera i diplomati come unità elementari

e le scuole come unità di livello superiore. Le analisi multilivello si sono enormemente sviluppate a partire dagli anni Ottanta, trovando applicazioni in numerosi ambiti, soprattutto fra le scienze sociali. Nel campo specifico della valutazione degli istituti scolastici, i modelli multilivello sono stati largamente usati, soprattutto nei paesi anglosassoni, per determinare misure di efficacia basate sui voti di diploma o sui punteggi forniti da test standardizzati (Goldstein & Spiegelhalter, 1996). In Italia le rare applicazioni di tali modelli hanno fatto riferimento ad altre misure di efficacia, come il completamento del ciclo di studi (Montagni, 1997), il successo nella prosecuzione degli studi (Gori, 1992) o l'ottenimento di un lavoro (Bini, 1999; Biggeri *et al.*, 1999), utilizzando sempre dati relativi a studenti universitari o laureati. Come detto, nella presente ricerca si affronta invece la questione relativamente ai diplomati della scuola secondaria superiore. Comunque il nostro lavoro si diversifica non soltanto per il campo di applicazione, certamente più complesso di quelli prima menzionati, ma anche per l'approfondimento di alcune questioni metodologiche e tecniche inerenti la specificazione e la stima dei modelli multilivello, l'analisi dei residui in presenza di una componente di varianza modesta e la correzione delle distorsioni delle stime indotte dal piano di campionamento.

La presente ricerca è così strutturata. Inizieremo (capitolo 1) proponendo una classificazione dei fattori che influenzano la transizione scuola-lavoro, facendo riferimento sia alla teoria che all'evidenza empirica. Quindi illustreremo brevemente gli studi di natura statistica sull'argomento e alcune apposite indagini condotte in Italia, rimandando all'appendice A per un quadro della situazione italiana in merito all'istruzione superiore e universitaria e al mercato del lavoro giovanile. Infine, metteremo in evidenza come un'analisi statistica adeguata dei fenomeni di interesse richieda l'impiego di appropriati modelli multilivello.

Per questo motivo la parte successiva del lavoro (capitolo 2), di carattere metodologico, sarà dedicata ai modelli multilivello. Al fine di presentare i concetti fondamentali inizieremo con la descrizione dei modelli lineari. Poi proporremo un'originale trattazione dei modelli non lineari per variabili di risposta categoriche (inclusi i modelli di sopravvivenza in tempo discreto) che useremo nelle analisi successive, evidenziando, anche ricorrendo ai concetti di variabile latente e soglia, gli aspetti teorici e computazionali che accomunano tutti questi modelli. I metodi di stima di cui ci avvarremo per le analisi sono la Quasi-Verosimiglianza Penalizzata (Goldstein & Rasbash, 1996) e la Massima Verosimiglianza con integrazione numerica (Hedeker & Gibbons, 1994): il primo metodo, computazionalmente efficiente, verrà usato per la selezione del modello, mentre il secondo, più preciso, servirà per convalidare le stime finali.

Collegato a questo capitolo è il materiale presentato nell'appendice B, contenente i risultati delle simulazioni che abbiamo condotto per valutare alcuni aspetti, sinora trascurati dalla letteratura, relativi ai residui del modello logit a intercetta casuale. In particolare si dimostra che, quando la componente di varianza è modesta, i residui sono scarsamente affidabili come misure di efficacia.

Nei capitoli successivi presenteremo la parte empirica di questa ricerca, dedicata a tre aspetti fondamentali dell'esperienza post-diploma: (a) l'analisi della probabilità di occupazione; (b) l'analisi dei tempi di ingresso al lavoro; (c) l'analisi della probabilità di immatricolazione all'università.

Il terzo capitolo si aprirà pertanto con una descrizione dettagliata dell'indagine da cui abbiamo attinto gran parte dei dati usati nel presente lavoro: si tratta dell'indagine sui Percorsi di Studio e Lavoro dei Diplomati (PSLD) del 1995 condotta dall'Istat nel 1998, che, a parte l'esperienza dell'indagine EVA dell'Isfol, è la prima indagine a livello nazionale relativa ai diplomati della scuola secondaria superiore. Il campione, selezionato con uno schema a due stadi stratificato, si compone di 18843 diplomati in 1563 scuole e fornisce una notevole quantità di informazioni sui diplomati relativamente a: notizie anagrafiche, background familiare, curriculum degli studi fino alla maturità, studi universitari, ricerca del lavoro, lavoro svolto al momento dell'intervista.

Sulla base delle informazioni disponibili, nella seconda parte del capitolo proporremo uno schema generale di analisi e definiremo le variabili che verranno impiegate nelle applicazioni presentate nei capitoli 4-6. A questo proposito va menzionata l'inclusione di variabili, tratte da altre fonti Istat, relative alla situazione del mercato del lavoro a livello regionale. L'inserimento di tali variabili, fortemente consigliato dalla teoria, è raro nelle applicazioni, in quanto richiede la disponibilità di una sufficiente disaggregazione dei dati a livello territoriale. La valutazione dell'eventuale importanza di fattori come il tasso di disoccupazione giovanile o la proporzione di occupati nel terziario ai fini della possibilità di trovare o meno un lavoro rappresenta dunque un aspetto degno di nota della nostra ricerca.

L'analisi della probabilità di svolgere un lavoro continuativo al momento dell'intervista costituirà l'oggetto del capitolo 4. Il primo problema affrontato riguarda la definizione di un appropriato sottocampione sul quale condurre l'analisi. Successivamente presenteremo i risultati ottenuti con un modello logit a intercetta casuale che include sia variabili relative al diplomato (caratteristiche anagrafiche, background familiare, esperienze di studio e di lavoro) che variabili relative alla scuola (tipo di scuola e tasso di disoccupazione della regione in cui è ubicata la scuola). La varianza residua attribuibile alle scuole risulterà piccola ma di rilevanza pratica, motivando l'uso dei residui a livello di scuola per la valutazione dell'efficacia. Vedremo, tuttavia, che

la variabilità legata alla stima dei residui è tale da rendere impossibili dei confronti statisticamente significativi.

L'analisi svolta nella prima parte del capitolo 4 si basa su stime che non tengono conto del fatto che i diplomati campione sono stati selezionati secondo un complesso schema a due stadi, in cui le scuole (unità di primo stadio) sono state stratificate per regione e tipo di insegnamento e selezionate con probabilità differenziate a seconda dello strato di appartenenza e del numero di maturi. Abbiamo perciò ritenuto importante cercare di verificare l'effetto del piano di campionamento sulle stime, che è un problema rilevante e poco trattato in letteratura, specialmente per quanto riguarda i modelli multilivello. Dopo aver descritto il piano di campionamento dell'indagine PSLD, sperimenteremo due procedure di stima pesata volte a correggere le distorsioni degli stimatori. La prima procedura consiste in un'applicazione diretta, ma parziale del principio della Pseudo-Massima Verosimiglianza, proposto da Skinner (1989) per i modelli di regressione ordinari. La seconda procedura consiste, invece, nell'estensione al modello logit a intercetta casuale del metodo approssimato proposto da Pfeffermann *et al.* (1998) per il modello multilivello lineare. Quest'ultima procedura è particolarmente interessante, poiché permette di inserire come pesi i reciproci delle probabilità di inclusione di entrambi gli stadi di campionamento; essendo questa la prima applicazione ad un modello non lineare, la sua validità verrà verificata confrontando i risultati con quelli forniti dall'altra procedura. Infine, vedremo che le differenze fra stime standard e stime pesate, sebbene rilevanti per alcuni parametri, non sono tali da modificare il quadro d'insieme.

Il capitolo 5 è dedicato all'analisi dell'aspetto temporale dell'inserimento lavorativo dei diplomati, assumendo come variabile di risposta il tempo in mesi intercorrente fra il conseguimento del diploma e l'ottenimento del primo lavoro continuativo. Nella prima parte esamineremo la stima non parametrica della funzione di rischio per alcune tipologie di diplomati, individuando alcuni andamenti caratteristici. Inoltre valuteremo il ruolo svolto dalla domanda di lavoro tramite la comparazione con una serie storica di ingressi al lavoro che l'Isfol ha derivato, su nostra richiesta, dai panel trimestrali dell'indagine Istat sulle Forze di Lavoro. La disponibilità di questa serie ci consentirà di affermare che la ripresa della probabilità di occupazione dopo due anni dal conseguimento del diploma, riscontrata nei dati PSLD, è dovuta solo in piccola parte alla domanda di lavoro e quindi è per lo più imputabile a dinamiche legate alla coorte dei diplomati.

Al fine di invidiuare un appropriato modello per i tempi di ingresso al lavoro che tenga conto della struttura gerarchica dei dati, nella seconda parte del capitolo 5 sperimenteremo una serie di modelli di sopravvivenza in tempo discreto a intercetta casuale, partendo dal modello di Cox a rischi

proporzionali per dati raggruppati, del quale esistono due versioni equivalenti, dette *grouped continuous* (McCullagh, 1980) e *continuation ratio* (Prentice & Gloeckler, 1978). Il confronto di questi due modelli, e delle loro varianti, in presenza di effetti casuali non è documentata in letteratura. La necessità di tener conto di rischi non proporzionali motiverà l'utilizzo di alcune varianti, che consisteranno nell'inserimento di interazioni fra parametri di soglia e covariate per il modello *grouped continuous* e nell'inclusione di covariate tempo-dipendenti per il modello *continuation ratio*. Nella nostra applicazione il secondo modello risulterà chiaramente preferibile, poiché è in grado di incorporare una gran varietà di andamenti del rischio con un numero di parametri relativamente modesto. La serie storica di ingressi al lavoro calcolata dall'Isfol, a cui abbiamo accennato poco sopra, è stata inserita come variabile tempo-dipendente, risultando non significativa e confermando quindi le impressioni tratte dall'analisi descrittiva della prima parte del capitolo.

Infine il cap. 6 sarà dedicato all'analisi della probabilità di immatricolazione all'università nei tre anni successivi al conseguimento del diploma. Rispetto alle analisi dei capitoli precedenti il numero di dati mancanti per alcune variabili di interesse è più rilevante, suggerendo un appropriato trattamento basato sull'imputazione casuale condizionata. Il modello logit selezionato presenterà un coefficiente casuale per la variabile indicatrice del sesso, ammettendo così una diversa variabilità delle scuole relativamente ai maschi e alle femmine. Nell'ultima parte del capitolo tracceremo un interessante confronto fra i risultati delle analisi della probabilità di occupazione e di immatricolazione.

Concluderemo il lavoro con alcune considerazioni sui principali risultati empirici e metodologici conseguiti.

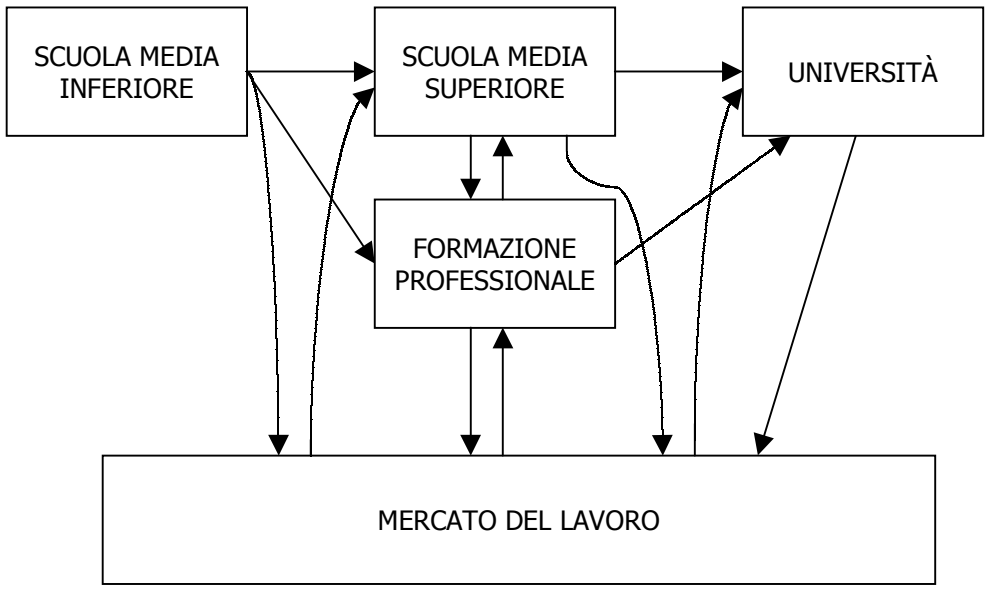
Capitolo 1

Scelte formative e inserimento professionale dei diplomati

Il cammino formativo dei giovani e la transizione verso il mondo del lavoro sono fenomeni estremamente complessi, che possono essere analizzati con i metodi e per i fini più diversi. Come mostra la fig. 1.1, una volta terminata la scuola dell'obbligo il giovane può decidere in qualunque momento di passare dallo studio al lavoro e viceversa. In molti casi l'ingresso nel mondo del lavoro si compie in un periodo più o meno lungo, caratterizzato da lavori precari e da ripresa dell'attività formativa. Inoltre, soprattutto quando il giovane diventa adulto, la permanenza nel sistema formativo può essere associata a un'attività lavorativa parallela.

In questo capitolo intendiamo fornire un quadro teorico generale per l'analisi delle scelte formative e dell'inserimento professionale dei giovani, considerando in particolare il caso dei diplomati della scuola secondaria superiore, che costituisce l'oggetto di studio della presente ricerca. Dopo una descrizione sintetica dei fattori che determinano le scelte e i successi dei giovani nell'ambito della formazione e del lavoro, presenteremo una breve rassegna degli studi di natura statistica sull'argomento, classificati in base alle finalità dell'analisi. Successivamente esamineremo gli aspetti metodologici delle indagini sulla transizione scuola-lavoro e forniremo alcune informazioni sulle indagini di questo tipo condotte in Italia. Concluderemo il capitolo con alcune considerazioni sulle motivazioni che ci hanno indotto ad effettuare le analisi impiegando i modelli multilivello.

Fig. 1.1 - Percorsi di studio e inserimento professionale



Fonte: Bini (1999)

1.1 I fattori che determinano le scelte formative e l'inserimento professionale

L'esame della vasta letteratura sui percorsi di studio e gli sbocchi occupazionali dei giovani, siano essi licenziati dalla scuola secondaria inferiore, diplomati o laureati, permette di fare un quadro generale dei fattori che influenzano le decisioni e i successi o insuccessi dei giovani in merito a formazione e lavoro. Di seguito presentiamo i pricipali fattori, distinguendoli in *individuali*, *micro-contestuali* e *macrocontestuali*[1]. Per ognuna di queste categorie proporremo un elenco commentato dei fattori legati all'ingresso nel mondo del lavoro, poiché è questo l'aspetto che più interessa nella presente tesi, concludendo con una nota sui fattori che determinano le scelte formative.

1.1.1 Fattori individuali

I principali fattori individuali che determinano il successo dei giovani nel mondo del lavoro sono:

- Sesso: la quasi totalità degli studi segnala una discriminazione a scapito del sesso femminile;
- Razza: questo fattore è rilevante nei Paesi multietnici, come gli Stati Uniti d'America, dove si rileva una condizione di vantaggio a favore dei bianchi (Coleman, 1984);
- Età: spesso l'inserimento professionale si rende più difficile per i più anziani;
- Condizioni fisiche e mentali: la presenza di un handicap o di cattive condizioni psico-fisiche costituisce un ostacolo all'ottenimento del lavoro;
- Abilità: la teoria suggerisce che questo fattore, solitamente rappresentato dal voto finale o dai risultati di test standardizzati, abbia un effetto

[1] Si dicono *contestuali* tutti quei fattori propri dell'ambiente in cui l'individuo vive e che hanno un effetto sui risultati della sua azione. Proponiamo una distinzione di tali fattori in *micro-* e *macrocontestuali*: i primi si riferiscono all'ambiente familiare, scolastico e socio-economico della zona di residenza e sono rilevanti in tutte le analisi sui comportamenti e i destini dei giovani; i secondi, invece, riguardano un contesto più ampio, regionale o nazionale, e sono importanti soprattutto nelle comparazioni internazionali.

positivo; in realtà in alcuni lavori si evidenzia un effetto positivo, seppur debole (Biggeri *et al.*, 1999), mentre in altri emerge addirittura un effetto negativo (Santoro & Pisati, 1996, cap. 3)[2];

- Titolo di studio: solitamente l'ingresso nel mondo del lavoro è più immediato per coloro che hanno una preparazione di tipo tecnico (Fiorito, 1981);

- Servizio militare: l'adempimento degli obblighi di leva dopo il conseguimento del titolo posticipa la ricerca del lavoro, generando delle conseguenze spesso durature (Biggeri *et al.*, 1999);

- Esperienze lavorative: le esperienze lavorative durante gli studi favoriscono l'ingresso nel mondo del lavoro, poiché, oltre ad essere valutate positivamente dalle imprese, permettono al giovane di instaurare dei contatti che risultano utili nel momento della ricerca (Coleman, 1984; Santoro & Pisati, 1996, cap. 3).

Quanto alla scelta di proseguire gli studi, questa è strettamente connessa al successo o all'aspettativa di successo nel mondo del lavoro, per cui molti fattori sono comuni. Un fattore importante nella scelta di continuare a studiare, ma che non influenza la probabilità di occupazione, è il livello di soddisfazione associato allo studio (O'Higgins, 1992).

1.1.2 Fattori microcontestuali

Per quanto riguarda l'inserimento nel mondo del lavoro da parte dei giovani, i principali fattori microcontestuali sono i seguenti:

- Background familiare (numerosità della famiglia di origine, livello di istruzione e status occupazionale dei genitori, . . .): la famiglia di origine è importante sia perché influenza le capacità, i valori e le aspettative dell'individuo, sia perché può aiutare fattivamente il giovane nella ricerca del lavoro;

- Scuola frequentata: a prescindere dalle materie insegnate, ogni scuola fornisce una preparazione diversa, che dipende dagli insegnanti, dalle strutture, dal clima scolastico, ecc.; particolarmente importante è il contatto della scuola con il mondo del lavoro;

[2] L'effetto negativo del voto può essere spiegato da due fattori: 1) i giovani con i voti migliori tendono a proseguire gli studi, anche se magari dichiarano di cercare lavoro; 2) l'acquisizione di un buon voto crea delle aspettative che inducono a rifiutare lavori giudicati non adeguati.

- Zona di residenza: le caratteristiche socio-economiche del quartiere, della città o della regione in cui il giovane risiede influiscono, spesso in modo notevole, sul processo di inserimento nel mondo del lavoro (si pensi all'importanza del tasso di disoccupazione locale o della facilità di spostamento).

Questi fattori sono rilevanti anche per la scelta di proseguire o meno gli studi. In questo caso si pone particolare enfasi sul background familiare, considerando anche variabili come il numero di fratelli e il reddito della famiglia. In effetti alcune analisi mostrano che all'aumentare del numero di fratelli si riduce la probabilità di continuare gli studi[3]. Per quanto riguarda il reddito, la teoria suggerisce che le difficoltà economiche possano costituire un vincolo finanziario e quindi ostacolare il proseguimento degli studi; tuttavia l'evidenza empirica a supporto di questa teoria è molto debole (Micklewright, 1989).

1.1.3 Fattori macrocontestuali

I fattori macrocontestuali che influenzano la transizione scuola-lavoro possono essere classificati nel seguente modo (cfr. Bennett, 1995):

- Fattori demografici (tasso di crescita della popolazione, distribuzione per età, ecc.);
- Fattori geografici (economie regionali, barriere alla mobilità, risorse naturali, ecc.):
- Fattori economici (tipo di sistema economico, livello di sviluppo, tasso di crescita, ecc.);
- Fattori politico-istituzionali (sistema legislativo e fiscale, presenza di autonomie locali, appartenenza a organismi sovranazionali, incentivi al lavoro giovanile, standardizzazione e stratificazione del sistema scolastico[4], ecc.);

[3] La spiegazione classica è che il tempo e le risorse che i genitori possono dedicare ad un figlio si riducono progressivamente all'aumentare del numero di figli (Micklewright, 1989).

[4] Un sistema scolastico si dice *standardizzato* quando tutte le scuole del Paese seguono procedure uniformi per l'insegnamento e la valutazione. Inoltre, un sistema scolastico si dice *stratificato* quando l'accesso ai livelli di istruzione superiore è limitato da regole selettive (prove di ingresso, numero chiuso, ecc.). Il sistema italiano può essere classificato come molto standardizzato e poco stratificato. Allmendinger (1989) dimostra che sia la standardizzazione che la stratificazione semplificano la transizione scuola-lavoro, riducendo i tempi di inserimento e il numero di cambiamenti di lavoro nella fase iniziale, oltre a

- Fattori socio-culturali (composizione etnica della popolazione, mobilità sociale, emancipazione femminile, ecc.).

Per quanto riguarda le scelte formative dei giovani, i fattori fondamentali sono il livello di sviluppo economico e culturale e il tipo di sistema scolastico.

1.2 Rassegna degli studi di natura statistica sulle scelte formative e sulla transizione scuola-lavoro

Le scelte e i destini dei giovani rispetto alla formazione e al lavoro interessano una molteplicità di soggetti, fra cui gli studenti e le loro famiglie, le imprese, i politici, gli amministratori locali, nonché i ricercatori in varie discipline dell'area socio-economica. Naturalmente ognuno di questi soggetti vede il problema dal proprio punto di vista ed è interessato a determinate risposte, contribuendo a generare una varietà di studi caratterizzati da tagli diversi.

Le metodologie di analisi impiegate vanno dal calcolo di semplici statistiche descrittive, a tecniche di analisi multivariata (ad es. Tronti & Mariani, 1994), a modelli di transizione a stati discreti (ad es. Trivellato & Bernardi, 1994; D'Agostino, 1998). Tuttavia lo strumento fondamentale è senz'altro il modello di regressione, nelle sue molte varianti: infatti oltre ai modelli di regressione classici, trovano largo uso i modelli per dati di durata e i modelli multilivello.

Di seguito presentiamo una breve rassegna delle analisi di natura statistica sulle scelte formative e sulla transizione scuola-lavoro, delineando una classificazione basata non sulla metodologia adottata ma sulle finalità dell'indagine. Poiché le finalità possono essere molteplici, molti lavori hanno una collocazione intermedia.

1.2.1 Gli studi politico-istituzionali

I percorsi di studio dei giovani e la transizione scuola-lavoro rivestono un notevole interesse per l'intera collettività, per cui molti organismi nazionali e internazionali si occupano della questione, sia per rendere pubblicamente noti i dati essenziali, sia per fornire strumenti informativi a supporto delle decisioni di politica economica.

favorire l'acquisizione da parte dei giovani con titolo elevato di posizioni prestigiose e remunerative.

A livello internazionale ricordiamo l'attività dell'Organizzazione per la Cooperazione e lo Sviluppo Economico (ad es. OECD, 1997), finalizzata alla produzione di statistiche che permettano il confronto fra gli stati membri.

A livello nazionale la questione è affrontata da molti organismi: oltre ai competenti Ministeri, segnaliamo l'attività di Istat, Isfol (Istituto per lo Sviluppo e la Formazione dei Lavoratori), FORMEZ, Fondazione G. Brodolini, Osservatorio per la Valutazione del Sistema Universitario. Inoltre rivestono un ruolo non secondario le molte analisi statistiche condotte dagli enti locali, in particolare dalle Regioni. Una rassegna completa degli studi sulla transizione scuola-lavoro in Italia è riportata da Zaccarin (1994) e Bini (1999).

Solitamente le analisi svolte da questi organismi sono di tipo descrittivo o si avvalgono di semplici modelli di regressione, al fine favorirne la fruizione da parte di un pubblico con limitate conoscenze tecniche.

1.2.2 Gli studi sociologici

Il comportamento dei giovani nel momento in cui terminano un ciclo di studi è un classico argomento di interesse per i sociologi, poiché l'azione del giovane va posta in relazione con le strutture sociali nelle quali opera. Un esempio tipico di questo approccio è dato dall'articolo di Layder *et al.* (1991), nel quale gli autori verificano empiricamente l'importanza relativa delle variabili strutturali e individuali nella fase di ricerca del lavoro da parte dei giovani. Le variabili vengono classificate secondo il seguente schema, in ordine crescente di controllo da parte dell'individuo:

1. Variabili strutturali (sesso, classe sociale, luogo di residenza, condizioni del mercato del lavoro locale);
2. Variabili individuali:
   (a) Credenziali educative (titolo, voto);
   (b) Attitudini (aspettative relative al lavoro, disponibilità a viaggiare);
   (c) Comportamento (storia lavorativa pregressa, modalità di ricerca del lavoro).

Gli autori effettuano la verifica per mezzo di un modello di Cox a rischi concorrenti, in modo da determinare l'influenza delle variabili sulle probabilità di ingresso nei vari segmenti del mercato del lavoro (dirigenti, impiegati,

operai, ecc.). La conclusione a cui giungono gli autori è che le variabili individuali hanno un peso relativamente maggiore nel determinare l'ingresso nei segmenti più alti.

Molti lavori si concentrano su particolari fattori che influenzano il successo nel mondo del lavoro, quali il sesso (Dex, 1989), la razza (Coleman, 1984), il background familiare (Rees & Gray, 1982), il sistema educativo (Allmendinger, 1989).

Per quanto riguarda le scelte formative dei giovani, sono pochi gli studi che includono un'analisi statistica (ad es. Mare, 1980).

Generalmente gli studi in ambito sociologico sono caratterizzati dall'analisi di dati individuali per mezzo di metodi descrittivi o di modelli di regressione, talvolta anche complessi (come il modello di Cox a rischi concorrenti utilizzato da Layder *et al.* (1991)).

1.2.3 Gli studi economici

Le scelte formative dei giovani e la transizione scuola-lavoro sono tipici argomenti di studio dei microeconomisti, che si traducono in verifiche empiriche di tipo microeconometrico.

I microeconomisti sono interessati ai fattori di tipo economico che sottendono le scelte dei giovani al momento in cui terminano un ciclo di studi. L'aspetto che caratterizza questi studi è il ruolo chiave giocato dalle aspettative: infatti si ipotizza che i giovani compiano le loro scelte massimizzando una funzione di utilità che dipende in modo decisivo dalle aspettative di occupazione e di reddito. Un semplice modello (ad es. Willis & Rosen, 1979) è quello che fa dipendere la decisione di continuare o meno gli studi dai flussi attualizzati di reddito che si otterrebbero nei due casi: indicando con s la durata degli studi addizionali e con $Y_t^{(p)}$ e $Y_t^{(i)}$ $(t = 0, 1, \ldots)$ i redditi al tempo t nel caso, rispettivamente, di prosecuzione e interruzione degli studi, si assume che l'individuo scelga di proseguire gli studi quando

$$\sum_{t=s}^{\infty} \frac{E(Y_t^{(p)})}{(1+r)^t} > \sum_{t=0}^{\infty} \frac{E(Y_t^{(i)})}{(1+r)^t},$$

dove $E(\cdot)$ rappresenta le aspettative dell'individuo, mentre r è il tasso di sconto. In sostanza, l'individuo resta nel sistema educativo solo se i maggiori redditi derivanti da un'istruzione di tipo superiore sono in grado di ripagare la mancaza di redditi durante il periodo degli ulteriori studi. Willis & Rosen incorporano questa idea in un modello strutturale per la determinazione dei redditi e la decisione di proseguire o meno gli studi. Il modello può essere

complicato in vari modi, tuttavia, ai fini della verifica empirica, il punto critico rimane la misurazione delle aspettative (Micklewright, 1988; O'Higgins, 1992).

Per quanto riguarda l'inserimento professionale, in ambito microeconomico spesso si pone l'attenzione sui fattori che sottendono le scelte del giovane, in particolare quelle riguardanti la partecipazione alla forza lavoro e l'accettazione di eventuali offerte di lavoro. Un approccio tipico consiste nello studio delle transizioni tra gli stati di occupato, disoccupato e inattivo (non forza lavoro) (Ordine, 1992). Altre analisi mirano a individuare i fattori che determinano la durata della ricerca del lavoro (Torelli & Trivellato, 1988), spesso puntando l'attenzione sul ruolo svolto dalla durata della permanenza nello stato di disoccupazione (Lynch, 1985).

Una caratteristica peculiare degli studi microeconomici risiede nel fatto che la formazione viene solitamente valutata in termini di reddito (ad es. Hartog *et al.*, 1989), piuttosto che in base alla probabilità di occupazione o al prestigio del lavoro.

I metodi di analisi impiegati in questo ambito sono i più diversi e includono anche modelli sofisticati, come i modelli markoviani a stati discreti e i modelli a equazioni simultanee.

1.2.4 Gli studi di valutazione

I destini di diplomati e laureati, oltre ad avere un interesse di per sé, possono anche essere visti come un indicatore della qualità del servizio offerto dalle scuole secondarie superiori e dall'università. L'idea della valutazione dell'istruzione in Italia sta muovendo i primi passi nell'ambito del sistema universitario, grazie all'Osservatorio per la Valutazione del Sistema Universitario del Ministero per l'Università e la Ricerca Scientifica e Tecnologica[5]. Poiché si tratta di un ambito di ricerca relativamente nuovo, riteniamo utile delinearne i tratti salienti.

In termini di teoria economica la questione può ess\ere rappresentata dal modello Principale-Agente-Utente (Fabbri *et al.*, 1996), che nel caso presente si identificano, rispettivamente, con il Ministero della Pubblica Istruzione, la scuola secondaria superiore e lo studente. Questi soggetti si trovano in una condizione di asimmetria informativa e necessitano ognuno di un diverso tipo di valutazione del servizio offerto dall'Agente. Il punto cruciale è che l'istruzione appartiene alla categoria degli *experience goods* (Fabbri *et al.*, 1996), cioè è un servizio per il quale non è possibile definire un prodotto

[5]Informazioni sull'attività dell'Osservatorio possono essere ottenute consultanto il sito Internet www.murst.it/osservatorio/.

indipendente dall'Utente e quindi la valutazione è strettamente connessa agli effetti che l'azione produce sull'Utente stesso: il problema è che, a parità di qualità del servizio offerto, gli effetti differiscono in modo notevole a seconda delle caratteristiche dell'Utente (si pensi alla diversità dei risultati ottenuti dagli studenti di una stessa classe).

Nell'ambito della valutazione dell'istruzione il concetto chiave è quello di *efficacia*, che concerne il contributo fornito dalla singola scuola al raggiungimento delle finalità proprie dell'istruzione. Naturalmente tali finalità sono molteplici: acquisizione di conoscenze, maturazione intellettuale ed emotiva, preparazione per gli studi successivi, ottenimento di un buon lavoro, percezione di un reddito soddisfacente, ecc. Ognuna di queste finalità conduce ad una diverso concetto di efficacia. Gran parte della ricerca è dedita allo studio dell'*efficacia interna*, che fa riferimento soprattutto ai livelli di apprendimento, solitamente misurati per mezzo di test standardizzati (si veda per tutti il classico lavoro di Aitkin & Longford, 1986). Meno studiata è invece l'*efficacia esterna*, cioè relativa al successo degli studenti nel prosieguo degli studi o nel mondo del lavoro (uno dei primi lavori sull'argomento è quello di Gori (1992), che propone una valutazione della scuola secondaria superiore basata sulla probabilità di laurea dei maturi che si iscrivono all'università).

Una distinzione importante è quella tra *efficacia assoluta* (valutazione di impatto) ed *efficacia relativa*. Nel primo caso si tratta di valutare l'azione dell'Agente confrontando la condizione dell'Utente dopo la fruizione del servizio con la sua ipotetica condizione in assenza del servizio. Solitamente si fa riferimento all'efficacia assoluta quando esiste un unico agente o comunque quando si voglia valutare l'impatto dell'azione (un esempio tipico è quello della valutazione dei corsi di formazione professionale: Mansky & Garfinkel, 1992). Si parla invece di *efficacia relativa* quando si vogliano confrontare più Agenti che offrono lo stesso servizio, come accade nella maggior parte dei casi.

Poiché, come abbiamo accennato, la valutazione dell'istruzione è inscindibilmente legata alle caratteristiche dell'Utente e ogni Agente può avere un bacino di utenza con caratteristiche peculiari, la comparazione degli Agenti necessita di un processo di aggiustamento, in modo tale che il confronto sia *ceteris paribus* (Goldstein & Spiegelhalter, 1996). Il metodo classico per effettuare l'aggiustamento è l'analisi di regressione. Tuttavia, come già rilevato da Aitkin & Longford (1986), il modello di regressione ordinario non è adatto allo scopo, poiché trascura la correlazione esistente fra i risultati degli Utenti serviti dallo stesso Agente. Una soluzione al problema è rappresentata dai modelli multilivello (cfr. cap. 2), che oltretutto consentono una modellizzazione esplicita del concetto di efficacia (Raudenbush & Willms, 1995; Goldstein & Spiegelhalter, 1996).

Un altro aspetto caratteristico della valutazione dell'istruzione sta nel fatto che i concetti di efficacia rilevanti per l'Utente e per il Principale sono diversi. Infatti l'Utente è semplicemente interessato a confrontare i risultati che può ottenere ricorrendo ai vari Agenti, indipendentemente da come questi risultati si originano, mentre il Principale è interessato anche al processo produttivo in sé, al fine di valutare la capacità degli Agenti di sfruttare le risorse che hanno a disposizione. Usando la terminologia di Willms (1992) possiamo quindi distinguere due tipi di efficacia:

A) Efficacia di Tipo A (rilevante per l'Utente): riguarda la performance degli Agenti aggiustata per le caratteristiche dell'Utente;

B) Efficacia di Tipo B (rilevante per il Principale): riguarda la performance degli Agenti aggiustata non solo per le caratteristiche dell'Utente, ma anche per quelle dell'Agente stesso (dimensioni, composizione del personale, procedure di lavoro, risorse, ecc.), oltre che per le condizioni ambientali (fattori socio-economici che influenzano il risultato).

I modelli multilivello sono lo strumento ideale per studiare entrambi i tipi di efficacia. Tuttavia è bene sottolineare che la valutazione dell'efficacia, specie di Tipo B, è un'operazione molto complessa, che richiede l'impiego di un gran numero di variabili spesso di difficile misurazione, il cui risultato deve essere interpretato con le dovute cautele (Raudenbush & Willms, 1995). Senza contare il fatto che i risultati della valutazione non suggeriscono di per sé quale azione il Principale debba intraprendere, lasciando sul campo una questione di difficile soluzione.

1.3 Le indagini sulla transizione scuola-lavoro

Concludiamo questo capitolo introduttivo esaminando gli aspetti metodologici delle indagini sulla transizione scuola-lavoro e fornendo alcune informazioni sulle indagini di questo tipo condotte in Italia.

1.3.1 Aspetti metodologici

La progettazione di un'indagine sulla transizione scuola-lavoro è un'operazione alquanto complessa perché il fenomeno di interesse ha una natura dinamica e molte sfaccettature. In particolare risulta difficile seguire il giovane in tutti i suoi cambiamenti di stato, non solo perché tali cambiamenti sono frequenti, ma anche perché gli stati hanno spesso i contorni sfumati (una situazione tipica è quella del giovane che studia e lavora). Inoltre, il fine dell'indagine

solitamente non è limitato alla descrizione dei percorsi, ma include anche la determinazione delle relazioni con il background familiare e l'iter formativo. Una tale quantità di informazioni può essere ottenuta solo con un'indagine *ad hoc* basata su un questionario da sottoporre direttamente al giovane.

Il raggiungimento dei fini dell'indagine presuppone la considerazione e la definizione di una serie di elementi relativi al disegno e alla conduzione dell'indagine (Zaccarin, 1994):

a) Popolazione oggetto di studio. La popolazione giovanile può essere individuata: 1) scegliendo una classe di età; oppure 2) scegliendo una leva diplomati. Nella maggior parte delle indagini viene preferita la seconda opzione, poiché l'anno di ottenimento del titolo è, per molti fini, più rilevante dell'età anagrafica.

b) Tipo di indagine. L'indagine può essere *completa* o *campionaria*. Naturalmente le indagini a livello nazionale, dato il costo e la numerosità dell'universo, sono sempre campionarie. Poiché gli elenchi dei diplomati vengono ottenuti direttamente dalle scuole, la procedura di campionamento solitamente è a due stadi con qualche forma di stratificazione.

c) Periodo di osservazione. Il raggiungimento da parte del giovane di uno stato relativamente stazionario richiede spesso molti anni. Infatti anche coloro che si mettono sul mercato del lavoro appena ottenuto il diploma tipicamente vanno incontro ad un periodo in cui si alternano disoccupazione, formazione e lavori precari. Molte indagini adottano un periodo di osservazione di tre anni, che rappresenta un compromesso fra l'esigenza di determinare con buona approssimazione il successo nel mondo del lavoro o negli studi universitari e l'esigenza di avere informazioni non obsolete[6].

d) Procedura di osservazione. La raccolta di dati relativi ad un intervallo temporale può avvenire mediante due procedure fondamentali: 1) osservazione *retrospettiva* mediante un'unica rilevazione; oppure 2) osservazioni *ripetute* in più istanti temporali sugli stessi individui (indagini *longitudinali* o *panels*). La procedura retrospettiva è più semplice e meno costosa, anche se presenta dei seri limiti: infatti la collocazione temporale degli eventi da parte dell'intervistato è un'operazione molto delicata, che riduce l'affidabilità dei dati. Inoltre, proprio per garantire un'affidabilità minima, il questionario deve contenere poche e semplici domande sulle dinamiche temporali degli eventi. Ciò rende impraticabile la costruzione di un *career record* individuale, cioè di una sequenza ordinata di *stati* e di *durate* necessaria per un'analisi

[6]Tutte le indagini sull'entrata nella vita attiva rischiano, per loro natura, di fornire semplicemente un quadro storico relativo alla leva esaminata. Infatti, la necessità di avere un lungo periodo di osservazione e la rapidità dei cambiamenti della realtà sociale ed economica possono far sì che i risultati dell'indagine siano di poca utilità per i giovani che ottengono il diploma nel momento in cui tali risultati sono disponibili.

dettagliata dei percorsi (Bernardi & Trivellato, 1986). Generalmente nelle indagini retrospettive si pongono solo alcune domande che mirano a conoscere lo *stato* dopo un certo intervallo, oppure la *durata* fra due eventi (ad esempio, fra il conseguimento del diploma e l'ottenimento del lavoro).

e) Metodo di rilevazione. Il questionario può essere sottoposto al giovane per mezzo di tre metodi principali: 1) intervista diretta; 2) invio del questionario tramite posta; 3) intervista telefonica. L'intervista diretta, adottata nell'indagine EVA, è certamente il metodo più affidabile, ma anche il più costoso, perché richiede l'addestramento e la retribuzione di un cospicuo numero di intervistatori. L'invio del questionario a mezzo posta è invece molto semplice ed economico, ma presenta generalmente elevati tassi di errore e di non risposta. Una soluzione intermedia è rappresentata dall'intervista telefonica, che, specie nella versione CATI[7] adottata nell'indagine PSLD, può fornire ottimi risultati in termini di rapporto qualità/costo.

f) Contenuti informativi. A seconda del fine dell'indagine la struttura del questionario può essere più o meno ampia. Generalmente un'indagine retrospettiva su una leva di diplomati si compone delle seguenti parti:

- *Informazioni sullo studio.* Tipicamente si rileva il tipo scuola frequentata, il voto di maturità, le eventuali ripetenze, ecc. Inoltre, per coloro che proseguono la formazione, si acquisiscono notizie relative a studi universitari, corsi di formazione professionale, stages, ecc.

- *Informazioni sul lavoro.* Sono costituite da un quadro dettagliato del lavoro eventualmente svolto al momento dell'intervista e da alcune indicazioni sulle variazioni dello stato occupazionale avvenute dopo il conseguimento del diploma.

- *Informazioni sulla ricerca del lavoro.* Includono il tempo di inizio della ricerca, le iniziative concrete poste in atto, la disponibilità a svolgere un certo tipo di lavoro, ecc.

- *Informazioni sulla famiglia di origine.* L'attenzione viene posta soprattutto sul livello di istruzione, lo stato occupazionale e l'eventuale condizione professionale dei genitori. Talvolta si rilevano anche informazioni sui fratelli o sui nonni. Invece, data l'età degli intervistati, viene generalmente trascurata l'eventuale famiglia del giovane.

[7] Il sistema CATI (Computer Assisted Telephone Interview) prevede che l'intervista venga effettuata telefonicamente da parte di un operatore che si avvale dell'aiuto di un computer per la formulazione dei quesiti e la registrazione delle risposte. La prima indagine sulla transizione scuola-lavoro che si è avvalsa di questo metodo di rilevazione è quella condotta nel 1991 dalla regione Veneto (Pedenzini & Zaccarin, 1992).

- *Informazioni anagrafiche.* Comprendono alcuni caratteri relativi al giovane, quali il sesso, l'età, la residenza, la posizione nei confronti della leva (per i soli maschi).

1.3.2 Alcune informazioni sulle indagini condotte in Italia

Le fonti statistiche sulla transizione scuola-lavoro in Italia sono numerose (una rassegna è contenuta in Micali, 1993a e 1993b). Tuttavia un'analisi approfondita dei percorsi di studio e lavoro richiede delle indagini *ad hoc*, costruite in modo tale da fornire degli elementi che permettano non solo la valutazione quantitativa del fenomeno, ma anche la comprensione dei processi casuali sottostanti. D'altra parte le indagini *ad hoc* sono relativamente rare, a causa del costo elevato e del fine circoscritto, tant'è che in Italia, fino a due anni fa, esisteva una sola indagine a livello nazionale sui destini dei diplomati della scuola secondaria superiore, quella condotta dall'Isfol nel periodo 1980-1985, nota con il nome di Entrata nella Vita Attiva (EVA).

L'indagine EVA riguardava non solo i diplomati dalla scuola secondaria superiore, ma anche i licenziati dalla scuola media inferiore e i qualificati dalla formazione professionale, tutti intervistati tre anni dopo il conseguimento del titolo. Il questionario era piuttosto articolato e presentava un interessante tentativo di ricostruire un *career record* individuale. Tuttavia la dimensione campionaria era esigua (6000 unità in totale, di cui 3500 riguardanti i diplomi di scuola secondaria superiore). I risultati delle indagini EVA sono pubblicati in Battistoni & Ruberto (1988) e Isfol (1989).

Adesso, a partire dal 1998, si rende disponibile la nuova indagine sui Percorsi di Studio e Lavoro dei Diplomati (PSLD) condotta dall'Istat con cadenza triennale, che si affianca all'ormai classica indagine sugli sbocchi professionali dei laureati (Istat, 1996) e alla nuova indagine sugli sbocchi professionali dei diplomati universitari. L'indagine PSLD, da cui trarremo gran parte dei dati per le elaborazioni, verrà descritta in dettaglio nel par. 3.1.

A livello locale sono state condotte numerose indagini occasionali, ognuna con caratteristiche peculiari, per le quali rimandiamo a Zaccarin (1994) e Bini (1999). Di queste la più imponente è senz'altro l'indagine Longitudinale sull'Entrata nella Vita Attiva della Regione Lombardia, relativa agli studenti di terza media (Ghellini, 1992; D'Agostino, 1998).

1.4 Motivazioni dell'analisi multilivello

Concludiamo questo capitolo intoduttivo con alcune considerazioni sull'opportunità dell'impiego dei modelli multilivello nell'analisi delle esperienze post-diploma.

Come discusso nel par. 1.1 le scelte formative e gli esiti occupazionali dei diplomati sono determinati sia da fattori individuali che da fattori di contesto, fra i quali rivestono particolare interesse quelli legati alla scuola frequentata. Ciò richiede l'utilizzo di una metodologia che consenta di studiare congiuntamente l'effetto dei fattori individuali e di contesto, anche al fine di evidenziare eventuali interazioni.

Il punto cruciale sta nel fatto il fenomeno di interesse può essere studiato per mezzo di dati caratterizzati da una struttura gerarchica, in cui i diplomati costituiscono le unità di primo livello (unità elementari) e le scuole le unità di secondo livello (gruppi). Nel caso dell'indagine PSLD, da cui attingeremo i dati per le analisi presentate in questa ricerca, la struttura gerarchica è rispecchiata nel piano di campionamento a due stadi, che vede le scuole come unità primarie e i diplomati come unità secondarie. In termini statistici l'effetto di questa struttura gerarchica è quello di indurre una correlazione positiva nelle variabili relative agli individui provenienti dalla medesima scuola, che costituiscono gruppi relativamente omogenei[8].

Come è noto, i modelli di regressione ordinari si basano sull'assunzione di indipendenza della variabile di risposta fra tutte le unità del campione. Quando questa ipotesi viene violata, gli stimatori dei parametri del modello sono ancora consistenti, ma non lo sono gli stimatori dei loro errori standard. In particolare, la correlazione positiva indotta dalla struttura gerarchica provoca la sottostima degli errori standard associati alle stime, rendendo inattendibili i risultati delle inferenze basate sui test di ipotesi (cfr. Kreft & DeLeeuw, 1998, par. 1.3.2).

[8]Nel caso dell'indagine PSLD, da cui attingeremo i dati per le analisi presentate in questa ricerca, la struttura gerarchica è rispecchiata nel piano di campionamento a due stadi, che vede le scuole come unità primarie e i diplomati come unità secondarie. Le conseguenze di un siffatto piano di campionamento possono essere viste in termini di *design effect* (cfr. Snijders & Bosker, 1999, par. 3.4): infatti, se in un piano di campionamento a due stadi sono state selezionate n unità secondarie per ogni unità primaria, con una dimensione campionaria totale di N, la dimensione campionaria effettiva è

$$N_{eff} = \frac{N}{1 + (n-1)\rho},$$

dove ρ è il coefficiente di correlazione intraclasse (definito nel par. 2.1.3). Pertanto, quando i gruppi sono molto numerosi o la correlazione intraclasse è elevata, la dimensione campionaria effettiva è sensibilmente inferiore a quella nominale.

Vale la pena di ricordare che l'introduzione nel modello di regressione di un'intercetta distinta per ogni gruppo (*modello di analisi della covarianza*) non risolve il problema della sottostima degli errori standard. Inoltre, il modello di analisi della covarianza presenta i seguenti problemi: (a) i risultati che si ottengono sono relativi ai gruppi presenti nel campione e quindi non sono generalizzabili all'universo dei gruppi; (b) in presenza di molti gruppi il numero di parametri del modello diviene eccessivo; (c) non è possibile, a causa della multicollinearità, inserire una variabile di contesto, cioè che assume un valore costante per tutte le unità dello stesso gruppo. Sulla scelta fra modello di analisi della covarianza e modello multilivello torneremo nel par. 2.1.4.

Qualora l'interesse principale riguardi le relazioni fra le variabili a livello di gruppo, una classica soluzione è quella di aggregare le variabili individuali, calcolando le medie di gruppo, e applicare un modello di regressione ordinario. Questa procedura però, oltre a trascurare una gran quantità di informazioni, produce risultati difficilmente interpretabili, in quanto le relazioni a livello di gruppo possono essere radicalmente diverse da quelle a livello individuale[9].

I modelli multilivello, come vedremo nel prossimo capitolo, rappresentano un'ottima soluzione, in quanto consentono di studiare il fenomeno delle esperienze post-diploma sia a livello di individuo che a livello di scuola, tenendo conto della correlazione intraclasse. E' bene ricordare che esiste un importante approccio alternativo all'analisi dei dati correlati, detto *marginale* o *population average* (Diggle *et al.*, 1994), sui cui spenderemo alcune parole nel par. 2.2.1. Da un punto di vista concettuale i due approcci differiscono nel ruolo assegnato alla correlazione infraclasse: mentre nell'approccio marginale la correlazione viene vista come un fattore di disturbo, nell'approccio multilivello essa è una parte fondamentale del modello, poiché deriva dall'esistenza di fattori a livello di gruppo, alcuni dei quali non osservati (i cosiddetti *effetti casuali*). Nel contesto della presente ricerca l'approccio multilivello è, da un punto di vista teorico, più adeguato, in quanto è naturale assumere che la correlazione fra i risultati conseguiti dai diplomati di una stessa scuola dipenda dal fatto che quegli individui hanno condiviso una serie di fattori relativi alla scuola, alcuni osservati (come il tipo di insegnamento, la natura pubblica o privata dell'istituto ecc.) e altri non osservati (come la qualità dell'insegnamento e delle infrastrutture, il clima scolastico ecc.). Inoltre, sotto opportune condizioni, gli effetti casuali, possono essere interpretati come misure di efficacia (cfr. par. 1.2.4) e stimati per mezzo dei residui di secon-

[9] Questo fenomeno è noto come *distorsione da aggregazione* o *ecological fallacy* (Robinson, 1950).

do livello, rendendo il modello multilivello uno strumento particolarmente idoneo alla valutazione di efficacia delle scuole.

In estrema sintesi, le ragioni della scelta dei modelli multilivello per le analisi della presente ricerca vanno ricercate nel fatto che tali modelli consentono di: (i) tener conto della correlazione (positiva) che interessa i diplomati di una stessa scuola; (ii) effettuare un'analisi sia a livello di individuo che a livello di scuola, valutando anche le eventuali interazioni scuola-individuo; (iii) inserire in modo esplicito, per mezzo degli effetti casuali, i fattori di contesto non osservabili (o non osservati); (iv) valutare le singole scuole, poiché gli effetti casuali, sotto oppurtune ipotesi, sono interpretabili come misure di efficacia.

Capitolo 2

I modelli multilivello

L'analisi dei fenomeni oggetto di studio della presente ricerca, come discusso nel par. 1.4, richiede l'impiego di appropriati modelli multilivello. Per tale motivo il secondo capitolo è dedicato ad una trattazione di carattere metodologico dei modelli multilivello, prima quelli lineari e poi quelli non lineari, ponendo particolare attenzione al caso dei dati categorici. Prima di iniziare ci pare opportuno ricordare che tali modelli si sono sviluppati parallelamente in vari ambiti: disegno degli esperimenti, biostatistica, econometria e, in particolare, scienze sociali. Negli anni Ottanta gli sviluppi specifici dei vari ambiti sono stati in parte ricondotti ad unitarietà da una teoria generale dei modelli multilivello; tuttavia alcune differenze rimangono e ciò trova riscontro nella babele di termini che caratterizza questo settore della statistica. Di seguito proponiamo un elenco di termini usati in ambiti diversi per fare riferimento a modelli identici o comunque simili: modelli multilivello, contestuali, misti (mixed), gerarchici lineari, a coefficienti casuali, a effetti casuali, a componenti di varianza. Nel presente lavoro abbiamo adottato il termine *modello multilivello*, poiché è molto generale ed è largamente usato nelle scienze sociali. La varietà terminologica viene rispecchiata anche dai titoli dei principali libri di testo sull'argomento, che sono (in ordine cronologico):

- Searle, Casella & McCulloch (1992): *Variance Components*;
- Bryk & Raudenbush (1992): *Hierarchical Linear Models*;
- Longford (1993): *Random Coefficient Models*;
- Goldstein (1995): *Multilevel Statistical Models*;
- Kreft & De Leeuw (1998): *Introducing Multilevel Models*;

- Snijders & Bosker (1999): *An Introduction to Basic and Advanced Multilevel Modeling.*

Nel primo paragrafo introdurremo il concetto di modello multilivello lineare, soffermandoci poi sugli apetti tecnici della struttura del modello e della stima dei parametri. Nel secondo paragrafo tratteremo i modelli multilivello non lineari, che costituiscono una estensione tutt'altro che immediata di quelli lineari; in particolare, considerate le finalità del presente lavoro, faremo riferimento ai modelli lineari generalizzati multilivello per dati categorici, inclusi i dati di sopravvivenza in tempo discreto.

2.1 I modelli multilivello lineari

2.1.1 Introduzione

I modelli multilivello nascono dall'esigenza di analizzare dati dotati di *struttura gerarchica*, cioè dati in cui le unità statistiche elementari fanno parte di unità statistiche di livello superiore. Dati di questo tipo sono molto frequenti soprattutto nelle scienze sociali, poiché gli uomini vivono in strutture sociali quali la famiglia, il quartiere, la città, la scuola, l'azienda ecc. La struttura gerarchica può essere composta anche da molti livelli: ad esempio, lo studente fa parte di una classe, che appartiene a una certa scuola, che è inserita in un dato distretto scolastico ecc.

I caratteri delle unità elementari sono influenzati, spesso in modo notevole, dalla gerarchia: ad esempio, uno studente può avere rendimenti ben diversi a seconda della scuola in cui è inserito. È importante notare che la gerarchia esercita il proprio effetto per il solo fatto di esistere, indipendentemente dalla sua genesi: infatti, anche se gli studenti non hanno scelto di frequentare quella data scuola, il fatto oggettivo di condividere strutture didattiche, insegnanti, programmi scolastici ecc. rende quel gruppo di studenti diverso da quello di un'altra scuola. Talvolta il piano di campionamento si basa esplicitamente sulla gerarchia, usando metodi a due o più stadi; tuttavia, l'esistenza della gerarchia non è legata al piano di campionamento, per cui anche i dati raccolti con il campionamento casuale semplice possono richiedere l'utilizzo di tecniche multilivello.

L'importanza delle strutture gerarchiche è ben nota agli statistici da lungo tempo. Tuttavia fino a pochi anni fa mancavano gli sviluppi metodologici necessari ad includere esplicitamente la gerarchia nell'analisi. Un esempio significativo è la discussione sull'*unità di analisi,* che negli anni Settanta ha animato la statistica educazionale. Il problema riguardava la scelta dell'unità di analisi negli studi sull'efficacia dell'insegnamento: ci si chiedeva se fosse

preferibile basare l'analisi su dati a livello di alunno o di insegnante. Infatti, se da un lato l'oggetto di interesse era l'insegnante, dall'altro le considerazioni sulla *distorsione da aggregazione* (nota anche come *ecological fallacy*: Robinson, 1950) consigliavano di condurre l'analisi a livello di alunno. I modelli multilivello rappresentano la risposta a questo problema, poichè consentono di effettuare un'analisi basata contemporaneamente su entrambi i livelli.

L'ambito di applicazione dei modelli multilivello è sorprendentemente vasto, poiché oltre all'analisi di dati *cross-section* con struttura gerarchica, si prestano bene anche all'analisi di dati *longitudinali* (Diggle *et al.*, 1994, cap. 9; Goldstein, 1995, cap. 6): infatti gli individui possono essere pensati come unità di secondo livello e le osservazioni ripetute come unità di primo livello. E se gli individui fanno parte di gruppi, questi rappresentano le unità di terzo livello. Inoltre, con opportune modifiche, i modelli multilivello si applicano anche ai casi di *strutture a classificazione incrociata* (Goldstein, 1995, cap. 8), nelle quali un individuo è classificato secondo più criteri e quindi appartiene contemporaneamente a più unità di secondo livello (ad esempio, uno studente può essere classificato in base al quartiere in cui vive e alla scuola che frequenta).

L'aggettivo *multilivello* sta ad indicare semplicemente che il modello include esplicitamente la gerarchia. Nella maggior parte dei casi ciò viene realizzato per mezzo di un *modello a coefficienti casuali*, anche se rientrano nella classe dei multilivello l'approccio noto come *slopes-as-outcomes* (Burnstein *et al.*, 1978) e alcune estensioni dei *modelli ad equazioni strutturali* (Muthén, 1994). Nel presente lavoro faremo esclusivo riferimento ai modelli a coefficienti casuali, usando come sinonimi i termini "multilivello" e "a coefficienti casuali".

2.1.2 Il modello lineare a due livelli

Supponiamo di avere delle osservazioni su N individui (unità statistiche di primo livello) facenti parte di J gruppi (unità statistiche di secondo livello) di numerosità $N_1, \ldots, N_J$ e indichiamo con ij l'i-mo individuo del j-mo gruppo. Il *modello lineare a due livelli (a coefficienti casuali)* è dato dalle seguenti equazioni:

$$\begin{cases} y_{ij} = \alpha_j + \beta_j x_{ij} + e_{ij} \\ \alpha_j = \gamma_{00} + \gamma_{01} w_j + u_{0j} \\ \beta_j = \gamma_{10} + \gamma_{11} w_j + u_{1j} \end{cases} \tag{2.1}$$

dove y_{ij} e x_{ij} sono, rispettivamente, la risposta e una variabile esplicativa dell'individuo i del gruppo j, w_j è una variabile esplicativa del gruppo j (*variabile di contesto*), e_{ij} è un termine di errore a livello individuale, mentre

u_{0j} e u_{1j} sono termini di errore a livello di gruppo detti *effetti casuali*. I parametri hanno la seguente interpretazione: γ_{st} rappresenta l'effetto della t-ma variabile di gruppo sul coefficiente della s-ma variabile individuale. Le assunzioni standard sui termini di errore sono le seguenti:

$$e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2) \tag{2.2}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix} \right) \tag{2.3}$$

$$e_{ij} \perp \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \tag{2.4}$$

dove *iid* sta per indipendenti e identicamente distribuiti e il simbolo $\perp$ indica indipendenza stocastica. A ciò va aggiunta l'ipotesi di indipendenza di tutti i termini di errore dalle variabili esplicative.

La prima equazione del modello (2.1) (detta equazione *micro*) è simile a quella di un modello di regressione ordinario, salvo il fatto che l'intercetta α e il coefficiente angolare β hanno l'indice j e quindi sono specifici del gruppo. Tuttavia, tali coefficienti non sono delle costanti ignote, ma sono delle realizzazioni di variabili casuali le cui distribuzioni sono specificate nella seconda e terza equazione del modello (2.1) (dette equazioni *macro*). Dunque ogni gruppo ha la propria retta di regressione, però le rette sono collegate dal fatto che i loro parametri scaturiscono da *iperdistribuzioni* comuni. Ciò significa assumere che i gruppi presenti nei dati siano un campione casuale semplice da una ipotetica popolazione di gruppi. Ad esempio l'interpretazione dell'espressione $\alpha_j = \gamma_{00} + \gamma_{01} w_j + u_{0j}$ è la seguente: $\gamma_{00} + \gamma_{01} w_j$ è il valore medio dell'intercetta nell'ipotetica popolazione di gruppi che hanno quel dato valore di w_j, mentre u_{0j} è lo scostamento da tale valore medio relativo al j-mo gruppo del campione.

Dunque un modello multilivello si pone a metà strada tra due approcci estremi:

a) applicare un unico modello di regressione ai dati individuali, ignorando la presenza dei gruppi;

b) applicare un insieme di modelli di regressione specifici per gruppo, riconoscendo esplicitamente i gruppi, ma trattandoli come entità del tutto autonome.

In effetti l'approccio multilivello si basa su modelli di regressione specifici per gruppo, ma riconosce l'esistenza di fattori comuni che rendono inadeguata un'analisi indipendente gruppo per gruppo. Questa caratteristica trova

riscontro nelle proprietà delle stime. I parametri del modello multilivello sono i γ della (2.1) e i σ^2 e σ delle (2.2)-(2.3), che possono essere stimati secondo varie procedure (cfr. par. 2.1.7). Tuttavia l'uso del modello presuppone anche la stima del valore assunto dagli effetti casuali in ogni gruppo, che permette la stima dei coefficienti di regressione specifici per gruppo, α_j e β_j, secondo le

$$\begin{cases} \hat{\alpha}_j = \hat{\gamma}_{00} + \hat{\gamma}_{01} w_j + \hat{u}_{0j} \\ \hat{\beta}_j = \hat{\gamma}_{10} + \hat{\gamma}_{11} w_j + \hat{u}_{1j} \end{cases}$$

Ebbene, come vedremo nel par. 2.1.9, queste stime hanno un valore compreso fra due valori estremi che corrispondono, grosso modo, agli approcci a) e b) descritti sopra. Inoltre, nei gruppi poco numerosi le stime sono vicine a quelle ottenibili con l'approccio a); viceversa, nei gruppi molto numerosi le stime sono simili a quelle dell'approccio b). Questo comportamento degli stimatori $\hat{\alpha}_j$ e $\hat{\beta}_j$ viene indicato con il termine *borrowing streght* (Kreft & De Leeuw, 1998, par. 1.3.6), poiché la stima per i gruppi più poveri di informazione viene "consolidata" attingendo informazione dagli altri gruppi.

Il modello (2.1) può essere riscritto in forma compatta nel seguente modo:

$$\begin{aligned} y_{ij} &= (\gamma_{00} + \gamma_{01} w_j + u_{0j}) + (\gamma_{10} + \gamma_{11} w_j + u_{1j}) x_{ij} + e_{ij} \\ &= [\gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} w_j + \gamma_{11} w_j x_{ij}] + [u_{0j} + u_{1j} x_{ij} + e_{ij}] \,. \end{aligned} \tag{2.5}$$

Questo modello è noto anche come *modello lineare misto (mixed linear model*, Harville, 1977)[1]. Il primo termine in parentesi quadra costituisce la *parte fissa*, il secondo termine la *parte casuale*, che può essere distinta in primo livello (e_{ij}) e secondo livello ($u_{0j} + u_{1j} x_{ij}$).

La parte fissa deve il proprio nome al fatto che non comprende variabili casuali, poiché i γ sono parametri da stimare e le variabili esplicative (in seguito dette anche covariate) sono considerate fisse[2]. Si noti che la parte fissa include un'intercetta, una covariata a livello individuale (livello 1), una covariata a livello di gruppo (livello 2) e un'interazione *cross-level* fra le due.

La parte casuale comprende il termine di errore di primo livello e gli effetti

[1] Naturalmente i modelli (2.1) e (2.5) sono formalmente equivalenti; tuttavia il primo è utile a fini interpretativi, mentre il secondo si presta meglio ad illustrare le proprietà statistiche.

[2] Le covariate possono essere fisse per natura (come nel caso di un trattamento controllato dallo sperimentatore) oppure aleatorie. Tuttavia in quest'ultimo caso è possibile condurre l'analisi condizionatamente ai valori osservati nel campione, per cui le covariate sono considerate fisse. Poiché nel presente lavoro ci porremo sempre in questa ottica di condizionamento, le covariate saranno sempre trattate come fisse e nelle formule il condizionamento verrà omesso. In realtà nel prosieguo si parlerà spesso di condizionamento, riferendoci però agli effetti casuali e non alle covariate.

casuali: u_{0j} è l'effetto casuale relativo all'intercetta, u_{1j} quello relativo alla covariata x_{ij}.

La separazione dell'equazione (2.5) in parte fissa e parte casuale spiega la classificazione dei parametri in *parametri fissi* (i γ della (2.5)) e *parametri casuali* (i σ^2 e σ delle (2.2)-(2.3)).

Le assunzioni (2.2)-(2.4) determinano la struttura di covarianza del modello:

$$var(y_{ij}) = \sigma_e^2 + (\sigma_{u0}^2 + \sigma_{u1}^2 x_{ij}^2 + 2\sigma_{u01} x_{ij})$$

$$cov(y_{ij}, y_{i'j'}) = \begin{cases} \sigma_{u0}^2 + \sigma_{u1}^2 x_{ij} x_{i'j} + \sigma_{u01}(x_{ij} + x_{i'j}) & \text{se } j = j' \text{ e } i \neq i' \\ 0 & \text{se } j \neq j' \end{cases}$$

L'aspetto più rilevante è che la covarianza è diversa da zero per gli individui che appartengono allo stesso gruppo. Inoltre, la varianza ha due componenti, una infra-gruppo o *within* (σ_e^2) e una tra gruppi o *between* ($\sigma_{u0}^2 + \sigma_{u1}^2 x_{ij}^2 + 2\sigma_{u01} x_{ij}$). Poiché la componente *between* dipende da x_{ij}, si ha *eteroschedasticità*.

2.1.3 La correlazione intraclasse

Quando è presente solo l'effetto casuale sull'intercetta (cioè quando $\sigma_{u1}^2 = \sigma_{u01} = 0$) le espressioni precedenti si semplificano in

$$var(y_{ij}) = \sigma_e^2 + \sigma_{u0}^2$$

$$cov(y_{ij}, y_{i'j'}) = \begin{cases} \sigma_{u0}^2 & \text{se } j = j' \text{ e } i \neq i' \\ 0 & \text{se } j \neq j' \end{cases}$$

In questo caso particolare, noto come *modello a componenti di varianza* o come *modello ad effetti casuali*, la componente di varianza tra gruppi è costante, così come la covarianza. Di conseguenza è costante anche la *correlazione intraclasse*[3]:

$$corr(y_{ij}, y_{i'j}) = \frac{\sigma_{u0}^2}{\sigma_e^2 + \sigma_{u0}^2}.$$

Il concetto di correlazione intraclasse gioca un ruolo fondamentale nella teoria dei modelli multilivello. L'origine di tale correlazione va ricercata nel

[3]Il *coefficiente di correlazione intraclasse* può essere definito come la proporzione di variabilità attribuibile ai gruppi o, equivalentemente, come la correlazione fra due unità dello stesso gruppo. Naturalmente quando è presente un effetto casuale relativo ad una covariata tale coefficiente non può più essere calcolato, ma il concetto che rappresenta continua ad essere valido.

fatto che gli individui appartenenti al medesimo gruppo condividono una stessa realtà e quindi hanno in comune dei fattori non osservabili che nel modello assumono la forma di effetti casuali.

I modelli di regressione ordinari assumono indipendenza fra le unità e quindi la correlazione intraclasse costituisce una violazione di tale assunzione dalle conseguenze potenzialmente molto gravi. Infatti, sebbene gli stimatori dei parametri siano corretti, gli stimatori degli errori standard sono distorti verso il basso: ciò è dovuto al fatto che, a parità di numerosità, un campione con unità correlate contiene meno informazione di un campione con unità indipendenti e quindi l'incertezza relativa alle stime è maggiore. La conseguenza più pericolosa della sottostima degli errori standard consiste nell'aumento del livello di significatività dei test statistici, che dipende dal livello di correlazione intraclasse e dalla numerosità dei gruppi (cfr. Kreft & De Leeuw, 1998, par. 1.3.2). Pertanto, a prescindere dalla necessità di condurre un'analisi su più livelli contemporaneamente, i modelli multilivello rappresentano una scelta obbligata nelle situazioni di rilevante correlazione intraclasse.

2.1.4 Effetti fissi o casuali?

Supponiamo di avere dei dati con struttura gerarchica e di voler usare un modello lineare per studiare le differenze esistenti tra i gruppi. Usando la notazione del modello (2.1) possiamo scrivere

$$y_{ij} = \alpha_j + \beta x_{ij} + e_{ij}.$$

Le opzioni possibili sono due:

1) Assumere che $\alpha_1, \ldots, \alpha_J$ siano delle costanti da stimare (*effetti fissi*). Ciò equivale ad usare un modello di analisi della covarianza (ANCOVA);

2) Assumere che $\alpha_1, \ldots, \alpha_J$ siano *effetti casuali*, cioè realizzazioni di una variabile aleatoria $N(\alpha, \sigma_u^2)$ di cui è possibile stimare media e varianza. Ciò equivale ad usare un modello a componenti di varianza.

La scelta fra le due opzioni dovrebbe basarsi sulla natura dei gruppi e sul fine dell'indagine, usando

- effetti fissi se nel campione sono presenti tutti i possibili gruppi (ad esempio, pazienti trattati con una serie farmaci alternativi) oppure se sono presenti tutti gruppi di interesse ai fini dell'analisi (ad esempio, se il fine è quello di studiare solo le scuole incluse nel campione senza voler estendere i risultati anche ad altre scuole);

- effetti casuali se nel campione sono presenti dei gruppi che rappresentano una popolazione di gruppi e il fine dell'analisi è quello di estendere i risultati a tale popolazione.

In teoria, per poter usare gli effetti casuali, la popolazione da cui provengono i gruppi dovrebbe avere un numero infinito o molto grande di elementi. In pratica, anche se questo requisito non è pienamente soddisfatto, in molte applicazioni si preferisce usare un modello a componenti di varianza, poiché tale modello ha due importanti vantaggi rispetto al modello ANCOVA:

a) tiene conto della correlazione intraclasse;

b) consente di limitare notevolmente il numero di parametri da stimare.

Inoltre, la scelta diventa obbligata quando si vogliano inserire delle variabili di contesto, poiché in tal caso usando il modello ANCOVA si avrebbe una matrice dei regressori singolare. Dunque il modello ANCOVA consente di rispondere solo alla domanda "i gruppi sono diversi?", mentre il modello a componenti di varianza è potenzialmente capace di rispondere anche alla domanda "perché i gruppi sono diversi?"[4].

2.1.5 La relazione tra le componenti di varianza e le variabili esplicative

In un modello di regressione ordinario la varianza del termine di errore ha il significato di *varianza residua*, cioè di varianza non spiegata dai regressori. Pertanto l'inserimento di un nuova variabile provoca una riduzione della varianza residua, la cui entità dipende dal suo potere esplicativo.

La situazione è più complessa in un modello a componenti di varianza, nel quale la varianza non spiegata dai regressori viene scomposta in due parti: la componente *between* (σ_u^2), ovvero la varianza non spiegata dai regressori e che è attribuibile agli effetti casuali, cioè alla struttura gerarchica; la componente *within* (σ_e^2), ovvero la varianza residua in senso stretto, che non è spiegata né dai regressori, né dall'appartenenza ai gruppi, ma che è legata alla variabilità individuale.

L'effetto dell'inserimento di nuove variabili sulle componenti di varianza dipende dal tipo di variabile (Longford, 1993, pp. 29-30):

[4] Un'interessante discussione sulle implicazioni della scelta del modello è contenuta nel classico articolo di Aitkin & Longford (1986).

- Variabile di contesto (livello 2): una variabile misurata a livello di gruppo contribuisce a spiegare le differenze tra i gruppi e quindi a ridurre la componente *between*, mentre non ha nessun effetto sulla componente *within*;

- Variabile individuale (livello 1): come è naturale attendersi, l'inserimento di una variabile individuale riduce la varianza *within*, ma la direzione del suo effetto sulla componente *between* non è determinabile a priori.

Per comprendere l'ultimo punto bisogna pensare che la componente *between* è una misura del grado di eterogeneità dei gruppi non spiegata dai regressori e che l'inserimento di una nuova variabile individuale può sia aumentare che diminuire tale eterogeneità non spiegata. Consideriamo ad esempio uno studio sulla mortalità dei degenti di un insieme di ospedali (unità di livello 2) e supponiamo di inserire una variabile che misura la gravità dei pazienti. Se i pazienti più gravi sono ricoverati negli ospedali più qualificati, l'inserimento di tale variabile provoca un aumento della componente *between*, poiché porta alla luce un'eterogeneità che in precedenza era mascherata dal modo in cui i pazienti sono assegnati agli ospedali.

2.1.6 Il modello multilivello lineare nella notazione matriciale

Al fine di illustrare le proprietà statistiche e le procedure di stima relativamente al modello più generale possibile è utile introdurre la seguente notazione matriciale, che generalizza il *modello lineare misto* (2.5). Posto

n = numero di unità elementari
H = numero di livelli
n_h = numero di unità di livello h ($h = 1, \ldots, H$)
$n_{h(j)}$ = numero di unità elementari appartenenti alla j-ma unità di livello h
p = numero di parametri fissi
q_h = numero di effetti casuali di livello h

il *modello multilivello lineare* può scriversi come

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{h=1}^{H} \mathbf{Z}_h \mathbf{u}_h \tag{2.6}$$

dove[5]

- $\underset{n\times 1}{\mathbf{y}}$ è il vettore delle risposte;
- $\underset{n\times p}{\mathbf{X}}$ è la matrice delle variabili esplicative;
- $\underset{p\times 1}{\boldsymbol{\beta}}$ è il vettore dei parametri fissi;
- $\underset{n\times(n_h q_h)}{\mathbf{Z}_h} = \oplus_{j=1}^{n_h}\{\underset{(n_{h(j)}\times q_h)}{\mathbf{Z}_{h(j)}}\}$ è la matrice diagonale a blocchi relativa agli effetti casuali di livello h;
- $\underset{(n_h q_h)\times 1}{\mathbf{u}'_h} = (\underset{q_h\times 1}{\mathbf{u}'_{h(1)}}, \ldots, \underset{q_h\times 1}{\mathbf{u}'_{h(n_h)}})$ è il vettore degli effetti casuali di livello h[6].

Le ipotesi sulla parte casuale del modello sono le seguenti:

1) $E(\mathbf{u}_h) = \mathbf{0}$ (gli effetti casuali hanno valore atteso nullo);

2) $cov(\mathbf{u}_h, \mathbf{u}_{h'}) = \mathbf{0}$ per $h \neq h'$ (gli effetti casuali relativi a unità appartenenti a livelli diversi sono incorrelati);

3) $var(\mathbf{u}_h) = \mathbf{S}_h = \mathbf{I}_{n_h} \otimes \boldsymbol{\Omega}_h$ (gli effetti casuali relativi a unità diverse appartenenti allo stesso livello sono incorrelati ed hanno la stessa matrice di covarianza).

Pertanto la matrice di covarianza di $\mathbf{y}$ (condizionata a $\mathbf{X}$) è data da

$$\begin{aligned} \mathbf{V}_H &= var\left(\sum_{h=1}^{H}\mathbf{Z}_h\mathbf{u}_h\right) = \sum_{h=1}^{H} var\left(\mathbf{Z}_h\mathbf{u}_h\right) = \sum_{h=1}^{H}\mathbf{Z}_h\mathbf{S}_h\mathbf{Z}'_h \qquad (2.7)\\ &= \sum_{h=1}^{H}\left(\oplus_{j=1}^{n_h}\{\mathbf{Z}_{h(j)}\boldsymbol{\Omega}_h\mathbf{Z}'_{h(j)}\}\right) = \sum_{h=1}^{H}\mathbf{V}_H^{(h)}, \end{aligned}$$

dove $\mathbf{V}_H^{(h)}$ è il contributo degli effetti casuali di livello h alla matrice di covarianza delle risposte in un modello a H livelli. Si noti che $\mathbf{V}_H^{(h)}$ è diagonale a blocchi, con blocchi corrispondenti alle unità di livello h.

[5] Le dimensioni delle matrici sono riportate sotto le stesse nella forma (numero righe) $\times$ (numero colonne). Per quanto riguarda gli operatori matriciali, l'apice indica la matrice trasposta, $\otimes$ è il prodotto di Kronecker, mentre $\oplus$ è la *somma diretta*, cioè $\oplus_{j=1}^{n}\mathbf{A}_j$ è la matrice diagonale a blocchi i cui blocchi sono, da sinistra verso destra, $\mathbf{A}_1, \ldots, \mathbf{A}_n$ (cfr. Searle *et al.*, 1992, Appendix M).

[6] Il termine di errore individuale è rappresentato da $\mathbf{u}_1$, che può essere pensato come effetto casuale di primo livello.

Per quanto riguarda la distribuzione degli effetti casuali, l'ipotesi usuale è quella di normalità, che risulta conveniente soprattutto in presenza di molti effetti casuali (Goldstein, 1995, p. 22)[7].

2.1.7 Stima dei parametri

La stima dei parametri di un modello multilivello lineare ha costituito per lungo tempo un problema proibitivo a causa della notevole mole di calcoli richiesta dagli algoritmi di stima. Fra i metodi proposti in letteratura ricordiamo:

- Massima verosimiglianza (ML) (Harville, 1977; Longford, 1987);
- Massima verosimiglianza vincolata (REML) (Patterson & Thompson, 1971);
- Minimi quadrati generalizzati iterati (IGLS) (Goldstein, 1986);
- Minimi quadrati generalizzati iterati vincolati (RIGLS) (Goldstein, 1989);
- Algoritmo EM (Aitkin *et al.*, 1981; Bryk & Raudenbush, 1992);
- Analisi bayesiana con metodi Markov Chain Monte Carlo (MCMC) (Gilks *et al.*, 1996).

Per quanto riguarda la verosimiglianza, osserviamo che l'ipotesi di normalità degli effetti casuali permette di determinare facilmente la distribuzione marginale della risposta. Con riferimento alla notazione del modello (2.6) si ottiene

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}_H(\boldsymbol{\theta})),$$

dove $\boldsymbol{\theta}$ è un vettore che raccoglie i parametri casuali contenuti nelle matrici $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_H$. Pertanto il logaritmo naturale della verosimiglianza marginale è dato da

$$\begin{aligned} &l(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}) = \\ &-\tfrac{1}{2}\left\{n\log(2\pi) + \log(\det \mathbf{V}_H(\boldsymbol{\theta})) + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}_H^{-1}(\boldsymbol{\theta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \end{aligned} \tag{2.8}$$

[7] Quando l'ipotesi di normalità non è soddisfatta, gli stimatori dei parametri sono consistenti, ma non efficienti, mentre gli stimatori degli errori standard non sono consistenti (Goldstein, 1995, p. 22). L'ipotesi di normalità viene solitamente controllata per mezzo dei *diagrammi quantile-quantile* (Goldstein, 1995, p. 28; Longford, 1993, cap. 3).

La massimizzazione della (2.8) comporta alcuni problemi computazionali, che sono stati risolti da Longford (1987), il quale ha proposto un algoritmo di massimizzazione di tipo *Fisher scoring*.

Di seguito illustriamo in dettaglio il metodo IGLS proposto da Goldstein (1986), poiché è quello implementato nel programma MLwiN (Goldstein *et. al*, 1998) di cui ci avvarremo per stimare i modelli utilizzati nel presente lavoro. In realtà, gli algoritmi *Fisher scoring* e IGLS sono formalmente equivalenti (Goldstein, 1995, p. 23).

Il metodo IGLS si basa sulla seguente osservazione: se i parametri fissi fossero noti si potrebbe usare il principio dei minimi quadrati generalizzati (GLS) per stimare i parametri casuali, e viceversa. Pertanto, partendo da una stima iniziale dei parametri fissi (ad esempio ottenuta con i minimi quadrati ordinari), l'algoritmo IGLS alterna la stima dei parametri casuali e fissi con il metodo GLS, fino a convergenza.

Usando la notazione del modello (2.6) e scrivendo $\mathbf{V}$ in luogo di $\mathbf{V}_H$, i due passi dell'algoritmo IGLS possono essere formalizzati come segue:

a) Stima GLS dei parametri fissi
Nota la matrice $\mathbf{V}$, lo stimatore GLS dei parametri fissi è

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{2.9}$$

con

$$cov(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.$$

b) Stima GLS dei parametri casuali
Noto il vettore $\boldsymbol{\beta}$, lo stimatore GLS dei parametri casuali (inclusi nella matrice $\mathbf{V}$ in base alla (2.7)) può essere ottenuto come segue. Indichiamo con

$$\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$

il vettore dei residui totali, per i quali vale la relazione $E(\tilde{\mathbf{y}}\tilde{\mathbf{y}}') = \mathbf{V}$. Poi definiamo un nuovo vettore $\mathbf{y}^*$ tale che

$$\mathbf{y}^* = vec(\tilde{\mathbf{y}}\tilde{\mathbf{y}}'), \tag{2.10}$$

dove *vec* è l'operatore che forma un vettore da una matrice impilando le sue colonne una sotto l'altra. Adesso, indicando i parametri casuali con il vettore $\boldsymbol{\theta}$, è possibile scrivere un modello lineare per i parametri casuali:

$$E(\mathbf{y}^*) = \mathbf{X}^*\boldsymbol{\theta}, \tag{2.11}$$

dove la matrice dei regressori $\mathbf{X}^*$ può essere determinata colonna per colonna in base alla seguente formula (Goldstein, 1986; Goldstein & Rasbash, 1992):

$$\mathbf{x}_k^* = vec\left(\frac{\partial \mathbf{V}}{\partial \theta_k}\right) = vec\left[\oplus_{j=1}^{n_{\bar{h}}}\left\{\mathbf{Z}_{\bar{h}(j)}\left(\frac{\partial \mathbf{\Omega}_{\bar{h}}}{\partial \theta_k}\right)\mathbf{Z}'_{\bar{h}(j)}\right\}\right],$$

dove $\mathbf{x}_k^*$ è la k-ma colonna di $\mathbf{X}^*$, mentre θ_k è il k-mo elemento di $\boldsymbol{\theta}$ che assumiamo essere un parametro casuale appartenente al livello arbitrario $\bar{h}$.

Il modello lineare (2.11) consente di usare il metodo GLS per stimare i parametri casuali:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^{*\prime}\mathbf{V}^{*-\mathbf{1}}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{V}^{*-1}\mathbf{y}^*, \tag{2.12}$$

dove $\mathbf{V}^* = \mathbf{V} \otimes \mathbf{V}$. Si noti che $\mathbf{V}^*$ non è esattamente la matrice di covarianza di $\mathbf{y}^*$, la quale è singolare e quindi non può essere usata nella stima GLS[8]. Dopo alcuni calcoli (Goldstein & Rasbash, 1992) si ottiene

$$cov(\hat{\boldsymbol{\theta}}) = 2(\mathbf{X}^{*\prime}\mathbf{V}^{*-\mathbf{1}}\mathbf{X}^*)^{-1}.$$

L'algoritmo IGLS itera tra la (2.9) e la (2.12) fino a convergenza, usando di volta in volta le stime correnti dei parametri fissi e casuali (le stime iniziali dei parametri fissi sono solitamente ottenute con i minimi quadrati ordinari). Goldstein (1986) dimostra che, sotto ipotesi di normalità, le stime così ottenute sono di massima verosimiglianza. In assenza di normalità, lo stimatore IGLS è comunque consistente, anche se non pienamente efficiente; tuttavia il corrispondente stimatore degli errori standard non è più consistente (Goldstein, 1995, p. 22).

Lo stimatore IGLS in generale è distorto e ciò può costituire un problema nei campioni di piccola numerosità. Pertanto è utile disporre anche di uno stimatore non distorto, che può essere ricavato apportando una piccola modifica alla procedura IGLS. Infatti, il passo dell'algoritmo deputato alla stima dei parametri casuali si basa sulla relazione

$$E\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\right] = \mathbf{V}.$$

Tuttavia tale relazione non è più vera se si sostituisce $\boldsymbol{\beta}$ con il suo stimatore GLS $\hat{\boldsymbol{\beta}}$, poiché in tal caso

$$E\left[(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\right] = \mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'. \tag{2.13}$$

[8] La matrice di covarianza di $\mathbf{y}^*$ è $\mathbf{V}^*(\mathbf{I} + \mathbf{S}_N)$, dove $\mathbf{S}_N$ è la cosiddetta *vec permutation matrix* (Searle *et al.*, 1992, par. 12.3).

Per correggere questo errore si può sommare il termine $\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'$ a $\tilde{\mathbf{y}}$ prima di calcolare $\mathbf{y}^*$ secondo la (2.10). In tal modo lo stimatore IGLS diviene corretto e viene indicato con l'acronimo RIGLS (Goldstein, 1989)[9].

Comunque, anche nei campioni di piccola numerosità, la scelta fra IGLS e RIGLS non è ovvia, poiché alcuni studi di simulazione mostrano che la correttezza del metodo vincolato viene pagata con una minore efficienza e che non esistono linee guida per risolvere tale conflitto in favore di un metodo o dell'altro (Kreft & De Leeuw, 1998, par. 5.4).

2.1.8 Stima degli effetti casuali (o residui)

Come suggerito nel par. 2.1.2, l'uso di un modello multilivello presuppone anche la stima degli effetti casuali (o residui). In realtà gli effetti casuali sono variabili aleatorie per cui ciò che si stima è la *realizzazione* di tali variabili aleatorie nei vari gruppi.

In un modello di regressione classico i termini di errore (che si riferiscono ad un unico livello) sono usualmente stimati dai residui di regressione. Invece in un modello multilivello i residui $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, detti *residui totali stimati*, devono essere opportunamente scomposti nelle loro componenti di primo, secondo, . . ., H-mo livello. Usando la notazione del par. 2.1.6 e supponendo per il momento noti tutti i parametri del modello, i residui di livello h possono essere stimati per mezzo del loro valore atteso condizionato ai *residui totali veri* $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$,

$$\hat{\mathbf{u}}_h = E(\mathbf{u}_h|\tilde{\mathbf{y}}). \tag{2.14}$$

Se si assume una distribuzione normale degli effetti casuali, il valore atteso (2.14) può essere calcolato tramite la distribuzione a posteriori $\mathbf{u}_h \,|\, \tilde{\mathbf{y}}$. In generale, in assenza di ipotesi distribuzionali specifiche (come nel caso dello stimatore IGLS), si può usare una semplice regressione lineare di $\mathbf{u}_h$ su $\tilde{\mathbf{y}}$ (Goldstein, 1995, app. 2.2). Poiché $cov(\tilde{\mathbf{y}}, \mathbf{u}_h) = \mathbf{R}_h$, dove

$$\mathbf{R}_h = \oplus_{j=1}^{n_h}\{\mathbf{Z}_{h(j)}\boldsymbol{\Omega}_h\},$$

la regressione fornisce

$$\hat{\mathbf{u}}_h = \mathbf{R}_h'\mathbf{V}^{-1}\tilde{\mathbf{y}}. \tag{2.15}$$

Sostituendo ai parametri incogniti il loro valore stimato, si ottiene uno stimatore consistente degli effetti casuali che, nell'ipotesi di normalità, coincide

[9] Nell'acronimo RIGLS la R sta per Restricted. L'origine di tale termine va ricercata nel fatto che lo stimatore RIGLS è equivalente allo stimatore di massima verosimiglianza vincolata (REML).

con lo stimatore bayesiano empirico. La sua matrice di covarianza, data la (2.13), è

$$var(\hat{\mathbf{u}}_h) = \mathbf{R}_h'\mathbf{V}^{-1}(\mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}')\mathbf{V}^{-1}\mathbf{R}_h. \tag{2.16}$$

La (2.16) è nota come *matrice di covarianza non condizionata* e i relativi errori standard vengono detti *diagnostici*, poiché vengono usati per standardizzare i residui a fini diagnostici (ad esempio per tracciare il *diagramma quantile-quantile* per la verifica dell'ipotesi di normalità).

Tuttavia, se il fine è quello di fare inferenza sul valore assunto dagli effetti casuali (ad esempio costruendo un intervallo di confidenza) è opportuno usare la *matrice di covarianza condizionata*

$$\mathbf{S}_h - \mathbf{R}_h'\mathbf{V}^{-1}(\mathbf{V} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}')\mathbf{V}^{-1}\mathbf{R}_h, \tag{2.17}$$

che può essere ottenuta come errore quadratico medio della regressione di $\mathbf{u}_h$ su $\tilde{\mathbf{y}}$, oppure, nell'ipotesi di normalità, come varianza della distribuzione a posteriori $\mathbf{u}_h \,|\, \tilde{\mathbf{y}}$. Gli errori standard ottenuti dalla (2.17) si dicono *comparativi* in quanto spesso vengono usati nei confronti fra effetti casuali relativi a unità diverse.

Si noti che nel calcolo della (2.16) e della (2.17) si tiene conto della variabilità campionaria dei coefficienti fissi, ma non di quella dei coefficienti casuali. Pertanto in campioni di piccola numerosità può essere opportuno stimare tali matrici di covarianza con procedure di tipo *bootstrap* (Goldstein, 1995, par. 3.5).

2.1.9 L'effetto shrinkage

Al fine di mostrare le proprietà dello stimatore dei residui (2.15) è utile considerare il seguente modello a componenti di varianza:

$$\begin{array}{l} y_{ij} = \alpha_j + \beta x_{ij} + e_{ij} \\ \left\{ \begin{array}{l} \alpha_j = \alpha + u_j \\ u_j \sim N(0, \sigma_u^2) \\ e_{ij} \sim N(0, \sigma_e^2) \end{array} \right. \end{array} \tag{2.18}$$

Nella notazione del par. 2.1.6 si ha $\mathbf{\Omega}_1 = \sigma_e^2$; $\mathbf{\Omega}_2 = \sigma_u^2$; $\mathbf{Z}_{1(j)} = 1$ per ogni $j = 1, \ldots, n_1$; $\mathbf{Z}_{2(j)} = \mathbf{1}_{n_{2(j)}}$ per ogni $j = 1, \ldots, n_2$ ($\mathbf{1}_k$ indica il vettore unitario di lunghezza k). Pertanto dalla (2.15) si ricava

$$\hat{u}_j = s(n_j, \tau) \cdot \breve{y}_j, \tag{2.19}$$

dove

- $\check{y}_j = (\bar{y}_j - \hat{\alpha} - \hat{\beta}\bar{x}_j)$ è il *residuo totale stimato medio* del j-mo gruppo ($\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ e analogamente $\bar{x}_j$);
- $s(n_j, \tau) = \frac{1}{1+\frac{1}{n_j \tau}}$ è il cosiddetto *shrinkage factor* (dove $\tau = \frac{\sigma_u^2}{\sigma_e^2}$ è il *rapporto delle componenti di varianza*).

Lo shrinkage factor è un numero compreso fra 0 e 1 che comprime il residuo totale stimato medio in modo differenziato a seconda della numerosità del gruppo n_j e del rapporto fra le componenti di varianza τ. In particolare, lo shrinkage sarà più forte nei gruppi poco numerosi che in quelli molto numerosi; inoltre, a parità di numerosità, lo shrinkage sarà più forte quando la componente di varianza *between* è piccola rispetto a quella *within*.

Lo shrinkage rende più affidabile la stima degli effetti casuali, poiché tende a riportare verso lo zero (cioè verso la media degli effetti casuali nella popolazione) la stima relativa ai gruppi poco numerosi, cioè che contengono poca informazione per la stima dell'effetto casuale. D'altra parte lo shrinkage ha delle conseguenze indesiderate quando si vogliano confrontare due gruppi sulla base dei residui stimati, poiché può accadere che un gruppo con un elevato valore dell'effetto casuale ma di scarsa numerosità abbia lo stesso residuo stimato di un gruppo con un piccolo valore dell'effetto casuale ma di grande numerosità.

L'effetto dello shrinkage si ripercuote anche sulla stima di α_j nella (2.18). Infatti

$$\begin{aligned} \hat{\alpha}_j &= \hat{\alpha} + \hat{u}_j \\ &= \hat{\alpha} + s(n_j, \tau) \cdot \check{y}_j \\ &= (1 - s(n_j, \tau))\, \hat{\alpha} + s(n_j, \tau) \left(\bar{y}_j - \hat{\beta}\bar{x}_j\right), \end{aligned}$$

per cui $\hat{\alpha}_j$ è un valore intermedio tra $\hat{\alpha}$ (stima del coefficiente medio nella popolazione) e $\bar{y}_j - \hat{\beta}\bar{x}_j$ (stima relativa al j-mo gruppo). Come accennato nel par. 2.1.2 questa proprietà viene indicata con il termine *borrowing strenght*.

I concetti di shrinkage e borrowing strenght, che abbiamo illustrato per il modello a componenti di varianza, valgono in generale (Bryk & Raudenbush, 1992) e rappresentano uno degli aspetti più caratteristici dell'analisi multilivello.

2.2 I modelli multilivello non lineari

I modelli multilivello sono stati inizialmente concepiti per lo studio di variabili di risposta quantitative con specificazione lineare del valore atteso e

delle varianze e covarianze. Il florido sviluppo di questa classe di modelli può essere spiegato in parte dal loro vasto campo di applicazione e in parte dalla relativa semplicità della trattazione matematica e dell'interpretazione statistica. Tuttavia le esigenze della ricerca scientifica, soprattutto in ambito sociale e bio-medico, hanno spinto verso un'estensione dei modelli multilivello tale da poter includere specificazioni non lineari del valore atteso della risposta oppure delle varianze e covarianze. In particolare, una specificazione non lineare del valore atteso si rende necessaria quando la risposta è di tipo qualitativo, come accade di frequente nelle indagini campionarie in ambito sociale.

In questo paragrafo, considerate le finalità del presente lavoro, concentreremo l'attenzione sui *modelli lineari generalizzati multilivello*, che costituiscono la scelta più conveniente per l'analisi di dati categorici con struttura gerarchica. Dopo un'introduzione generale, esamineremo in dettaglio i modelli per dati binari, politomici, ordinali e di sopravvivenza in tempo discreto, concludendo con la descrizione di alcune procedure di stima.

2.2.1 Definizione e interpretazione

Prima di parlare di modelli multilivello non lineari è opportuno soffermarci su alcune proprietà degli analoghi modelli lineari che a prima vista sembrano ovvie, ma che in realtà sono fondamentali per capire le implicazioni della non linearità. A fini illustrativi consideriamo il seguente modello lineare a due livelli[10]:

$$y_{ij} = \alpha + \beta x_{ij} + u_{0j} + u_{1j} x_{ij} + e_{ij}, \tag{2.20}$$

dove

$$e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2);$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix} \right);$$

$$e_{ij} \perp \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}.$$

Una conseguenza di queste ipotesi distribuzionali è che sono normali sia la distribuzione della risposta condizionata agli effetti casuali, sia la sua distribuzione marginale:

$$y_{ij} | \; u_{0j}, u_{1j} \sim N(\alpha + \beta x_{ij} + u_{0j} + u_{1j} x_{ij} \; , \; \sigma_e^2)$$

[10] La notazione è la stessa del par. 2.1.2, salvo sostituire γ_{00} con α e γ_{10} con β. Ricordiamo inoltre che "iid" sta per "indipendenti e identicamente distribuiti" e che il simbolo $\perp$ indica indipendenza stocastica.

$$y_{ij} \sim N(\alpha + \beta x_{ij}\ ,\ \sigma_e^2 + (\sigma_{u0}^2 + \sigma_{u1}^2 x_{ij}^2 + 2\sigma_{u01} x_{ij})).$$

Osservando la distribuzione marginale notiamo che rispetto ad un modello di regressione classico la struttura di covarianza è più complessa, ma i valori attesi sono identici; in altre parole, gli effetti casuali modificano solo la struttura di covarianza. Come vedremo tra breve, ciò non è più vero nei modelli non lineari.

Un'altra considerazione riguarda l'effetto delle covariate sulla risposta: infatti si ha

$$\frac{\partial}{\partial x_{ij}} E(y_{ij}|\ u_{0j}, u_{1j}) = \beta + u_{1j} \text{ (effetto nel gruppo } j)$$

$$\frac{\partial}{\partial x_{ij}} E(y_{ij}) = \beta \text{ (effetto medio nella popolazione)},$$

per cui l'effetto di x su y dipende dal gruppo, ma nel gruppo medio (cioè, quello per il quale $u_{0j} = u_{1j} = 0$) tale effetto coincide con quello medio nella popolazione. Anche questa fondamentale proprietà è peculiare dei modelli lineari.

Dopo queste osservazioni prendiamo in esame la classe più importante di modelli multilivello non lineari, cioè i *modelli lineari generalizzati multilivello,* che, per una struttura a due livelli, possono essere definiti come segue:

1) in ogni gruppo le risposte sono indipendenti condizionatamente agli effetti casuali del gruppo e seguono un modello lineare generalizzato (GLM, Generalised Linear Model: McCullagh & Nelder, 1989);

2) gli effetti casuali dei vari gruppi sono un campione casuale semplice da una distribuzione comune multivariata, solitamente gaussiana.

Questo modo di definire il modello non consente, in generale, di determinare in forma analitica la distribuzione marginale, salvo il caso del modello lineare con distribuzione normale. In alternativa è possibile definire direttamente la distribuzione marginale (Diggle *et al.*, 1994), ma, come vedremo, ciò risulta meno conveniente per gli sviluppi teorici.

Per quanto riguarda la specificazione della distribuzione condizionata, l'uso dei GLM costituisce la regola, poiché tali modelli possiedono un elevato grado di generalità e poggiano su fondamenti teorici ormai consolidati. Anche l'assunzione di normalità degli effetti casuali è largamente diffusa, sebbene non manchino proposte in senso contrario[11].

[11] Generalmente l'assunzione di normalità viene abbandonata nel caso in cui la teoria

In un GLM multilivello le risposte sono marginalmente indipendenti tra gruppi diversi, mentre all'interno di uno stesso gruppo l'indipendenza non è marginale, ma è condizionata agli effetti casuali. Ciò significa che la dipendenza esistente tra le risposte di un certo gruppo è interamente attribuibile agli effetti casuali, cioè a quei fattori non osservabili comuni a tutte le unità del gruppo.

Formalmente, la versione *GLM multilivello* del modello lineare (2.20) può scriversi come segue:

1) dati (u_{0j}, u_{1j}), le risposte $y_{1j}, \ldots, y_{n_j j}$ sono mutualmente indipendenti e seguono un GLM con densità

$$f(y_{ij}|\ u_{0j}, u_{1j}) = \exp\left\{\left[y_{ij}\theta_{ij} - \psi(\theta_{ij})\right]/\phi + c(y_{ij}, \phi)\right\},$$

dove θ_{ij} è il *parametro naturale*, ϕ è il *parametro di dispersione*, mentre $\psi(\cdot)$ e $c(\cdot)$ sono funzioni note; il valore atteso e la varianza condizionati soddisfano:

$$\mu^u_{ij} = E(y_{ij}|\ u_{0j}, u_{1j}) = g^{-1}(\alpha + \beta x_{ij} + u_{0j} + u_{1j}x_{ij}), \tag{2.21}$$

$$v^u_{ij} = Var(y_{ij}|\ u_{0j}, u_{1j}) = v(\mu^u_{ij})\phi, \tag{2.22}$$

dove $g^{-1}(\cdot)$ è l'inversa della *funzione link* $g(\cdot)$, mentre $v(\cdot)$ è la *funzione di varianza*;

2) gli effetti casuali $\{(u_{0j}, u_{1j}) : j = 1, \ldots, J\}$ sono un campione casuale semplice da una distribuzione comune multivariata, solitamente gaussiana:

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}\right).$$

A differenza del modello lineare, nel GLM multilivello non compare il termine di errore di primo livello, che viene implicitamente specificato con la distribuzione della risposta. L'unico parametro relativo alla variabilità di primo livello è il parametro di dispersione ϕ, che però in alcune importanti

suggerisca la presenza di effetti casuali con distribuzione discreta (McDonald, 1994) oppure quando si voglia ottenere la distribuzione della risposta in forma chiusa (Conaway, 1990). L'assunzione di normalità resta comunque la scelta classica, poiché è conveniente da un punto di vista teorico ed è difficilmente confutabile dall'evidenza empirica, in quanto gli effetti casuali sono quantità non osservabili. Inoltre, Gibbons *et al.* (1994) e Gibbons & Hedeker (1997) mostrano che, nelle applicazioni da loro discusse, la scelta di una distribuzione uniforme degli effetti casuali conduce a risultati del tutto simili a quelli ottenibili con la distribuzione normale.

distribuzioni (es. binomiale, Poisson) risulta fissato a priori[12]. Si noti inoltre che, ad eccezione del modello lineare normale, la varianza (2.22) dipende dal valore atteso condizionato e quindi dalla realizzazione degli effetti casuali.

Il valore atteso marginale della risposta è dato da

$$E(y_{ij}) = \int_{-\infty}^{+\infty} g^{-1}(\alpha + \beta x_{ij} + u_0 + u_1 x_{ij})\varphi(u_0, u_1)du_0 du_1, \qquad (2.23)$$

dove $\varphi(\cdot, \cdot)$ è la densità di probabilità degli effetti casuali. Tale valore atteso è diverso da quello condizionato relativo al gruppo medio

$$E(y_{ij}|\ u_{0j} = 0,\ u_{1j} = 0) = g^{-1}(\alpha + \beta x_{ij}).$$

Una conseguenza di questo fatto è che

$$\frac{\partial}{\partial x_{ij}} E(y_{ij}|\ u_{0j} = 0,\ u_{1j} = 0) \neq \frac{\partial}{\partial x_{ij}} E(y_{ij})$$

cioè, l'effetto di x su y nel gruppo medio non coincide con l'effetto medio nella popolazione. Questa osservazione è fondamentale per interpretare correttamente il coefficiente β, che ha un significato diverso rispetto al β di un analogo modello senza effetti casuali. A questo proposito giova ricordare che in letteratura sono presenti due approcci alternativi per l'analisi di dati gerarchici (Zeger *et al.*, 1988; Goldestein & Rasbash, 1996):

a) L'approccio *unit specific* o *condizionato* (nel quale rientrano i modelli multilivello), che consiste nel rendere esplicita l'influenza della gerarchia per mezzo degli effetti casuali, specificando la distribuzione della risposta in modo condizionato; in questi modelli i coefficienti si riferiscono all'*effetto delle covariate per ogni dato gruppo* (unità di secondo livello).

b) L'approccio *population average* o *marginale*, che si basa sulla specificazione della distribuzione marginale della risposta, considerando la correlazione generata dalla gerarchia come un fattore di disturbo; in questi modelli i coefficienti si riferiscono all'*effetto medio delle covariate nella popolazione*.

[12] Se si effettua la stima con un metodo di *quasi-verosimiglianza* (Wedderburn, 1974) il parametro ϕ può diventare comunque oggetto di stima, qualora si voglia modellare una *extra-variabilità* (Williams, 1982). Tuttavia nei GLM multilivello il problema dell'extra-variabilità è più raro che nei GLM ordinari, poiché gli effetti casuali contribuiscono a modellare in modo migliore la variabilità. Sul ruolo dell'extra-variabilità nei modelli multilivello cfr. Goldstein (1995), pp. 98-99.

I coefficienti di regressione nei due casi possono essere diversi in modo sistematico. Ad esempio, in un modello logit ad intercetta casuale i coefficienti di regressione del modello *unit specific* $(\beta_1^{US}, \ldots, \beta_p^{US})$ e del modello *population average* $(\beta_1^{PA}, \ldots, \beta_p^{PA})$ sono tali che (Neuhaus *et al.*, 1991):
1) $\left|\beta_k^{PA}\right| \leq \left|\beta_k^{US}\right|$ per ogni $k = 1, \ldots, p$;
2) l'uguaglianza vale se e solo se $\beta_k^{US} = 0$;
3) la differenza fra β_k^{PA} e β_k^{US} aumenta all'aumentare della varianza dell'effetto casuale.

La scelta fra i due approcci è dettata dalle finalità dell'indagine: se la struttura gerarchica ha un interesse di per sé è opportuno usare un modello *unit specific*, altrimenti si possono usare entrambi. In effetti, il modello *unit specific* è più generale, poiché può essere usato anche per studiare l'effetto medio delle covariate nella popolazione. L'unica difficoltà è che il valore atteso marginale, come mostrato dalla (2.23), è dato da un integrale che spesso non ha soluzione analitica: tuttavia il problema può essere facilmente risolto per mezzo di un'approssimazione analitica o di una simulazione Monte Carlo (Goldstein, 1995, par. 5.3).

2.2.2 Modelli per dati binari

Quando i dati sono binari, cioè $y_{ij} \in \{0, 1\}$, il valore atteso $E(y_{ij})$ coincide con la probabilità di successo $P\{y_{ij} = 1\}$, e ciò vale anche condizionatamente agli effetti casuali. Pertanto, il GLM multilivello viene solitamente scritto sostituendo $\mu_{ij}^u = E(y_{ij} |\, u_{0j}, u_{1j})$ con[13]

$$\pi_{ij} = P\{y_{ij} = 1 |\, u_{0j}, u_{1j}\}. \tag{2.24}$$

Quanto alla *funzione link* del GLM, le tre scelte più comuni sono

- *probit:* $g(\pi) = \Phi^{-1}(\pi)$, dove è $\Phi(\cdot)$ la funzione di ripartizione della distribuzione normale standard;
- *logit:* $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$, la cui inversa $g^{-1}(x) = \frac{1}{1+\exp(-x)}$ è la funzione di ripartizione della distribuzione logistica standard.
- *complementary log-log:* $g(\pi) = \log\left[-\log(1-\pi)\right]$, la cui inversa $g^{-1}(x) = 1 - \exp\left[-\exp(x)\right]$ è la funzione di ripartizione di una distribuzione di tipo "extreme-value".

[13] Nel prosieguo, per non appesantire la notazione, ometteremo di indicare esplicitamente la dipendenza della probabilità di successo dagli effetti casuali, scrivendo π_{ij} in luogo di π_{ij}^u.

Ricordiamo che la distribuzione logistica standard ha media nulla, varianza $\pi^2/3$ e una forma molto simile a quella di una normale di pari varianza, rispetto alla quale, però, ha le code leggermente più "pesanti". Pertanto i risultati ottenibili con i link probit e logit sono praticamente identici, a meno che non si abbiano probabilità molto vicine a 0 oppure a 1. In effetti la scelta solitamente dipende dall'impostazione teorica che sottende il modello: il logit, essendo il *link canonico*, semplifica le proprietà del modello (McCullagh & Nelder, 1989) e, inoltre, ha il vantaggio di consentire un'interpretazione dei risultati in termini di *odds ratio* (Agresti, 1990); d'altra parte, come vedremo di seguito, il probit rappresenta la scelta più naturale nel caso di un modello a soglia con variabile latente (Winship & Mare, 1983).

Il link *complementary log-log* si distingue dagli altri due per la sua asimmetria e per la varianza della relativa distribuzione "extreme-value", che è pari a $\pi^2/6$. Questo link trova importanti applicazioni nei modelli per dati ordinali e per dati di sopravvivenza in tempo discreto (cfr. parr. 2.2.4 e 2.2.5).

Versione con variabile latente e soglia

I modelli a soglia sono interessanti perché consentono di derivare certe proprietà del modello non lineare da quelle del modello lineare per la variabile latente; inoltre possono essere facilmente estesi al caso di variabili di risposta ordinali. Usando la notazione del GLM multilivello definito nel par. 2.2.1, un *modello multilivello a soglia per dati binari* può essere costruito a partire da una variabile latente (non osservabile) y^*_{ij} che segue un modello lineare a due livelli

$$y^*_{ij} = \alpha + \beta x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij},$$

dove i termini di errore di primo livello sono un campione casuale semplice da una normale, da una logistica o da una distribuzione di tipo "extreme-value"[14]. Anche se y^*_{ij} non è direttamente osservabile, supponiamo che sia possibile sapere se la sua realizzazione supera oppure no un certo valore, detto *soglia*, che solitamente viene posto uguale a zero[15], e definiamo la variabile

[14] Tutte queste distribuzioni vengono assunte nella forma standard, cioè con la varianza fissata (1 nel caso della normale, $\pi^2/3$ nel caso della logistica, $\pi^2/6$ nel caso della "extreme-value"). La scelta arbitraria della varianza del termine di errore non causa una perdita di generalità poiché è una condizione necessaria per l'identificabilità del modello (Winship & Mare, 1983; Hedeker & Gibbons, 1994).

[15] Per l'identificabilità del modello è necessario porre un vincolo sulla soglia oppure sulla costante della variabile latente. L'opzione più comune è quella di porre a zero la soglia (Winship & Mare, 1983; Hedeker & Gibbons, 1994).

binaria osservabile y_{ij} come

$$y_{ij} = I\{y^*_{ij} > 0\},$$

dove $I\{\cdot\}$ è la funzione indicatrice che vale 1 quando l'evento in parentesi è vero. Pertanto, usando la notazione introdotta con la (2.24), si ottiene

$$\pi_{ij} = P\{y^*_{ij} > 0 \mid u_{0j}, u_{1j}\} = g^{-1}(\alpha + \beta x_{ij} + u_{0j} + u_{1j}x_{ij}),$$

ovvero

$$g(\pi_{ij}) = \alpha + \beta x_{ij} + u_{0j} + u_{1j}x_{ij},$$

dove, a seconda della distribuzione del termine di errore di primo livello, la funzione di *link* g è *probit*, *logit* oppure *complementary log-log*.

Nel caso del modello probit è possibile calcolare in forma chiusa il valore atteso marginale della risposta. Infatti, poiché

$$y^*_{ij} \sim N(\alpha + \beta x_{ij}\ ,\ 1 + (\sigma^2_{u0} + \sigma^2_{u1}x^2_{ij} + 2\sigma_{u01}x_{ij})),$$

si ottiene

$$E(y_{ij}) = P\{y^*_{ij} > 0\} = \Phi\left(\frac{\alpha + \beta x_{ij}}{\left[1 + (\sigma^2_{u0} + \sigma^2_{u1}x^2_{ij} + 2\sigma_{u01}x_{ij})\right]^{\frac{1}{2}}}\right),$$

da cui risulta evidente che i coefficienti del modello marginale sono *attenuati* rispetto a quelli del modello multilivello[16].

Un vantaggio della specificazione tramite variabile latente riguarda il calcolo del coefficiente di correlazione intraclasse (cfr. par. 2.1.3). Infatti, con l'eccezione del modello lineare normale, in un GLM multilivello il coefficiente di correlazione intraclasse non è definito nemmeno quando è presente un unico effetto casuale sull'intercetta, poiché, a causa della dipendenza dalla media, la varianza marginale della risposta non è costante. Tuttavia è possibile calcolare il coefficiente sulla variabile latente: infatti, se

$$y^*_{ij} = \alpha + \beta x_{ij} + u_{0j} + e_{ij},$$

il coefficiente di correlazione intraclasse è

$$\frac{\sigma^2_{u0}}{\sigma^2_e + \sigma^2_{u0}},$$

dove σ^2_e è fissato a 1 nel modello *probit*, $\pi^2/3$ nel modello *logit* e $\pi^2/6$ nel modello *complementary log-log*.

[16] Per il modello logit il calcolo del valore atteso marginale in forma chiusa non è possibile, ma Zeger *et al.* (1988) hanno derivato una formula approssimata che conferma il fenomeno dell'attenuazione.

2.2.3 Modelli per dati politomici

I dati si dicono politomici quando le risposte appartengono ad un insieme non ordinato di tre o più categorie, ad esempio "bianco", "nero", "rosso". Per l'individuo i del gruppo j, la risposta politomica può essere rappresentata da un vettore multinomiale

$$\mathbf{y}'_{ij} = (y_{ij}^{(1)}, \ldots, y_{ij}^{(m)}),$$

dove $y_{ij}^{(s)} \in \{0, 1\}$ è una variabile di Bernoulli che vale 1 quando l'osservazione cade nella categoria s $(s = 1, \ldots, m)$. Poiché le categorie sono mutualmente esclusive, si ha $\sum_{s=1}^{m} y_{ij}^{(s)} = 1$. Analogamente al caso binario poniamo

$$\pi_{ij}^{(s)} = P\{y_{ij}^{(s)} = 1 |\ u_{0j}, u_{1j}\}, \qquad s = 1, \ldots, m, \tag{2.25}$$

dove u_{0j} e u_{1j} sono, come al solito, gli effetti casuali relativi all'intercetta e al coefficiente angolare. Le probabilità (2.25) sono legate dal vincolo $\sum_{s=1}^{m} \pi_{ij}^{(s)} = 1$.

Poiché le variabili che compongono il vettore $\mathbf{y}'_{ij}$ sono linearmente dipendenti, una di esse può essere esclusa. Convenzionalmente la variabile esclusa è $y_{ij}^{(m)}$, dove m è la cosiddetta *categoria di base*, scelta arbitrariamente. In questo modo il vettore $(y_{ij}^{(1)}, \ldots, y_{ij}^{(m-1)})$ ha una distribuzione multinomiale con matrice di covarianza non singolare, i cui elementi sono

$$var(y_{ij}^{(s)}) = \pi_{ij}^{(s)}(1 - \pi_{ij}^{(s)}), \qquad s = 1, \ldots, m-1,$$

$$cov(y_{ij}^{(s)}, y_{ij}^{(r)}) = -\pi_{ij}^{(s)}\pi_{ij}^{(r)}, \qquad s \neq r.$$

Il modello più comune per l'analisi di dati politomici si basa sul *link logit multivariato* (Fahrmeir & Tutz, 1994). Nel caso di m categorie, si ottengono $m-1$ modelli logistici, ognuno dei quali confronta la probabilità di una delle prime $m-1$ categorie con quella di base:

$$\log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(m)}}\right) = \alpha^{(s)} + \beta^{(s)}x_{ij} + u_{0j}^{(s)} + u_{1j}^{(s)}x_{ij}, \qquad s = 1, \ldots, m-1. \tag{2.26}$$

Si noti che i parametri fissi e gli effetti casuali sono specifici di ogni equazione e che i parametri casuali modellano non solo la variabilità degli effetti casuali all'interno delle singole equazioni, ma anche la variabilità degli effetti casuali appartenenti a equazioni diverse.

Scrivendo il secondo membro delle equazioni (2.26) come $\eta_{ij}^{(s)}$, le probabilità delle singole categorie sono date da (Fahrmeir & Tutz, 1994)

$$\begin{aligned}\pi_{ij}^{(s)} &= \frac{\exp\left(\eta_{ij}^{(s)}\right)}{1+\sum_{h=1}^{m-1}\exp\left(\eta_{ij}^{(h)}\right)}, \qquad s=1,\ldots,m-1,\\ \pi_{ij}^{(m)} &= \frac{1}{1+\sum_{h=1}^{m-1}\exp\left(\eta_{ij}^{(h)}\right)}.\end{aligned}$$

I parametri del modello logit multivariato possono essere facilmente interpretati osservando che, per due categorie arbitrarie diverse da quella di base, si ha

$$\begin{aligned}\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(r)}} &= \exp\left(\eta_{ij}^{(s)}-\eta_{ij}^{(r)}\right)\\ &= \exp\left(\alpha^{(s)}-\alpha^{(r)}\right)\exp\left((\beta^{(s)}-\beta^{(r)})x_{ij}\right)\times\\ &\quad \exp\left(u_{0j}^{(s)}-u_{0j}^{(r)}\right)\exp\left((u_{1j}^{(s)}-u_{1j}^{(r)})x_{ij}\right).\end{aligned}$$

2.2.4 Modelli per dati ordinali

I dati ordinali sono caratterizzati dal fatto che le risposte appartengono ad un insieme ordinato di categorie, ad esempio "basso", "medio", "alto". Solitamente le categorie vengono contrassegnate con i numeri naturali in modo da rifletterne l'ordinamento; tuttavia la numerazione delle categorie è una semplice convenzione, da non confondere con l'assegnazione di punteggi da utilizzare nei modelli[17]. L'interesse per i dati ordinali deriva dalla loro ampia diffusione, considerando che possono rientrare in questa categoria anche i dati di sopravvivenza in tempo discreto.

I dati ordinali possono essere rappresentati in modo del tutto analogo ai dati politomici, cioè assumendo per ogni individuo un vettore multinomiale

$$\mathbf{y}_{ij}' = (y_{ij}^{(1)},\ldots,y_{ij}^{(m)}),$$

dove $y_{ij}^{(s)} \in \{0,1\}$ è una variabile di Bernoulli che vale 1 quando l'osservazione cade nella categoria s ($s=1,\ldots,m$). Come nel caso politomico poniamo

$$\pi_{ij}^{(s)} = P\{y_{ij}^{(s)}=1|\ u_{0j},u_{1j}\}, \qquad s=1,\ldots,m,$$

[17] In questo lavoro eviteremo di fare ricorso ai modelli basati sull'assegnazione di punteggi, che sono suscettibili di numerose critiche (cfr. Agresti, 1984).

con il vincolo $\sum_{s=1}^{m} \pi_{ij}^{(s)} = 1$.

L'ordinamento delle categorie, che differenzia i dati ordinali da quelli politomici, può essere sfruttato basando i modelli sulle variabili cumulate

$$z_{ij}^{(s)} = \sum_{l=1}^{s} y_{ij}^{(l)} \qquad s = 1, \ldots, m-1;$$

l'ultima variabile cumulata, $z_{ij}^{(m)}$, è identicamente uguale a 1 e quindi non viene presa in considerazione.

Le variabili cumulate hanno valore atteso pari a

$$E(z_{ij}^{(s)}) = \sum_{l=1}^{s} \pi_{ij}^{(l)} = \gamma_{ij}^{(s)} \qquad s = 1, \ldots, m-1,$$

dove $\gamma_{ij}^{(s)}$ è la probabilità che la risposta cada in una categoria non superiore a s. Inoltre, assumendo una distribuzione multinomiale delle risposte, le variabili cumulate hanno varianze e covarianze pari a

$$var(z_{ij}^{(s)}) = \gamma_{ij}^{(s)}(1 - \gamma_{ij}^{(s)}), \qquad s = 1, \ldots, m-1,$$

$$cov(z_{ij}^{(s)}, z_{ij}^{(r)}) = \gamma_{ij}^{(s)}(1 - \gamma_{ij}^{(r)}), \qquad s \leq r.$$

I due modelli fondamentali per l'analisi dei dati ordinali, basati sulle probabilità cumulate, sono il modello con il link *logit*, detto *proportional odds,*

$$\log\left(\frac{\gamma_{ij}^{(s)}}{1 - \gamma_{ij}^{(s)}}\right) = \alpha^{(s)} + \beta x_{ij} + u_{0j} + u_{1j} x_{ij}, \qquad s = 1, \ldots, m-1, \quad (2.27)$$

e il modello con il link *complementary log-log*, detto *proportional hazards,*

$$\log\left[-\log(1 - \gamma_{ij}^{(s)})\right] = \alpha^{(s)} + \beta x_{ij} + u_{0j} + u_{1j} x_{ij}, \qquad s = 1, \ldots, m-1. \quad (2.28)$$

In entrambe le specificazioni il coefficiente β è identico per tutti gli s, mentre le intercette sono ordinate in modo non decrescente, $\alpha^{(1)} \leq \alpha^{(2)} \ldots \leq \alpha^{(m-1)}$. Gli effetti casuali sono ipotizzati comuni a tutti gli s, anche se questa non è un'assunzione necessaria.

Il modello (2.27) viene detto *proportional odds* perché il rapporto fra gli odds di due individui non dipende da s:

$$\frac{\frac{\gamma_{ij}^{(s)}}{1-\gamma_{ij}^{(s)}}}{\frac{\gamma_{i'j'}^{(s)}}{1-\gamma_{i'j'}^{(s)}}} = \exp\left[\beta(x_{ij} - x_{i'j'}) + (u_{0j} - u_{0j'}) + (u_{1j} x_{ij} - u_{1j'} x_{i'j'})\right].$$

Invece il modello (2.28) è noto come *proportional hazards* perché, pensando i dati ordinali come tempi di sopravvivenza, rappresenta una versione discreta del *modello a rischi proporzionali di Cox* che si ottiene raggruppando le osservazioni in intervalli. Questa versione discreta, dovuta a McCullagh (1980), si aggiunge alla versione discreta di Prentice & Gloeckler (1978). Nel prossimo paragrafo introdurremo l'analisi di sopravvivenza e discuteremo brevemente le proprietà delle versioni discrete del modello di Cox. Per il momento ci limitiamo ad osservare che nel modello (2.28) la *funzione di sopravvivenza* discreta è data da

$$1 - \gamma_{ij}^{(s)} = \left\{ \exp\left[-\exp(\alpha^{(s)}) \right] \right\}^{\exp(\beta x_{ij} + u_{0j} + u_{1j} x_{ij})}, \tag{2.29}$$

dove $\exp\left[-\exp(\alpha^{(s)})\right]$ è la *funzione di sopravvivenza* di base, cioè relativa ad un individuo con covariate ed effetti casuali nulli. Inoltre il *rischio* o *hazard* al tempo s è dato da

$$\frac{\gamma_{ij}^{(s)} - \gamma_{ij}^{(s-1)}}{1 - \gamma_{ij}^{(s-1)}} = 1 - \exp\left\{ \left[\exp(\alpha^{(s-1)}) - \exp(\alpha^{(s)}) \right] \exp(\beta x_{ij} + u_{0j} + u_{1j} x_{ij}) \right\}.$$

Versione con variabile latente e soglie

I modelli (2.27) e (2.28) possono essere derivati anche per mezzo di un modello a soglia, in modo del tutto analogo a quanto visto per i dati binari. Data la variabile latente

$$y_{ij}^* = \alpha + \beta x_{ij} + u_{0j} + u_{1j} x_{ij} + e_{ij}$$

e un insieme di valori di soglia $-\infty = \nu_0 \leq \nu_1 \leq \nu_2 \ldots \leq \nu_{m-1} \leq \nu_m = \infty$, ponendo

$$y_{ij}^{(s)} = I\{\nu_{s-1} < y_{ij}^* \leq \nu_s\} \qquad s = 1, \ldots, m,$$

si ottiene

$$\begin{aligned} \pi_{ij}^{(s)} &= P\{\nu_{s-1} < y_{ij}^* \leq \nu_s \mid u_{0j}, u_{1j}\} \\ &= P\{y_{ij}^* \leq \nu_s \mid u_{0j}, u_{1j}\} - P\{y_{ij}^* \leq \nu_{s-1} \mid u_{0j}, u_{1j}\} \\ &= g^{-1}(\nu_s - (\alpha + \beta x_{ij} + u_{0j} + u_{1j} x_{ij})) \\ &\quad - g^{-1}(\nu_{s-1} - (\alpha + \beta x_{ij} + u_{0j} + u_{1j} x_{ij})), \end{aligned}$$

dove, a seconda della distribuzione del termine di errore di primo livello, il link g è il *logit* oppure il *complementary log-log*. In termini di probabilità cumulate ciò equivale a

$$\gamma_{ij}^{(s)} = g^{-1}(\nu_s - (\alpha + \beta x_{ij} + u_{0j} + u_{1j} x_{ij})), \qquad s = 1, \ldots, m-1,$$

ovvero

$$g(\gamma_{ij}^{(s)}) = \nu_s - (\alpha + \beta x_{ij} + u_{0j} + u_{1j}x_{ij}), \qquad s = 1, \ldots, m-1. \tag{2.30}$$

Il modello (2.30), a parte una differenza nella parametrizzazione, è lo stesso delle equazioni (2.27) e (2.28), cioè è il modello *proportional odds* se g è il link *logit* ed è il modello *proportional hazards* se g è il link *complementary log-log*. A proposito di parametrizzazioni osserviamo che:

a) Il modello (2.30) necessita di un vincolo, poiché ha un parametro di troppo per l'intercetta. Nei modelli a soglia il vincolo di solito è $\nu_1 = 0$ (Hedeker & Gibbons, 1994), altrimenti si può porre $\alpha = 0$, ottenendo così la stessa parametrizzazione dei modelli (2.27) e (2.28).

b) Il coefficiente β e gli effetti casuali hanno il segno invertito rispetto alle equazioni (2.27) e (2.28). Infatti, una covariata che ha un effetto positivo sulla variabile latente (nel senso che un incremento di x determina un incremento di $E(y^*)$) ha, allo stesso tempo, un effetto negativo sulle probabilità cumulate, come risulta evidente dalla definizione del modello a soglia.

2.2.5 Modelli per dati di sopravvivenza in tempo discreto

In questo paragrafo presentiamo i principali modelli dell'analisi di sopravvivenza in tempo discreto, esaminando poi l'estensione al caso multilivello.

Introduzione

I dati di *sopravvivenza* scaturiscono da indagini di tipo longitudinale finalizzate all'osservazione del tempo intercorrente fra due eventi, il secondo dei quali viene convenzionalmente chiamato *morte*[18]. L'oggetto di interesse è dunque il tempo di attesa T, che, a seconda dei casi, si ipotizza essere una variabile aleatoria continua o discreta. Tuttavia l'osservazione di T per tutti gli individui del campione è generalmente impossibile, sia perché ciò può richiedere un tempo di osservazione estremamente lungo e non pianificabile, sia perché alcuni individui possono sottrarsi all'osservazione prima di aver

[18] I termini sopravvivenza e morte traggono origine dalle indagini demografiche e mediche in cui l'evento finale è la morte della persona (l'evento iniziale può essere la nascita, la diagnosi di una certa malattia ecc.). Naturalmente l'evento finale può essere di qualunque tipo e può avere connotati positivi (ad esempio, trovare lavoro), ma la terminologia in uso è quella relativa alle indagini sulla sopravvivenza.

sperimentato l'evento di interesse. Questo fenomeno, tipico dei dati di sopravvivenza, è noto con il termine di *censura a destra*[19]. Pertanto i dati di sopravvivenza sono solitamente costituiti da coppie di variabili aleatorie (X, δ), dove X è il tempo osservato e δ è un indicatore che vale 1 se l'osservazione si è conclusa con l'evento di interesse e 0 se si è conclusa con la censura. I metodi dell'analisi di sopravvivenza sono costruiti in modo tale da fare inferenza su T a partire dalle osservazioni su (X, δ).

I metodi dell'analisi di sopravvivenza sono usualmente basati sulla *funzione di sopravvivenza* e sulla *funzione di rischio* o *hazard*, che, nel caso di una variabile aleatoria continua T, sono definite rispettivamente da

$$S(t) = P(T > t);$$

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

Dunque $S(t)$ è la probabilità di sopravvivere oltre il tempo t, mentre $\lambda(t)$, che assume valori nell'intervallo $[0, \infty)$, è il rischio istantaneo di morte al tempo t per un individuo sopravvissuto fino a quell'istante. Le funzioni di sopravvivenza e di rischio non sono altro che modi alternativi di caratterizzare la distribuzione di T che risultano utili per la definizione dei modelli e l'interpretazione dei risultati. Si può passare da una funzione all'altra per mezzo delle seguenti relazioni (Kalbfleisch & Prentice, 1980):

$$\lambda(t) = -\frac{\partial}{\partial t} \log S(t);$$

$$S(t) = \exp\{-\int_0^t \lambda(s) ds\}. \tag{2.31}$$

Se la variabile aleatoria T è discreta, la definizione della funzione di sopravvivenza rimane invariata, mentre quella della funzione di rischio diviene

$$\lambda(t) = P(T = t \mid T \geq t),$$

per cui nel caso discreto l'hazard è una vera e propria probabilità. La relazione fra le funzioni di sopravvivenza e di rischio in tempo discreto è

$$S(t) = \prod_{s=1}^{t} (1 - \lambda(s)). \tag{2.32}$$

[19] Nei dati di sopravvivenza esiste un'ampia casistica di osservazioni incomplete, di cui la censura a destra rappresenta il caso di gran lunga più frequente (Kalbfleisch & Prentice, 1980). Un'ipotesi fondamentale che sta alla base dell'analisi di sopravvivenza è che il meccanismo di censura sia indipendente dal processo che governa il succedersi degli eventi: questa è una condizione necessaria per poter riferire le conclusioni dell'analisi al tempo sottostante T, che non è direttamente osservabile per tutti gli individui (Kalbfleish & Prentice, 1980).

Inoltre, mentre in tempo continuo $P(T = t) = 0$, in tempo discreto si ha

$$P(T = t) = \lambda(t)S(t-1) = \lambda(t)\prod_{s=1}^{t-1}(1 - \lambda(s)). \tag{2.33}$$

Alcuni modelli classici

Consideriamo innanzitutto il *modello a rischi proporzionali di Cox* (Cox, 1972), il modello in tempo continuo più ampiamente usato, che rappresenta un punto di partenza per gli sviluppi di alcuni modelli in tempo discreto. Dato un campione casuale di individui $i = 1, \ldots, n$, il modello di Cox si basa sulla seguente specificazione della funzione di rischio:

$$\lambda(t \mid \mathbf{x}_i) = \lambda_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta}), \qquad t \in [0, \infty),$$

dove $\mathbf{x}_i$ è un vettore di *covariate fisse*[20] per l'individuo i, $\boldsymbol{\beta}$ è un vettore di parametri e $\lambda_0(\cdot)$ è una funzione non specificata, detta *funzione di rischio di base*, che rappresenta l'andamento del rischio per un individuo con $\mathbf{x}_i = \mathbf{0}$. Il modello è semiparametrico poiché, nonostante la presenza di un vettore di parametri, la distribuzione del tempo T non è completamente specificata[21]: proprio questa è la maggiore virtù del modello di Cox, che permette di studiare il rischio relativo fra gli individui senza bisogno di fare troppe ipotesi sulla distribuzione di T. Il modello viene detto a rischi proporzionali perché fra le funzioni di rischio di due individui qualunque esiste un rapporto di proporzionalità costante nel tempo:

$$\frac{\lambda(t \mid \mathbf{x}_i)}{\lambda(t \mid \mathbf{x}_{i'})} = \exp\left[(\mathbf{x}_i - \mathbf{x}_{i'})'\boldsymbol{\beta}\right] \qquad \text{per ogni } t \in [0, \infty).$$

Il modello di Cox viene usualmente specificato per mezzo della funzione di rischio, ma, per la relazione (2.31), può anche essere visto in termini della funzione di sopravvivenza:

$$S(t \mid \mathbf{x}_i) = S_0(t)^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}, \tag{2.34}$$

dove $S_0(t)$ è la funzione di sopravvivenza di base, cioè per un individuo con $\mathbf{x}_i = \mathbf{0}$. Inoltre il modello di Cox può essere esteso al caso di covariate tempo-dipendenti: in tal caso però, oltre ad avere problemi teorici e computazionali, si perde la proprietà di rischi proporzionali (Kalbfleisch & Prentice, 1980).

[20] Nell'analisi di sopravvivenza il termine *covariata fissa* viene usato in contrapposizione al termine *covariata tempo-dipendente*, per indicare che la covariata assume, per ogni individuo, un unico valore che non cambia nel tempo.

[21] La distribuzione del tempo T può essere completamente specificata solo specificando la funzione di rischio di base $\lambda_0(\cdot)$: ad esempio, se tale funzione è costante, T ha una distribuzione esponenziale.

Mentre il tempo di per sé è continuo, la sua misurazione avviene necessariamente ad intervalli discreti. Quando tali intervalli non sono abbastanza piccoli l'uso dei modelli in tempo continuo presenta dei seri problemi dovuti alla presenza di *parità*, cioè individui che sperimentano l'evento di interesse nello stesso intervallo di misura[22]. Inoltre ci sono delle situazioni in cui l'evento di interesse può verificarsi solo in determinati momenti, per cui il tempo deve considerarsi discreto (si pensi a indagini riguardanti le dichiarazioni dei redditi). Pertanto in molte applicazioni è opportuno, se non necessario, ricorrere a modelli in tempo discreto.

I modelli di sopravvivenza in tempo discreto possono essere definiti seguendo due diversi approcci (Allison, 1982): il primo consiste nel trattare il tempo come se fosse effettivamente discreto (Myers *et al.*, 1973), mentre il secondo assume l'esistenza di un modello sottostante in tempo continuo con osservazioni raccolte in determinati intervalli temporali (Holford, 1976).

Nell'ambito del primo approccio il modello più largamente usato si basa sul link *logit*. Dato un campione casuale di individui $i = 1, \ldots, n$ e indicando con k è l'ultimo tempo osservato nel campione si ha (Cox, 1972)

$$\log \left[\frac{\lambda(t \mid \mathbf{x}_{it})}{1 - \lambda(t \mid \mathbf{x}_{it})} \right] = \alpha_t + \mathbf{x}'_{it}\boldsymbol{\beta}, \qquad t = 1, \ldots, k, \tag{2.35}$$

dove $\mathbf{x}_{it}$ è un vettore di covariate per l'individuo i al tempo t. I parametri $(\alpha_1, \ldots, \alpha_k)$ modellano il rischio di base, svolgendo un ruolo analogo a quello della funzione $\lambda_0(\cdot)$ nel modello di Cox, mentre i parametri del vettore $\boldsymbol{\beta}$ misurano l'effetto delle covariate sul logit del rischio. Per due individui i e i' si ha

$$\frac{\frac{\lambda(t \mid \mathbf{x}_{it})}{1-\lambda(t \mid \mathbf{x}_{it})}}{\frac{\lambda(t \mid \mathbf{x}_{i't})}{1-\lambda(t \mid \mathbf{x}_{i't})}} = \exp\left[(\mathbf{x}_{it} - \mathbf{x}_{i't})'\boldsymbol{\beta}\right],$$

per cui, se le covariate sono fisse (cioè $\mathbf{x}_{it} = \mathbf{x}_i$ per ogni t), gli odds dei rischi sono proporzionali.

Se invece si assume che i dati siano generati da un modello di Cox, il corrispondente modello in tempo discreto per osservazioni raggruppate in intervalli si basa sul link *complementary log-log* (Prentice & Gloeckler, 1978):

$$\log[-\log(1 - \lambda(t \mid \mathbf{x}_{it}))] = \alpha_t + \mathbf{x}'_{it}\boldsymbol{\beta}, \qquad t = 1, \ldots, k, \tag{2.36}$$

dove il vettore di parametri $\boldsymbol{\beta}$ è identico a quello del modello di Cox sottostante[23]. Si noti che nella versione discreta la funzione di sopravvivenza ha

[22] In un modello in tempo continuo una parità è un evento di probabilità nulla e quindi la presenza di molte parità rende il modello inadeguato.

[23] Per questo motivo nel modello con link complementary log-log, a differenza del modello

la stessa specificazione del modello di Cox: infatti dalla (2.32) si ottiene

$$S(t) = \exp\left[-\sum_{s=1}^{t} \exp(\alpha_s) \exp(\mathbf{x}'_{is}\boldsymbol{\beta})\right],$$

che, nel caso di covariate fisse, fornisce la (2.34):

$$S(t \mid \mathbf{x}_i) = S_0(t)^{\exp(\mathbf{x}'_i\boldsymbol{\beta})},$$

dove $S_0(t) = \exp\left[-\sum_{s=1}^{t} \exp(\alpha_s)\right]$ è la funzione di sopravvivenza di base. Si noti, però, che nella versione discreta i rischi non sono proporzionali[24].

La specificazione (2.34) della funzione di sopravvivenza caratterizza anche la versione discreta proposta da McCullagh (1980) che abbiamo introdotto nel paragrafo precedente direttamente per il caso multilivello (vedi formula (2.29)). In entrambe le versioni discrete il vettore di parametri $\boldsymbol{\beta}$ è identico a quello del modello di Cox sottostante (in sostanza le due versioni discrete differiscono solo per la parametrizzazione della funzione di sopravvivenza di base: cfr. Läärä & Matthews, 1985).

A differenza dei modelli in tempo continuo, nei modelli in tempo discreto l'inclusione di covariate tempo-dipendenti è del tutto naturale. Ciò risulta utile anche quando i dati contengano esclusivamente covariate fisse: infatti, il modo più semplice per consentire ad una covariata fissa di avere un effetto variabile nel tempo è quello di costruire una covariata tempo-dipendente fittizia definita dall'interazione fra il tempo e la covariata fissa di interesse.

Sia nel modello logit che in quello complementary log-log si possono imporre delle restrizioni sui parametri α_t che risultano particolarmente utili quando l'insieme di tali parametri sia molto numeroso. Ad esempio, Mantel & Hankey (1978) propongono una specificazione attraverso un polinomio in t:

$$\alpha_t = \sum_{r=0}^{R} \delta_r t^r,$$

per cui i parametri $(\alpha_1, \ldots, \alpha_k)$ vengono sostituiti dai parametri $(\delta_0, \ldots, \delta_R)$.

con link logit, il vettore di parametri $\boldsymbol{\beta}$ è invariante rispetto alla suddivisione del tempo in intervalli (Allison, 1982). Comunque, ai fini pratici, la differenza fra i due modelli è spesso irrilevante e si annulla quando la partizione del tempo in intervalli è molto fine, poiché il modello con link logit converge al modello di Cox al tendere a zero dell'ampiezza degli intervalli (Thompson, 1977).

[24] In tempo discreto la proporzionalità dei rischi è resa impossibile dal fatto che i rischi sono compresi nell'intervallo $[0, 1]$. Quello che si può imporre è la proporzionalità degli odds dei rischi, come accade nel modello con il link logit.

Rappresentazione per mezzo di variabili indicatrici

Prima di passare alla versione multilivello di questi modelli è opportuno esaminare brevemente una rappresentazione alternativa dei dati di sopravvivenza utile a fini computazionali, che estenderemo poi alla versione multilivello. Per cominciare osserviamo che, per un qualsiasi modello di sopravvivenza in tempo discreto, la verosimiglianza è

$$L = \prod_{i=1}^{n} \left\{ [P(T_i = t_i)]^{\delta_i} \, [P(T_i > t_i)]^{1-\delta_i} \right\}, \tag{2.37}$$

dove t_i è il tempo osservato per l'individuo i e δ_i è l'indicatore di non censura[25]. Pertanto, in base alle (2.32)-(2.33), la verosimiglianza può essere scritta in termini di hazard nel seguente modo:

$$L = \prod_{i=1}^{n} \left\{ [\lambda(t_i \mid \mathbf{x}_{it_i})]^{\delta_i} \, [1 - \lambda(t_i \mid \mathbf{x}_{it_i})]^{1-\delta_i} \prod_{s=1}^{t_i-1} [1 - \lambda(s \mid \mathbf{x}_{is})] \right\}. \tag{2.38}$$

Adesso, seguendo un'idea di Brown (1975), definiamo un'insieme di variabili *dummy* y_{is} tali che $y_{is} = 1$ se e solo se l'individuo i sperimenta l'evento di interesse al tempo s ($s = 1, 2, ..., t_i$). In questo modo, per ogni individuo del campione, la coppia (t_i, δ_i) viene sostituita da un vettore $(y_{i1}, y_{i2}, \ldots, y_{it_i})$ che assume i valori $(0, 0, \ldots, 0, 1)$ se $\delta_i = 1$ oppure $(0, 0, \ldots, 0, 0)$ se $\delta_i = 0$. Pertanto la verosimiglianza (2.38) può riscriversi come

$$L = \prod_{i=1}^{n} \prod_{s=1}^{t_i} [\lambda(s \mid \mathbf{x}_{is})]^{y_{is}} \, [1 - \lambda(s \mid \mathbf{x}_{is})]^{1-y_{is}}. \tag{2.39}$$

La (2.39) non è altro che la verosimiglianza di un campione casuale

$$\{y_{is} : i = 1, \ldots, n; \;\; s = 1, 2, ..., t_i\}$$

di variabili casuali Bernoulli con probabilità di successo

$$P(y_{is} = 1 \mid \mathbf{x}_{is}) = \lambda(s \mid \mathbf{x}_{is}).$$

Pertanto i modelli di sopravvivenza in tempo discreto possono essere visti come modelli di regressione per dati binari applicati ad un campione esteso che si ottiene sostituendo ad ogni record i contributi relativi alle singole

[25] Nella (2.37) si assume che un individuo censurato venga osservato fino all'unità temporale t_i inclusa. Nel caso di dati osservati ad intervalli ciò equivale ad assumere che la censura intervenga al termine dell'intervallo. Sulle implicazioni di questa assunzione cfr. Allison (1982), p.71.

unità temporali in cui l'individuo è stato osservato. Pertanto la stima dei parametri dei modelli (2.35) e (2.36) può essere effettuata applicando al campione opportunamente esteso la procedura di stima relativa ad un modello di regressione per dati binari con link logit o complementary log-log[26].

Come risulta evidente dalla (2.39) le osservazioni del campione esteso sono statisticamente indipendenti, anche quelle che si riferiscono a tempi diversi di uno stesso individuo. Questa indipendenza può apparire strana, ma è necessaria per garantire l'equivalenza dei modelli. In effetti abbandonare l'ipotesi di indipendenza per le osservazioni relative ad un individuo significa ammettere la presenza di *eterogeneità non osservabile* (Allison, 1982, p. 82).

Versione multilivello

I modelli di sopravvivenza in tempo discreto presentati in questo paragrafo possono essere facilmente estesi al caso multilivello. Consideriamo una struttura gerarchica a due livelli, indicando con ij l'individuo i del gruppo j ($i = 1, \ldots, n_j$; $j = 1, \ldots, J$). Una semplice versione multilivello dei modelli (2.35) e (2.36) è data da

$$g(\lambda(t \mid \mathbf{x}_{ijt}, u_{0j})) = \alpha_t + \mathbf{x}'_{ijt}\boldsymbol{\beta} + u_{0j}, \qquad t = 1, \ldots, k, \tag{2.40}$$

dove $g(\cdot)$ è la funzione logit o complementary log-log, mentre $u_{0j} \overset{iid}{\sim} N(0, \sigma^2_{u0})$. In questo modello l'effetto casuale provoca una traslazione della funzione di rischio di base nella scala indotta dalla trasformazione $g(\cdot)$. Il modello (2.40) gode delle proprietà dei modelli (2.35) e (2.36) condizionatamente agli effetti casuali, ma non marginalmente (si ricordi la distinzione fra modelli *unit-specific* e *polulation-average* delineata nel par. 2.2.1). Ad esempio, nel caso di link logit e covariate fisse si ha

$$\frac{\frac{\lambda(t \mid \mathbf{x}_{ij}, u_{0j})}{1-\lambda(t \mid \mathbf{x}_{ij}, u_{0j})}}{\frac{\lambda(t \mid \mathbf{x}_{i'j'}, u_{0j'})}{1-\lambda(t \mid \mathbf{x}_{i'j'}, u_{0j'})}} = \exp\left[(\mathbf{x}_{ij} - \mathbf{x}_{i'j'})\boldsymbol{\beta} + (u_{0j} - u_{0j'})\right],$$

per cui, condizionatamente agli effetti casuali, gli odds dei rischi sono proporzionali[27].

[26] Ovviamente il limite di questa strategia sta nella numerosità del campione esteso, che può essere enorme nel caso che gli individui vengano osservati per una lunga sequenza di tempi. D'altra parte, quando il tempo sottostante è continuo e gli intervalli di osservazione sono sufficientemente piccoli si possono usare direttamente i modelli in tempo continuo.

[27] In questo caso, incidentalmente, anche gli odds marginali sono proporzionali, poiché in un modello logit a intercetta casuale i parametri dei modelli marginale e condizionato sono legati da un fattore di proporzionalità (Zeger *et al.*, 1988).

Il modello (2.40) può includere anche dei coefficienti casuali per modellare un effetto delle covariate differenziato nei gruppi. Ad esempio, se è presente una sola covariata ed il suo coefficiente è casuale, il modello (2.40) diviene

$$g(\lambda(t \mid x_{ijt}, u_{0j}, u_{1j})) = \alpha_t + \beta x_{ijt} + u_{0j} + u_{1j} x_{ijt}, \qquad t = 1, \ldots, k, \quad (2.41)$$

con

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix} \right).$$

Le considerazioni sull'equivalenza fra un modello di sopravvivenza in tempo discreto e un modello per dati binari applicato ad un campione opportunamente esteso possono essere ripetute anche nel caso multilivello, trattando tutte le variabili condizionatamente agli effetti casuali. Pertanto un modello come il (2.40) o il (2.41) è equivalente ad un modello multilivello per dati binari (con link logit o complementary log-log) applicato al campione esteso che si ottiene sostituendo ad ogni record i contributi relativi alle singole unità temporali in cui l'individuo è stato osservato. L'unica accortezza riguarda la definizione della struttura gerarchica. Infatti il campione esteso ha, formalmente, una struttura a tre livelli, in cui le osservazioni sulle singole unità temporali di un individuo costituiscono il primo livello, gli individui il secondo livello e i gruppi il terzo livello. Tuttavia l'equivalenza di cui si discute richiede l'eliminazione del secondo livello, quello relativo agli individui, poiché si assume che tutte le osservazioni sulle singole unità temporali, anche quelle riferite ad uno stesso individuo, siano indipendenti condizionatamente agli effetti casuali relativi al gruppo. Pertanto nella specificazione del modello la variabilità di secondo livello deve essere vincolata a zero, oppure la struttura gerarchica deve essere ridotta a due livelli accorpando il primo e il secondo livello. Per analogia a quanto discusso nel caso dei modelli ordinari, l'inclusione di effetti casuali a livello di individuo comporterebbe una eterogeneità non osservabile, nel senso che due individui con le stesse covariate ed appartenenti allo stesso gruppo avrebbero rischi diversi.

2.2.6 Stima

La stima dei parametri dei GLM multilivello è un problema piuttosto complesso, innanzitutto perché la verosimiglianza marginale in generale non è esprimibile in forma chiusa. Infatti tali modelli sono definiti condizionatamente agli effetti casuali e quindi è immediatamente determinabile solo la verosimiglianza condizionata, dalla quale si ottiene poi quella marginale integrando rispetto alla distribuzione degli effetti casuali: in generale questa operazione di integrazione non è fattibile per via analitica, salvo il caso di

distribuzioni coniugate (Lee & Nelder, 1996). Il caso più importante di coniugatezza si ha quando sia la distribuzione condizionata della risposta che la distribuzione degli effetti casuali sono normali, cosa che accade nel modello multilivello lineare. Negli altri casi per sfruttare la coniugatezza occorre di volta in volta assumere per gli effetti casuali una distribuzione coniugata con quella della risposta, il che spesso costituisce un vincolo inaccettabile (Longford, 1996).

Pertanto i numerosi metodi proposti in letteratura rappresentano possibili soluzioni al problema della stima in presenza di una verosimiglianza non esprimibile in forma chiusa. Fra i principali metodi ricordiamo:

- Massima verosimiglianza marginale con integrazione numerica di Gauss-Hermite (Anderson & Aitkin, 1985; Hedeker & Gibbons, 1994; Rampichini, 1994);
- Metodi di quasi-verosimiglianza, fra cui Quasi-Verosimiglianza Marginale (Marginal Quasi-Likelihood, MQL) e Quasi-Verosimiglianza Penalizzata (Penalized Quasi-Likelihood, PQL) (Goldstein, 1991; Breslow & Clayton, 1993; Goldstein & Rasbash, 1996);
- Metodi basati sulle equazioni di stima generalizzate (Liang & Zeger, 1986);
- Metodi bayesiani basati su simulazioni di tipo Markov Chain Monte Carlo (MCMC) (Zeger & Karim, 1991);
- Metodi classici basati su simulazioni (Mealli & Rampichini, 1999; Calzolari *et al.*, 1999).

Descriveremo in dettaglio i seguenti metodi, di cui ci siamo avvalsi per le elaborazioni di questa tesi: 1) Massima verosimiglianza marginale con integrazione numerica di Gauss-Hermite, implementato nei programmi MIXOR (Hedeker & Gibbons, 1996) e MIXNO (Hedeker, 1998); b) MQL e PQL, implementati nel programma MLwiN (Goldstein *et al.*, 1998).

Massima verosimiglianza marginale con integrazione numerica di Gauss-Hermite

Consideriamo un modello multilivello per dati ordinali nella specificazione basata su variabile latente e soglie, così come discusso nel par. 2.2.4. Questo tipo di modello consente di illustrare in modo naturale il metodo di stima, includendo inoltre come caso particolare i modelli a risposta binaria. Successivamente, accenneremo all'estensione ai modelli a risposta politomica.

Per semplificare la notazione scriviamo la variabile latente del par. 2.2.4 come

$$y^*_{ij} = z_{ij} + e_{ij},$$

dove

$$z_{ij} = \alpha + \beta x_{ij} + u_{0j} + u_{1j} x_{ij}.$$

Assumiamo m categorie e un insieme di valori di soglia

$$-\infty = \nu_0 \leq \nu_1 \leq \nu_2 \ldots \leq \nu_{m-1} \leq \nu_m = \infty,$$

in cui la soglia ν_1 viene posta uguale a 0 per motivi di identificabilità. Pertanto, la probabilità di un'arbitraria categoria s è (cfr. par. 2.2.4)

$$\pi^{(s)}_{ij} = g^{-1}(\nu_s - z_{ij}) - g^{-1}(\nu_{s-1} - z_{ij}),$$

dove g è la funzione link (probit, logit o complementary log-log).

Raccogliendo i parametri di soglia nel vettore $\boldsymbol{\nu}' = (\nu_2, \ldots, \nu_{m-1})$ e ponendo

$$\mathbf{y}'_j = (y_{1j}, \ldots, y_{n_j j}), \quad \mathbf{u}'_j = (u_{0j}, u_{1j}),$$

la verosimiglianza condizionata agli effetti casuali può scriversi, relativamente al j-mo gruppo, nel seguente modo:

$$L(\alpha, \beta, \sigma^2_{u0}, \sigma^2_{u1}, \sigma_{u01}, \boldsymbol{\nu} \mid \mathbf{y}_j, \mathbf{u}_j) = \prod_{i=1}^{n_j} \prod_{s=1}^{m} \left[g^{-1}(\nu_s - z_{ij}) - g^{-1}(\nu_{s-1} - z_{ij}) \right]^{d_{ijs}},$$

dove d_{ijs} è un indicatore che vale 1 se e solo se $y_{ij} = s$.

Adesso consideriamo la scomposizione di Cholesky della matrice di covarianza degli effetti casuali, indicando con $\boldsymbol{\Psi}$ quella matrice sottotriangolare tale che

$$\boldsymbol{\Psi}\boldsymbol{\Psi}' = \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}.$$

Ciò consente di riparametrizzare il modello, sostituendo $\mathbf{u}_j$ con $\boldsymbol{\Psi}\mathbf{w}_j$. Assumendo per $\mathbf{u}_j$ una distribuzione normale multivariata, segue che $\mathbf{w}_j$ ha una distribuzione normale multivariata *standard*[28]. Perciò i tre parametri casuali $(\sigma^2_{u0}, \sigma^2_{u1}, \sigma)$ vengono sostituiti dai tre parametri di $\boldsymbol{\Psi}$, che indichiamo con il vettore $\boldsymbol{\psi}$.

[28] Questa riparametrizzazione è utile per l'implementazione dell'integrazione numerica. Tuttavia presenta anche il vantaggio di consentire una stima più stabile dei parametri casuali nel caso che questi siano prossimi a zero (infatti la scomposizione di Cholesky è una versione matriciale della radice quadrata).

Tenuto conto della riparametrizzazione, la verosimiglianza marginale del j-mo gruppo è data da

$$L(\alpha, \beta, \boldsymbol{\psi}, \boldsymbol{\nu} \mid \mathbf{y}_j) = \int_{-\infty}^{\infty} L(\alpha, \beta, \boldsymbol{\psi}, \boldsymbol{\nu} \mid \mathbf{y}_j, \mathbf{w}) \phi(\mathbf{w}) d\mathbf{w}, \tag{2.42}$$

dove $\phi(\cdot)$ denota la densità della distribuzione normale multivariata standard. Poiché le osservazioni relative a gruppi diversi sono marginalmente indipendenti (cfr. par. 2.2.1), la verosimiglianza complessiva è data dal prodotto di verosimiglianze come la (2.42) e quindi il suo logaritmo naturale è esprimibile come una somma di contributi, uno per ogni gruppo:

$$\log L = \sum_{j=1}^{J} \log L(\alpha, \beta, \boldsymbol{\psi}, \boldsymbol{\nu} \mid \mathbf{y}_j), \tag{2.43}$$

dove $L = L(\alpha, \beta, \boldsymbol{\psi}, \boldsymbol{\nu} \mid \mathbf{y}_1, \ldots, \mathbf{y}_J)$.

La log-verosimiglianza marginale (2.43) può essere massimizzata con il classico algoritmo *Fisher Scoring*: indicando con $\boldsymbol{\theta}$ il vettore di tutti i parametri del modello e con $\boldsymbol{\theta}_t$ il valore che esso assume alla t-ma iterazione, si ha

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \left[E\left(-\frac{\partial^2 \log L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \right) \right]^{-1} \left[\frac{\partial \log L}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \right],$$

dove la matrice di informazione attesa è data da (Hedeker & Gibbons, 1994)

$$E\left(-\frac{\partial^2 \log L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right) = \sum_{j=1}^{J} \left(L(\boldsymbol{\theta} \mid \mathbf{y}_j) \right)^{-2} \left(\frac{\partial L(\boldsymbol{\theta} \mid \mathbf{y}_j)}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial L(\boldsymbol{\theta} \mid \mathbf{y}_j)}{\partial \boldsymbol{\theta}} \right)'.$$

Le espressioni delle derivate di log$L$ rispetto ai vari tipi di parametro sono riportate in Hedeker & Gibbons (1994). Ognuna di queste derivate include un integrale rispetto alla densità della distribuzione normale multivariata standard. Poiché l'algoritmo *Fisher Scoring* prevede il calcolo, ad ogni iterazione, del valore delle derivate di log$L$ nel punto $\boldsymbol{\theta}_t$, si rende necessario approssimare in qualche modo gli integrali presenti, poiché tali integrali non sono risolvibili per via analitica. Una soluzione semplice ed efficace consiste nell'integrazione numerica secondo il *metodo di quadratura di Gauss-Hermite*.

Il metodo di Gauss-Hermite consiste nell'approssimare l'integrale con la somma *ponderata* dei valori della funzione integranda calcolati in una serie di punti, detti *punti di quadratura*. Nel caso di un integrale ad una dimensione si ha

$$\int_{-\infty}^{\infty} f(s)\phi(s)ds \simeq \sum_{q=1}^{Q} f(x_q) p_{x_q},$$

dove $f(\cdot)$ è una funzione arbitraria, $\phi(\cdot)$ è la densità della distribuzione normale univariata standard, Q è il numero di punti di quadratura e $\{(x_q, p_{x_q}) : q = 1, \ldots, Q\}$ sono, rispettivamente, i punti di quadratura e i pesi associati, che vengono scelti in base a criteri di ottimalità (Stroud & Sechrest, 1966). La scelta fondamentale riguarda il valore di Q, al crescere del quale aumenta la bontà dell'approssimazione, ma anche la mole di calcoli. In molti casi valori compresi tra 5 e 10 rappresentano un buon compromesso. Nel caso di un integrale a r dimensioni ogni punto di quadratura diviene un vettore r-dimensionale

$$\mathbf{x}_q = (x_{q1}, \ldots, x_{qr}),$$

il cui peso (scalare) associato è dato dal prodotto dei corrispondenti pesi univariati:

$$p_{\mathbf{x}_q} = \prod_{h=1}^{r} p_{x_{qh}}.$$

Poiché i punti di quadratura r-dimensionali si ottengono incrociando in tutti i modi possibili i punti unidimensionali, si ha un totale di Q^r punti. Ciò costituisce il limite di questa procedura, perché al crescere di r la mole di calcoli diviene presto insostenibile, anche se Hedeker & Gibbons (1994) suggeriscono che al crescere di r si può comunque ridurre Q senza incidere troppo sulla bontà dell'approssimazione[29].

La quadratura di Gauss-Hermite viene usata per approssimare gli integrali che compaiono nelle espressioni delle derivate della log-verosimiglianza che servono per implementare l'algoritmo *Fisher Scoring*. In tal caso la dimensione degli integrali è pari al numero di effetti casuali presenti nel modello, per cui la quantità di calcoli rimane accettabile solo per modelli relativamente semplici. La quadratura consente inoltre di approssimare la log-verosimiglianza marginale: infatti dalle (2.42) e (2.43) si ottiene[30]

$$\begin{aligned} \log L &= \sum_{j=1}^{J} \log L(\alpha, \beta, \boldsymbol{\psi}, \boldsymbol{\nu} \mid \mathbf{y}_j) \\ &\quad \sum_{j=1}^{J} \log \sum_{q=1}^{Q^r} L(\alpha, \beta, \boldsymbol{\psi}, \boldsymbol{\nu} \mid \mathbf{y}_j, \mathbf{x}_q) p_{\mathbf{x}_q}. \end{aligned}$$

La log-verosimiglianza così calcolata può essere usata, nel modo convenzionale, per il test χ^2 del rapporto di verosimiglianza (Hedeker & Gibbons,

[29] Ad esempio, in un'applicazione con $r = 5$ può talora essere sufficiente usare $Q = 3$, limitando così il numero totale di punti a $3^5 = 243$ (si noti che già con $Q = 5$ si avrebbero ben 3125 punti).

[30] Il modello che abbiamo preso ad esempio ha due effetti casuali, per cui $r = 2$.

1994)[31]. Ciò costituisce un vantaggio del presente metodo di stima rispetto ai metodi di quasi-verosimiglianza che discuteremo tra breve, per i quali non esiste un'approssimazione affidabile della verosimiglianza.

Il programma MIXOR (Hedeker & Gibbons, 1996) usa la procedura appena descritta per la stima dei parametri di *modelli a risposta ordinale a due livelli*, con quattro possibili link (probit, logit, log-log, complementary log-log). La possibilità di scegliere il numero di punti di quadratura consente di ottenere il livello di approssimazione desiderato. Il programma MIXNO (Hedeker, 1998) estende la procedura al *modello logistico politomico a due livelli* (cfr. par. 2.2.3). In tal caso, poiché gli effetti casuali sono diversi per ogni equazione (cfr. espressione (2.26)), la riparametrizzazione del modello avviene ponendo

$$\mathbf{u}_j^{(s)} = \mathbf{\Psi}^{(s)}\mathbf{w}_j, \qquad s = 1, \ldots, m-1,$$

dove $\mathbf{\Psi}^{(s)}$ è la matrice sotto-triangolare che si ottiene dalla scomposizione di Cholesky della matrice di covarianza degli effetti casuali relativi alla categoria s, mentre $\mathbf{w}_j$ è un vettore aleatorio con distribuzione normale multivariata standard.

Quasi-Verosimiglianza Marginale (MQL) e Penalizzata (PQL)

I metodi di quasi-verosimiglianza MQL e PQL si basano su un'approssimazione lineare del modello di cui si vogliono stimare i parametri, in modo tale da poter usare gli algoritmi IGLS e RIGLS per i modelli lineari (cfr. par. 2.1.7).

Per illustrare i metodi MQL e PQL consideriamo il caso di dati binari a due livelli (cfr. par. 2.2.2), scrivendo il modello come

$$\begin{aligned} y_{ij} &= \pi_{ij} + e_{ij} \\ &= h(\alpha + \beta x_{ij} + u_{0j}) + e_{ij}, \end{aligned} \tag{2.44}$$

dove $h(\cdot) = g^{-1}(\cdot)$ è l'inversa della funzione link e e_{ij} è il termine di errore di primo livello, con valore atteso nullo e varianza $\pi_{ij}(1-\pi_{ij})$. La rappresentazione (2.44) richiama alla mente il modello multilivello lineare, ma esistono

[31] Tuttavia, nel caso si voglia verificare l'ipotesi che una componente di varianza sia nulla, bisogna tener presente che tale ipotesi cade sulla frontiera dello spazio parametrico e che, quindi, la distribuzione asintotica del rapporto di verosimiglianza non è χ^2 con 1 grado di libertà (cfr. Self & Liang, 1987). Il problema può essere risolto considerando uno spazio parametrico esteso in cui la componente di varianza può assumere anche valori negativi (Longford, 1993, p. 172; in effetti, la maggior parte degli algoritmi consente una stima non vincolata dei parametri casuali). In alternativa, l'ipotesi di componente di varianza nulla può essere verificata per mezzo di un test *score* (Lin, 1997).

due differenze fondamentali: 1) la funzione h in generale non è lineare; 2) il termine di errore di primo livello non è indipendente dall'effetto casuale, poiché la sua varianza dipende da π_{ij} e quindi da u_{0j}.

Data un generica funzione di due variabili $h(\eta + \vartheta)$, consideriamo la seguente approssimazione in serie di Taylor intorno al punto $(\eta_0 + \vartheta_0)$, arrestando lo sviluppo al primo termine per η e al secondo termine per ϑ:

$$\begin{aligned} h(\eta + \vartheta) &\simeq h(\eta_0 + \vartheta_0) + h'(\eta_0 + \vartheta_0) \cdot (\eta - \eta_0) \\ &\quad + h'(\eta_0 + \vartheta_0) \cdot (\vartheta - \vartheta_0) + \frac{1}{2} h''(\eta_0 + \vartheta_0) \cdot (\vartheta - \vartheta_0)^2, \end{aligned} \tag{2.45}$$

dove l'apice indicata la derivata. Questa approssimazione può essere applicata al modello (2.44) ponendo η e ϑ uguali rispettivamente alla parte fissa e alla parte casuale del predittore lineare:

$$\begin{cases} \eta = \alpha_{t+1} + \beta_{t+1} x_{ij} \\ \vartheta = u_{0j,t+1} \end{cases}$$

dove l'indice $t + 1$ significa che stiamo considerando i valori da stimare alla iterazione $t + 1$. I possibili metodi di stima differiscono in base al punto $(\eta_0 + \vartheta_0)$ dell'approssimazione e in base alla presenza o meno del termine di secondo ordine. Per quanto rigarda il punto $(\eta_0 + \vartheta_0)$, due possibili scelte sono

$$\begin{cases} \eta_0 = \hat{\alpha}_t + \hat{\beta}_t x_{ij} \\ \vartheta_0 = 0 \end{cases} \qquad \text{Marginal Quasi-Likelihood (MQL)}$$

$$\begin{cases} \eta_0 = \hat{\alpha}_t + \hat{\beta}_t x_{ij} \\ \vartheta_0 = \hat{u}_{0j,t} \end{cases} \qquad \text{Penalized Quasi-Likelihood (PQL)}$$

dove la notazione $\hat{\alpha}_t$ indica il valore stimato alla iterazione t. Inoltre i due metodi possono essere implementati limitatamente ai termini di primo ordine dell'espressione (2.45) oppure includendo anche il termine di secondo ordine: pertanto sia MQL che PQL vengono distinti in *di primo ordine* e *di secondo ordine*. In seguito indicheremo con MQL1 e MQL2 i metodi MQL rispettivamente di primo e secondo ordine, e analogo significato avranno PQL1 e PQL2.

Adesso esaminiamo in dettaglio la stima con PQL2 (Goldstein & Rasbash, 1996). Indichiamo con $\hat{h}_{ij,t}$ la quantità $h(\hat{\alpha}_t + \hat{\beta}_t x_{ij} + \hat{u}_{0j,t})$ e analogo significato attibuiamo a $\hat{h}'_{ij,t}$ e $\hat{h}''_{ij,t}$. In base alla (2.45), il modello (2.44) può essere approssimato alla iterazione $t + 1$ dal modello

$$\begin{aligned} y_{ij} &= \hat{h}_{ij,t} + \hat{h}'_{ij,t} \cdot (\alpha_{t+1} + \beta_{t+1} x_{ij} - \hat{\alpha}_t - \hat{\beta}_t x_{ij}) \\ &\quad + \hat{h}'_{ij,t} \cdot (u_{0j,t+1} - \hat{u}_{0j,t}) + \frac{1}{2} \hat{h}''_{ij,t} \cdot (u_{0j,t+1} - \hat{u}_{0j,t})^2 + e_{ij}. \end{aligned}$$

Ora sostituiamo il termine quadratico con il suo valore atteso, $\frac{1}{2}\hat{h}''_{ij,t} \cdot \hat{c}_{j,t}$, dove $\hat{c}_{j,t}$ è la stima della varianza condizionata del residuo all'iterazione t (cfr. par. 2.1.8). Poi scriviamo e_{ij} come $\hat{z}_{ij,t} e^{\#}_{ij}$, dove

$$\hat{z}_{ij,t} = \sqrt{\hat{h}_{ij,t} \cdot (1 - \hat{h}_{ij,t})}, \tag{2.46}$$

mentre $e^{\#}_{ij}$ è un termine di errore di media nulla e varianza unitaria[32]. La varianza di $e^{\#}_{ij}$ è dunque fissata e non è oggetto di stima, a meno che non si voglia stimare una componente di *extravariabilità binomiale* (Goldstein, 1995, pp. 98-99).

Pertanto il modello (2.44) può essere riscritto come

$$y^{\#}_{ij,t} = \pi^{\#}_{ij,t} + \hat{z}_{ij,t} e^{\#}_{ij}, \tag{2.47}$$

dove

$$y^{\#}_{ij,t} = y_{ij} - \hat{h}_{ij,t} + \hat{h}'_{ij,t} \cdot (\hat{\alpha}_t + \hat{\beta}_t x_{ij} + \hat{u}_{0j,t}) - \frac{1}{2}\hat{h}''_{ij,t} \cdot \hat{c}_{j,t},$$

$$\pi^{\#}_{ij,t} = \hat{h}'_{ij,t} \cdot (\alpha_{t+1} + \beta_{t+1} x_{ij} + u_{0j,t+1}).$$

La risposta modificata $y^{\#}_{ij,t}$ si ottiene sottraendo dalla risposta originaria y_{ij} un *offset* calcolato sulla base delle stime all'iterazione t, mentre il valore atteso modificato lineare $\pi^{\#}_{ij,t}$ non è altro che il predittore lineare moltiplicato per la costante $\hat{h}'_{ij,t}$ determinata all'iterazione t. Dunque, noti i risultati dell'iterazione t, il modello (2.47) è un modello multilivello lineare che può essere stimato con l'algoritmo IGLS o RIGLS.

Riassumendo, ogni iterazione è composta dai seguenti passi: 1) calcolo delle *variabili modificate* (dette anche *variabili di lavoro*) presenti nel modello (2.47), basandosi sulle stime ottenute all'iterazione precedente; 2) stima dei parametri casuali e fissi per mezzo di un'iterazione dell'algoritmo IGLS o RIGLS (cfr. par. 2.1.7); 3) stima degli effetti casuali e della loro varianza condizionata secondo la procedura per il modello lineare (cfr. par. 2.1.8). Le iterazioni vengono ripetute fino a convergenza.

L'uso dei nomi *Quasi-Verosimiglianza Marginale* e *Quasi-Verosimiglianza Penalizzata* è divenuto comune in seguito ad un fondamentale articolo di Breslow & Clayton (1993). Il termine *Quasi-Verosimiglianza* sta ad indicare che questi metodi si basano solo sui valori attesi e sulla funzione di varianza, senza specificare l'intera distribuzione; il termine *Marginale* si riferisce al fatto che la relativa procedura approssima il modello multilivello con il modello

[32] La sostituzione di e_{ij} con $\hat{z}_{ij,t} e^{\#}_{ij}$ ha il fine di eliminare la dipendenza dell'errore di primo livello dall'effetto casuale, scaricando la dipendenza di e_{ij} da π_{ij} su una covariata fittizia.

marginale (cfr. par. 2.2.1); infine, il termine *Penalizzata* è motivato dall'analogia della relativa procedura con la Quasi-Verosimiglianza Penalizzata usata da Green (1987) nell'ambito della regressione semiparametrica.

Goldstein (1991) ha inizialmente proposto il metodo MQL, che si è ben presto rivelato insufficiente per una stima accurata (Breslow & Clayton, 1993; Rodriguez & Goldman, 1995). Le deficienze del metodo MQL vanno ricercate nel fatto che i parametri del modello marginale usato per l'approssimazione sono sistematicamente minori (in valore assoluto) di quelli del modello multilivello, come dimostra il risultato di Neuhaus *et al.* (1991) riportato nel par. 2.2.1. Dunque il metodo MQL produce, per i parametri fissi, stime distorte verso il basso, con una distorsione che cresce con la varianza degli effetti casuali. Le simulazioni Monte Carlo hanno inoltre mostrato forti distorsioni verso il basso delle stime dei parametri casuali[33].

Successivamente Goldstein & Rasbash (1996) hanno messo a punto il metodo PQL, mostrando, per via di simulazioni, che funziona assai meglio del metodo MQL, producendo stime solo leggermente distorte verso il basso. Naturalmente il metodo PQL è più complesso, poiché prevede l'uso dei residui ad ogni iterazione, e quindi presenta maggiori problemi di convergenza e maggiore variabilità degli stimatori. Per quanto rigurada la scelta fra PQL1 e PQL2, le simulazioni mostrano che PQL2 ha una performance leggermente migliore e qualche problema di convergenza in più.

Le stime ottenute con i metodi MQL e PQL possono eventualmente essere corrette per mezzo delle procedure di bootstrap parametrico iterato implementate in MLwiN (Goldstein *et al.*, 1998, cap. 7). Tuttavia occorre sottolineare che non è comunque possibile ottenere approssimazioni affidabili della verosimiglianza, per cui l'uso del test del rapporto di verosimiglianza è precluso.

I metodi MQL e PQL, che abbiamo illustrato nel caso di dati binari a due livelli, possono essere estesi all'intera classe dei GLM multilivello con un numero arbitrario di livelli. In particolare, Goldstein (1995, p. 105) descrive le modifiche necessarie per la stima dei modelli per dati politomici e ordinali.

[33] Nei GLM per dati binari è difficile dare delle indicazioni sul tasso di distorsione, poiché ci sono molti fattori che entrano in gioco. Particolare importanza sembrano rivestire la struttura gerarchica (numero di gruppi e numerosità dei gruppi) e la variabilità della risposta nei gruppi. La situazione più sfavorevole si verifica quando i gruppi sono pochi ed al loro interno hanno risposte con poca variabilità.

Capitolo 3

L'Indagine sui Percorsi di Studio e Lavoro dei Diplomati: caratteristiche generali e analisi preliminari

In questo capitolo descriveremo innanzitutto l'Indagine sui Percorsi di Studio e Lavoro dei Diplomati, condotta dall'Istat nel 1998, soffermandoci in particolare sul disegno campionario e sul contenuto informativo. Successivamente porremo le basi per le analisi riportate nei prossimi capitoli, presentando nell'ordine: (i) uno schema generale di analisi; (ii) le variabili utilizzate nei modelli, sia quelle tratte dall'indagine PSLD che quelle tratte da altre fonti; (iii) le strategie di selezione dei modelli e i metodi di stima adottati. I successivi tre capitoli saranno poi dedicati ognuno alla presentazione delle specifiche analisi effettuate:

- Cap. 4: probabilità di svolgere un lavoro continuativo al momento dell'intervista;
- Cap. 5: tempo di ingresso al primo lavoro continuativo;
- Cap. 6: probabilità di immatricolazione all'università.

3.1 Caratteristiche dell'Indagine

Negli ultimi anni il complesso delle indagini Istat sul sistema formativo è stato oggetto di un profondo processo di ristrutturazione (Cariani, 1998). In particolare, all'indagine sull'inserimento professionale dei laureati, condotta

sin dal 1989, sono state affiancate due nuove indagini, quella sui percorsi di studio e lavoro dei diplomati della scuola secondaria superiore (PSLD) e quella sull'inserimento professionale dei diplomati universitari. Le tre indagini costituiscono un sistema integrato di rilevazioni, i cui questionari sono stati resi il più possibile uniformi.

L'indagine PSLD, che è quella che interessa ai nostri fini, si propone di approfondire due temi principali: la carriera universitaria da un lato, la ricerca del lavoro e l'inserimento professionale dall'altro (Micali & Ungaro, 1998). La speciale attenzione rivolta all'esperienza universitaria deriva dal fatto che in Italia, in parte a causa degli elevati tassi di disoccupazione giovanile, i diplomati si riversano in massa sull'università, con risultati spesso non soddisfacenti (circa una matricola su tre riesce a laurearsi). Pertanto l'indagine mira a delineare le motivazioni e i comportamenti degli studenti nei primi anni dei corsi di laurea, quelli tipicamente più difficili e che fanno registrare la maggior parte degli abbandoni. Inoltre, il processo di selezione ed espulsione operato dall'università viene esaminato in termini di estrazione sociale degli studenti, tentando nel contempo di valutare l'efficacia delle politiche del diritto allo studio.

Per quanto riguarda la ricerca del lavoro, l'indagine è volta a delineare le modalità con cui questa si esplica, nonché le aspirazioni dei giovani in merito al lavoro. Particolare attenzione viene posta alle motivazioni che possono aver portato a rifiutare un lavoro e alle condizioni che si pongono per accettare un lavoro: ciò al fine di valutare quanto la disoccupazione giovanile sia una situazione subita oppure scelta.

Sul versante del lavoro, a parte l'esame di alcuni aspetti della storia lavorativa utili per comprendere le modalità di inserimento professionale, l'indagine intende fornire un quadro dettagliato del lavoro svolto al momento dell'intervista, al fine di capire quanto questo sia soddisfacente e rispondente al percorso formativo effettuato.

Una caratteristica di questa indagine è la particolare attenzione riservata alle determinanti sociali dei processi di studio e lavoro, attraverso una disamina approfondita della famiglia di origine.

Le caratteristiche tecniche dell'indagine PSLD possono essere descritte seguendo lo schema delineato nel par. 1.3.1:

a) Popolazione di riferimento. Comprende i diplomati delle scuole medie superiori italiane, sia pubbliche che private, che hanno ottenuto il titolo nel 1995, cioè circa tre anni prima dell'intervista, che è stata effettuata nel periodo settembre-dicembre 1998.

b) Tipo di indagine. Si basa su un campione stratificato a due stadi. Le unità di primo stadio sono le 7144 scuole medie superiori italiane, stratificate

per *tipo di insegnamento*[1] e *regione* (circa 350 strati). Le unità di secondo stadio sono i 490348 maturi dell'anno 1995 (Istat, 1999b).

La dimensione del campione di maturi è stata fissata in 20000 unità. Tuttavia, al fine di effettuare sostituzioni in caso di mancata intervista, sono stati selezionati altri 20000 nominativi suppletivi, per un totale di 40000 maturi. Questo numero di maturi è stato ripartito fra gli strati secondo la seguente formula (Rinaldelli, 1997):

$$m_h = 40000 \frac{1}{2} \left(\frac{1}{H} + \frac{M_h}{M} \right),$$

dove h è l'indice di strato, H è il numero di strati, M_h è il numero di maturi nello strato h, M è il numero totale di maturi[2]. Il numero di scuole campione nello strato h è stato poi determinato secondo la formula

$$n_h = \frac{m_h}{m^*},$$

dove n_h è il numero di scuole campione nello strato h, mentre m^* è il numero di maturi campione da rilevare in ogni scuola campione; il valore m^* è stato fissato pari a 30 (15 nominativi di base e 15 nominativi suppletivi). La ripartizione per regione dell'universo e del campione di scuole e maturi è incluso nella tab. 3.2, riportata al termine del par. 3.3.

Successivamente, all'interno di ogni strato, le scuole campione sono state selezionate con probabilità proporzionale al numero di maturi, secondo il metodo di Madow, ottenendo un campione di 1321 scuole. Ogni scuola ha poi provveduto a selezionare in maniera sistematica i 30 maturi campione, secondo le istruzioni fornite dall'Istat. Per le scuole con meno di 30 maturi la lista dei nominativi è stata integrata attingendo a scuole suppletive appartenenti allo stesso strato (complessivamente sono state selezionate 280 scuole suppletive).

Una caratteristica interessante dell'indagine consiste nel fatto che le scuole, oltre a costituire il tramite attraverso cui vengono reperiti i nominativi dei giovani da intervistare, forniscono direttamente delle informazioni sui propri diplomati: infatti dagli archivi scolastici si possono agevolmente ottenere dati come il giudizio riportato all'esame di licenza media, i passaggi tra scuole pubbliche e private, il voto di maturità, ecc.

[1] Sono stati considerati singolarmente tutti i tipi di insegnamento, con l'eccezione dei seguenti accorpamenti: 1) Istituti tecnico Nautico ed Aeronautico; 2) Istituto professionale femminile e Scuola magistrale; 3) Istituti professionale Industriale e Marinaro.

[2] Il criterio di allocazione dei maturi da intervistare che è stato adottato in questa indagine rappresenta un compromesso fra l'allocazione proporzionale e quella di uguale dimensione in tutti gli strati.

Delle 1321 scuole selezionate per l'indagine, 1287 hanno effettivamente risposto, con un tasso di caduta del 2.6%. Considerando anche le scuole suppletive, le scuole presenti nel campione sono risultate 1563.

Le interviste utili realizzate sono state 18843, a fronte di 30881 nominativi che si è cercato di contattare. Nel 35.9% dei casi il contatto non è riuscito (numero telefonico errato, nessuna risposta alle telefonate, ecc.); nel 3.3% dei casi l'intervista è stata rifiutata o interrotta; infine, nel 60.8% dei casi l'intervista è stata completata[3]. È interessante osservare che le persone che hanno rifiutato o interrotto l'intervista sono appena il 5.1% delle persone contattate. Il tasso di sostituzione degli individui dell'elenco base è risultato del 38.9%[4].

c) Periodo di osservazione. Poiché le interviste sono state effettuate tra settembre e dicembre 1998, il periodo di osservazione è di poco più di tre anni.

d) Procedura di osservazione. L'indagine è di tipo retrospettivo. La ricostruzione della dinamica temporale dei percorsi di studio e lavoro è molto frammentaria, poiché le informazioni retrospettive su studio e lavoro vengono raccolte in modo indipendente e con quesiti di tipo diverso. In particolare, i quesiti che richiedono una collocazione temporale di eventi riguardano i seguenti aspetti (l'intervistato è tenuto a rispondere solo se si trova o si è trovato in una determinata condizione diversa per ogni quesito):

- anno di iscrizione all'università (quesiti 32-33-48-49);
- mese e anno in cui si è superato l'ultimo esame, per coloro che sono iscritti all'università al momento dell'intervista (quesito 23);
- mese e anno di inizio del primo lavoro intrapreso dopo il conseguimento del diploma e interrotto in un momento antecedente l'intervista (quesito 59);
- mese e anno di inizio del lavoro svolto al momento dell'intervista, purché si tratti di un lavoro di tipo *continuativo* iniziato *dopo* il conseguimento del diploma (quesito 68).

[3]Gli individui *fuori target*, cioè coloro che al momento dell'intervista hanno dichiarato di non aver conseguito il diploma nell'anno 1995, sono risultati solo lo 0.3% delle persone con cui si è tentato il contatto.

[4]Le sostituzioni di nominativi della lista di base sono avvenute estraendo un nominativo dall'elenco suppletivo in modo da lasciare inalterati lo strato di appartenenza (tipo di scuola e regione geografica) e altri caratteri quali il sesso e le ripetenze. Il tasso di sostituzione presenta marcate differenze a livello territoriale, passando dal 25.2% del Veneto al 54.1% della Puglia. Esistono inoltre differenze di minore entità legate al sesso, al tipo di scuola, al voto di maturità e alle ripetenze (cfr. Istat, 1999b, Prospetto 6).

e) Metodo di rilevazione. Il questionario è stato sottoposto al giovane tramite intervista telefonica con il sistema CATI (Computer Assisted Telephone Interview)[5]. Ciò ha permesso di aumentare il numero di rispondenti e di ridurre gli errori non campionari[6].

f) Contenuto informativo. Il questionario comprende 8 sezioni, che possono essere raggruppate nelle seguenti macro aree (Micali & Ungaro, 1998; tra parentesi sono riportati i numeri dei relativi quesiti):

Gli studi:

Sez. 1 - Curriculum scolastico (1-9);

Sez. 2 - Corsi di formazione post-secondaria (10-15);

Sez. 3 - Studi universitari (16-45);

Sez. 4 - Interruzione degli studi (46-54);

Il lavoro:

Sez. 5 - Lavoro (55-91);

Sez. 6 - Ricerca di lavoro (92-98);

La condizione socio-demografica:

Sez. 7 - Notizie sulla famiglia (99-119);

Sez. 8 - Notizie anagrafiche (120-127).

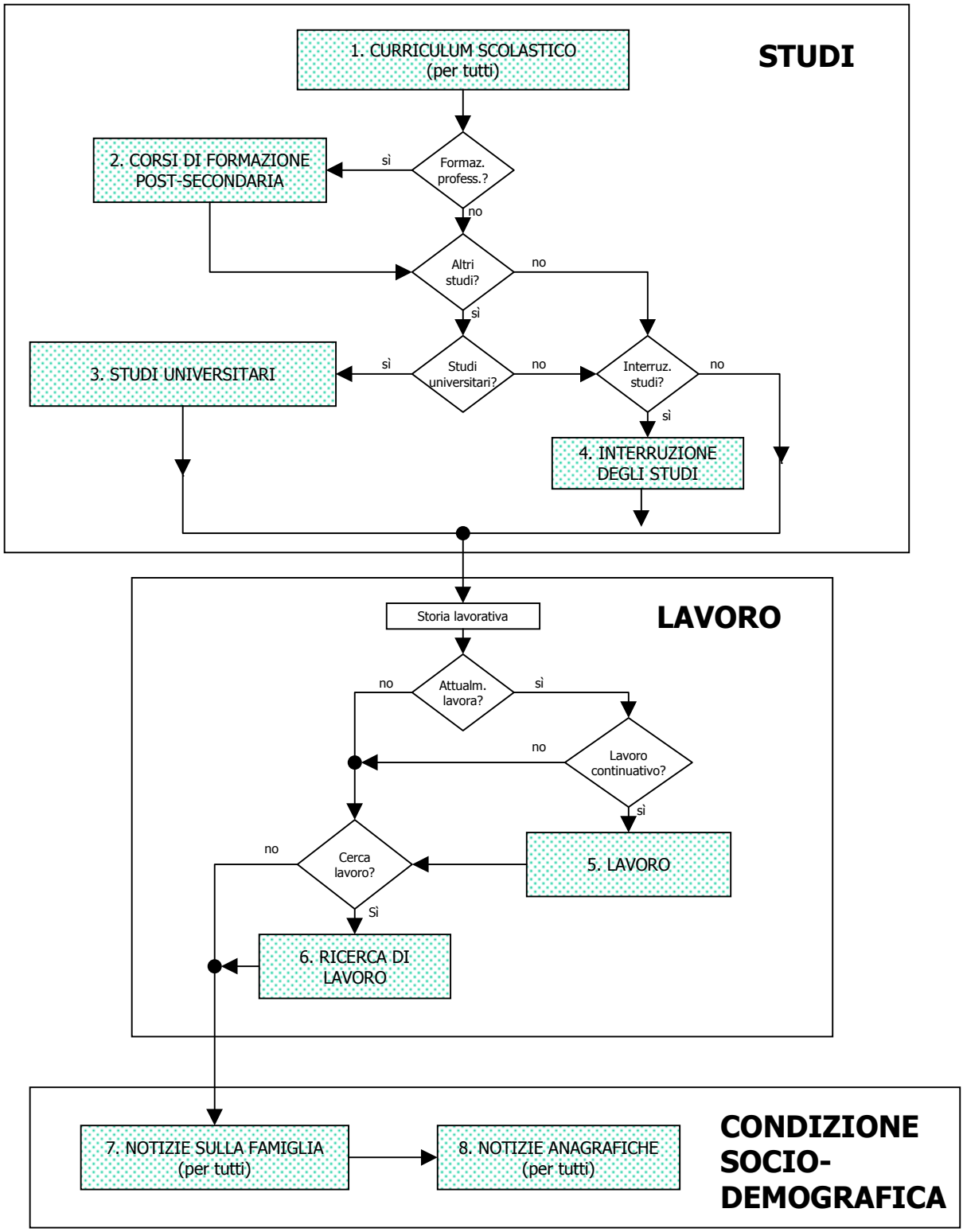
La struttura del questionario è rappresentata in fig. 3.1. Si noti, in particolare, che il corpo centrale della sezione 5 viene proposto solo a coloro che al momento dell'intervista svolgono un lavoro *continuativo*, mentre la sezione 6 viene proposta a tutti coloro che cercano lavoro, *anche se già occupati*[7].

[5]Durante l'indagine pilota è stato sperimentato anche un sistema misto, che consiste nel far precedere l'intervista telefonica dall'invio del questionario cartaceo a mezzo posta (Istat, 1999b). Tuttavia questa variante del sistema CATI ha fornito solo una modesta riduzione degli errori e della durata dell'intervista, a fronte di un sensibile aumento dei costi e dei tempi organizzativi, per cui non è stata adottata nell'indagine definitiva.

[6]Infatti il software che supporta l'intervista effettua un controllo delle risposte in tempo reale, consentendo all'intervistatore di riproporre il quesito al quale sia stata fornita una risposta non ammissibile o non coerente con altre risposte.

[7]Questa scelta, che costituisce una novità della presente indagine, è motivata dal fatto che la transizione scuola-lavoro viene tipicamente portata a termine per aggiustamenti successivi, per cui spesso il lavoro svolto a distanza di tre anni dal diploma non è ancora quello che caratterizzerà la vita del diplomato.

Fig. 3.1 - Struttura del questionario dell'indagine PSLD



Fonte: Micali & Ungaro (1998)

3.2 Schema generale di analisi

L'indagine PSLD esamina numerosi aspetti dell'esperienza di studio e lavoro dei neodiplomati, per cui fornisce informazioni utili allo studio di fenomeni quali:

- la performance scolastica (voto di maturità);
- la decisione di continuare gli studi o di iniziare la ricerca del lavoro;
- l'esito della prosecuzione degli studi (probabilità di interruzione);
- il successo nella ricerca del lavoro (probabilità e tempi di ingresso al lavoro);
- la differenze nel reddito da lavoro.

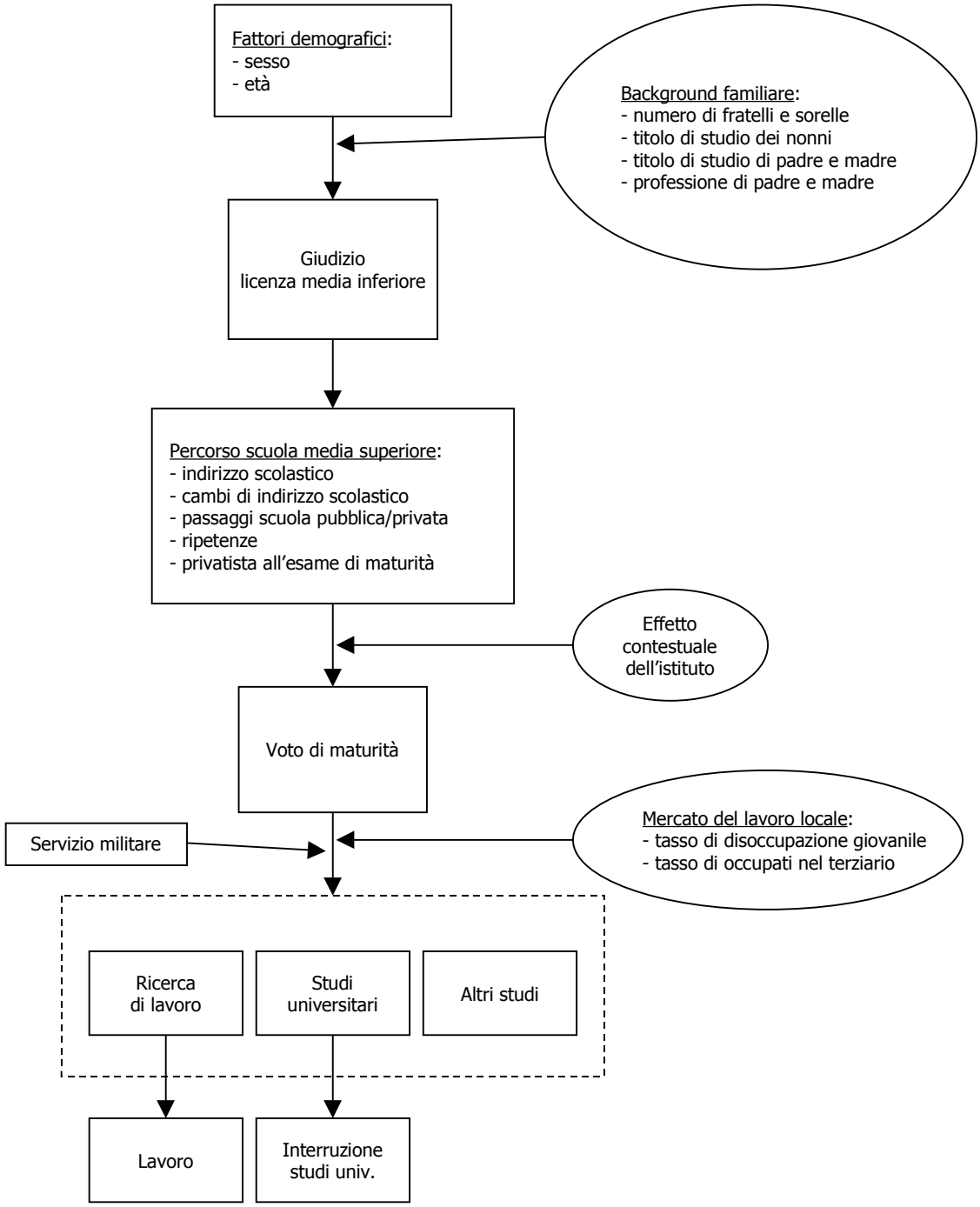
Un possibile schema di analisi di tali fenomeni consiste nel disporre le variabili di interesse in sequenza temporale, studiandone gli effetti per mezzo di regressioni ricorsive[8]. Il tempo a disposizione per il completamento della tesi non ci ha consentito di sviluppare compiutamente tale approccio, che tuttavia rimane utile come schema di riferimento.

In effetti, un attento esame del questionario PSLD evidenzia la possibilità costruire un ordinamento temporale delle variabili piuttosto articolato, che consente di analizzare, oltre alle esperienze post-diploma, anche il percorso di studio fino al conseguimento del diploma. La fig. 3.2 rappresenta uno schema di analisi basato sull'ordinamento temporale delle variabili, secondo i seguenti criteri:

- le variabili sono disposte dall'alto verso il basso secondo un ideale asse temporale;
- i rettangoli contengono *variabili individuali* e gli ovali *variabili contestuali* (cfr. par. 1.1);
- ogni variabile ha un potenziale effetto su tutte le variabili che la seguono nell'asse temporale;
- il rettangolo tratteggiato che segue il voto di maturità sta ad indicare che lo studio ed il lavoro non sono mutualmente esclusivi;
- il diagramma si basa sulle informazioni rese disponibili dall'indagine PSLD, ad eccezione delle informazioni sulla situazione del mercato del lavoro locale, che devono essere tratte da altre fonti.

[8] Una classica metodologia che nel presente contesto potrebbe fornire dei risultati interessanti è la *path analysis* (Duncan, 1966).

Fig. 3.2 - Diagramma della modellizzazione dei percorsi di studio e lavoro sulla base dell'indagine PSLD



I percorsi di studio e l'inserimento professionale possono essere analizzati per mezzo di una serie di modelli di regressione, uno per ogni variabile individuale, in cui le variabili esplicative sono tutte quelle che in fig. 3.2 precedono la variabile di risposta.

Per quanto riguarda le variabili contestuali, quelle relative al background familiare e al mercato del lavoro possono essere inserite come variabili esplicative. Invece, l'effetto contestuale delle scuole secondarie superiori (che intende catturare le differenze fra gli istituti imputabili alla composizione degli studenti, alla qualità dell'insegnamento, alla dotazione di infrastrutture ecc.) si presta ad essere studiato per mezzo dei modelli multilivello (cfr. cap. 2). Pertanto le analisi relative a voto di maturità, ricerca del lavoro e studi universitari dovrebbero far ricorso a modelli in cui i diplomati costituiscono le unità di primo livello e i singoli istituti le unità di secondo livello.

Nel presente lavoro abbiamo analizzato solo alcuni aspetti, ricercando i fattori che determinano:

- la probabilità di svolgere un lavoro continuativo al momento dell'intervista (cap. 4);
- il tempo necessario per ottenere il primo lavoro continuativo (cap. 5);
- la probabilità di immatricolazione all'università (cap. 6).

Inoltre, assumendo il successo nella ricerca del lavoro come indicatore dell'output del processo educativo, abbiamo tentato una valutazione dell'*efficacia relativa* degli istituti (cfr. par. 1.2.4).

3.3 Le variabili utilizzate nelle analisi

3.3.1 Variabili desunte dall'indagine PSLD

Definizione e codifica delle variabili

Il questionario dell'indagine PSLD ha consentito la raccolta di una quantità enorme di informazioni, che solo in parte sono state utilizzate nella presente tesi. La tab. 3.1 fornisce un elenco delle principali variabili che sono state prese in considerazione, anche se non tutte sono entrate a far parte dei modelli che saranno presentati nei prossimi capitoli.

La tab. 3.1 riporta, per ogni variabile, un nome convenzionale (che verrà usato nelle tabelle dei capp. 4-6), una breve descrizione, il quesito o i quesiti dell'Indagine che la definiscono, le modalità originarie e le modalità ricodificate. La maggior parte dei quesiti prevede risposte con un numero finito

Tab. 3.1 – Definizione delle variabili dell'indagine PSLD utilizzate nella presente ricerca

Avvertenze: (1) le definizioni delle variabili, se non diversamente specificato, si riferiscono alla condizione del diplomato al momento dell'intervista; (2) salvo diversa specificazione, le variabili devono intendersi dicotomiche con modalità 0 e 1 e la loro descrizione si riferisce alla modalità 1: ad es., "PROFPA-B: tecnico/impiegato" significa che PROFPA-B è una variabile binaria che assume il valore 1 se il padre è un tecnico o un impiegato e il valore 0 in tutti gli altri casi.

N.	Nome variabile	Descrizione	Num. quesito	Modalità originarie	Definizione della variabile
Variabili anagrafiche					
1	FEMM	Sesso	125.1	1: maschio 2: femmina	FEMM: femmina
2	LEVA-IN LEVA-POI LEVA-ESO	Posizione nei confronti degli obblighi di leva	125.2	1: già assolti 2: in corso 3: ancora da assolvere 4: esonerato	LEVA-IN: mod. 1 LEVA-POI: mod. 2,3 LEVA-ESO: mod. 4
Background familiare					
3	FRAT>1	Numero di fratelli e sorelle	99-100	0; 1; 2; 3; 4 e oltre	FRAT>1: 2 o più
4	TITPA TITMA	Titolo di studio del padre/della madre quando l'intervistato aveva 14 anni	107 108	1: analfabeta/senza titolo 2: licenza elementare 3: licenza media/avv. prof. 4: qualifica prof. (2-3 anni) 5: dipl. scuola media sup. (4-5 anni) 6: dipl. univers. o ex Scuole parauniv. 7: laurea o dottorato di ricerca	TITPA: mod. 5-7 TITMA: mod. 5-7
5	TIT	Titolo di studio dei genitori quando l'intervistato aveva 14 anni		Variabile derivata: TIT=TITPA+TITMA	
6	PROFPA-A PROFPA-B PROFPA-C PROFPA-D PROFPA-E	Condizione professionale del padre quando l'intervistato aveva 14 anni (se non lavorava si fa riferimento all'ultimo lavoro svolto)	112-114	PROFPA-A: operaio/lavoratore dipend. senza qualif. PROFPA-B: tecnico/impiegato PROFPA-C: dirigente/quadro/insegnante PROFPA-D: lavoratore indip., escluso imprenditore e libero professionista PROFPA-E: imprenditore/libero professionista	
7	PROFMA-0 PROFMA-A PROFMA-B PROFMA-C PROFMA-D PROFMA-E	Condizione professionale della madre quando l'intervistato aveva 14 anni (se non lavorava si fa riferimento all'ultimo lavoro svolto)	117-119	PROFMA-0: casalinga PROFMA-A: operaio/lavoratore dipend. senza qualif. PROFMA-B: tecnico/impiegato PROFMA-C: dirigente/quadro/insegnante PROFMA-D: lavoratore indip., escluso imprenditore e libero professionista PROFMA-E: imprenditore/libero professionista	
8	PROF-A PROF-B PROF-C PROF-D PROF-E PROF-AB PROF-AC ecc.	Condizione professionale dei genitori quando l'intervistato aveva 14 anni	117-119	Variabili derivate: PROF-X=(PROFPA-X)+(PROFMA-X) per X∈{A,B,C,D,E} PROF-XY=(PROF-X)+(PROF-Y) per X,Y∈{A,B,C,D,E}, X≠Y	

Tab. 3.1 – Definizione delle variabili dell'indagine PSLD utilizzate nella presente ricerca (segue)

N.	Nome variabile	Descrizione	Num. quesito	Modalità originarie	Definizione della variabile
Curriculum scolastico					
9	GIUD4	Giudizio licenza media inf.	Scheda scuola	1: sufficiente 2: buono 3: distinto 4: ottimo	GIUD4: ottimo
10	ETA<14 ETA>14	Età di prima iscrizione ad una scuola secondaria superiore	4	12-52	ETA<14: età 12-13 ETA>14: età 15-52
11	CAMBIND	Cambi di indirizzo scolastico	1	1: almeno un cambio 2: nessun cambio	CAMBIND: mod. 1
12	PRIV	Privatista all'esame di maturità	Scheda scuola	1: privatista 2: non privatista	PRIV: privatista
13	RIP	Ripetenze durante gli studi superiori	6	1: no 2: sì	RIP: almeno una ripetenza
14	VM36 VM37-42 VM43-49 VM50-59 VM60	Voto di maturità	Scheda scuola	36-60	VM36: voto 36 VM37-42: voto 37-42 VM43-49: voto 43-49 VM50-59: voto 50-59 VM60: voto 60
Corsi di formazione professionale					
15	CFP	Iscrizione ad un Corso Regionale di Formazione Prof. dopo il conseguimento del diploma	10	1: mai iscritto 2: iscritto	CFP: mod. 2
16	CFP-ORA CFP-CONC CFP-INT	Esito del CFP	11	1: iscritto, ma non ancora iniziato 2: in corso 3: concluso 4: interrotto	CFP-ORA: mod. 1 e 2 CFP-CONC: mod. 3 CFP-INT: mod 4
Studi universitari e altri corsi					
17	IMM IMMSUB	Immatricolazione ad un corso di studi universitario (anche se poi interrotto)	32-33 48-49	Vedi quesionario	IMM: immatricolato nel periodo tra il conseguim. del diploma e l'intervista IMMSUB: immatricolato subito dopo il conseguim. del diploma
18	UNIV	Frequenza di un corso di studi universitario al momento dell'intervista	16.2	5: Corso di Laurea 6: Corso di Diploma Univ. 7: Scuola Dir. a Fini Spec.	UNIV: impegnato in studi universitari (mod. 5-7)
19	INTUNIV	Interruzione di un corso di studi universitario	46	5: Corso di Laurea 6: Corso di Diploma Univ. 7: Scuola Dir. a Fini Spec.	INTUNIV: interrotti studi universitari (mod. 5-7)
20	INTUNIV-5 INTUNIV-0	Interruzione di un corso di studi universitario, distinta per motivo dell'interruzione	47	5: per lavoro 1-4,6-8: altri motivi	INTUNIV5: interrotti studi universitari per motivi di lavoro INTUNIV$: interrotti studi universitari per motivi diversi dal lavoro

Tab. 3.1 – Definizione delle variabili dell'indagine PSLD utilizzate nella presente ricerca (segue)

N.	Nome variabile	Descrizione	Num. quesito	Modalità originarie	Definizione della variabile
21	ALTRICOR	Iscrizione ad un corso non universitario (esclusi i Corsi di Formazione Professionale), anche se poi interrotto	16.2,46	1: Accademia di belle arti 2: Isef 3: Conservatorio 4: Scuole di FF.AA./Polizia	ALTRICOR: Iscritto ad un corso non universitario
22	ALTRIORA	Frequenza di un corso di studi non universitario al momento dell'intervista	16.2	1: Accademia di belle arti 2: Isef 3: Conservatorio 4: Scuole di FF.AA./Polizia	ALTRIORA: impegnato in un corso di studi non universitario al momento dell'intervista
Lavoro					
23	LAVC	Lavoro *continuativo* svolto al momento dell'intervista	65	1: lavoro occasionale 2: lavoro stagionale 3: lavoro continuativo	LAVC: svolge un lavoro continuativo
24	LAVCINT	Lavoro *continuativo* iniziato dopo il conseguimento del diploma e poi interrotto	58	1: lavoro continuativo 2: lavoro stagionale 3: lavoro occasionale	LAVCINT: dopo il diploma ha iniziato e poi interrotto un lavoro continuativo
25	LAVCPRIM	Lavoro *continuativo* iniziato prima del conseguimento del diploma e svolto al momento dell'intervista	67	1: prima 2: dopo	LAVCPRIM: il lavoro continuativo svolto al momento dell'intervista è iniziato prima del conseguimento del diploma
26	TEMPO	Tempo in mesi, arrotondato per eccesso, intercorso tra il conseguimento del diploma e l'inizio del primo lavoro *continuativo* (escludendo coloro per i quali il lavoro continuativo svolto al momento dell'intervista è iniziato prima del conseguim. del diploma e attribuendo il tempo massimo a coloro che non hanno ancora svolto un lavoro continuativo)	58-59; 67-68	Variabile derivata dal mese e anno di inizio del lavoro (alcuni dettagli del calcolo sono descritti nel par. 3.3)	TEMPO $\in \{1,2, \ldots ,40\}$ 1 ↔ luglio 1995 2 ↔ agosto 1995 . . . 40 ↔ ottobre 1998 Il valore 40 è attribuito anche a coloro che non hanno ancora svolto un lavoro continuativo (tempo *censurato a destra*)
27	EVENTO	Indicatore di evento da associare alla variabiel TEMPO	-	-	EVENTO: vale 1 quando TEMPO è il tempo di un effettivo ingresso al lavoro, mentre vale 0 quando TEMPO è un tempo censurato a destra
Caratteristiche delle scuole					
28	ISTPROF ISTTEC ISTMAG LICEI ALTRIIST ALTRIIST2	Tipo di istituto	Scheda scuola	11-16: Ist. Profess. 20-28: Ist. Tecnici 32: Ist. Magistrale 41-43: Licei 31, 51-52: Altri Istituti	ISTPROF: Ist. Profess. ISTTEC: Ist. Tecnici ISTMAG: Ist. Magistrale LICEI: Licei ALTRIIST: Altri Istituti ALTRIIST2: Altri Istituti, incluso l'Ist. Magistrale

Tab. 3.1 – Definizione delle variabili dell'indagine PSLD utilizzate nella presente ricerca (segue)

N.	Nome variabile	Descrizione	Num. quesito	Modalità originarie	Definizione della variabile
29	ITP-AGR ITP-IND ITP-COMM ITP-TUR ITP-ALT ITP-ALT2	Indirizzi scolastici degli Istituti Tecnici e Professionali	Scheda scuola	11,20: Agrario 12,21: Industriale 14,24,27: Commerciale/Aziendale 15,26: Turistico-alberghiero 16,22,23,25,28: altri indir.	ITP-AGR: Agrario ITP-IND: Industriale ITP-COMM: Commerciale/Aziendale ITP-TUR: Turist.-albergh. ITP-ALT: altri indirizzi ITP-ALT2: altri indirizzi, inclusi agrario e turistico
30	SCPR	Scuola privata	-	Variabile derivata (cfr. par. 3.3.1)	SCPR: scuola privata
Variabili relative al contesto socio-economico					
31	NORD CENTRO SUD	Circoscrizione geografica di appartenenza della scuola	-	Variabili derivate dalla regione di appartenenza della scuola (cfr. par. 3.3.1)	NORD: Piemonte, Valle d'Aosta, Lombardia, Trentino A.A., Veneto, Friuli V.G., Liguria, Emilia Romagna. CENTRO: Toscana, Umbria, Marche, Lazio. SUD: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna.
32	DISOCC	Tasso di disoccupazione (%) a livello regionale dei giovani in età 15-24 nel 1995	-	Variabile tratta dall'Indagine sulle Forze di Lavoro (Istat)	DISOCC∈[-16.1, 39.6] Variabile centrata rispetto al valore della Toscana (25.1)
33	±DISOCC	Variazione del tasso di disoccupazione (%) a livello regionale dei giovani in età 15-24 tra il 1995 e il 1998	-	Variabile tratta dall'Indagine sulle Forze di Lavoro (Istat)	±DISOCC∈[-9.5, 7.0]
34	%TERZ	Occupati nel terziario (%) a livello regionale nel 1995	-	Variabile tratta dall'Indagine sulle Forze di Lavoro (Istat)	%TERZ∈[-8.3, 14.1] Variabile centrata rispetto al valore della Toscana (60.6)
35	GRANDIAZ	Numero di grandi aziende (20 e più addetti) ogni 1000 aziende a livello regionale nel 1995	-	Variabile tratta dai Conti Economici delle Imprese (Istat)	GRANDIAZ∈[-9.6, 9.1] Variabile centrata rispetto al valore della Toscana (18.9)

di categorie; in tali casi le trasformazioni effettuate rispondono ai seguenti criteri:

- Le risposte con due categorie (dicotomiche) sono codificate nel questionario con i numeri 1 e 2; tuttavia, ai fini dell'utilizzo nei modelli di regressione, è preferibile una codifica 0-1 in modo tale che quando la variabile assume la prima modalità (detta *modalità di base*) il suo contributo al predittore lineare sia nullo[9].

- Le risposte con più di due categorie (politomiche) sono codificate nel questionario con i numeri naturali $(1, 2, 3, \ldots, n)$, dove n è il numero di categorie. Il metodo più generale per l'inserimento di una variabile con n modalità in un modello di regressione consiste nell'effettuare una trasformazione in $n-1$ variabili binarie (dette *variabili dummy*). Tuttavia, in presenza di numerose variabili con n elevato, questo metodo può causare un'inaccettabile proliferazione dei regressori. La semplice alternativa che abbiamo adottato in gran parte dei casi consiste nel raccogliere le modalità originarie in gruppi, ottenendo così una serie di variabili binarie derivate con codifica 0-1[10].

Particolarmente complicata è risultata la codifica delle informazioni relative alla *professione dei genitori*, caratterizzata da oltre 20 categorie. Sulla base delle distribuzioni doppie rispetto alle variabili relative al lavoro (LAVC) o agli studi universitari (IMM), le professioni sono state raccolte in 5 categorie (più la categoria "casalinga" per la madre). Per l'inserimento delle variabili nei modelli due soluzioni standard sono: a) utilizzare separatamente le professioni del padre e della madre; b) utilizzare la professione del solo padre. Nell'implementazione dei modelli abbiamo constatato che la soluzione a non è praticabile (poiché comporta l'inserimento di troppe variabili, molte delle quali non statisticamente significative), mentre la soluzione b spreca delle importanti informazioni (infatti, dai dati a nostra disposizione emerge che ciò che conta è che uno dei genitori svolga quella certa professione, non importa se si tratta del padre o della madre). Pertanto abbiamo usato

[9] In tal modo la costante del modello è direttamente interpretabile come il valore previsto per l'*individuo-base*, cioè per quell'ipotetico individuo caratterizzato dall'avere un valore nullo in tutte le variabili.

[10] La suddivisione delle modalità in gruppi è un'operazione alquanto delicata, che dipende da: 1) le finalità dell'analisi; 2) le conoscenze a priori; 3) la distribuzione della variabile nel caso in esame. Nelle situazioni più complesse abbiamo effettuato la scelta confrontando le stime di modelli alternativi che differiscono solo per il tipo di partizione della variabile in esame.

una terza soluzione: c) utilizzare congiuntamente le professioni di entrambi i genitori, assumendo che gli effetti di padre e madre siano identici[11].

Alcune statistiche descrittive dei dati

La fase preliminare delle nostre analisi si è basata sul calcolo di una serie di statistiche descrittive dei dati campionari. Per brevità riportiamo solo alcune informazioni essenziali (ricordiamo che il campione dell'indagine PSLD è composto da 18843 maturi in 1563 scuole):

- gli individui sono per il 44.6% di sesso maschile, di cui il 50.5% ha già assolto gli obblighi di leva al momento dell'intervista, mentre il 19.6% ne è stato esonerato;
- il 93.3% degli intervistati ha un età non superiore a 24 anni;
- la distribuzione del voto di maturità presenta un'asimmetria positiva, con una netta prevalenza dei voti pari e un'elevata frequenza dei voti estremi (36 e 60). Ciò ha suggerito di non usare il voto come variabile quantitativa. In fase di implementazione dei modelli la ripartizione del voto in classi che ha dato i migliori risultati in termini di adattamento è risultata la seguente:

36	37-41	42-49	50-59	60
10.3%	23.9%	38.7%	21.6%	5.4%

- per quanto riguarda la formazione post-diploma, il 12.9% dei maturi si è iscritto ad un Corso di Formazione Regionale (il 77.6% dei quali lo ha già concluso al momento dell'intervista), mentre il 49.6% si è iscritto ad un corso universitario[12] (il 18.6% dei quali ha già interrotto gli studi al momento dell'intervista);
- al momento dell'intervista i maturi che lavorano sono 9311, di cui 7020 (37.3% del totale) svolgono un lavoro *continuativo*;
- il 45.5% di coloro che non lavorano dichiara di essere in cerca di lavoro.

[11] Come riportato nella tab. 3.1, ciò equivale a definire le variabili relative alle professioni come somma delle variabili *dummy* relative ai due genitori. Si noti che le variabili così definite possono assumere il valore 0 (nessun genitore svolge quella professione), 1 (un genitore svolge quella professione), 2 (entrambi i genitori svolgono quella professione).

[12] Sono considerati corsi universitari i Corsi di Laurea e di Diploma Universitario e le Scuole Dirette a Fini Speciali.

Dati mancanti

Alcune delle variabili che abbiamo utilizzato nelle analisi presentano delle lacune dovute alla mancata risposta, come risulta dal seguente prospetto (ricordiamo che la dimensione campionaria totale è 18843):

VARIABILE	dati mancanti
Posizione nei confronti degli obblighi di leva	11
Giudizio di scuola media inferiore	1236
Titolo di studio del padre	469
Titolo di studio della madre	306
Professione del padre	301
Professione della madre	166

Per quanto riguarda il giudizio della scuola media inferiore, questo è un dato fornito direttamente dagli istituti, per cui è lecito attendersi che l'assenza del dato sia correlata con le caratteristiche degli istituti e non con quelle del maturo. Ciò viene confermato dal confronto delle distribuzioni delle principali variabili nei due sottoinsiemi, quello con dati mancanti e quello con dati presenti. In particolare, gli istituti che più frequentemente non riportano il giudizio della scuola media inferiore sono quelli professionali e magistrali.

Le lacune nelle altre variabili (obblighi di leva, titolo e professione dei genitori) sono invece dovute alla mancata risposta del soggetto, per cui è opportuno delineare il *profilo del non rispondente*, inteso come colui che non ha risposto ad una o più delle domande riguardanti gli obblighi di leva e il titolo e la professione dei genitori. A tal fine mostriamo le distribuzioni percentuali del tipo di scuola e del voto di maturità per i due sottoinsiemi di coloro con risposta presente (P) e con risposta mancante (M):

Tipo scuola	P	M
Ist. Profess.	19.1	25.8
Ist. Tecnici	42.0	38.3
Ist. Magistr.	7.1	6.9
Licei	24.6	18.6
Altri istituti	7.1	10.5
	100.0	100.0

Voto matur.	P	M
36	10.1	13.7
37-41	23.7	27.8
42-49	38.7	39.8
50-59	21.9	15.8
60	5.6	2.9
	100.0	100.0

E' evidente che la mancata risposta è più frequente per coloro che provengono da istituti professionali e hanno ottenuto un voto di maturità basso. Inoltre, i non rispondenti hanno un lavoro continuativo nel 41.3% dei casi (contro il 37.1% degli altri) e si sono immatricolati all'università nel 40.4% dei casi (contro il 50.1% degli altri).

Ai fini dello studio della probabilità di svolgere un lavoro continuativo (cap. 4) e dell'analisi dei tempi di ingresso al lavoro (cap. 5), i dati mancanti, per i motivi che descriveremo in seguito, non rappresentano un serio problema e quindi non necessitano di alcun trattamento particolare. Invece, nello studio della probabilità di immatricolazione all'università (cap. 6) si è ritenuto opportuno sostituire i dati mancanti con dei valori imputati, secondo quanto descritto sarà nel par. 6.1.

Struttura gerarchica dei dati e variabili a livello di gruppo

Come risulta evidente dal disegno campionario, i dati dell'indagine PSLD presentano una struttura gerarchica a due livelli, in cui i 18843 maturi intervistati sono le unità di livello 1 e le 1563 scuole di appartenenza sono le unità di livello 2 (gruppi). La distribuzione del numero di maturi per scuola presenta una forte asimmetria negativa ed è caratterizzata dai seguenti valori[13]:

Minimo	1° Quartile	Mediana	3° Quartile	Massimo
1	8	13	16	26

Le informazioni sulle scuole che abbiamo ottenuto dall'Istat si compongono del codice che indica il *tipo di istituto* e di un semplice codice numerico che consente l'abbinamento dei maturi alle scuole, da cui non è possibile, per ovvi motivi di riservatezza, risalire al nome e all'ubicazione dell'istituto. Tuttavia è stato possibile determinare la *regione di appartenenza dell'istituto* a partire dalla regione di residenza dei maturi al momento dell'intervista[14], assegnando l'istituto alla regione in cui risiede la maggior parte dei propri maturi. Con buona probabilità l'assegnazione che abbiamo effettuato è esente da errori, poiché all'interno dei gruppi definiti dalle scuole la variabilità della regione di residenza è molto bassa. Inoltre, al termine dell'assegnazione, abbiamo constatato che la ripartizione delle scuole per regione coincide con quella dichiarata dall'Istat (Istat, 1999).

Un'altra informazione che abbiamo desunto in modo indiretto dal questionario riguarda la natura pubblica o privata dell'istituto[15].

[13]La presenza di alcune scuole con un solo maturo campione è dovuta al fatto che in alcuni casi le sostituzioni dei maturi hanno richiesto l'utilizzo delle scuole suppletive.

[14]In realtà la domanda formulata nel questionario (quesito 122) riguarda la *provincia* di residenza; tuttavia, per motivi di riservatezza, i dati in nostro possesso contengono solo il codice di regione.

[15]In questo caso abbiamo usato le informazioni sul tipo di scuola a cui l'intervistato si è iscritto dopo la licenza media (quesito 3) e sul numero di passaggi fra scuola pubblica e privata (scheda compilata dalla scuola). L'assegnazione presenta un certo margine di

3.3.2 Variabili tratte da altre fonti

Ai fini delle analisi che intendiamo sviluppare i dati dell'indagine PSLD necessitano di alcune integrazioni, poiché è opportuno disporre anche di informazioni sul contesto socio-economico in cui il giovane si è formato e, in gran parte dei casi, ha iniziato la ricerca del lavoro. Poiché è stato possibile determinare solo la regione di appartenenza dell'istituto, la dimensione geografica è necessariamente quella regionale.

Le variabili che abbiamo preso in esame, tutte disaggregate per regione, si riferiscono alle condizioni del mercato del lavoro che, come è noto, presentano marcate differenze territoriali e influenzano pesantemente non solo gli esiti della ricerca del lavoro, ma anche la decisione stessa di cercare lavoro e/o di proseguire la formazione. In dettaglio, le variabili che abbiamo tratto da altre fonti sono le seguenti:

- Tasso di disoccupazione dei giovani in età 15-24 anni negli anni 1995 e 1998, tratto dall'Indagine sulle Forze di Lavoro dell'Istat (Istat, 1996a, 1999a);
- Tasso di occupati nel terziario nel 1995, anch'esso tratto dall'Indagine sulle Forze di Lavoro;
- Proporzione di aziende con 20 e più addetti nel 1995, tratto dai Conti Economici delle Imprese (Istat, 1998).

La tab. 3.2 sintetizza le principali informazioni a livello regionale, sia per quanto riguarda l'universo e il campione PSLD dei maturi che per quanto riguarda alcune delle variabili sopra elencate.

3.4 Strategie di selezione dei modelli e metodi di stima

Concludiamo il capitolo descrivendo le linee guida che abbiamo adottato per l'implementazione dei modelli riportati nei prossimi capitoli.

Innanzitutto premettiamo che gran parte delle operazioni sui dati sono state condotte per mezzo del programma MLwiN (Goldstein *et al.*, 1998), il cui foglio di lavoro consente di gestire in modo conveniente archivi di grosse

errore, poiché l'informazione sul numero di passaggi fra scuola pubblica e privata è poco affidabile. Per 18 scuole l'assegnazione sulla base delle due variabili descritte è risultata impossibile, per cui in 9 casi si fatto ricorso solo alla prima variabile, mentre nei restanti si è proceduto ad una imputazione casuale sulla base della proporzione di scuole private.

Tab. 3.2 - Composizione regionale dell'universo e del campione PSLD e tassi di disoccupazione giovanile regionali

	Scuole		Maturi		Percent. campion. maturi	Giovani 15-24 anni - Media 1995			Giovani 15-24 anni - Media 1998			Differ. disocc. 95-98
	Universo	Campione	Universo	Campione		In cerca di occupaz.	Forze di lavoro	Tasso di disoccupaz.	In cerca di occupaz.	Forze di lavoro	Tasso di disoccupaz.	
Piemonte	510	101	30220	1196	4.0	67	256	26.2	63	238	26.5	0.3
Valle d'Aosta	23	18	685	212	30.9	1	7	14.3	1	6	16.7	2.4
Lombardia	966	148	66116	1943	2.9	107	573	18.7	87	533	16.3	-2.4
Trentino-A.A.	119	41	6207	454	7.3	6	67	9.0	4	64	6.3	-2.7
Veneto	497	106	35473	1325	3.7	42	304	13.8	36	294	12.2	-1.6
Friuli V.G.	145	53	8920	598	6.7	16	66	24.2	9	61	14.8	-9.5
Liguria	213	63	11458	661	5.8	30	73	41.1	26	69	37.7	-3.4
Emilia Rom.	400	90	30921	1149	3.7	42	239	17.6	33	223	14.8	-2.8
Toscana	400	86	28529	1099	3.9	49	195	25.1	39	177	22.0	-3.1
Umbria	110	42	7377	493	6.7	12	37	32.4	9	34	26.5	-6.0
Marche	187	57	12658	675	5.3	14	76	18.4	13	73	17.8	-0.6
Lazio	697	124	51608	1569	3.0	106	227	46.7	91	209	43.5	-3.2
Abruzzo	161	53	12232	670	5.5	16	54	29.6	14	50	28.0	-1.6
Molise	54	31	3103	375	12.1	7	14	50.0	7	15	46.7	-3.3
Campania	694	129	55918	1620	2.9	194	300	64.7	194	310	62.6	-2.1
Puglia	491	106	39613	1197	3.0	100	217	46.1	118	240	49.2	3.1
Basilicata	119	45	6101	499	8.2	13	27	48.1	13	26	50.0	1.9
Calabria	325	76	21496	904	4.2	57	95	60.0	65	97	67.0	7.0
Sicilia	786	130	46088	1457	3.2	144	245	58.8	156	260	60.0	1.2
Sardegna	247	64	15625	747	4.8	51	99	51.5	50	93	53.8	2.2
ITALIA	7144	1563	490348	18843	3.8	1074	3172	33.9	1028	3075	33.4	-0.4

dimensioni. Il miglior metodo di stima disponibile in MLwiN è la Quasi-Verosimiglianza Penalizzata del Secondo Ordine (PQL2, cfr. par. 2.2.6), che abbiamo adottato in tutte le nostre applicazioni. Una volta selezionato il modello, abbiamo verificato la validità delle stime PQL2 confrontandole con quelle di Massima Verosimiglianza con integrazione numerica fornite dal programma MIXOR (Hedeker&Gibbons, 1996).

Nel caso di variabili di risposta di tipo binario abbiamo usato la distribuzione Bernuolli[16] e il link *logit*, che rappresenta la scelta standard e consente l'interpretazione dei risultati in termini di *odds ratios* (Agresti, 1990). Per i modelli di sopravvivenza si è invece preferito il link *complementary log-log*, per le ragioni che verranno discusse nel cap. 5.

La strategia cha abbiamo adottato per la selezione del modello si compone dei seguenti passi:

1. Si implementa un modello a *intercetta casuale* con un insieme iniziale di variabili scelte in base a:
   (a) i suggerimenti della teoria (cfr. cap. 1);
   (b) i precedenti lavori empirici (cfr. cap. 1);
   (c) l'esame delle distribuzioni doppie che coinvolgono la variabile di risposta.
2. Si tolgono le variabili non significative, se ne provano altre, si testano le interazioni e i termini quadratici, il tutto usando come metro di giudizio il test di Wald generalizzato (Goldstein, 1995, par. 2.11) al livello di confidenza 95%[17];
3. Si testa l'ipotesi che qualche covariata abbia un coefficiente casuale anziché fisso, usando anche in questo caso il test di Wald generalizzato[18].

[16] Vale a dire, abbiamo fissato a 1 il parametro di variabilità extra-binomiale stimabile con i metodi di quasi-verosimiglianza. Nelle prove di stima che abbiamo fatto tale parametro risultava di poco inferiore a 1.

[17] Il test del rapporto di verosimiglianza non può essere usato quando la stima viene effettuata tramite un metodo di quasi-verosimiglianza come PQL2. Infatti, in tal caso la verosimiglianza non viene nemmeno definita e può essere calcolata solo per mezzo di approssimazioni del tutto inaffidabili (cfr. Goldstein, 1995, par. 5.1.3).

[18] Nelle nostre applicazioni, data la dimensione e la struttura del campione, le ipotesi alla base del test di Wald sono verosimilmente soddisfatte anche per i parametri casuali, per i quali, notoriamente, le approssimazioni asintotiche sono accettabili solo per un elevato numero di unità di secondo livello. L'ipotesi di nullità di una componente di varianza può essere sottoposta a verifica anche con il test del rapporto di verosimiglianza; tuttavia, poiché l'ipotesi nulla cade sulla frontiera dello spazio parametrico, il valore del *p-value* del test di significatività deve essere dimezzato (cfr. Snijders & Bosker, 1999, par. 6.2.1). In alternativa si possono usare test basati sullo *score* (Mealli, 1998).

Capitolo 4

Analisi della probabilità di occupazione

La prima analisi che abbiamo condotto riguarda gli sbocchi professionali dei diplomati, in particolare la probabilità di svolgere un *lavoro continuativo* al momento dell'intervista. I lavori di tipo occasionale o stagionale sono stati ignorati, in quanto rappresentano situazioni transitorie legate a fattori contingenti[1]. L'arco temporale di oltre tre anni considerato dall'indagine sui Percorsi di Studio e Lavoro dei Diplomati (PSLD) è sufficiente per una prima valutazione del ruolo giocato dai fattori individuali e contestuali (cfr. par. 1.1), anche se è bene ricordare che l'inserimento nel mondo del lavoro è caratterizzato da un lungo periodo di aggiustamenti successivi e che molti diplomati rimandano l'inserimento di alcuni anni per proseguire gli studi.

Nel primo paragrafo discuteremo le modalità di ottenimento del sottocampione sul quale abbiamo condotto l'analisi. Successivamente presenteremo il modello finale con le relative stime. Il terzo paragrafo sarà dedicato all'esame dei residui e alla valutazione dell'efficacia delle scuole. Infine, concluderemo il capitolo discutendo i possibili effetti del piano di campionamento sulle stime e presentando i risultati ottenuti dall'applicazione di alcune procedure di stima che utilizzano i pesi campionari.

[1] Va detto che sempre più spesso i lavori occasionali/stagionali rappresentano una tappa obbligata nel percorso che porta al lavoro continuativo. Tuttavia è impossibile sapere se il lavoro occasionale/stagionale si trasformerà in un lavoro continuativo oppure se rappresenterà solo un episodio isolato. Pertanto riteniamo opportuno restringere l'attenzione ai soli lavori continuativi (che, lo ricordiamo, includono anche i contratti a tempo determinato).

4.1 Determinazione del sottocampione di interesse

Il fine della nostra analisi è lo studio dell'effetto dei fattori individuali e contestuali sulla probabilità di occupazione, in un ottica di valutazione dell'efficacia (cfr. par. 1.2.4). Pertanto la popolazione di interesse non è costituita da tutti i diplomati, ma solo da quelli che, negli anni immediatamente successivi al conseguimento del titolo, hanno cercato lavoro. Nel nostro caso effettuare l'analisi su tutti i diplomati porterebbe a dei risultati del tutto fuorvianti: ad esempio, poiché i liceali si iscrivono in massa all'università e quindi hanno una bassa propensione a cercare lavoro, le stime di un modello riferito all'intero campione indicherebbero una pessima performance dei licei.

Sebbene necessaria, l'individuazione del sottoinsieme di coloro che cercano lavoro presenta dei notevoli problemi sia teorici che pratici:

- innanzitutto, in generale la definizione di "persona in cerca di lavoro" è ambigua[2], poiché il desiderio di lavorare può essere più o meno forte e può rimanere latente oppure tradursi in una serie di azioni concrete;
- inoltre, la ricerca di lavoro è un fenomeno variabile nel tempo per lo stesso soggetto.

Dall'indagine PSLD si possono trarre molte informazioni sulle modalità della ricerca di lavoro; tuttavia quasi tutte le informazioni si riferiscono al momento dell'intervista ed è ben difficile ricostruire una storia della ricerca di lavoro[3].

Tenendo presenti queste considerazioni generali, abbiamo seguito i seguenti criteri per individuare il sottocampione sul quale effettuare le analisi:

1. Innanzitutto abbiamo eliminato dal campione originale (composto da 18843 individui) tutti coloro che al momento dell'intervista svolgono un lavoro continuativo iniziato *prima* dell'ottenimento del diploma (si tratta di 743 individui). Infatti per questi soggetti è impossibile valutare l'effetto del diploma sulle possibilità occupazionali.

[2] A livello di statistiche ufficiali, ciò ha generato un acceso dibattito sulla definizione delle "forze di lavoro", che ha portato ad una revisione delle modalità di calcolo dei tassi di disoccupazione (cfr. Isfol, 1998).

[3] La ricerca di lavoro al momento dell'intervista è oggetto dei quesiti 63, 64, 91 e 93-98. Invece i quesiti relativi alla ricerca di lavoro in passato sono: (i) il quesito 92, che chiede di indicare a quando risale l'ultima iniziativa concreta per cercare lavoro; (ii) i quesiti 13 e 14, che indagano sulla ricerca di lavoro *al momento dell'iscrizione* a Corsi di Formazione Professionale; (iii) i quesiti 44 e 45, che indagano sulla ricerca di lavoro *al momento dell'iscrizione* a corsi universitari.

2. Poi, per i motivi sopra accennati, abbiamo tolto coloro che al momento dell'intervista non hanno un lavoro continuativo e non cercano lavoro (6821 individui). La condizione di "persona in cerca di lavoro" viene valutata in base ad un quesito diretto riferito al momento dell'intervista (il n. 63). Naturalmente in questo modo non si tiene conto della condizione passata, in particolare c'è il rischio di escludere i cosiddetti "lavoratori scoraggiati", cioè coloro che, dopo un periodo di ricerca senza successo, abbandonano ogni speranza e si dichiarano non più in cerca. Fra i diplomati esclusi dal campione ci sono anche coloro che, pur dichiarando di cercare lavoro, non hanno ancora preso nessuna iniziativa concreta (quesito 92): questi sono per lo più individui entrati da poco nel mercato del lavoro e quindi non confrontabili con coloro che hanno cercato attivamente lavoro per gran parte del periodo di osservazione[4].

3. Infine abbiamo escluso anche i diplomati che al momento dell'intervista frequentano corsi universitari, ad eccezione di coloro che hanno dichiarato di essersi iscritti a tali corsi pur desiderando iniziare stabilmente un lavoro (quesito 45). Ciò ha comportato l'eliminazione di 1569 individui. In questo modo gli studenti universitari che rimangono nel campione sono quelli che hanno manifestato il desiderio di lavorare sia al momento dell'iscrizione che al momento dell'intervista e che, pertanto, possono essere identificati come coloro che utilizzano l'università come "seconda scelta" o come "parcheggio" in attesa di tempi migliori.

Al termine di queste operazioni il sottocampione risulta composto da 9710 unità, di cui 225 (il 2.3%) hanno dati mancanti in alcune delle covariate da includere nel modello[5]. Essendo poche, queste unità sono state eliminate, per cui l'analisi è stata effettuata sui casi completi[6].

Il sottocampione così ottenuto si compone di 9485 maturi in 1448 scuole, con un valore mediano di 6 maturi per scuola. Naturalmente, poiché l'eliminazione degli individui non è stata casuale, alcune caratteristiche sono assai

[4]Gli individui che si dichiarano in cerca di lavoro, ma che non hanno ancora preso alcuna iniziativa sono 697 e costituiscono il 9.7% delle persone in cerca di lavoro.

[5]Si tratta delle covariate riguardanti la posizione nei confronti degli obblighi di leva e la professione dei genitori. Alcune analisi preliminari condotte sugli individui con record completi hanno permesso di scartare in partenza alcune altre covariate con dati mancanti, come quelle relative al giudizio di licenza media e al titolo di studio dei genitori.

[6]Ricordiamo che l'analisi sui casi completi (*Complete-Case Analysis*, Little, 1998) nei modelli di regressione fornisce inferenze valide a patto che: (i) il modello sia correttamente specificato; (ii) i dati mancanti non dipendano dalla variabile di risposta. Nel nostro caso la relazione fra dati mancanti e condizione lavorativa risulta debole (hanno dati mancanti il 2.47% di coloro che lavorano e il 2.17% di coloro che non lavorano).

diverse rispetto al campione originale. Ad esempio, la percentuale di coloro che lavorano sale dal 37% al 60%, mentre gli istituti tecnici e professionali sono più rappresentati (si passa dal 61% al 78%).

La tab. 4.1 riporta alcune statistiche descrittive delle variabili impiegate nel modello finale.

4.2 Analisi dei risultati delle stime del modello

Utilizzando il campione determinato secondo i criteri descritti nel precedente paragrafo, abbiamo assunto come variabile di risposta LAVC ("l'intervistato svolge un lavoro continuativo al momento dell'intervista", cfr. tab. 3.1) e implementato un modello logit a intercetta casuale (cfr. par. 2.2.2):

$$\begin{aligned} y_{ij} \,|\, u_j &\sim \text{Bernoulli}(\pi_{ij}), \\ \text{logit}(\pi_{ij}) &= \alpha + u_j + \boldsymbol{\beta}' \mathbf{x}_{ij}, \\ u_j &\sim N(0, \sigma_u^2), \end{aligned}$$

dove $i = 1, \ldots, 9485$ indica i maturi (unità di livello 1) e $j = 1, \ldots, 1448$ indica le scuole (unità di livello 2 o gruppi).

Seguendo la strategia delineata nel par. 3.4 abbiamo ottenuto un modello finale con 27 variabili esplicative, più un'intercetta e un effetto casuale sull'intercetta (vedi tab. 4.2):

- Le differenze fra i risultati ottenuti con i metodi PQL2 e ML (cfr. par. 2.2.6) sono nel complesso trascurabili e riguardano soprattutto la componente di varianza e gli errori standard. L'unica differenza di rilievo consiste nel fatto che con ML il coefficiente di ITP-AGR, a causa di un notevole aumento della stima dell'errore standard, non è significativo.

- La componente di varianza risulta ampiamente significativa, ma di valore modesto ($\hat{\sigma}_u^2 = 0.126$)[7]. Ciò va imputato all'elevato potere esplicativo delle variabili di livello 2 (tipo di scuola, tassi di disoccupazione, indicatori della circoscrizione geografica): infatti usando lo stesso modello, ma senza variabili esplicative si ottiene $\hat{\sigma}_u^2 = 1.470$, per cui le variabili di livello 2 spiegano il 91.4% della variabilità tra gruppi. In

[7] Assumendo un modello a soglia con variabile latente (cfr. par. 2.2.2), la correlazione (residua) infragruppo è 0.037, un valore modesto, ma tutt'altro che inusuale nei modelli a risposta binaria.

Tab. 4.1 - Statistiche descrittive delle variabili impiegate nel modello per la probabilità di svolgere un lavoro continuativo al momento dell'intervista

Unità livello 1 (diplomati)	9485
Unità livello 2 (scuole)	1448

Variabile	Descrizione sintetica*	Min	Max	Media	Err. Std.
LAVC	Attualmente svolge lavoro continuativo	0	1	0.60	0.49
FEMM	Femmina	0	1	0.54	0.50
LEVA-POI	Servizio di leva ancora da svolgere	0	1	0.03	0.16
VM36	Voto di maturità 36	0	1	0.13	0.34
VM37-42	Voto di maturità 37-42	0	1	0.40	0.49
VM50-59	Voto di maturità 50-59	0	1	0.16	0.37
VM60	Voto di maturità 60	0	1	0.02	0.15
PROF-D	Prof. genitore: altri indipendenti	0	2	0.31	0.55
PROF-E	Prof. genitore: imprenditore/lib.prof.	0	2	0.08	0.28
LAVCINT	Lavori continuativi interrotti	0	1	0.14	0.35
UNIV	Attualmente iscritto università	0	1	0.06	0.24
INTUNIV-0	Interrotta università motivi diversi da lavoro	0	1	0.10	0.30
INTUNIV-5	Interrotta università motivi lavoro	0	1	0.04	0.20
CFP-ORA	Attualmente iscritto Corso Formaz. Prof.	0	1	0.02	0.13
CFP-CONC	Concluso Corso Formaz. Prof.	0	1	0.15	0.36
ALTRIORA	Attualmente iscritto altri corsi	0	1	0.02	0.13
ITP-AGR	Ist. Tec./Prof. agrario	0	1	0.06	0.23
ITP-IND	Ist. Tec./Prof. industriale	0	1	0.20	0.40
ITP-TUR	Ist. Tec./Prof. turistico	0	1	0.07	0.25
ITP-ALTR	Ist. Tec./Prof. altri	0	1	0.14	0.35
ISTMAG	Istituto magistrale	0	1	0.07	0.25
LICEI	Licei	0	1	0.07	0.25
ALTRIIST	Altri istituti	0	1	0.08	0.27
DISOCC	Tasso disocc. giovanile region. 1995	-16.2	39.5	11.13	17.66
±DISOCC	Variaz. tasso disocc. giov. regionale 1995-1998	-9.5	7	-1.14	3.17
NORD	Circoscrizione Nord	0	1	0.40	0.49
SUD	Circoscrizione Sud	0	1	0.40	0.49

* Le descrizioni dettagliate delle variabili sono riportate nella tab. 3.1

Tab. 4.2 - Risultati delle stime relative al modello per la probabilità di svolgere un lavoro continuativo al momento dell'intervista

	MLwiN		**MIXOR**		
	Link	Logit	Link	Logit	
	Distr. EC	Gaussiana	Distr. EC	Gaussiana	
	Metodo	PQL2	Metodo	ML	
	Variaz. EB	no	Punti quad.	10	
	Iterazioni	5	Iterazioni	15	
	Devianza	-	Devianza	9849.2	
Variabili	**Effetti fissi**		**Effetti fissi**		**Differenze fra ML e PQL2**
	Stima	Err. Std.	Stima	Err. Std.	Stima	Err. Std.
Intercetta	1.427	0.107	1.428	0.110	0.044%	3.100%
FEMM	-0.428	0.062	-0.428	0.060	0.035%	-3.583%
LEVA-POI	-0.763	0.174	-0.764	0.166	0.151%	-4.618%
VM36	-0.234	0.084	-0.233	0.089	-0.098%	5.318%
VM37-42	-0.149	0.062	-0.149	0.065	0.087%	5.173%
VM50-59	0.169	0.081	0.169	0.083	0.397%	2.288%
VM60	0.503	0.188	0.505	0.199	0.338%	5.622%
PROF-D	0.133	0.047	0.134	0.049	0.188%	2.962%
PROF-E	0.454	0.097	0.455	0.093	0.167%	-3.864%
LAVCINT	0.258	0.077	0.258	0.081	-0.062%	4.887%
UNIV	-2.160	0.132	-2.162	0.135	0.077%	1.662%
INTUNIV-0	-0.914	0.084	-0.915	0.081	0.086%	-4.135%
INTUNIV-5	0.495	0.133	0.494	0.124	-0.151%	-6.469%
CFP-ORA	-1.475	0.209	-1.476	0.213	0.052%	2.228%
CFP-CONC	-0.540	0.069	-0.541	0.070	0.115%	2.001%
ALTRIORA	-1.246	0.207	-1.247	0.198	0.083%	-4.386%
ITP-AGR	-0.268	0.125	-0.268	0.153	0.093%	22.472%
ITP-IND	0.234	0.089	0.234	0.089	0.163%	0.079%
ITP-TUR	-0.266	0.114	-0.266	0.111	0.105%	-3.006%
ITP-ALTR	-0.304	0.088	-0.304	0.092	0.059%	4.731%
ISTMAG	-0.512	0.113	-0.512	0.115	0.107%	2.405%
LICEI	-0.281	0.116	-0.281	0.114	0.046%	-2.339%
ALTRIIST	-0.597	0.108	-0.598	0.105	0.167%	-2.167%
DISOCC	-0.043	0.005	-0.043	0.005	0.094%	2.345%
DISOCC^2	0.00030	0.00013	0.00030	0.00014	-0.365%	8.359%
±DISOCC	-0.042	0.011	-0.042	0.012	0.024%	7.756%
NORD	0.418	0.088	0.418	0.091	0.136%	3.732%
SUD	-0.319	0.105	-0.319	0.103	0.022%	-1.656%
	Varianza effetto casuale		**Errore Std. effetto casuale**		
	Stima	Err. Std.	Stima	Err. Std.		
Intercetta	0.126	0.034	0.366	0.051		
	Var. residua	pi*pi/3	Var. residua	pi*pi/3		
	Var. di gruppo	0.126	Var. di gruppo	0.134	6.294%	
	Corr. infragruppo	0.037	Corr. infragruppo	0.039		

Nota: le definizioni delle variabili sono riportate nella tab. 3.1

termini pratici, ciò significa che eventuali sforzi per introdurre ulteriori variabili di livello 2 (indicatori della qualità dell'insegnamento, delle risorse disponibili ecc.) non produrrebbero grandi risultati in termini di riduzione della variabilità fra gruppi[8]. L'impatto degli effetti casuali è comunque considerevole: infatti data la distribuzione $N(0, 0.126)$, il 2.5% delle scuole ha un effetto superiore a 0.695, che è maggiore di tutti i coefficienti stimati.

- Nella ricerca del lavoro risultano svantaggiate le femmine e, soprattutto, i maschi che devono ancora svolgere il servizio militare; non sembra esservi invece un vantaggio significativo dei maschi esonerati dal servizio rispetto a quelli che lo hanno svolto negli anni immediatamente successivi al conseguimento del diploma.

- Per quanto riguarda il curriculum scolastico, solo il voto di maturità ha un effetto significativo (in particolare, si distinguono coloro che hanno ottenuto il massimo dei voti). Il giudizio di licenza media, le ripetenze e i cambi di indirizzo scolastico non hanno effetti apprezzabili.

- Tra le professioni dei genitori, quelle di tipo indipendente hanno un effetto positivo. Non si registra invece l'effetto negativo che ci si poteva attendere per la madre casalinga. Inoltre non sono significativi i coefficienti delle variabili relative al titolo di studio dei genitori.

- Coloro che hanno avuto qualche esperienza di lavoro continuativo hanno maggiori probabilità di lavorare al momento dell'intervista. Ciò conferma il ruolo esercitato dall'esperienza lavorativa (cfr. par. 1.1).

- Il coefficiente di UNIV, -2.160, è il più grande in valore assoluto[9]. Pertanto coloro che usano l'università come "seconda scelta" o come "parcheggio" hanno una bassa probabilità di occupazione nell'immediato. Ciò può essere dovuto al fatto che gli studi universitari riducono le motivazioni e le risorse impiegate nella ricerca del lavoro. Tuttavia non bisogna dimenticare che coloro per i quali l'università è una "seconda scelta" o un "parcheggio" sono soggetti deboli sul mercato del lavoro, per cui è difficile pensare che i loro esiti occupazionali siano attribuibili

[8] Si consideri anche che parte della variabilità fra gli istituti è imputabile a differenze infraregionali nelle condizioni del mercato del lavoro. Tale variabilità potrebbe essere rimossa se fosse possibile conoscere la provincia di appartenenza dei singoli istituti e quindi inserire i tassi di disoccupazione a livello provinciale.

[9] Ricordiamo che, per come abbiamo selezionato il campione (cfr. par. 4.1), UNIV=1 individua coloro che al momento dell'intervista cercano lavoro e frequentano corsi universitari a cui si sono iscritti pur desiderando iniziare stabilmente un lavoro.

interamente all'impegno universitario. L'effetto negativo si attenua per coloro che gli studi universitari li hanno interrotti (INTUNIV-0)[10].

- Anche la frequenza di corsi di formazione professionale (CFP-ORA) e di altri corsi non universitari (ALTRIORA) riduce la probabilità di occupazione nell'immediato, per motivi analoghi a quelli discussi al punto precedente[11]. Parzialmente inatteso è il valore negativo del coefficiente relativo a coloro che hanno già concluso un corso di formazione professionale (CFP-CONC): una possibile spiegazione è che l'ottenimento del titolo rilasciato dal corso non riesce a compensare la perdita di opportunità di lavoro che si verifica durante la frequenza del corso stesso oppure non riesce a colmare quelle lacune che impedivano al giovane di entrare nel mondo del lavoro.

- Per quanto riguarda i tipi di scuola, poiché la distinzione fra istituti professionali e tecnici non risultava statisticamente significativa, per questi istituti abbiamo effettuato una classificazione in base all'indirizzo dei corsi (la categoria di base è quella degli istituti tecnici e professionali a indirizzo commerciale/aziendale). I valori dei coefficienti sono in linea con le attese: le migliori prospettive occupazionali sono fornite dagli istituti tecnici e professionali, in particolare quelli a indirizzo industriale e commerciale/aziendale, mentre le peggiori sono fornite dagli istituti di tipo magistrale e artistico. Non è risultata invece significativa la variabile indicatrice delle scuole private, per cui, ai fini degli esiti occupazionali, non esiste alcuna differenza di rilievo fra scuole pubbliche e private.

- Le variabili relative alla situazione del mercato del lavoro che sono risultate significative sono il tasso di disoccupazione giovanile regionale del 1995 (DISOCC), il suo termine quadratico (DISOCC^2) e la variazione dello stesso tasso fra il 1995 e il 1998 (±DISOCC). Tuttavia, poiché sono risultate significative anche le variabili relative alle circoscrizioni

[10] Ricordiamo che INTUNIV-0=1 individua coloro che hanno interrotto gli studi universitari per motivi diversi dal lavoro e INTUNIV-5=1 coloro che li hanno interrotti per motivi di lavoro. La seconda variabile, che ovviamente ha un coefficiente positivo, presenta un potenziale problema di "circolarità" con la variabile di risposta del modello. Una soluzione radicale può essere quella di escludere dal campione tutti gli individui con INTUNIV-5=1. Tuttavia, non tutti questi individui svolgono poi un lavoro continuativo al momento dell'intervista, per cui abbiamo preferito lasciarli nel campione e segnalarli con una variabile *ad hoc*.

[11] Tuttavia, a differenza dei corsi universitari, CFP-ORA e ALTRIORA includono tutti i frequentanti, indipendentemente dalle motivazioni al momento dell'iscrizione, per cui l'interpretazione è leggermente divesa da quella di UNIV.

geografiche (Nord, Centro e Sud)[12], i tassi di disoccupazione impiegati non sono indicatori esaustivi delle difficoltà di inserimento nel mondo del lavoro[13]. Comunque è interessante notare che senza le variabili NORD e SUD i coefficienti dei tassi di disoccupazione sono maggiori (in valore assoluto) di circa il 50% ($\hat{\beta}_{\text{DISOCC}} = 0.060$, $\hat{\beta}_{\pm\text{DISOCC}} = 0.062$).

- Non hanno un effetto significativo la percentuale regionale di occupati nel terziario e la proporzione regionale di grandi aziende (20 e più addetti).

4.2.1 Impatto delle variabili esplicative e degli effetti casuali sulla probabilità di occupazione

L'impatto delle variabili esplicative sulla probabilità di occupazione può essere apprezzato con l'ausilio della tab. 4.3, nella quale i coefficienti del predittore lineare vengono tradotti in variazioni di probabilità rispetto all'*individuo-base*, cioè quell'individuo per il quale tutte le variabili esplicative e gli effetti casuali assumono il valore zero. Nel nostro caso il profilo dell'individuo-base è il seguente:

- maschio;
- ha già svolto il servizio di leva oppure è stato esonerato;
- ha frequentato una scuola con le seguenti caratteristiche:
    - istituto tecnico o professionale di tipo commerciale/aziendale;
    - istituto ubicato in Toscana;
    - istituto medio (nel senso che il suo effetto casuale è nullo);
- ha ottenuto un voto di maturità compreso tra 43 e 49;
- i genitori sono lavoratori dipendenti (oppure la madre è casalinga);
- non ha interrotto lavori continuativi iniziati dopo il diploma;
- non si è iscritto all'università, né ad altri corsi di studio o formazione professionale.

Nella tabella vengono inoltre presentate le probabilità di occupazione per due profili estremi, cioè per due ipotetici individui con caratteristiche

[12] Abbiamo provato anche a suddividere ulteriormente le circoscrizioni (ad es., il Nord in Nord-Ovest e Nord-Est), ma i coefficienti delle partizioni non presentavano differenze significative.

[13] In effetti, i tassi di disoccupazione che abbiamo usato non sono del tutto adeguati, poiché si riferiscono ai *giovani in età 15-24, indipendentemente dal titolo di studio*, mentre nel campione PSLD sono presenti esclusivamente diplomati, per lo più in età 21-24. Di conseguenza i tassi impiegati non rendono conto appieno del divario Nord-Sud esistente nel campione se tale divario cambia con il titolo di studio e con l'età.

Tab. 4.3 - Impatto delle variabili esplicative e degli effetti casuali sulla probabilità di occupazione - Stime PQL2

Variabili	Stima	Variaz. Prob. rispetto all' individuo-base
Intercetta	1.427	/
FEMM	-0.428	-0.075
LEVA-POI	-0.763	-0.146
VM36	-0.234	-0.039
VM37-42	-0.149	-0.024
VM50-59	0.169	0.025
VM60	0.503	0.067
PROF-D	0.133	0.020
PROF-E	0.454	0.062
LAVCINT	0.258	0.038
UNIV	-2.160	-0.481
INTUNIV-0	-0.914	-0.180
INTUNIV-5	0.495	0.066
CFP-ORA	-1.475	-0.318
CFP-CONC	-0.540	-0.098
ALTRIORA	-1.246	-0.261
ITP-AGR	-0.268	-0.045
ITP-IND	0.234	0.034
ITP-TUR	-0.266	-0.044
ITP-ALTR	-0.304	-0.051
ISTMAG	-0.512	-0.092
LICEI	-0.281	-0.047
ALTRIIST	-0.597	-0.110
DISOCC	-0.043	-0.074 *
DISOCC^2	0.00030	/
±DISOCC	-0.042	-0.006
NORD	0.418	0.058
SUD	-0.319	-0.054

* Per quanto riguarda il tasso di disoccupazione giovanile, si è considerata una variazione di 10 punti percentuali, includendo nel calcolo anche il termine quadratico

Probabilità individuo-base 0.806

Esempio di profilo favorevole:

	Coeff.	Pred. Lin.	Prob.
Trentino A.A.		2.728	0.939
PROF-E	0.454	3.183	0.960
PROF-E	0.454	3.637	0.974
ITP-IND	0.234	3.871	0.980
VM60	0.503	4.374	0.988

Esempio di profilo sfavorevole:

	Coeff.	Pred. Lin.	Prob.
Calabria		-0.310	0.423
ALTRIIST	-0.597	-0.907	0.288
VM36	-0.234	-1.141	0.242
LEVA-POI	-0.763	-1.904	0.130
UNIV	-2.160	-4.064	0.017

Valutazione degli effetti casuali:

Ipotetico valore dell' effetto casuale	Variaz. Prob. rispetto all' individuo-base
-0.710	-0.134
-0.355	-0.061
0.355	0.050
0.710	0.088

particolarmente favorevoli o sfavorevoli[14]. Si noti che gran parte dell'enorme differenza esistente fra i due profili estremi è imputabile alla regione di appartenenza della scuola.

Infine la tab. 4.3 presenta una valutazione dell'importanza degli effetti casuali, calcolando le variazioni della probabilità rispetto all'individuo-base indotte dalle seguenti realizzazioni dell'effetto casuale: $(-2\hat{\sigma}_u, -\hat{\sigma}_u, \hat{\sigma}_u, 2\hat{\sigma}_u)$. La conclusione è che, *ceteris paribus*, frequentare un istituto piuttosto che un altro modifica sensibilmente le probabilità di occupazione[15].

4.3 Analisi dei residui ed efficacia delle scuole

I residui di livello 2 sono delle stime delle realizzazioni degli effetti casuali nei singoli gruppi e quindi possono essere usati per controllare alcune assunzioni del modello e per effettuare confronti fra i gruppi.

La fig. 4.1 riporta il diagramma quantile-quantile dei residui standardizzati relativi alle scuole, da cui si conclude che la loro distribuzione è approssimativamente normale. Pertanto l'ipotesi di normalità degli effetti casuali è plausibile[16].

Per quanto riguarda la valutazione delle scuole, osserviamo innanzitutto che il modello discusso nel precedente paragrafo include, oltre alle variabili individuali, anche delle variabili relative al tipo di istituto e alle condizioni socio-economiche dell'ambiente in cui le scuole operano. Pertanto gli effetti casuali delle singole scuole si prestano ad essere interpretati come misure di *efficacia di Tipo B* (cfr. par. 1.2.4). Tale interpretazione richiede, però, molta cautela, sia per i problemi di carattere generale discussi da Raudenbush & Willms (1995), sia perché nel nostro caso l'aggiustamento per le condizioni socio-economiche non è ottimale (per le ragioni delineate nel paragrafo precedente).

La fig. 4.2 mostra i residui in ordine crescente, insieme ad alcuni intervalli di confidenza costruiti secondo il metodo proposto da Goldstein&Healy

[14]Le probabilità per i profili favorevole e sfavorevole sono calcolate in modo sequenziale apportando di volta in volta una variazione rispetto al profilo-base. Il coefficiente della regione è stato calcolato combinando i tassi di disoccupazione e gli indicatori della circoscrizione geografica.

[15]Tuttavia non va dimenticato che parte delle differenze attribuibili agli istituti sono dovute alla variabilità infraregionale dei tassi di disoccupazione.

[16]L'importanza della scelta della distribuzione degli effetti casuali è stata verificata confrontando le stime che si ottengono con MIXOR passando dalla distribuzione normale a quella uniforme. Le differenze nei parametri fissi e nella devianza sono risultate quasi nulle, suggerendo che le stime sono robuste rispetto alla scelta della distribuzione degli effetti casuali.

Fig. 4.1 - Diagramma quantile-quantile dei residui standardizzati relativi alle scuole

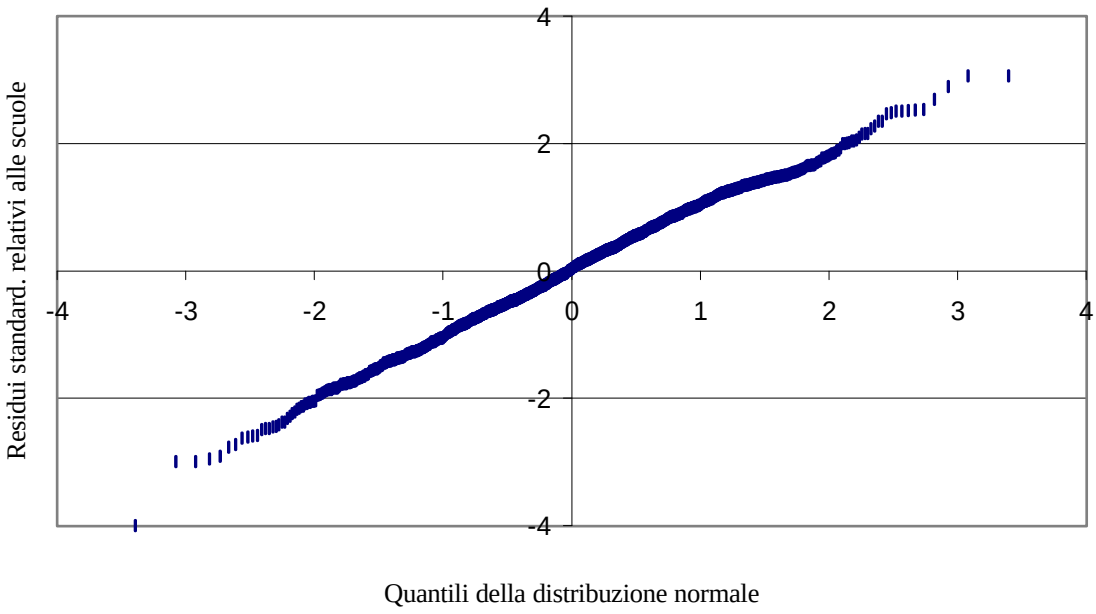
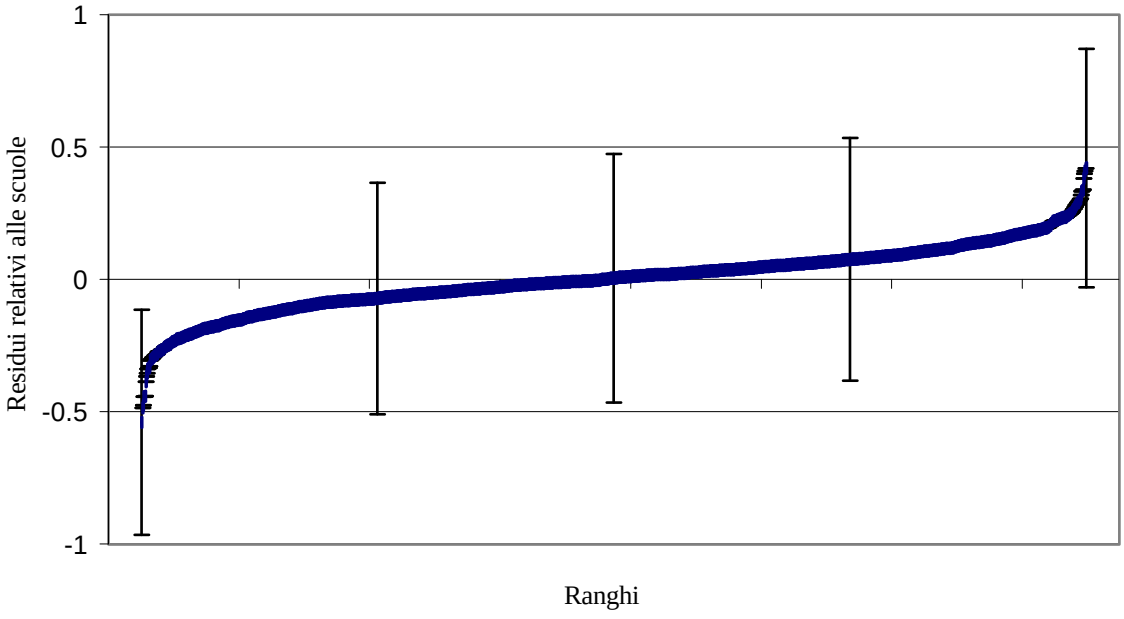


Fig. 4.2 - Residui relativi alle scuole in ordine crescente, con alcuni intervalli di confidenza al 95% per i confronti a coppie



(1994) per i confronti a coppie[17]. Dal grafico risulta evidente che le uniche coppie di intervalli disgiunti sono quelle formate da scuole con residui estremi: in altri termini, solo le prime tre-quattro scuole sono significativamente migliori dell'ultima. In pratica, ciò significa che nella presente applicazione i residui non possono essere usati per confrontare l'efficacia delle scuole (questa è una limitazione ben nota in letteratura: cfr. Goldstein&Spiegelhalter, 1996).

A questo proposito ricordiamo anche i risultati delle simulazioni dell'appendice B, che possono essere ragionevolmente estesi al presente caso in quanto basati su gruppi di analoga numerosità. Nella presente applicazione il valore stimato della componente di varianza è 0.126, per cui possiamo far riferimento alla simulazione con $\sigma_u^2 = 0.15$ (cfr. tab. 2.1), che suggerisce due osservazioni:

a) La copertura media percentuale dell'intervallo di confidenza al 95% per gli effetti casuali è dell'85.4%. Ciò significa che gli errori standard comparativi dei residui sono distorti verso il basso e che, quindi, gli intervalli della fig. 4.2 sono più stretti del dovuto. La conseguenza è che il livello di confidenza effettivo dei confronti a coppie è inferiore al 95%.

b) L'indice di cograduazione di Spearman medio fra i ranghi dei residui e degli effetti casuali è di 0.38, per cui la graduatoria delle scuole basata sui residui è scarsamente affidabile.

Questi risultati rafforzano la conclusione già raggiunta.

Le comparazioni di efficacia di Tipo B fra le scuole sono dunque un'utopia? Nel nostro caso certamente sì, poiché le differenze fra le scuole sono troppo piccole per poter essere rivelate. Tuttavia, le differenze verrebbero alla luce se i dati contenessero maggiori informazioni, ad esempio:

- un maggior numero di studenti per scuola: infatti in questo modo si ridurrebbe l'*effetto shrinkage* (par. 2.1.9), cioè l'appiattimento dei residui verso lo zero (a questo proposito si veda la simulazione n. 2 dell'app. B);

[17] Gli intervalli sono costruiti aggiungendo e sottraendo al residuo 1.39 volte l'errore standard comparativo (cfr. par. 2.1.8). In questo modo, sotto l'ipotesi di normalità, le realizzazioni degli effetti casuali in due gruppi qualsiasi sono significativamente diverse, al livello di confidenza 95%, se e solo se gli intervalli dei due gruppi sono disgiunti. Per rendere leggibile il grafico, nella fig. 4.2 non abbiamo riportato gli intervalli di tutti i residui, ma solo di alcuni (comunque, essendo gli errori standard comparativi poco variabili, l'ampiezza degli intervalli è praticamente identica per tutti i residui).

- una variabile di risposta più informativa rispetto a quella dicotomica usata (esaminando le applicazioni dei modelli multilivello appare evidente che i migliori risultati, in termini di discriminazione fra i gruppi, si ottengono quando la variabile di risposta è continua).

4.4 Effetto del piano di campionamento e stime pesate

I piani di campionamento che prevedono probabilità di inclusione differenziate possono essere *informativi* e quindi avere degli effetti sugli stimatori dei parametri del modello di regressione (Skinner, 1989). Ciò accade quando, condizionatamente alle covariate inserite nel modello, le probabilità di inclusione dipendono dalla variabile di risposta. In tal caso gli stimatori dei parametri del modello di regressione possono essere *non consistenti*, per cui si rendono necessarie delle modifiche alle procedure di stima che tengano conto delle probabilità di inclusione. Un approccio generale al problema è quello noto come Pseudo-Massima Verosimiglianza (PML: Skinner, 1989), che consiste nel derivare gli stimatori non dalla verosimiglianza campionaria, ma da una stima della verosimiglianza dell'intera popolazione (*verosimiglianza censuaria*). In pratica, si tratta di stimare la log-verosimiglianza censuaria per mezzo di una somma pesata delle log-verosimiglianze individuali campionarie, dove i pesi sono i reciproci delle probabilità di inclusione. Lo stimatore che massimizza questa log-verosimiglianza censuaria è lo *stimatore PML* che, sotto condizioni generali, è consistente per il corrispondente parametro del modello.

Nel caso dei modelli multilivello la situazione è più complessa, poiché le probabilità di inclusione rilevanti non sono solo quelle relative alle unità elementari, ma anche quelle relative alle unità di livello superiore (gruppi); inoltre, le unità statistiche non sono indipendenti. Pertanto l'estensione del metodo PML ai modelli multilivello è tutt'altro che immediata. Pfeffermann *et al.* (1998), seguendo la logica dell'approccio PML, propongono una modifica dell'algoritmo IGLS per la stima del modello lineare a due livelli, indicata con l'acronimo PWIGLS (Probability-Weighted IGLS). Un punto particolarmente interessante è che i risultati forniti dall'algoritmo PWIGLS possono essere approssimati, generalmente in modo soddisfacente, da quelli che si ottengono applicando l'algoritmo IGLS standard ai dati opportunamente trasformati. Ciò consente di estendere questa metodologia ai modelli non lineari, come suggerito dallo stesso Goldstein nella guida del programma MLwiN.

In questo paragrafo, dopo una breve descrizione del piano di campionamento dell'indagine PSLD, effettueremo un'applicazione parziale del metodo PML con i soli pesi di livello 2, utilizzando l'algoritmo per la stima di massima verosimiglianza del programma MIXOR; successivamente mostreremo i risultati della stima pesata che si ottiene applicando l'algoritmo PQL2 ai dati opportunamente trasformati.

4.4.1 Il piano di campionamento dell'indagine PSLD

Il piano di campionamento dell'indagine PSLD, illustrato nel par. 3.1, è a due stadi, con stratificazione delle unità di primo stadio (le scuole) in base alla regione di appartenenza e alla tipologia della scuola. Le probabilità di inclusione delle scuole dipendono:

1. dallo strato di appartenenza. Infatti il numero di scuole selezionate nel singolo strato è per metà fisso e per metà proporzionale alla numerosità (in termini di maturi) dello strato, cosicché le scuole appartenenti a piccoli strati sono sovrarappresentate nel campione;

2. dalla dimensione in termini di maturi della scuola (il campionamento è proporzionale alla dimensione).

Le probabilità di inclusione dei maturi sono invece inversamente proporzionali alla dimensione della scuola, poiché il numero di maturi campione per scuola è costante. A tutto ciò vanno aggiunte le complicazioni derivanti dalle non risposte e dalle sostituzioni, sia per le scuole che per i maturi, rendendo il quadro estremamente complesso.

Il modello per la probabilità di occupazione che abbiamo sviluppato in questo capitolo include tra le covariate solo alcuni dei fattori che determinano le probabilità di inclusione. Infatti, gli strati sono individuati in modo parziale, poiché nel modello le tipologie di scuola sono state raggruppate, mentre le regioni sono rappresentate dal relativo tasso di disoccupazione giovanile. E soprattutto è assente, perché non inclusa nel file standard, la dimensione della scuola, che presumibilmente esercita un qualche effetto sulla probabilità di occupazione. Dunque nella nostra applicazione è possibile che il piano di campionamento sia di tipo informativo e quindi è interessante verificare quali risultati si ottengono con una procedura di stima pesata[18].

[18]La stima pesata non può essere adottata indiscriminatamente, perché comporta una perdita di efficienza. Nel caso di un piano di campionamento *non informativo* questa perdita di efficienza non è compensata da alcun guadagno in termini di riduzione della distorsione, per cui la stima pesata è sconsigliabile. Purtroppo nel caso dei modelli multilivello non esistono, al momento, strumenti diagnostici in grado di valutare se il piano di campionamento è informativo oppure no (cfr. Pfeffermann *et al.*, 1998).

Il file standard che abbiamo ottenuto dall'Istat non contiene le probabilità di inclusione ai vari stadi, ma solamente il *peso finale individuale*, calcolato come reciproco della probabilità di inclusione dell'unità (corretto per un fattore che consente di soddisfare, in ogni regione e tipo di scuola, la condizione di uguaglianza tra i totali noti della popolazione e le corrispondenti stime campionarie: cfr. Istat (1999b)). Indicando con p_{ij} il peso finale del maturo i della scuola j si ha

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} p_{ij} = N_{univ},$$

dove J è il numero totale di scuole incluse nel campione, n_j è il numero di maturi campione appartenenti alla j-ma scuola e N_{univ} è la numerosità dell'*universo* dei maturi.

Si noti che nell'analisi della probabilità di occupazione si è utilizzato un sottocampione selezionato secondo i criteri indicati nel par. 4.1. Per tale sottocampione i pesi andrebbero ricalcolati usando le probabilità di inclusione relative alla corrispondente sottopopolazione. Comunque, nel presente contesto, l'uso dei pesi originali p_{ij} in luogo dei pesi ricalcolati appare una ragionevole approssimazione.

Per la procedura di stima è opportuno scalare il peso delle unità di livello 1 (cfr. Pfeffermann *et al.*, 1998) in modo tale che il peso totale coincida con la dimensione campionaria N:

$$w_{ij} = \frac{N}{N_{univ}} \cdot p_{ij}.$$

Purtroppo i pesi di livello 2, che nel nostro caso sono sconosciuti, non possono essere determinati in modo univoco da quelli di livello 1. In mancanza di altre informazioni un possibile metodo consiste nell'attribuire alla j-ma scuola (unità di livello 2) un peso ottenuto come media artimetica dei pesi finali dei relativi maturi, scalata in modo tale che la somma dei pesi delle scuole coincida con il numero di scuole:

$$w_j = c \cdot \left(\frac{1}{n_j} \sum_{i=1}^{n_j} p_{ij} \right),$$

dove $c = J \cdot \left(\sum_{j=1}^{J} \frac{1}{n_j} \sum_{i=1}^{n_j} p_{ij} \right)^{-1}$. I pesi w_{ij} e w_j vengono detti *pesi standardizzati*.

4.4.2 Stima di Massima Verosimiglianza con pesi di livello 2

Come sottolineato in precedenza, l'applicazione del metodo PML ai modelli multilivello è tutt'altro che immediata. Tuttavia, qualora i pesi di interesse siano solo quelli di livello 2, il metodo PML può essere implementato con una semplice modifica dell'algoritmo di stima.

A questo proposito possiamo sfruttare la proposta di Hedeker&Gibbons (1994) relativa al trattamento di più unità di livello 2 con identico *pattern*, cioè con la stessa numerosità e gli stessi valori in tutte le variabili: si tratta di riscrivere la log-verosimiglianza (2.43) come

$$\log L = \sum_{k=1}^{K} s_k \log L(\alpha, \beta, \boldsymbol{\psi}, \boldsymbol{\nu} \mid \mathbf{y}_k),$$

dove k denota un certo pattern e s_k il numero di unità di livello 2 che presentano quel pattern. Il numero totale di pattern è ovviamente non superiore al numero di unità di livello 2, cioè $K \leq J$. Questo metodo è stato implementato nel programma MIXOR e usato dagli stessi Autori nel caso di pattern ripetuti. Tuttavia i numeri s_k non sono vincolati ad essere interi, per cui il metodo può essere usato anche per inserire i pesi di livello 2. In questo caso ogni unità ha un pattern distinto, per cui la log-verosimiglianza pesata può essere scritta come

$$\sum_{j=1}^{J} w_j \log L(\alpha, \beta, \boldsymbol{\psi}, \boldsymbol{\nu} \mid \mathbf{y}_j). \tag{4.1}$$

Naturalmente lo stimatore che massimizza la (4.1) non risente del fattore di scala dei pesi, che influenza invece la varianza dello stimatore: infatti la matrice di informazione attesa è data da (Hedeker&Gibbons, 1994, p. 937)

$$\sum_{j=1}^{J} w_j \left(L_j\right)^{-2} \left(\frac{\partial L_j}{\partial \boldsymbol{\Theta}}\right) \left(\frac{\partial L_j}{\partial \boldsymbol{\Theta}}\right)',$$

dove $\boldsymbol{\Theta}$ è il vettore che include tutti i parametri del modello e L_j sta per $L(\boldsymbol{\Theta} \mid \mathbf{y}_j)$. Per rendere comparabili i risultati delle stime con e senza pesi, in questa applicazione abbiamo moltiplicato i pesi di livello 2 per $N \cdot \left(\sum_{j=1}^{J} n_j w_j\right)^{-1}$.

Le stime di Massima Verosimiglianza con pesi di livello 2, riportate nella tab. 4.4, sono utili per valutare l'affidabilità della procedura approssimata che utilizza l'algoritmo PQL2 con i dati trasformati.

Tab. 4.4 - Risultati delle stime con e senza pesi relative al modello per la probabilità di svolgere un lavoro continuativo al momento dell'intervista

	Senza pesi (MLwiN)		**Pesi livelli 1 e 2 (MLwiN)**		**Pesi livello 2 (MLwiN)**		**Pesi livello 2 (MIXOR)**	
	Link	Logit	Link	Logit	Link	Logit	Link	Logit
	Distr. EC	Gauss.	Distr. EC	Gauss.	Distr. EC	Gauss.	Distr. EC	Gauss.
	Metodo	PQL2	Metodo	PQL2	Metodo	PQL2	Metodo	ML
	Var. EB	no	Var. EB	no	Var. EB	no	Pt. quad.	10
	Iterazioni	5	Iterazioni	6	Iterazioni	6	Iterazioni	24
	Devianza	-	Devianza	-	Devianza	-	Devianza	9747.4
Variabili	**Effetti fissi**		**Effetti fissi**		**Effetti fissi**		**Effetti fissi**	
	Stima	Err. Std.	Stima	Err. Std.	Stima	Err. Std.	Stima	Err. Std.
Intercetta	1.427	0.107	1.453	0.109	1.446	0.109	1.452	0.112
FEMM	-0.428	0.062	-0.437	0.063	-0.442	0.063	-0.443	0.062
LEVA-POI	-0.763	0.174	-0.789	0.171	-0.753	0.171	-0.756	0.165
VM36	-0.234	0.084	-0.260	0.085	-0.264	0.085	-0.265	0.091
VM37-42	-0.149	0.062	-0.130	0.064	-0.137	0.063	-0.137	0.068
VM50-59	0.169	0.081	0.214	0.085	0.196	0.084	0.197	0.086
VM60	0.503	0.188	0.612	0.187	0.605	0.184	0.608	0.202
PROF-D	0.133	0.047	0.136	0.048	0.131	0.048	0.131	0.049
PROF-E	0.454	0.097	0.701	0.099	0.682	0.100	0.684	0.097
LAVCINT	0.258	0.077	0.183	0.078	0.178	0.078	0.179	0.084
UNIV	-2.160	0.132	-2.074	0.126	-2.077	0.125	-2.086	0.127
INTUNIV-0	-0.914	0.084	-0.978	0.085	-0.982	0.085	-0.986	0.081
INTUNIV-5	0.495	0.133	0.658	0.134	0.654	0.133	0.657	0.128
CFP-ORA	-1.475	0.209	-1.577	0.214	-1.539	0.212	-1.545	0.228
CFP-CONC	-0.540	0.069	-0.462	0.071	-0.454	0.070	-0.455	0.070
ALTRIORA	-1.246	0.207	-1.245	0.226	-1.253	0.226	-1.258	0.229
ITP-AGR	-0.268	0.125	-0.395	0.184	-0.397	0.182	-0.399	0.207 *
ITP-IND	0.234	0.089	0.143	0.083 *	0.121	0.083 *	0.121	0.084 *
ITP-TUR	-0.266	0.114	-0.318	0.172 *	-0.297	0.171 *	-0.299	0.157 *
ITP-ALTR	-0.304	0.088	-0.318	0.095	-0.310	0.095	-0.312	0.098
ISTMAG	-0.512	0.113	-0.433	0.108	-0.448	0.107	-0.450	0.116
LICEI	-0.281	0.116	-0.287	0.109	-0.273	0.109	-0.275	0.106
ALTRIIST	-0.597	0.108	-0.876	0.139	-0.874	0.138	-0.878	0.136
DISOCC	-0.043	0.005	-0.045	0.005	-0.043	0.005	-0.043	0.005
DISOCC^2	0.00030	0.00013	0.00030	0.00014	0.00028	0.00014	0.00028	0.00015 *
±DISOCC	-0.042	0.011	-0.024	0.012 *	-0.022	0.012 *	-0.022	0.013 *
NORD	0.418	0.088	0.501	0.100	0.538	0.099	0.541	0.101
SUD	-0.319	0.105	-0.281	0.116	-0.281	0.115	-0.283	0.112
	Var. Eff. Cas.		**Var. Eff. Cas.**		**Var. Eff. Cas.**		**Err. Std. Eff. Cas.**	
	Stima	Err. Std.	Stima	Err. Std.	Stima	Err. Std.	Stima	Err. Std.
Intercetta	0.126	0.034	0.141	0.030	0.131	0.029	0.404	0.052
	Var. gruppo	0.126	Var. gruppo	0.141	Var. gruppo	0.131	Var. gruppo	0.164
	Corr. infragr	0.037	Corr. infragr	0.041	Corr. infragr	0.038	Corr. infragr	0.047

Nota: l'asterisco indica che il coefficiente non è statisticamente significativo al livello 95%

4.4.3 Stima pesata approssimata con PQL2

Estendendo il metodo proposto da Pfeffermann *et al.* (1998), così come suggerito dalla guida di MLwiN, la stima PQL2 pesata dei parametri di un modello logit a intercetta casuale può essere approssimata dalla stima PQL2 che si ottiene trasformando i dati nel seguente modo:

1. Il denominatore binomiale viene moltiplicato per i pesi finali standardizzati di livello 1, cosicché nell'espressione (2.46) si ha

$$\hat{z}_{ij,t} = \sqrt{\frac{1}{w_{ij}} \cdot \hat{h}_{ij,t} \cdot (1 - \hat{h}_{ij,t})}.$$

2. La variabile esplicativa dell'effetto casuale, cioè il vettore unitario, viene divisa per la radice quadrata dei pesi standardizzati di livello 2, per cui il predittore lineare diviene

$$\alpha + \beta x_{ij} + \frac{1}{\sqrt{w_j}} u_{0j}.$$

Come detto, questo metodo è solo un'approssimazione della stima pesata che si otterrebbe applicando integralmente il principio della Pseudo Massima Verosimiglianza. Le simulazioni di Pfeffermann *et al.* (1998) per il modello lineare indicano che in generale l'approssimazione è molto buona, salvo il caso della stima della varianza di livello 1 quando le numerosità dei gruppi dipendono dai pesi di livello 2 (questa circostanza non si verifica nella nostra applicazione, poiché il piano di campionamento prevede un numero costante di maturi per ogni scuola). L'estensione dei risultati dai modelli lineari a quelli a risposta binaria richiede molta cautela (sarebbero necessarie ulteriori simulazioni).

Tuttavia un modo per valutare l'affidabilità del nuovo metodo è quello di confrontare i risultati che si ottengono utilizzando i soli pesi di livello 2 [19] con i quelli forniti dall'applicazione diretta del principio PML che abbiamo presentato nel sottoparagrafo precedente. Come illustrato dalla tab. 4.4, l'unica differenza fra i due metodi risiede nel valore stimato della componente

[19] Ciò significa assumere che i pesi *condizionati* di livello 1, $w_{i\,|\,j}$, siano tutti uguali a 1. Tuttavia per la procedura di stima occorrono i pesi *finali*, w_{ij}, calcolati secondo la

$$w_{ij} = c \cdot w_{i\,|\,j} \cdot w_j,$$

dove $c = N \cdot \left(\sum_{j=1}^{J} \sum_{i=1}^{n_j} w_{i\,|\,j} \cdot w_j \right)^{-1}$. Si noti che quando $w_{i\,|\,j}$ è costante, si ha $w_{ij} \propto w_j$, cioè i pesi di livello 1 si ottengono semplicemente riproporzionando quelli di livello 2.

di varianza, che passa da 0.164 a 0.131. Gli scostamenti nei coefficienti di regressione sono una mera conseguenza della diversa stima della componente di varianza: in effetti i coefficienti forniti da PQL2, rispetto a quelli forniti da ML, sono attenuati in modo pressoché uniforme, di circa lo 0.4% (sul fenomeno dell'attenuazione cfr. par. 2.2.1). E' interessante notare come con la procedura di stima pesata la differenza fra PQL2 e ML si sia accentuata[20].

Una volta accertata l'affidabilità della procedura approssimata basata su PQL2, abbiamo utilizzato tale procedura per effettuare una stima pesata completa, cioè che include sia i pesi di livello 2 che quelli di livello 1 (cfr. tab. 4.4). L'aggiunta dei pesi di livello 1 cambia di poco i risultati, a testimonianza del fatto che i pesi rilevanti sono quelli di livello 2.

In generale, l'utilizzo dei pesi non modifica in modo sostanziale i risultati ottenuti con la procedura di stima standard, anche se per alcune variabili i coefficienti presentano differenze di rilievo. Come era prevedibile, l'impatto maggiore riguarda la stima della componente di varianza, che passa da 0.126 a 0.141, e le variabili di livello 2. In particolare è interessante notare che, mentre $\hat{\beta}_{\text{DISOCC}}$ è praticamente invariato, $\hat{\beta}_{\pm\text{DISOCC}}$ risulta quasi dimezzato e non più significativo. Ciò mette in dubbio il notevole effetto che la stima senza pesi attribuisce al trend del mercato del lavoro. Altre variabili di livello 2 che con la stima pesata divengono non significative sono quelle relative agli istituti industriali e turistici (ITP-IND, ITP-TUR). Fra le variabili di livello 1 va segnalato l'incremento di oltre il 50% di $\hat{\beta}_{\text{PROF-E}}$.

In conclusione, l'utilizzo delle stime pesate produce alcune modifiche apprezzabili, facendo pensare ad un piano di campionamento moderatamente informativo. Tuttavia, tali modifiche non sono tali da inficiare la validità dei risultati ottenuti con le stime standard.

[20]Ricordiamo la stima non pesata produceva $\hat{\sigma}_u^2 = 0.126$ con PQL2 e $\hat{\sigma}_u^2 = 0.134$ con ML.

Capitolo 5

Analisi dei tempi di ingresso al lavoro

La situazione occupazionale dei diplomati al momento dell'intervista è solo uno degli aspetti che caratterizzano la transizione scuola-lavoro. Un altro aspetto rilevante è rappresentato dal tempo di ingresso al lavoro, che nella presente applicazione deve intendersi come tempo intercorrente fra il conseguimento del diploma e l'inizio del primo lavoro *continuativo*[1]. In questo capitolo esamineremo l'andamento temporale degli ingressi al lavoro relativamente ai maturi dell'indagine sui Percorsi di Studio e Lavoro dei Diplomati (PSLD), prima con metodi descrittivi e poi per mezzo di modelli multilivello di sopravvivenza in tempo discreto.

Prima di procedere oltre è bene precisare che i tempi di ingresso non sono inclusi nel dataset originale, ma sono stati calcolati. Nel corso di questa operazione abbiamo dovuto affrontare alcuni problemi:

- Il mese di conseguimento del diploma non è noto. Tuttavia, visto che gli esami di maturità iniziano nella seconda metà di giugno, abbiamo ritenuto naturale assumere che si tratti per tutti del luglio 1995.

- Il mese dell'intervista non è noto. Ciò causa un problema nel calcolo del tempo censurato per coloro che al momento dell'intervista non avevano ancora ottenuto il primo lavoro continuativo. A tal fine, poiché le interviste sono state condotte tra settembre e dicembre 1998 e considerando che la distribuzione del numero di interviste realizzate per

[1] Mentre nell'analisi presentata nel cap. 4 si poteva scegliere se considerare tutti i lavori o solo quelli continuativi, nel presente caso la scelta è obbligata, poiché l'informazione sui tempi di inizio del lavoro svolto al momento dell'intervista è disponibile solo se si tratta di un lavoro continuativo.

giorno di lavoro presenta una forte asimmetria positiva (Istat, 1999), abbiamo deciso di riferire tutte le interviste al mese di ottobre 1998.

- Per coloro che dopo il diploma hanno iniziato e successivamente interrotto più lavori, il quesito sul tempo di inizio viene posto in riferimento al *primo* di tali lavori, anche se non continuativo. Pertanto risultano occultati gli eventuali lavori continuativi iniziati dopo la cessazione di un lavoro stagionale o occasionale e interrotti prima dell'intervista. Possiamo comunque ritenere che questo problema abbia un effetto trascurabile ai fini della nostra analisi.

- Alcuni maturi svolgono, al momento dell'intervista, un lavoro continuativo che hanno iniziato prima del conseguimento del diploma (si tratta di 743 individui, il 10.6% di quelli che hanno un lavoro continuativo). Tali individui sono stati ovviamente esclusi dal calcolo del tempo di ingresso.

Con riferimento ai mesi degli anni 1995-1998, i tempi di ingresso al lavoro sono indicati nel seguente prospetto:

	G	F	M	A	M	G	L	A	S	O	N	D
1995	-	-	-	-	-	-	1	2	3	4	5	6
1996	7	8	9	10	11	12	13	14	15	16	17	18
1997	19	20	21	22	23	24	25	26	27	28	29	30
1998	31	32	33	34	35	36	37	38	39	40	-	-

Il tempo relativo ai maturi che al momento dell'intervista non avevano ancora ottenuto un lavoro continuativo è uguale a 40 e risulta *censurato a destra*.

5.1 L'andamento temporale degli ingressi al lavoro

In questo paragrafo presentiamo alcune analisi preliminari sugli ingressi, finalizzate ad una prima conoscenza del fenomeno in vista dell'applicazione dei modelli di sopravvivenza. Pertanto l'attenzione sarà ristretta al sottoinsieme di individui che verrà impiegato per la stima di tali modelli, che coincide con quello utilizzato nel cap. 4 (cfr. par. 4.1). L'unica differenza è dovuta alla presenza di alcuni dati mancanti nei tempi di ingresso, che ha portato all'esclusione di altri 81 individui. Il sottocampione oggetto di studio in questo capitolo è dunque costituito da 9404 maturi, di cui 5938 (63.1%) hanno

ottenuto un lavoro continuativo prima dell'intervista e quindi per essi si può calcolare un effettivo tempo di ingresso, mentre 3466 non hanno ancora iniziato un lavoro continuativo e quindi il loro tempo, uguale a 40, è censurato a destra.

Il primo esame si basa sulla serie mensile degli ingressi al lavoro (fig. 5.1), che suggerisce le seguenti osservazioni:

- Esiste un picco di ingressi nel mese di settembre successivo al diploma, che può essere dovuto a varie situazioni: ad esempio, coloro che entrano in un'attività familiare già avviata, oppure coloro che avevano concluso accordi informali con il datore di lavoro e che attendevano il termine degli studi per iniziare l'attività lavorativa.

- Gli ingressi sono soggetti alla stagionalità, come risulta evidente dai mesi di agosto e settembre, che presentano l'uno valori particolarmente bassi e l'altro particolarmente alti.

- La serie degli ingressi non mostra nessun particolare trend.

Tuttavia ciò che interessa non sono gli ingressi in sè, ma l'andamento della probabilità di ingresso, cioè la *funzione di rischio*[2]. Una semplice stima *non parametrica* della funzione di rischio discreta $\lambda(t)$, dove $t = 1, \ldots, T$ è una successione di tempi, è data da

$$\hat{\lambda}(t) = \frac{e(t)}{r(t)}, \tag{5.1}$$

dove $e(t)$ è il numero di individui che sperimentano l'evento di interesse al tempo t, mentre $r(t)$ è la numerosità dell'*insieme di rischio al tempo* t, cioè il numero di individui che al tempo $t-1$ non avevano ancora sperimentato l'evento. Nel nostro caso t rappresenta i mesi, con $T = 40$, e l'evento di interesse è l'ingresso al lavoro.

La funzione di rischio stimata con questa procedura, riportata in fig. 5.2, consente di fare ulteriori osservazioni:

- Vengono confermati il picco di settembre 1995 e la stagionalità.

- L'andamento del rischio è sostanzialmente stazionario nel periodo compreso fra il mese 4 (ottobre 1995) e il mese 27 (settembre 1997), dopodiché manifesta un brusco cambiamento di livello verso l'alto, mantenuto fino al termine del periodo di osservazione.

[2] Cfr. par. 2.2.5. Nel presente capitolo, per non discostarci dalla terminologia classica dell'analisi di sopravvivenza, useremo il termine *rischio* per indicare la probabilità di ingresso al lavoro.

L'aumento della funzione di rischio negli ultimi mesi è in parte sorprendente, in quanto la teoria e i lavori empirici suggeriscono che la probabilità di occupazione decresce all'aumentare del tempo speso nella ricerca del lavoro. Tuttavia bisogna considerare che, nel nostro caso, il tempo è semplicemente il numero di mesi dal conseguimento del diploma, poiché, come sottolineato nel par. 4.1, non si hanno sufficienti informazioni per determinare l'andamento temporale della ricerca del lavoro. D'altra parte, il pattern mostrato dalla fig. 5.2 potrebbe derivare dal fatto che molti diplomati rimandano di uno-due anni l'inizio della ricerca del lavoro, ad esempio perché impegnati nel servizio militare o in corsi di formazione professionale o in corsi universitari successivamente interrotti.

La fig. 5.3 mostra le funzioni di rischio stimate separatamente per i 3108 maschi che hanno svolto il servizio militare durante il periodo di osservazione ("Maschi leva-in") e per i 958 maschi esonerati dal servizio ("Maschi leva-eso"). Il momento di inizio del servizio militare (che all'epoca aveva durata annuale) non è noto, ma, almeno per coloro che non si sono immatricolati all'università (circa l'80%), si colloca in gran parte nei primi 6-12 mesi. Ciò spiega perché coloro che hanno svolto il servizio si trovino inizialmente in una situazione di svantaggio, che si capovolge intorno al 24° mese, come si evince dal rapporto fra i rischi riportato in fig. 5.4. Tuttavia è interessante notare che i maschi esonerati, che pure hanno un andamento del rischio più stabile nel tempo, presentano comunque un aumento del livello del rischio a partire da ottobre 1997. La conclusione è che il servizio militare spiega solo in parte la ripresa del rischio nei mesi finali del periodo di osservazione.

La fig. 5.5 presenta le funzioni di rischio stimate separatamente per le 5087 femmine ("Femm") e per i 958 maschi esonerati dal servizio militare ("Maschi leva-eso"). Come si può osservare, anche le femmine manifestano un incremento del rischio nella fase finale, seppure attenuato. La fig. 5.6 mostra il rapporto fra i rischi delle femmine e dei maschi esonerati: tale rapporto, salvo i primi quattro mesi, è sostanzianzialmente costante nel tempo.

La situazione per le femmine e i maschi, distinti a seconda del servizio militare, è sintetizzata nel seguente prospetto[3]

[3] "TI-Mediana" è la mediana del tempo di ingresso, mentre "TI-Media C." è la media condizionata dei tempi di ingresso, calcolata sul sottoinsieme dei maturi che hanno effettivamente sperimentato l'ingresso (cioè, vengono esclusi i tempi censurati). *Leva-eso*, *Leva-in* e *Leva-poi* indicano, rispettivamente, i maschi che sono stati esonerati dal servizio militare, quelli che lo hanno svolto nel periodo di osservazione e quelli che lo hanno rinviato (o che lo stanno facendo al momento dell'intervista).

Fig. 5.1 - Serie mensile degli ingressi al lavoro

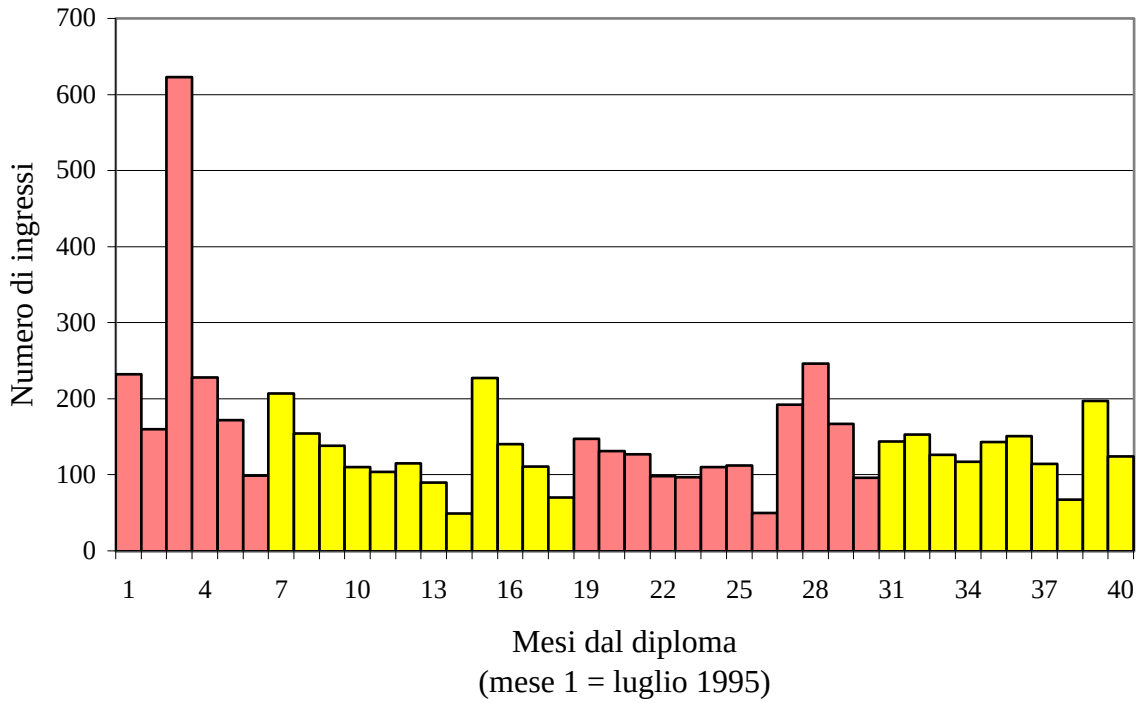


Fig. 5.2 - Funzione di rischio degli ingressi al lavoro

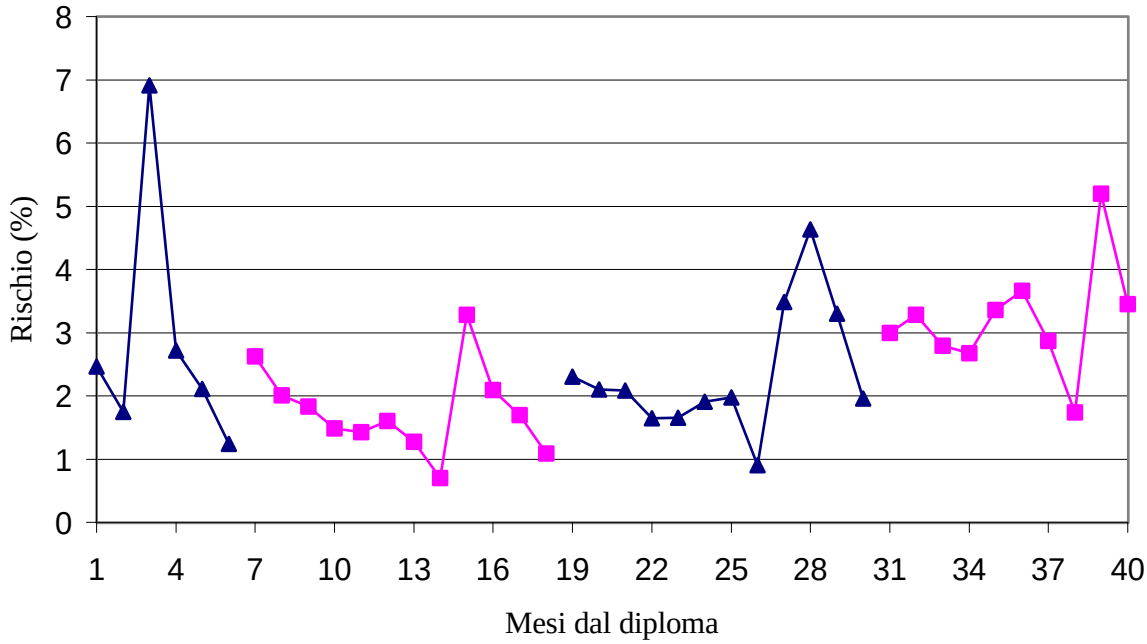


Fig. 5.3 - Funzioni di rischio dell'ingresso al lavoro

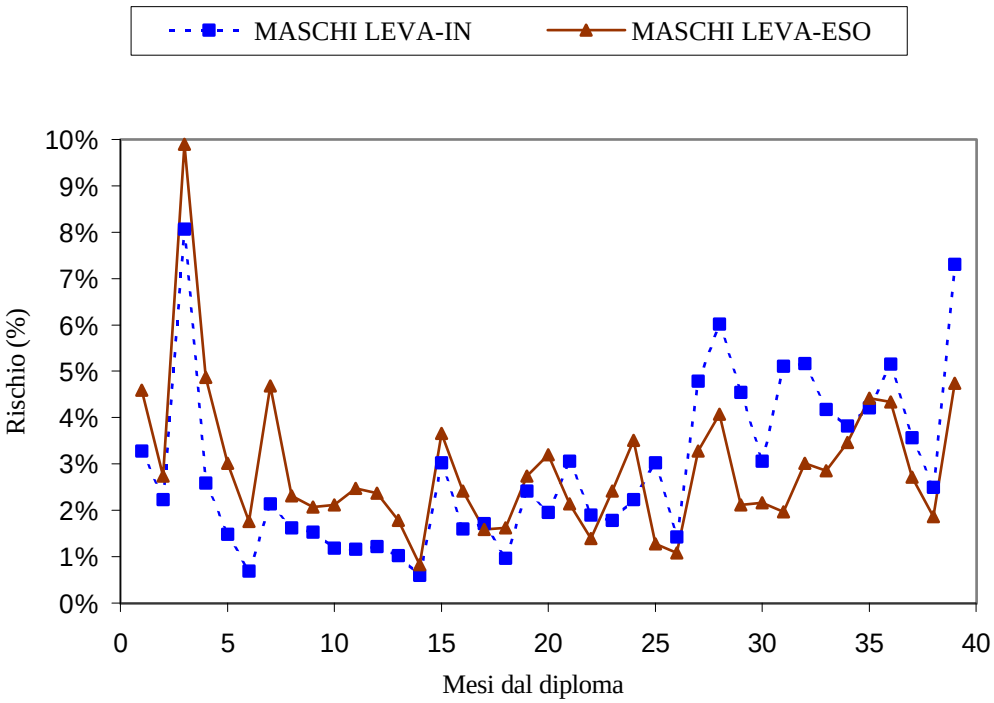


Fig. 5.4 - Rapporto fra i rischi
(MASCHI LEVA-IN) / (MASCHI LEVA-ESO)

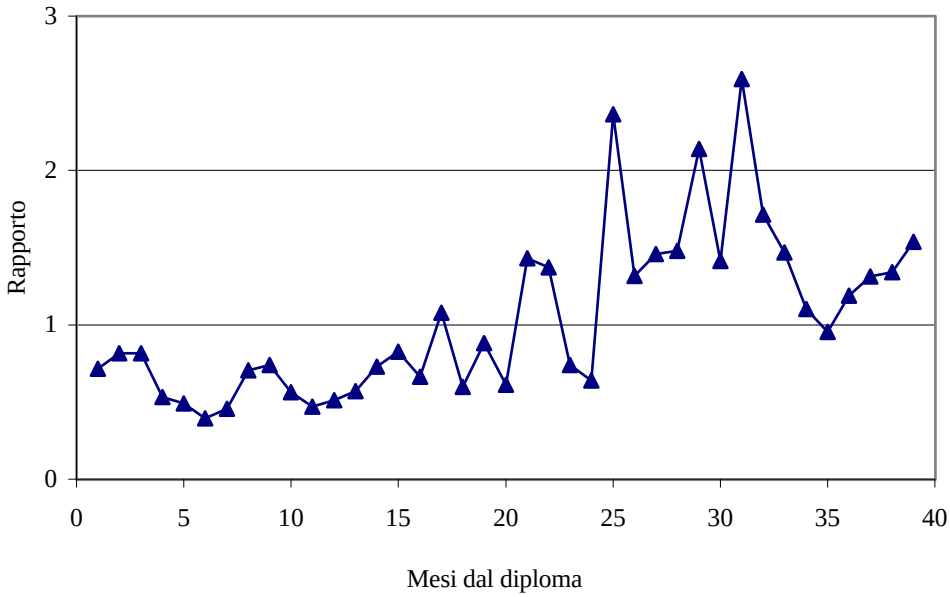


Fig. 5.5 - Funzioni di rischio dell'ingresso al lavoro

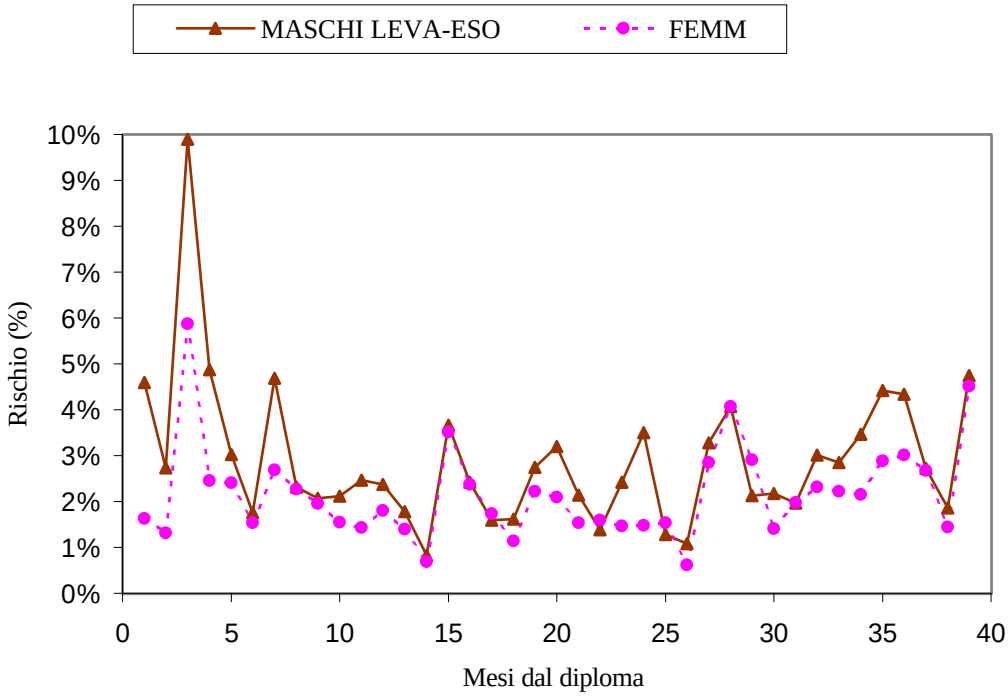
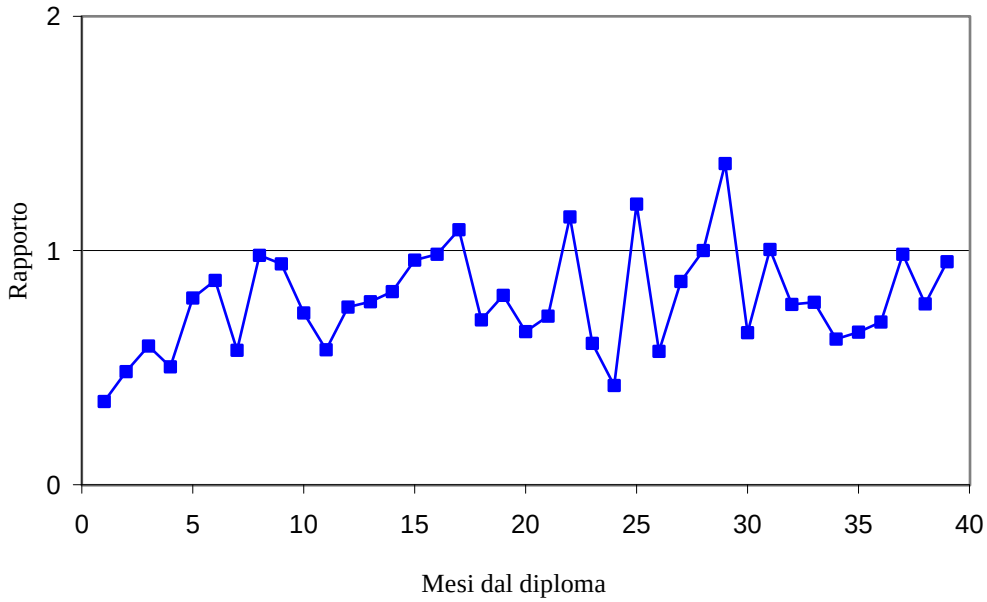


Fig. 5.6 - Rapporto fra i rischi
(FEMM) / (MASCHI LEVA-ESO)



	Numero	Ingressi	%Ingr.	TI-Mediana	TI-Media C.
Femmine	5087	2997	58.9	34	18.1
Maschi Leva-eso	958	667	69.6	24	15.4
Maschi Leva-in	3108	2177	70.0	29	19.7
Maschi Leva-poi	251	97	38.6	40	19.6
TOTALE	9404	5938	63.1	31	18.4

Si noti che la percentuale di ingressi e il tempo di ingresso medio (condizionato) rappresentano due diversi aspetti del fenomeno: ad esempio, i maschi che hanno svolto il servizio di leva nel periodo di osservazione hanno un tempo di ingresso medio nettamente maggiore dei maschi esonerati, ma la percentuale di ingressi è praticamente identica (ciò significa che, nel giro di un paio di anni, lo svantaggio del servizio militare viene annullato). A causa di queste differenze il modello per la probabilità di occupazione, che abbiamo usato nel cap. 4 [4], non è in grado di cogliere tutti gli aspetti rilevanti dell'inserimento professionale e si rende necessario l'utilizzo di un modello di sopravvivenza.

5.1.1 Il ruolo della domanda di lavoro

Prima di passare all'implementazione dei modelli si rende necessario un ultimo approfondimento. Infatti osservando i grafici delle funzioni di rischio rimane un dubbio: in che misura l'andamento degli ingressi dei maturi del nostro campione è legato a fattori congiunturali del mercato del lavoro? In particolare, l'innalzamento del rischio di ingresso a partire da ottobre 1997 dipende dal comportamento dei maturi del campione o piuttosto da un aumento della domanda di lavoro?

Per risolvere il problema è necessario disporre di una serie storica che rappresenti l'andamento della domanda di lavoro. Allo scopo abbiamo utilizzato una serie storica di ingressi al lavoro derivata dai panel trimestrali dell'Indagine sulle Forze di Lavoro dell'Istat. Tale serie, che è stata costruita ad hoc dall'Isfol su nostra richiesta[5], si riferisce agli ingressi trimestrali al *lavoro dipendente di tipo permanente* (inclusi contratti di tirocinio, apprendistato e Formazione e Lavoro), relativamente ai giovani in età 19-22 anni e in possesso di diploma, su tutto il territorio nazionale.

Per ragioni di confrontabilità si è reso necessario raggruppare in trimestri anche la serie degli ingressi dei maturi del campione. Peraltro questa opera-

[4] Si noti che la probabilità di occupazione al momento dell'intervista è diversa dalla probabilità di ingresso al lavoro nel periodo di osservazione. Tuttavia abbiamo verificato che lo studio dell'una o dell'altra probabilità porta alle stesse conclusioni.

[5] A questo proposito ringraziamo il dott. Marco Centra dell'Isfol, che si è interessato al problema ed ha realizzato le elaborazioni necessarie.

zione è necessaria anche per l'implementazione dei modelli di sopravvivenza, poiché l'utilizzo dei tempi in mesi comporta un carico computazionale non sostenibile con i software e i computer a nostra disposizione. Pertanto, d'ora in avanti il tempo sarà espresso in trimestri, da 1 a 13, come mostrato dal seguente prospetto:

	1° Trim.	2° Trim.	3° Trim.	4° Trim.
1995	-	-	1	2
1996	3	4	5	6
1997	7	8	9	10
1998	11	12	13	-

In questo modo restano esclusi i 124 maturi che hanno iniziato il lavoro dopo il settembre 1998, ai quali viene assegnato un tempo di 13 trimestri *con censura a destra*[6]. Le figg. 5.7 e 5.8 mostrano, rispettivamente, la serie trimestrale degli ingressi e la stima non parametrica della funzione di rischio: il raggruppamento in trimestri ha oscurato la stagionalità, mentre il pattern discusso in precedenza risulta ancora più evidente.

Per valutare il ruolo della domanda di lavoro, nella fig. 5.9 abbiamo posto a confronto la serie nazionale degli ingressi al lavoro dipendente con la stima non parametrica della funzione di rischio per le femmine del campione (abbiamo usato i dati delle femmine per evitare i problemi legati al servizio di leva). Il confronto richiede molta cautela, poiché la funzione di rischio riguarda una *coorte* di individui, mentre la serie nazionale degli ingressi si basa su *campioni trasversali*[7]. Possiamo comunque osservare che:

- La serie nazionale degli ingressi non mostra particolari tendenze, salvo un leggero innalzamento negli ultimi tre trimestri, mentre la funzione di rischio delle femmine presenta una fase discendente iniziale, una fase stazionaria centrale e una fase ascendente finale.

[6] Ciò equivale ad assumere che il periodo di osservazione termini per tutti il 30 settembre 1998. Così facendo si ignorano alcune informazioni (quelle relative agli ingressi successivi al 30 settembre), il cui impiego sarebbe stato comunque molto problematico. Infatti, poiché la maggior parte delle interviste è stata completata entro ottobre, il quarto trimestre del 1998 risulta incompleto. Inoltre per gli ingressi al lavoro registrati ad ottobre la definizione dell'*insieme di rischio* è molto problematica, poiché molti di coloro che erano a rischio all'inizio di ottobre sono stati intervistati nel corso del mese, uscendo così dall'insieme di rischio.

[7] Inoltre le due serie sono leggermente sfasate. Infatti le interviste dell'indagine sulle Forze di Lavoro si svolgono nei mesi di Gennaio, Aprile, Luglio e Ottobre, per cui, per fare un esempio, gli ingressi del 1° trimestre in realtà si riferiscono ad un periodo che, grosso modo, va dalla metà di Gennaio alla metà di Aprile.

Fig. 5.7 - Serie trimestrale degli ingressi al lavoro

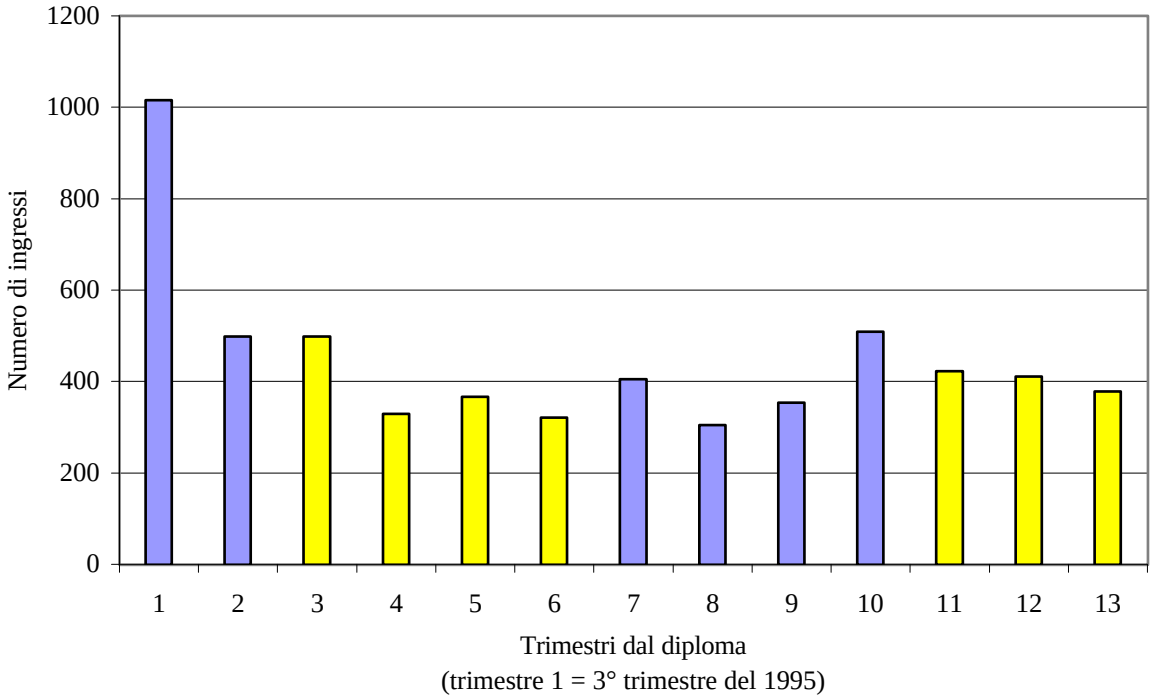


Fig. 5.8 - Funzione di rischio degli ingressi trimestrali al lavoro

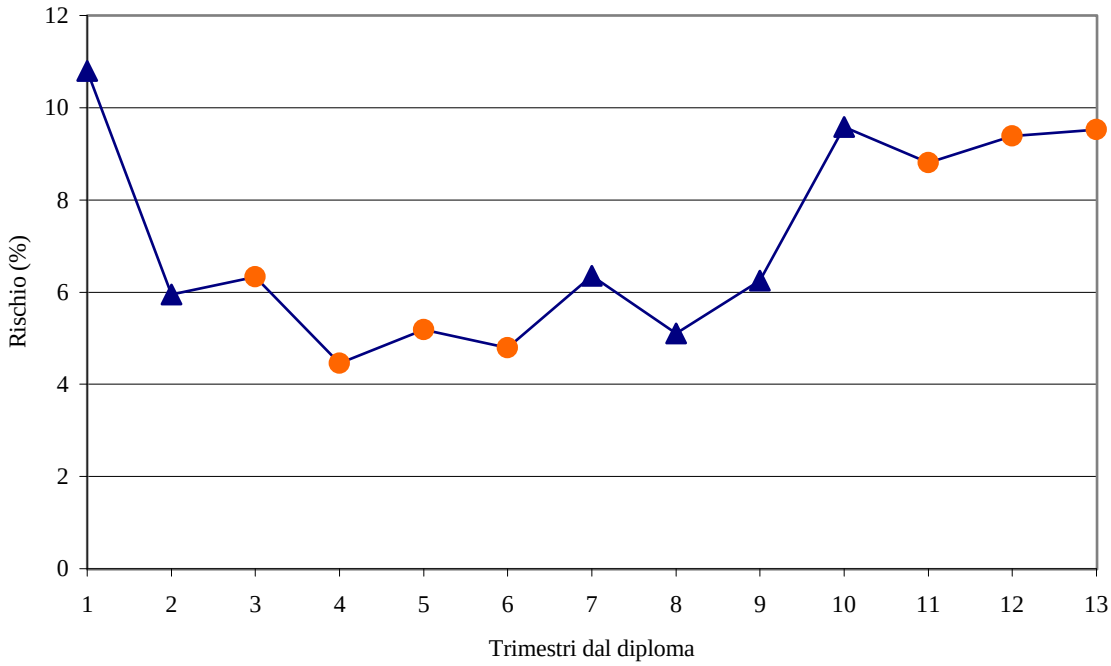
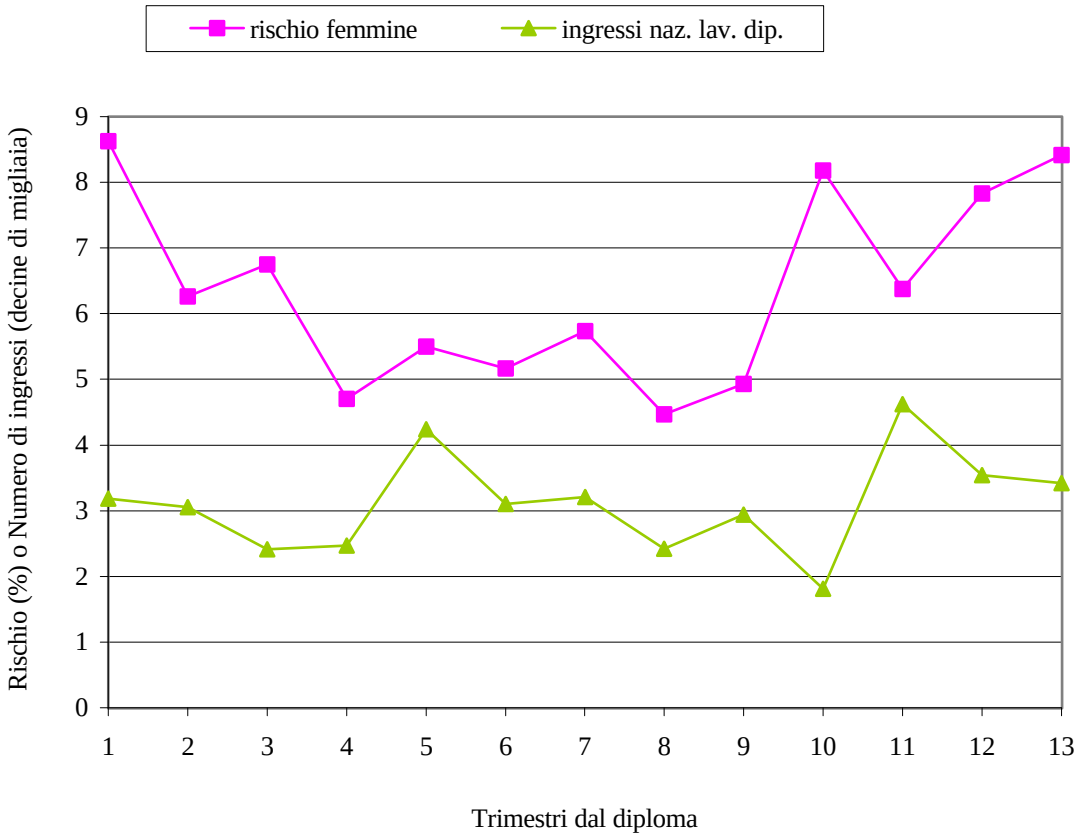


Fig. 5.9 - Funzione di rischio di ingresso al lavoro delle femmine (dati PSLD) e serie nazionale degli ingressi al lavoro dipendente (dati FdL)



- Le due serie hanno un andamento molto simile nella fase centrale (trimestri 4-9), mentre nei primi e negli ultimi trimestri sono addirittura discordanti, cioè hanno variazioni di segno opposto.

Queste osservazioni indicano che il pattern degli ingressi al lavoro dei maturi del nostro campione dipende più da dinamiche legate alla coorte di appartenenza che dall'andamento della domanda di lavoro. Alcune delle ipotesi che possono spiegare il pattern osservato sono le seguenti:

- il picco del primo trimestre, come già accennato, può essere originato dall'inserimento in attività familiari o dall'inizio di rapporti di lavoro definiti prima della conclusione degli studi;

- la ripresa degli ultimi trimestri è imputabile all'entrata sul mercato del lavoro di coloro che hanno svolto il servizio militare oppure che hanno terminato la formazione post-diploma (corsi di formazione professionale, corsi universitari interrotti)[8];

- infine, poiché i dati riguardano gli ingressi al lavoro *continuativo* e poiché è noto che l'inserimento professionale dei giovani spesso passa attraverso una serie di lavori provvisori, è plausibile che all'incremento degli ultimi trimestri contribuiscano anche i passaggi dal lavoro occasionale al lavoro continuativo.

Con un'estrema semplificazione l'inserimento lavorativo dei diplomati potrebbe essere descritto nel seguente modo: i diplomati migliori e quelli aiutati da familiari o conoscenti iniziano subito un lavoro continuativo, gli altri trovano notevoli difficoltà e quindi accettano lavori provvisori e/o continuano l'attività formativa, fin tanto che acquisiscono un bagaglio di esperienze e conoscenze che consente loro di ottenere un lavoro continuativo.

5.2 Specificazione di alcuni modelli di sopravvivenza multilivello in tempo discreto e analisi dei risultati delle stime

L'andamento degli ingressi al lavoro dei diplomati si presta ad essere studiato con un modello di sopravvivenza in tempo discreto, che nel nostro caso,

[8] A sostegno di questa tesi osserviamo che gli individui con LEVA-IN=1 costituiscono il 32.2% degli ingressi del periodo dal 2° al 9° trimestre e il 42.6% di quelli dal 10° trimestre in poi. Per quelli con INTUNIV-0=1 le percentuali sono, rispettivamente, 7.2% e 11.2%, mentre per quelli con CFP=1 le percentuali sono, rispettivamente, 15.7% e 22.0%.

vista la struttura gerarchica dei dati, dovrebbe includere anche degli effetti casuali. I modelli di sopravvivenza in tempo discreto, classici e multilivello, sono stati presentati nel par. 2.2.5. In questo paragrafo utilizzeremo due versioni discrete del modello a rischi proporzionali di Cox, la *grouped continuous* di McCullagh (1980) e la *continuation ratio* di Prentice&Gloeckler (1978), entrambe estese al caso di rischi non proporzionali.

5.2.1 Due versioni discrete del modello di Cox

Come discusso nel par. 2.2.5, il modello a rischi proporzionali di Cox, basato sul tempo continuo, può essere adattato anche al caso di tempi osservati ad intervalli discreti. Le due versioni che abbiamo presentato, che qui riportiamo con un effetto casuale sull'intercetta, sono la *grouped continuous* di McCullagh (1980), che indicheremo con la lettera M,

$$\log[-\log(1-F(t \mid \mathbf{x}_{ij}, u_{0j}))] = \alpha_t^{(M)} + \mathbf{x}'_{ij}\boldsymbol{\beta}^{(M)} + u_{0j}, \qquad t = 1, \ldots, T, \quad (5.2)$$

e la *continuation ratio* di Prentice&Gloeckler (1978), che indicheremo con la lettera P,

$$\log[-\log(1-\lambda(t \mid \mathbf{x}_{ijt}, u_{0j}))] = \alpha_t^{(P)} + \mathbf{x}'_{ijt}\boldsymbol{\beta}^{(P)} + u_{0j}, \qquad t = 1, \ldots, T, \quad (5.3)$$

dove $F(\cdot)$ è funzione di ripartizione della variabile aleatoria che rappresenta il tempo[9] e $\lambda(\cdot)$ è la funzione di rischio. I principali punti da sottolineare sono i seguenti (i dettagli sono forniti nel par. 2.2.5):

- Quando le covariate sono tutte *fisse* (cioè, non dipendono dal tempo) si ha $\boldsymbol{\beta}^{(M)} = \boldsymbol{\beta}^{(P)}$, interpretabile come il vettore degli effetti delle covariate nel modello di Cox latente; invece i parametri $\alpha_t^{(M)}$ e $\alpha_t^{(P)}$ sono distinti, in quanto rappresentano quantità diverse.

- Il modello P ammette anche covariate tempo-dipendenti; ciò consente di utilizzare una parametrizzazione più parsimoniosa del rischio di base, sostituendo gli $\alpha_t^{(P)}$ con un polinomio del tempo, e di modellare i rischi non proporzionali per mezzo di interazioni fra le covariate fisse ed il tempo. Invece nel modello M le covariate sono necessariamente fisse; il caso di rischi non proporzionali può comunque essere trattato inserendo delle interazioni fra i parametri $\alpha_t^{(M)}$ e alcune covariate.

[9] Si noti che $1-F(\cdot)$ è la funzione di sopravvivenza, che indichiamo con $S(\cdot)$. Pertanto, nel modello (5.2) il link è il *complementary log-log* se si fa riferimento a $F(\cdot)$ oppure il *log-log* se si fa riferimento a $S(\cdot)$.

- Per quanto riguarda gli aspetti computazionali, i parametri del modello M possono essere stimati con i metodi per i modelli multilivello a risposta ordinale (salvo una piccola modifica per la censura), mentre per il modello P si può espandere il dataset nel modo indicato nel par. 2.2.5 e usare i metodi per i modelli multilivello a risposta binaria. Nelle nostre applicazioni abbiamo usato il programma MIXOR per il modello M e il programma MLwiN per il modello P.

Per iniziare abbiamo usato esclusivamente covariate fisse, nel qual caso, come detto, i due modelli sono equivalenti. Naturalmente in questo modo si fa un'assunzione di rischi proporzionali, che nel nostro caso è poco giustificata (si veda, ad esempio, la fig. 5.6 relativa all'effetto del servizio militare). Tuttavia l'uso parallelo dei modelli M e P con covariate fisse è utile, oltre che per una prima selezione delle variabili, per una verifica del grado di accostamento delle stime che si ottengono con le due procedure.

Il sottoinsieme del campione PSLD che si è utilizzato è costituito da 9404 maturi (cfr. par. 5.1); l'espansione del dataset impiegata per la stima del modello P ha prodotto 83302 record. La procedura di selezione è partita dall'insieme delle variabili che compongono il modello per la probabilità di occupazione del cap. 4, con alcune piccole modifiche[10]. La tab. 5.1 mostra i risultati delle stime per il modello M e per il modello P:

- Le stime dei coefficienti di regressione e della componente di varianza sono praticamente identiche ed anche le stime degli errori standard sono molto vicine.

- Per quanto riguarda femmine, maschi e servizio militare, nel precedente modello per la probabilità di occupazione (par. 4.2) la categoria di base era costituita dall'unione di due categorie di maschi, quelli esonerati e quelli che avevano svolto il servizio prima dell'intervista. In questo modello per gli ingressi al lavoro le due categorie hanno effetti significativamente diversi, come c'era da attendersi sulla base dei tempi medi di ingresso (si veda il prospetto del paragrafo precedente). Adesso la categoria di base è quella dei maschi esonerati.

[10] Le modifiche derivano dal fatto che, mentre il modello per la probabilità di occupazione si basa sulla condizione ad una certa data (quella dell'intervista), il modello per gli ingressi fa riferimento ad un arco temporale (dal conseguimento del diploma all'intervista). Ciò ha suggerito di sostituire le variabili che indicano le attività formative svolte al momento dell'intervista (UNIV, CFP-ORA, ALTRIORA) con le corrispondenti variabili che indicano le attività formative iniziate nell'arco temporale (IMM, CFP, ALTRICOR). Si noti che in questo modo l'effetto per chi interrompe gli studi universitari è dato dalla somma algebrica dei coefficienti di IMM e di INTUNIV-0 (o INTUNIV-5). Infine si è eliminata, per ovvi motivi, la variabile LAVCINT (lavori continuativi interrotti).

Tab. 5.1 - Stime relative ai modelli *M* e *P* per i tempi di ingresso

Modello *M* Dataset ordinario (9404 record)	
Programma	MIXOR
Metodo	ML
Distr. EC	Gaussiana
Punti quad.	10
Iterazioni	12
Devianza	38549.5

Modello *P* Dataset esteso (83302 record)	
Programma	MLwiN
Metodo	PQL2
Distr. EC	Gaussiana
Variaz. EB	no
Iterazioni	7
Devianza	-

Effetti fissi (le definizioni delle variabili sono riportate nella tab. 3.1)

	Stima	Err. Std.	Stima	Err. Std.	Modello *P* /Modello *M* Stima	Err. Std.
Intercetta	-1.583	0.073	-1.580	0.073	1.00	1.00
FEMM	-0.345	0.045	-0.346	0.048	1.00	1.08
LEVA-IN	-0.257	0.044	-0.258	0.047	1.01	1.05
LEVA-POI	-0.458	0.116	-0.459	0.116	1.00	1.00
VM36	-0.144	0.047	-0.144	0.047	1.00	0.99
VM37-42	-0.114	0.033	-0.114	0.033	1.00	1.00
VM50-59	0.091	0.040	0.092	0.042	1.01	1.04
VM60	0.257	0.099	0.258	0.096	1.00	0.96
PROF-D	0.087	0.025	0.088	0.025	1.00	1.00
PROF-E	0.257	0.043	0.258	0.047	1.00	1.10
IMM	-1.666	0.105	-1.669	0.105	1.00	1.00
INTUNIV-0	0.956	0.119	0.957	0.114	1.00	0.96
INTUNIV-5	1.720	0.118	1.722	0.119	1.00	1.01
CFP	-0.440	0.042	-0.441	0.037	1.00	0.89
ALTRIORA	-0.666	0.112	-0.669	0.114	1.00	1.02
DISOCC	-0.026	0.003	-0.026	0.002	1.00	0.94
DISOCC^2	0.00019	0.00008	0.00019	0.00008	0.98	0.97
±DISOCC	-0.022	0.007	-0.022	0.006	1.00	0.88
ITP-IND	0.135	0.048	0.135	0.048	1.00	0.99
ITP-ALT2	-0.211	0.044	-0.211	0.042	1.00	0.97
LICEI	-0.256	0.072	-0.255	0.071	1.00	0.99
ALTRIIST2	-0.443	0.053	-0.444	0.053	1.00	1.01
NORD	0.130	0.051	0.131	0.047	1.01	0.92
SUD	-0.246	0.069	-0.246	0.066	1.00	0.96
	Soglie:		Variabili indicatrici dei trimestri:			
2	0.453	0.019	-0.559	0.055		
3	0.793	0.024	-0.451	0.055		
4	0.982	0.025	-0.776	0.064		
5	1.171	0.027	-0.589	0.062		
6	1.323	0.028	-0.640	0.065		
7	1.500	0.029	-0.318	0.060		
8	1.626	0.029	-0.506	0.066		
9	1.767	0.029	-0.261	0.063		
10	1.963	0.031	0.237	0.055		
11	2.122	0.031	0.209	0.059		
12	2.277	0.032	0.341	0.060		
13	2.422	0.033	0.422	0.062		

Errore std. dell'effetto casuale

Stima	Err. Std.
0.257	0.023

Varianza dell'effetto casuale

Stima	Err. Std.
0.068	0.011

Calcolo della correlazione infragruppo

	pi*pi/6	pi*pi/6
Varianza residua		
Varianza di gruppo	0.066	0.068
Correlaz. infragruppo	0.039	0.040

- La procedura di selezione delle variabili ha portato, rispetto al modello del cap. 4, all'accorpamento di alcune categorie relative alle scuole: la categoria residuale degli istituti tecnici e professionali (ITP-ALT2) adesso include anche quelli a indirizzo agrario e turistico, mentre la categoria residuale degli altri istituti (ALTRIIST2) include anche gli istituti magistrali.

Tuttavia, come detto, questo modello si basa su un'assunzione di rischi proporzionali, che nel nostro caso è poco giustificata; si rende pertanto necessaria l'estensione al caso di rischi non proporzionali. A questo punto, però, i modelli M e P prendono strade diverse, come vedremo nei prossimi due sottoparagrafi.

5.2.2 Il modello M

Nel modello M (McCullagh, 1980; Hedeker *et al.*, 1999) l'effetto non proporzionale di una covariata x_{ijh} può essere inserito per mezzo di interazioni γ_t con i parametri di soglia α_t, trasformando il predittore lineare dell'equazione (5.2) in

$$\alpha_t + \gamma_t x_{ijh} + \beta_h x_{ijh} + \mathbf{x}'_{ij(-h)}\boldsymbol{\beta}_{(-h)} + u_{0j}, \tag{5.4}$$

dove $(-h)$ indica l'insieme delle covariate esclusa la h-ma. Se x_{ijh} è una variabile dicotomica, come nella nostra applicazione, la (5.4) equivale ad assumere che l'insieme delle soglie sia $(\alpha_1, \ldots, \alpha_T)$ quando $x_{ijh} = 0$ e $(\alpha_1 + \gamma_1, \ldots, \alpha_T + \gamma_T)$ quando $x_{ijh} = 1$.

La tab. 5.2 riporta i risultati relativi al modello che include le interazioni della variabile relativa al servizio militare (LEVA-IN) con le soglie (le variabili sono le stesse del modello precedente). Possiamo notare che:

- L'aggiunta dei 12 termini di interazione[11] è statisticamente significativa (rispetto al modello precedente la devianza è inferiore di 172.8).
- I termini di interazione sono tutti negativi; in valore assoluto aumentano fino a raggiungere un massimo nel trimestre 6, per poi diminuire fino a valori prossimi allo zero (negli ultimi due trimestri le interazioni non sono significativamente diverse da zero). Questo pattern segnala che lo svolgimento del servizio militare ha un effetto negativo sul rischio di ingresso al lavoro, che raggiunge il massimo nel trimestre 6 (corrispondente a 18 mesi dal conseguimento del diploma) e che si annulla al termine del periodo di osservazione (39 mesi).

[11] Il termine di interazione relativo alla prima soglia è nullo per motivi di identificabilità del modello.

Tab. 5.2 - Stime relative al modello *M* con interazioni delle soglie

Programma	MIXOR
Distr. EC	Gaussiana
Punti quad.	10
Iterazioni	12
Devianza	38376.7

Effetti fissi

	Stima	Err. Std.	p-value
Intercetta	-1.654	0.077	0.000
FEMM	-0.336	0.046	0.000
LEVA-IN	-0.074	0.072	0.303
LEVA-POI	-0.451	0.119	0.000
VM36	-0.149	0.047	0.002
VM37-42	-0.113	0.034	0.001
VM50-59	0.091	0.041	0.027
VM60	0.248	0.101	0.014
PROF-D	0.088	0.025	0.000
PROF-E	0.259	0.043	0.000
IMM	-1.648	0.106	0.000
INTUNIV-0	0.928	0.120	0.000
INTUNIV-5	1.690	0.120	0.000
CFP	-0.442	0.043	0.000
ALTRICOR	-0.656	0.113	0.000
DISOCC	-0.026	0.003	0.000
DISOCC^2	0.00018	0.00008	0.022
±DISOCC	-0.022	0.007	0.002
ITP-IND	0.146	0.048	0.002
ITP-ALT2	-0.208	0.044	0.000
LICEI	-0.250	0.073	0.001
ALTRIIST2	-0.435	0.054	0.000
NORD	0.125	0.051	0.015
SUD	-0.254	0.069	0.000

Nota: le definizioni delle variabili sono riportate nella tab. 3.1.

Soglie

	Stima	Err. Std.	p-value
2	0.538	0.027	0.000
3	0.918	0.033	0.000
4	1.128	0.035	0.000
5	1.329	0.037	0.000
6	1.489	0.038	0.000
7	1.654	0.038	0.000
8	1.774	0.038	0.000
9	1.887	0.039	0.000
10	2.055	0.040	0.000
11	2.180	0.041	0.000
12	2.322	0.041	0.000
13	2.456	0.042	0.000

Interazioni delle soglie con LEVA-IN

	Stima	Err. Std.	p-value
2	-0.227	0.039	0.000
3	-0.342	0.049	0.000
4	-0.405	0.052	0.000
5	-0.437	0.055	0.000
6	-0.462	0.057	0.000
7	-0.424	0.059	0.000
8	-0.403	0.060	0.000
9	-0.323	0.060	0.000
10	-0.245	0.062	0.000
11	-0.147	0.063	0.021
12	-0.108	0.065	0.095
13	-0.074	0.065	0.254

Errore std. dell'effetto casuale

Estimate	Std. Error	p-value
0.255	0.023	0.000

Calcolo della correlazione infragruppo

Varianza residua	pi*pi/6
Varianza di gruppo	0.065
Correlaz. infragruppo	0.038

Fig. 5.10 - Funzione di rischio stimata per l'individuo di base al variare di LEVA-IN e FEMM - Modello *M*

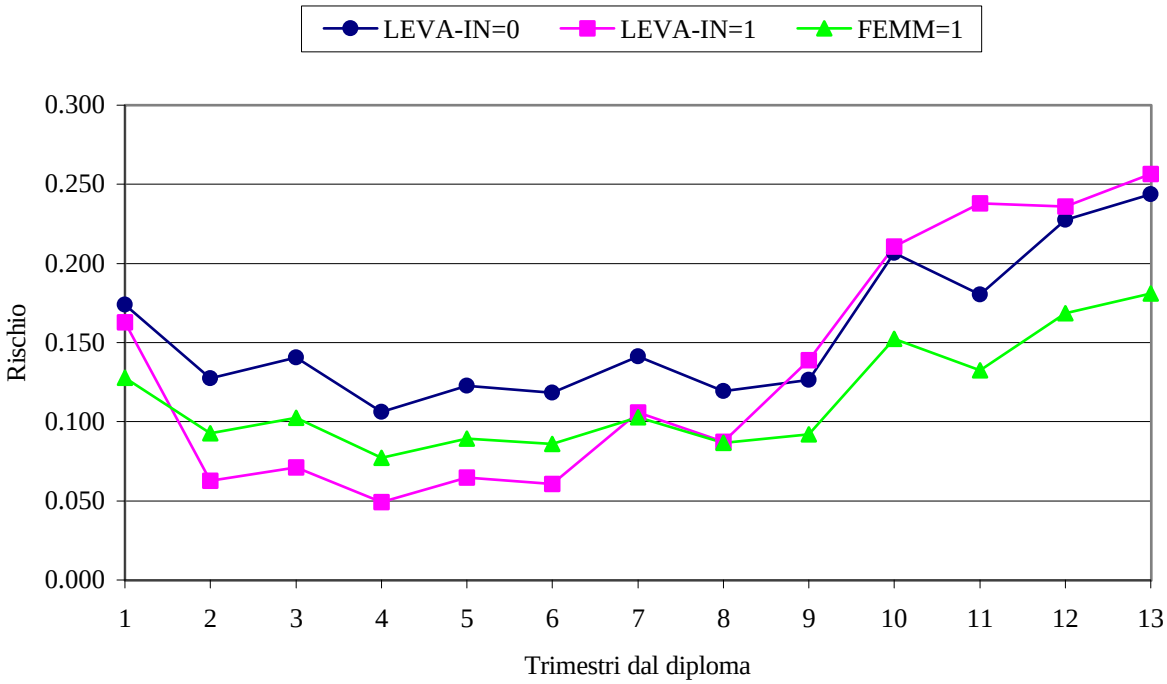
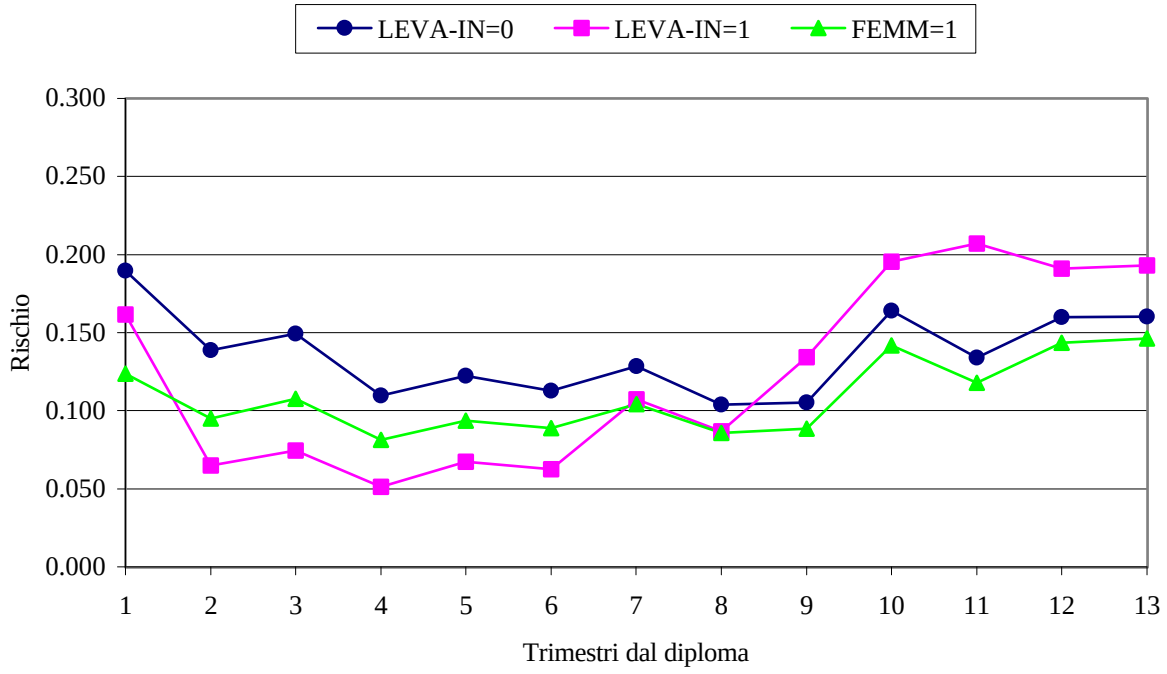


Fig. 5.11 - Funzione di rischio stimata per l'individuo di base al variare di LEVA-IN e FEMM - link LOGIT



- A parte LEVA-IN, le stime degli altri parametri cambiano di poco rispetto al modello senza interazioni.

La fig. 5.10 mostra la funzione di rischio stimata[12] per l'individuo di base[13] al variare di LEVA-IN e di FEMM. Per effetto delle interazioni, le funzioni con LEVA-IN=0 e LEVA-IN=1 si incrociano in più punti. Il pattern del rischio è simile a quello ottenuto con il metodo non parametrico (figg. 5.8 e 5.9); tuttavia le funzioni basate sul modello tendono ad essere più basse nei primi trimestri e più alte negli ultimi. Questo fatto potrebbe dipendere dall'asimmetria del link *complementary log-log.* Per verificare questa ipotesi abbiamo utilizzato un modello che differisce dal modello M per il tipo di link, il *logit*[14], e ricalcolato le funzioni di rischio (fig. 5.11), che in effetti risultano molto più simili a quelle delle figg. 5.8 e 5.9.

L'utilizzo del modello M non è stato ulteriormente approfondito, in quanto è evidente che l'estensione al caso di rischi non proporzionali per altre covariate, come quelle relative alle attività formative, comporta un'enorme crescita del numero di parametri. D'altra parte la specificazione non parametrica del rischio di base che caratterizza il modello M potrebbe essere non necessaria. Nel prossimo sottoparagrafo sfrutteremo le possibilità di parametrizzazione del rischio di base offerte dal modello P.

5.2.3 Il modello P

Il modello P (Prentice&Gloekler, 1978; Allison, 1982) consente l'inclusione di covariate tempo-dipendenti. Pertanto la specificazione non parametrica del rischio di base può essere sostituita da un polinomio in t, trasformando il predittore lineare dell'equazione (5.3) in

$$\sum_{r=0}^{R} \delta_r t^r + \mathbf{x}'_{ijt}\boldsymbol{\beta} + u_{0j}.$$

[12] La funzione di rischio è stata ottenuta attraverso i seguenti passi: 1. calcolo dei valori assunti dal predittore lineare nei 13 trimestri; 2. calcolo della funzione di sopravvivenza $S(t)$ (invertendo il link *log-log*); 3. calcolo della funzione di rischio per mezzo della relazione $\lambda(t) = 1 - \frac{S(t)}{S(t-1)}$.

[13] Ricordiamo che l'individuo di base è quell'ipotetico individuo che ha covariate ed effetti casuali identicamente nulli. Il profilo dell'individuo di base coincide con quello riportato nella tab. 4.3, salvo la condizione rispetto agli obblighi di leva (nel modello attuale la categoria di base è costituita da coloro che sono stati esonerati).

[14] Sostituendo il link cambiano le proprietà del modello (cfr. Hedeker *et al.*, 1999). Contrariamente a quanto avviene di solito, nella presente applicazione la sostituzione del link ha effetti apprezzabili sull'adattamento del modello, poiché la devianza aumenta di 60.0.

Inoltre il caso di rischi non proporzionali per una certa covariata fissa x^*_{ij} può essere trattato inserendo nel vettore $\mathbf{x}_{ijt}$ le interazioni di tale covariata con il tempo, cioè $\left(x^*_{ij} \cdot t\right)$, $\left(x^*_{ij} \cdot t^2\right)$, ... fino all'interazione di ordine R.

La selezione del modello prende le mosse dai risultati ottenuti con sole covariate fisse (tab. 5.1) e consiste fondamentalmente nella scelta dell'ordine del polinomio, R, nell'individuazione delle interazioni delle covariate fisse con il tempo e nell'inserimento dell'unica covariata tempo-dipendente, cioè la serie nazionale degli ingressi al lavoro dipendente (cfr. par. 5.1.1). Per quanto riguarda i risultati delle stime, riportati nella tab. 5.3, osserviamo quanto segue[15]:

- Per la modellizzazione del tempo è risultato sufficiente un polinomio di ordine 3, con l'aggiunta di un termine apposito per il primo trimestre (l'opportunità di tenere distinto il primo trimestre nasce dalle considerazioni riportate nel par. 5.1.1). Le covariate che interagiscono con il tempo sono cinque, ognuna con un proprio numero di interazioni. Il modello risulta dunque molto ricco e al tempo stesso relativamente parsimonioso, in quanto ha 10 parametri in meno del modello M con l'interazione per la sola LEVA-IN (tab. 5.2).

- Confrontando le tabb. 5.2 e 5.3 si può notare che i coefficienti delle covariate che non interagiscono con il tempo presentano differenze trascurabili, così come la stima della componente di varianza. Pertanto i due modelli differiscono essenzialmente nella rappresentazione dell'andamento temporale del rischio.

- La serie nazionale degli ingressi al lavoro dipendente, inserita come covariata tempo-dipendente, ha un coefficiente non statisticamente significativo. Ciò conferma le conclusioni del par. 5.1.1.

- La fig. 5.12 mostra la funzione di rischio stimata per l'individuo di base[16] al variare di LEVA-IN e di FEMM. Come prevedibile, coloro che svolgono il servizio di leva nel periodo di osservazione presentano un

[15]Nella tab. 5.3 t indica il tempo in trimestri e $t\hat{}r$ le sue potenze; la notazione NOMEVARIABILE*($t\hat{}r$) indica le interazioni delle covariate fisse con il tempo. E' bene precisare che t non assume i valori $\{1, \ldots, 13\}$, ma $\{0, \ldots, 12\}$, per consentire l'interpretazione dell'intercetta come valore relativo al primo trimestre. Tuttavia nel modello finale è presente anche una variabile indicatrice del primo trimestre, TRIM1, per cui l'intercetta non è direttamente interpretabile nemmeno per il primo trimestre.

[16]Il profilo dell'individuo di base è analogo a quello delineato nel par. 4.2.1, con due eccezioni: 1) rispetto agli obblighi di leva, si tratta di un individuo esonerato; 2) i lavori continuativi interrotti non vengono presi in considerazione.

Tab. 5.3 - Stime relative al modello *P* con covariate tempo-dipendenti

Programma	MLwiN
Metodo	PQL2
Distr. EC	Gaussiana
Variaz. EB	no
Iterazioni	7
Devianza	-

Dataset esteso (83302 record)

Variabili che indicano il tempo o che interagiscono con il tempo

	Stima	Err. Std.
Intercetta	-1.627	0.108
TRIM1	0.242	0.083
t	-0.160	0.056
t^2	0.016	0.010
t^3	-0.00019	0.00055
FEMM	-0.479	0.068
FEMM*t	0.030	0.011
LEVA-IN	-0.265	0.079
LEVA-IN*t	-0.497	0.058
LEVA-IN*(t^2)	0.114	0.012
LEVA-IN*(t^3)	-0.006	0.001
INTUNIV-5	1.335	0.148
INTUNIV-5*t	0.061	0.015
INTUNIV-0	-0.569	0.246
INTUNIV-0*t	0.707	0.143
INTUNIV-0*(t^2)	-0.096	0.026
INTUNIV-0*(t^3)	0.004	0.001
CFP	-1.097	0.095
CFP*t	0.162	0.036
CFP*(t^2)	-0.006	0.003

Variabili con effetti costanti nel tempo

	Stima	Err. Std.
LEVA-POI	-0.419	0.115
VM36	-0.149	0.047
VM37-42	-0.114	0.033
VM50-59	0.085	0.042
VM60	0.245	0.095
PROF-D	0.089	0.025
PROF-E	0.269	0.047
IMM	-1.622	0.105
ALTRICOR	-0.640	0.114
DISOCC	-0.026	0.002
DISOCC^2	0.00019	0.00008
±DISOCC	-0.022	0.006
ITP-IND	0.144	0.047
ITP-ALT2	-0.204	0.041
LICEI	-0.251	0.071
ALTRIIST2	-0.432	0.052
NORD	0.124	0.046
SUD	-0.252	0.065

Varianza dell'effetto casuale

Stima	Err. Std.
0.059	0.011

Calcolo della correlazione infragruppo

Varianza residua	pi*pi/6
Varianza di gruppo	0.059
Correlaz. infragruppo	0.035

Nota: le definizioni delle variabili sono riportate nella tab. 3.1.

Fig. 5.12 - Funzione di rischio stimata per l'individuo di base al variare di LEVA-IN e FEMM - Modello *P*

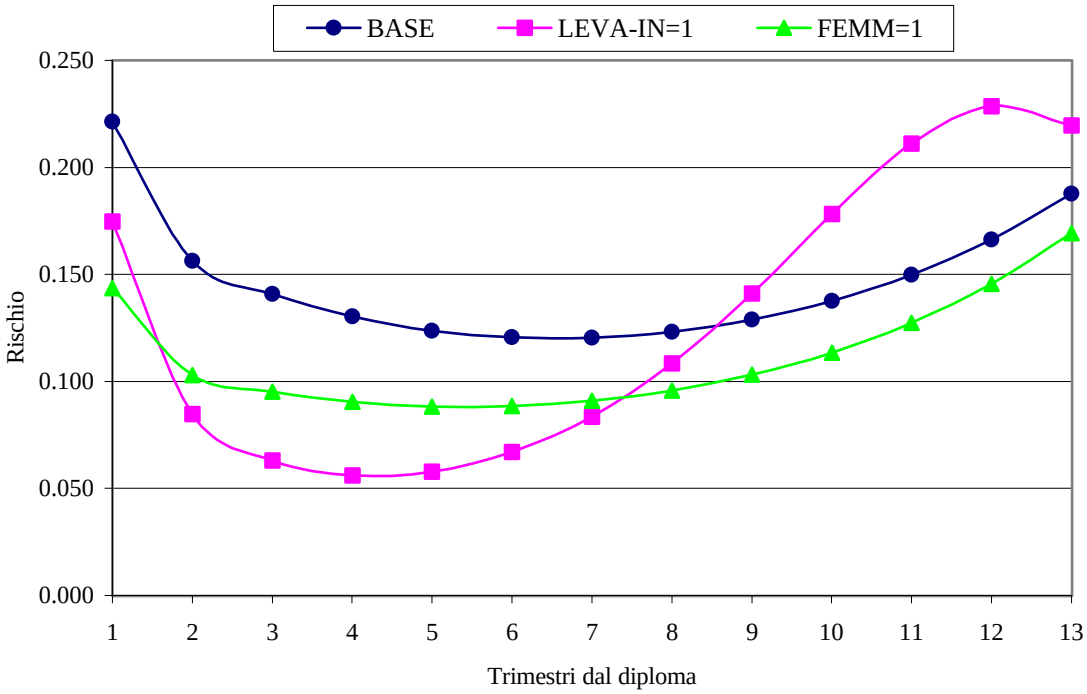


Fig. 5.13 - Funzione di rischio stimata per l'individuo di base al variare di INTUNIV-0 e CFP - Modello *P*

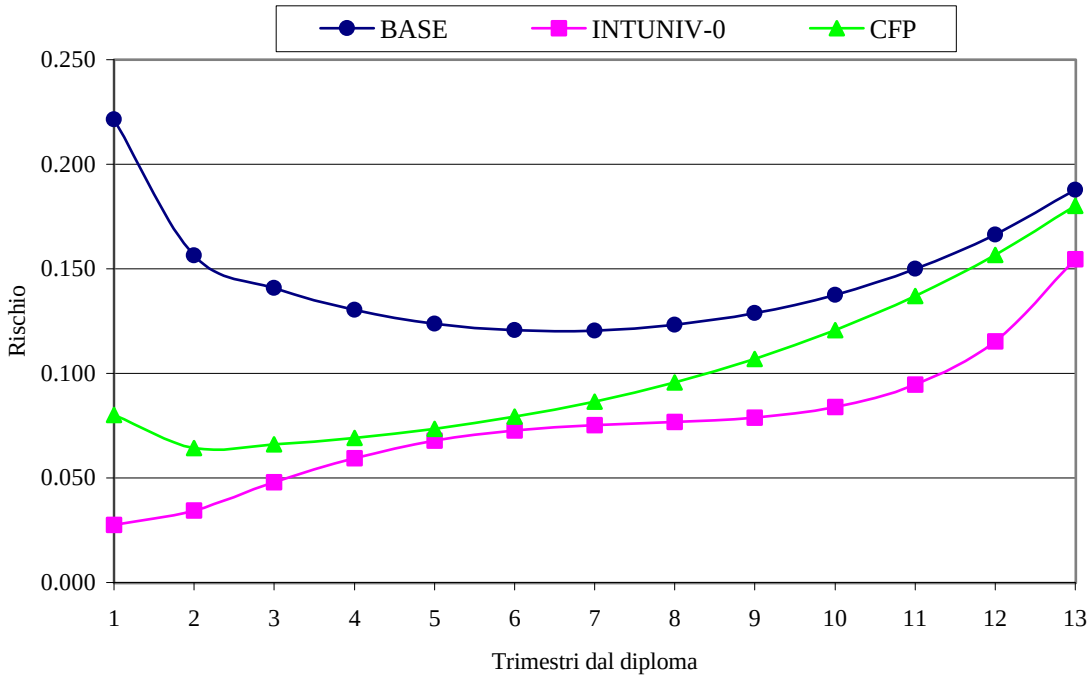
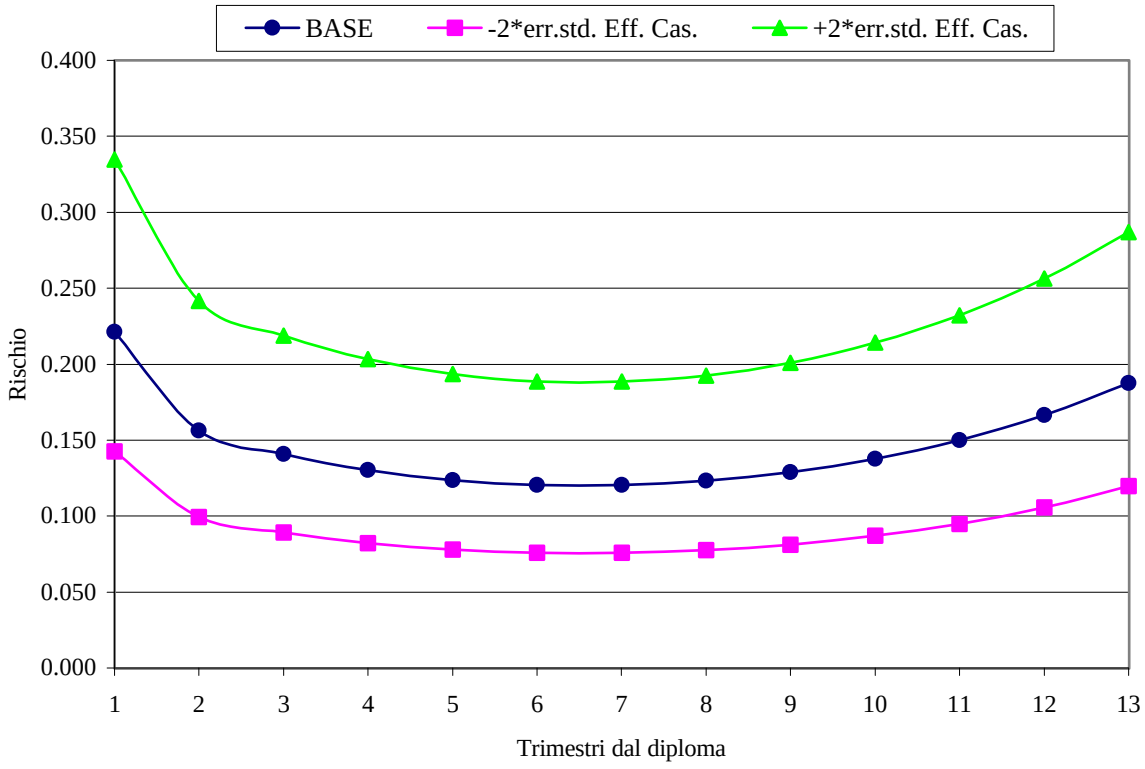


Fig. 5.14 - Funzione di rischio stimata per l'individuo di base al variare della scuola frequentata - Modello *P*



pattern tutto particolare. Inoltre è interessante notare che il divario fra i sessi si riduce con il passare dei mesi.

- La fig. 5.13 mostra la funzione di rischio stimata per l'individuo di base al variare di INTUNIV-0 e di CFP. Le funzioni sono difficilmente interpretabili, poiché il periodo di frequenza dell'università o del corso di formazione varia molto per momento d'inizio e per durata. Tuttavia forniscono una spiegazione intuitiva del perché la funzione di rischio dell'individuo di base presenti un andamento diverso rispetto a quello mostrato nella fig. 5.10 per il modello M, superiore nei trimestri iniziali e inferiore in quelli finali[17].

- Il ruolo giocato dalle singole scuole può essere valutato aggiungendo e sottraendo al predittore lineare dell'individuo di base due volte l'errore standard stimato dell'effetto casuale, ottenendo così il grafico della fig. 5.14, da cui risulta evidente che, *ceteris paribus*, frequentare un istituto piuttosto che un altro fa la sua bella differenza (tuttavia per l'interpretazione dell'effetto casuale in termini di efficacia valgono le stesse avvertenze del par. 4.3). Nel modello finale è presente un unico effetto casuale, per cui le differenze fra le scuole producono effetti costanti nel tempo. Per verificare la possibilità di effetti variabili nel tempo abbiamo testato, con esito negativo, l'ipotesi che la variabile tempo t abbia un coefficiente casuale[18].

Un'ulteriore possibilità offerta dal modello P riguarda l'inserimento di una componente di eterogeneità individuale non osservabile, anche se nella nostra applicazione, dato l'elevato numero di covariate, è improbabile che siano presenti fattori non osservabili di grande rilievo. Come descritto nel par. 2.2.5, l'inclusione dell'eterogeneità individuale non osservabile si ottiene semplicemente aggiungendo un livello gerarchico: nel nostro caso il modello diviene a 3 livelli, con il terzo livello costituito dalle scuole, il secondo dai maturi e il primo dalle osservazioni trimestrali sui singoli maturi. Come nel caso della stima in presenza di un coefficiente casuale per il tempo, anche in questa situazione i limiti di calcolo ci hanno costretto ad effettuare la stima su un sottocampione di 41065 record, ottenuto eliminando casualmente la metà delle scuole. Il risultato comunque è che l'algoritmo PQL2 non converge:

[17] Infatti il modello M a cui si riferisce la fig. 5.10 include solo le interazioni relative a LEVA-IN, per cui l'effetto non proporzionale di variabili come INTUNIV-0 e CFP altera la stima del rischio di base.

[18] Il realtà, per ovviare ai limiti di calcolo in cui ci siamo imbattuti, il test è stato eseguito su un sottocampione di 41065 record, ottenuto eliminando casualmente la metà delle scuole.

ciò significa, verosimilmente, che il modello con eterogeneità individuale non osservabile è troppo complesso rispetto ai dati a disposizione[19].

[19]Uno dei motivi che rende particolarmente difficile la stima va ricercato nel fatto che all'interno delle unità di livello 2 (che corrispondono ai maturi) sono possibili due soli pattern: $(0, 0, \ldots, 0, 1)$ se l'individuo ha iniziato un lavoro continuativo, $(0, 0, \ldots, 0)$ altrimenti.

Capitolo 6

Analisi della probabilità di immatricolazione all'università

La prosecuzione degli studi dopo il diploma è un fenomeno estremamente interessante, nel quale i fattori di contesto svolgono un ruolo centrale (cfr. par. 1.1). Infatti la decisione di proseguire la formazione dipende non solo dalle inclinazioni personali, ma anche dal background familiare, dall'ambiente scolastico frequentato, dalla vicinanza o meno di una sede universitaria[1] e dalle caratteristiche socio-economiche dell'area di residenza. In particolare, la teoria economica suggerisce che le scelte formative dipendano dalle prospettive occupazionali e reddituali immediate e future (cfr. par. 1.2.3).

In questo capitolo presenteremo un'analisi volta ad individuare e valutare i fattori che determinano l'immatricolazione all'università nei tre anni successivi al conseguimento del diploma. Poiché l'indagine sui Percorsi di Studio e Lavoro dei Diplomati (PSLD) presenta, per le variabili che in questo contesto sono rilevanti, un numero consistente di non risposte, inizieremo il capitolo illustrando i criteri che abbiamo seguito per il trattamento dei dati mancanti. Nel successivo paragrafo mostreremo i risultati dell'analisi e, infine, nel paragrafo 6.3, evidenzieremo analogie e differenze della presente analisi rispetto a quella relativa alla probabilità di occupazione (capitolo 4).

[1] La distanza (in termini di tempo di percorrenza) fra il luogo di residenza del diplomato e le sedi universitarie, e la tipologia di corsi attivati in tali sedi, esercitano certamente un importante effetto sulla decisione di immatricolazione. Tuttavia la portata di questo effetto non può essere indagata nella presente ricerca, in quanto i dati a nostra disposizione non contengono il comune di residenza dei diplomati, ma solo la regione.

6.1 Trattamento dei dati mancanti

Lo studio dell'immatricolazione all'università, a differenza di quello dell'occupazione, non richiede l'individuazione di una particolare sottopopolazione di interesse, poiché in questo caso tutti i diplomati sono potenziali candidati all'iscrizione a corsi universitari (salvo alcune limitazioni relative ai diplomi ottenuti in 4 anni: cfr. app. A). Tuttavia, la presente analisi, a differenza delle precedenti, comporta l'inclusione di diverse covariate con un numero non trascurabile di dati mancanti. Per questo motivo non è opportuno eliminare i record incompleti, ma è certamente preferibile sostituire i dati mancanti con valori adeguatamente imputati.

Allo scopo abbiamo utilizzato una procedura di *imputazione casuale condizionata singola* (Little, 1998), che consiste nel sostituire il dato mancante nella variabile h dell'individuo i, X_{ih}, con il valore estratto da una distribuzione di probabilità $f\left(X_{ih} \,|\, \mathbf{Z}_{ih} = \mathbf{z}_{ih}\right)$, dove $\mathbf{Z}_{ih}$ è un vettore di variabili scelte in modo opportuno e $\mathbf{z}_{ih}$ è il corrispondente vettore delle realizzazioni. I momenti della distribuzione f vengono stimati sulla base delle relative statistiche campionarie calcolate sui dati osservati. Nel nostro caso le variabili con dati mancanti sono tutte categoriche, per cui si tratta di generare valori da distribuzioni di tipo multinomiale. Ricordiamo che l'imputazione casuale si basa sull'ipotesi *MAR (Missing At Random)*, cioè che la presenza o l'assenza del dato non dipenda dai valori assunti dalle variabili non osservate.

L'imputazione di tipo casuale è senz'altro preferibile all'imputazione diretta del valore atteso, che produce stime sostanzialmente distorte delle misure di variabilità. Tuttavia l'imputazione di un singolo valore comporta in ogni caso una sottostima delle misure di variabilità, poiché i valori imputati vengono trattati allo stesso modo dei valori noti, trascurando l'incertezza relativa all'imputazione. Questo problema può essere risolto con una procedura di *imputazione multipla* (Rubin, 1987). Comunque nella presente applicazione, data la percentuale non eccessivamente elevata di dati mancanti, riteniamo che una semplice procedura di imputazione casuale singola produca risultati sufficientemente attendibili.

Consideriamo adesso le variabili con dati mancanti dell'indagine PSLD che sono rilevanti nello studio dell'immatricolazione. La variabile che presenta il maggior numero di dati mancanti è quella relativa al giudizio della scuola media inferiore (1236 su 18843). Questo dato viene fornito direttamente dall'istituto in cui l'individuo ha ottenuto la maturità, per cui è lecito attendersi che l'assenza del dato non dipenda da caratteristiche individuali, ma piuttosto da caratteristiche dell'istituto (ad esempio, l'assenza del dato è molto più frequente negli istituti Professionali e Magistrali). Le analisi preliminari hanno suggerito di inserire il giudizio di scuola media inferiore

come variabile binaria (GIUD4) che assume il valore 1 quando il giudizio è massimo (corrispondente a "ottimo"). Inoltre l'esame delle distribuzioni doppie ha consigliato di basare l'imputazione sulle proporzioni di individui con giudizio "ottimo" condizionatamente al sesso del diplomato (FEMM) e al titolo di studio del padre (TITPA), secondo il seguente prospetto:

		TITPA	
		0	1
FEMM	0	0.094	0.214
	1	0.155	0.315

La procedura di imputazione è stata ripetuta tre volte, in modo da poter effettuare un'analisi di sensibilità.

Le altre variabili che presentano dati mancanti sono quelle relative al titolo e alla professione del padre e della madre. Questo problema interessa 878 individui, di cui 300 con un dato mancante in due o più di tali variabili. In questo caso, vista la forte relazione esistente fra titolo di studio e professione, si è proceduto nel seguente modo:

- sono stati eliminati 123 individui per i quali non sono noti né il titolo di studio né la professione di uno dei genitori;
- negli altri casi si è proceduto a imputare, separatamente per padre e madre, il titolo di studio usando le proporzioni stratificate per professione, oppure la professione usando le proporzioni stratificate per titolo di studio.

I record trattati sono stati 755, per un totale di 932 valori imputati: 194 per PROFPA, 132 per PROFMA, 360 per TITPA e 246 per TITMA[2].

Ricordiamo infine che fra le variabili del modello per la probabilità di immatricolazione è presente anche quella che indica la natura privata dell'istituto (SCPR), i cui valori sono stati desunti da altre variabili oppure, in piccolissima parte, imputati con una procedura casuale (cfr. par. 3.3.1).

6.2 Specificazione del modello e analisi dei risultati

Utilizzando i dati ottenuti con le procedure descritte nel precedente paragrafo, abbiamo selezionato un modello logit multilivello per la variabile IMM

[2]E' bene precisare che anche nel modello per la probabilità di immatricolazione (cfr. tab. 6.2) le variabili relative ai due genitori sono state aggregate secondo le modalità discusse nel par. 3.3.1.

("immatricolazione all'università nei tre anni successivi al conseguimento del diploma"). Va detto che nel 91.4% dei casi l'iscrizione è stata perfezionata il primo anno, il che lascia supporre che in molti casi tale decisione sia stata presa prima di un'eventuale esperienza nel mercato del lavoro[3]. La tab. 6.1 riporta alcune statistiche descrittive delle variabili impiegate nel modello finale.

A differenza del modello per la probabilità di occupazione, nel presente caso la procedura di selezione ha individuato una variabile con coefficiente casuale (si tratta di FEMM, "sesso femminile"), per cui il modello finale è del seguente tipo:

$$y_{ij} \mid u_{0j}, u_{1j} \sim \text{Bernoulli}(\pi_{ij}),$$

$$\text{logit}(\pi_{ij}) = (\alpha + u_{0j}) + (\beta_1 + u_{1j}) \cdot x_{1ij} + \beta_2 x_{2ij} + \ldots + \beta_p x_{pij},$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix} \right)$$

dove $i = 1, \ldots, 18720$ indica i diplomati (unità di livello 1) e $j = 1, \ldots, 1561$ indica le scuole (unità di livello 2 o gruppi).

L'esame dei risultati delle stime (cfr. tab. 6.2) suggerisce le seguenti considerazioni:

- Le due procedure di stima adottate, PQL2 e ML, forniscono risultati praticamente identici, anche se la prima ha tempi di calcolo nettamente inferiori;

- Il coefficiente di FEMM ha una variabilità considerevole: ad esempio, $\hat{\beta}_1 \pm 2\sqrt{\hat{\sigma}^2_{u1}}$ fornisce l'intervallo $(-1.391, 0.959)$. Ciò significa che, sebbene l'effetto medio di FEMM sia negativo (-0.216), in alcune scuole tale effetto è largamente positivo. Il parametro σ^2_{u0} è invece interpretabile come variabilità fra le scuole *relativamente ai maschi*.

- L'interpretazione dei parametri casuali può essere agevolata dall'esame dei risultati che si ottengono con una diversa parametrizzazione del modello, sostituendo l'intercetta con la variabile binaria che indica il sesso maschile, MASC=(1-FEMM). Ciò equivale ad assumere che maschi e

[3] Il modello descritto in questo paragrafo è stato implementato anche con la variabile di risposta IMMSUB ("immatricolazione nel *primo anno* successivo al conseguimento del diploma"), senza che si registrassero differenze di rilievo rispetto a quanto ottenuto con IMM.

Tab. 6.1 - Statistiche descrittive delle variabili impiegate nel modello per la probabilità di immatricolazione all'università nei tre anni successivi al conseguimento del diploma

Unità livello 1 (diplomati)	18720
Unità livello 2 (scuole)	1561

Variabile	Descrizione sintetica*	Min	Max	Media	Err. Std.
IMM	Immatricolato all'università	0	1	0.50	0.50
FEMM	Femmina	0	1	0.54	0.50
ETA>14	Età>14 all'iscrizione scuola secondaria	0	1	0.10	0.30
GIUD4**	Giudizio licenza media "ottimo"	0	1	0.17	0.38
RIP	Almeno una ripetenza scuola second.	0	1	0.25	0.43
VM36	Voto di maturità 36	0	1	0.10	0.30
VM37-42	Voto di maturità 37-42	0	1	0.33	0.47
VM50-59	Voto di maturità 50-59	0	1	0.22	0.41
VM60	Voto di maturità 60	0	1	0.05	0.23
LAVCPRIM	Lavoro continuativo prima del diploma	0	1	0.04	0.19
FRAT>1	Più di un fratello	0	1	0.35	0.48
TIT**	Titolo genitore: almeno diploma maturità	0	2	0.62	0.81
PROF-0**	Prof. madre: casalinga	0	1	0.57	0.49
PROF-B**	Prof. genitore: dipendente	0	2	0.38	0.59
PROF-CE**	Prof. genitore: dirig./insegn./imprend./lib.prof.	0	2	0.30	0.57
PROF-D**	Prof. genitore: altri indipendenti	0	2	0.30	0.54
SCPR**	Scuola privata	0	1	0.14	0.34
ISTPROF	Istituto professionale	0	1	0.19	0.39
ISTMAG	Istituto magistrale	0	1	0.07	0.26
LICEI	Licei	0	1	0.24	0.43
ALTRIIST	Altri istituti	0	1	0.07	0.26
DISOCC	Tasso disocc. giovanile region. 1995	-16.2	39.5	11.23	17.67
SUD	Circoscrizione Sud	0	1	0.40	0.49

* Le descrizioni dettagliate delle variabili sono riportate nella tab. 3.1

** Variabile con alcuni valori imputati (cfr. par. 6.1)

Tab. 6.2 - Risultati delle stime relative al modello per la probabilità di immatricolazione all'università nei tre anni successivi al conseguimento del diploma

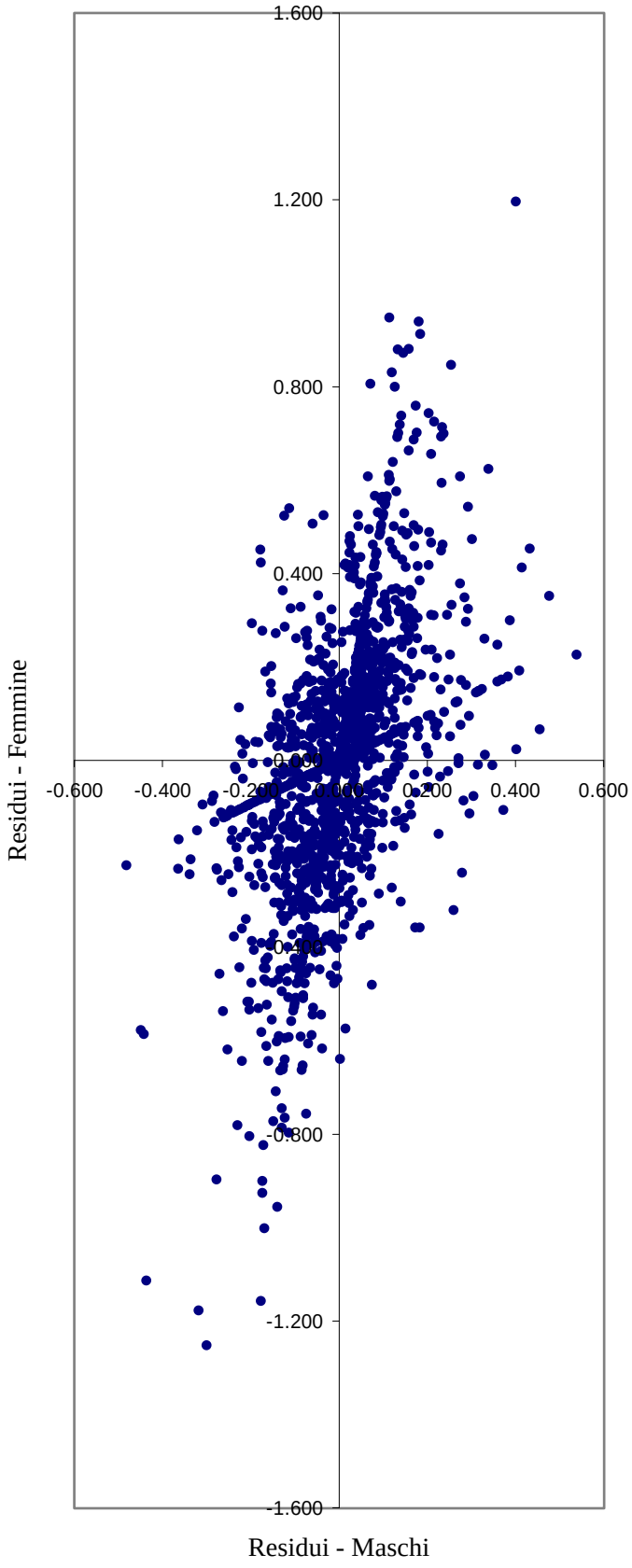
	MLwiN		**MIXOR**		
	Link	Logit	Link	Logit	
	Distr. EC	Gaussiana	Distr. EC	Gaussiana	
	Metodo	PQL2	Metodo	ML	
	Variaz. EB	no	Punti quad.	10	
	Iterazioni	7	Iterazioni	10	
	Devianza	-	Devianza	17710.1	
Variabili*	**Effetti fissi**		**Effetti fissi**		**Differenze fra ML e PQL2**
	Stima	Err. Std.	Stima	Err. Std.	Stima	Err. Std.
Intercetta	-0.389	0.074	-0.388	0.074	-0.069%	-0.524%
FEMM	-0.216	0.058	-0.215	0.060	-0.153%	3.697%
ETA>14	-0.306	0.067	-0.306	0.065	-0.016%	-3.622%
GIUD4**	0.279	0.065	0.280	0.066	0.075%	1.422%
RIP	-0.401	0.047	-0.401	0.047	0.130%	-0.618%
VM36	-1.036	0.074	-1.037	0.072	0.107%	-3.546%
VM37-42	-0.629	0.048	-0.629	0.049	0.137%	2.383%
VM50-59	0.673	0.054	0.673	0.056	0.080%	3.427%
VM60	1.424	0.108	1.424	0.106	0.007%	-1.854%
LAVCPRIM	-0.762	0.109	-0.762	0.106	-0.039%	-2.889%
FRAT>1	-0.169	0.043	-0.169	0.044	0.077%	2.417%
TIT**	0.360	0.032	0.361	0.034	0.083%	3.549%
PROF-0**	-0.168	0.047	-0.169	0.047	0.190%	0.042%
PROF-B**	0.300	0.043	0.300	0.043	-0.017%	0.680%
PROF-CE**	0.445	0.052	0.445	0.053	0.058%	1.871%
PROF-D**	0.152	0.040	0.152	0.039	0.158%	-2.208%
SCPR**	-0.196	0.068	-0.197	0.066	0.540%	-2.608%
ISTPROF	-1.032	0.082	-1.032	0.082	0.011%	1.005%
ISTPROF*FEMM	0.244	0.115	0.242	0.117	-0.796%	1.745%
LICEI	2.633	0.076	2.639	0.074	0.215%	-1.990%
ISTMAG	0.539	0.091	0.539	0.100	0.009%	10.233%
ALTRIIST	-0.360	0.086	-0.361	0.087	0.436%	0.266%
DISOCC	0.005	0.003	0.005	0.003	-0.162%	-2.025%
DISOCC^2	-0.00024	0.00010	-0.00024	0.00011	-0.125%	8.374%
DISOCC*FEMM	0.005	0.002	0.005	0.003	0.601%	3.184%
SUD	0.309	0.080	0.308	0.086	-0.246%	8.414%
	Effetti casuali		**Effetti casuali *** **			
	Stima	Err. Std.	Stima	Err. Std.		
Intercetta	0.138	0.040	0.144	-		
FEMM	0.345	0.101	0.362	-		
Covar.	-0.073	0.057	-0.078	-		
(Correlaz.)	-0.337		-0.341			

* Le definizioni delle variabili sono riportate nella tab. 3.1. La notazione VAR1*VAR2 indica interazione fra VAR1 e VAR2

** Variabile con alcuni valori imputati (cfr. par. 6.1)

*** L'algoritmo implementato in MIXOR stima la scomposizione di Cholesky, dalla quale si ricavano varianze e covarianze degli effetti casuali

Fig. 6.1 - Diagramma dei residui relativi alle scuole distintamente per maschi e femmine



femmine abbiano effetti casuali separati. I risultati della stima sono (in parentesi l'errore standard stimato):

$$\begin{bmatrix} \hat{\sigma}^2_{uMASC} = 0.137(0.040) & \\ \hat{\sigma}_{uMASC/FEMM} = 0.065(0.043) & \hat{\sigma}^2_{uFEMM} = 0.335(0.044) \end{bmatrix}$$

E' evidente che la variabilità fra le scuole in merito alla probabilità di passaggio all'università è molto maggiore per le femmine. Inoltre, trasformando la covarianza in correlazione si ottiene 0.3, per cui le probabilità di immatricolazione di maschi e femmine della stessa scuola presentano una moderata correlazione positiva[4]. La struttura di varianza e covarianza trova riscontro nei residui a livello di scuola, riportati in fig. 6.1. Si noti che, nonostante la correlazione positiva, in un consistente numero di scuole i residui per i maschi e per le femmine hanno segno opposto.

- L'età del diplomato è inserita per mezzo di una variabile binaria che assume il valore 1 se la prima iscrizione alla scuola secondaria superiore è avvenuta ad un'età maggiore di 14 anni (ETA>14), con coefficiente stimato pari a -0.306 (si è preferita l'età all'iscrizione per evitare interferenze con la variabile che indica le ripetenze, RIP).

- Per quanto riguarda il background familiare, osserviamo innanzitutto che la probabilità di immatricolazione si riduce se il diplomato ha più di un fratello ($\hat{\beta}_{FRAT>1} = -0.169$), confermando le indicazioni della letteratura (cfr. par. 1.1.2). La presenza di genitori con titolo di studio non inferiore al diploma ha, come atteso, un effetto ampiamente positivo ($\hat{\beta}_{TIT} = 0.360$)[5]. Anche la professione dei genitori ha un effetto rilevante: a questo proposito l'aspetto chiave sembra essere la componente intellettuale del lavoro[6]. Purtroppo il reddito della famiglia non è disponibile; comunque l'impressione complessiva che si trae dalle variabili relative al background familiare è che l'iscrizione all'università dipenda più da fattori culturali che da fattori economici (questa indicazione è in linea con i precedenti lavori empirici: cfr. Micklewright, 1989).

[4] Il termine di covarianza è statisticamente non significativo. Tuttavia non è opportuno vincolarlo a zero, poiché il suo valore dipende dalla parametrizzazione adottata (cfr. Snijders & Bosker, 1999, par. 5.1.2).

[5] Separando gli effetti del titolo del padre e della madre si ottiene $\hat{\beta}_{TITPA} = 0.419$ e $\hat{\beta}_{TITMA} = 0.289$. Tuttavia la differenza non è statisticamente significativa (Wald test = 2.366).

[6] Questo aspetto risulta evidente dall'esame del tasso di passaggio all'università per professione del padre. Ad esempio, per "insegnante di scuola media" il tasso di passaggio è 81.9%, mentre per "libero professionista" è 69.5% e per "imprenditore" è 53.9%.

- L'andamento scolastico ha una forte influenza sulla probabilità di immatricolazione: oltre al voto di maturità hanno effetto anche le ripetenze ($\hat{\beta}_{RIP} = -0.401$) e il giudizio della scuola media inferiore ($\hat{\beta}_{GIUD4} = 0.279$) [7]. E' interessante notare che togliendo le variabili relative ai voti e alle ripetenze il coefficiente di FEMM diviene significativamente positivo: ciò significa che il maggior tasso di passaggio all'università da parte delle femmine (che nel campione PSLD è del 52.3% contro il 46.5% dei maschi) è verosimilmente attribuibile al miglior andamento scolastico.

- L'effetto del tipo di scuola sulla probabilità di immatricolazione è in linea con le attese (cfr. app. A). Si noti che negli istituti Professionali, per effetto del termine di interazione, la differenza fra maschi e femmine è trascurabile. Va segnalato, inoltre, l'effetto negativo relativo alle scuole private ($\hat{\beta}_{SCPR} = -0.196$): questo risultato probabilmente è dovuto alla presenza di scuole private finalizzate al recupero degli anni scolastici.

- Per quanto riguarda il contesto socio-economico, le variabili utilizzate sono il tasso di disoccupazione giovanile a livello regionale, con il termine quadratico e l'interazione con il sesso, e la variabile che indica la circoscrizione geografica meridionale[8]. Tuttavia quest'ultima variabile, che ha un effetto rilevante ($\hat{\beta}_{SUD} = 0.309$), impedisce una corretta valutazione del ruolo svolto dal tasso di disoccupazione. Pertanto abbiamo effettuato anche la stima in assenza della variabile SUD: in tal caso si ottiene $\hat{\beta}_{DISOCC} = 0.00552$ e $\hat{\beta}_{DISOCC*FEMM} = 0.00505$ (il termine quadratico è non significativo). Dunque l'effetto del tasso di disoccupazione sulla probabilità di immatricolazione è positivo, così come suggerito dalla teoria economica (anche se l'evidenza empirica spesso è molto debole: cfr. O'Higgins, 1992). Per valutare l'effetto del tasso di disoccupazione giovanile regionale osserviamo che un'aumento di 50 punti percentuali (che si riscontra, ad esempio, fra Veneto e Campania) produce un incremento nel predittore lineare di 0.276 per i maschi e di 0.528 per le femmine, cosicché un individuo con probabilità stimata pari al 50% passerebbe a 56.9% se maschio o 62.9% se femmina.

[7] Poiché circa il 6.5% dei valori di GIUD4 sono stati imputati (cfr. par. 6.1), per valutare la stabilità della stima del coefficiente abbiamo effettuato due nuove imputazioni e stime, ottenendo 0.323 e 0.298: la sensibilità della stima ai valori imputati sembra quindi modesta.

[8] A causa del termine quadratico e delle interazioni, la selezione dei quattro regressori relativi alle condizioni socio-economiche è stata effettuata usando le stime ML di MIXOR con il relativo test del rapporto di massima verosimiglianza.

6.3 Confronto con i risultati relativi alla probabilità di occupazione

Concludiamo il capitolo con un breve confronto fra i risultati relativi alla probabilità di immatricolazione all'università (par. 6.2) e alla probabilità di occupazione (par. 4.2). In entrambi i casi la risposta è dicotomica, ma esistono alcune importanti differenze.

Infatti, nel caso dell'immatricolazione l'evento di interesse ha le seguenti caratteristiche:

a) si colloca nell'arco temporale che intercorre fra il conseguimento del diploma e l'intervista (anche se la maggioranza delle iscrizioni viene effettuata nel corso del primo anno);

b) è una decisione dell'individuo, pur in presenza di vincoli di varia natura (distanza dalle sedi universitarie, eventuale esame di ammissione ecc.);

c) tutti i diplomati sono potenziali candidati all'immatricolazione.

Invece, nel caso dell'occupazione l'evento di interesse ha le seguenti caratteristiche:

a') si colloca al momento dell'intervista;

b') dipende in gran parte da fattori non direttamente controllabili da parte dell'individuo;

c') non tutti i diplomati sono interessati a lavorare nel breve periodo.

Queste differenze hanno importanti conseguenze sull'analisi: ad esempio, il punto a') spiega perché nel modello per l'occupazione siano presenti una serie di variabili che si riferiscono alle esperienze di studio e lavoro post-diploma; mentre il punto c') spiega perché il modello per l'occupazione abbia richiesto l'individuazione di un adeguato sottocampione (cfr. par. 4.1).

Tenendo presenti queste distinzioni, dal confronto dei risultati (tabb. 4.2 e 6.2) scaturiscono le seguenti considerazioni:

- Le *differenze fra le scuole* sono più marcate nel caso dell'immatricolazione: infatti, gli effetti relativi ai vari tipi di scuola sono più forti e la variabilità residua è più grande (la stima del modello ad intercetta casuale - non riportata - fornisce $\hat{\sigma}_{u0} = 0.203$, contro $\hat{\sigma}_{u0} = 0.126$ del modello per l'occupazione). Questo risultato sembra indicare che

la decisione di continuare gli studi dipende in modo sostanziale, oltre che dal tipo di scuola, da fattori non osservabili come l'influenza degli insegnanti e dei compagni di classe. Inoltre, il modello per l'immatricolazione, a differenza di quello per l'occupazione, evidenzia una diversa variabilità tra le scuole relativamente ai maschi e alle femmine.

- Il *background familiare* e l'*andamento scolastico* hanno molta più influenza sulla scelta di iscriversi all'università che sulle possibilità occupazionali. Ciò conferma che le scelte formative sono fortemente condizionate dall'ambiente familiare e scolastico.

- Viceversa, il *contesto socio-economico* è più rilevante per l'occupazione che non per l'immatricolazione, a conferma delle difficoltà incontrate dai giovani nell'inserimento lavorativo.

Conclusioni

Gli sbocchi occupazionali e le scelte formative dei diplomati sono un fenomeno di estrema rilevanza non solo per gli studenti e le loro famiglie, ma anche per i datori di lavoro e gli amministratori pubblici, interessati a valutare la capacità del sistema scolastico di formare adeguatamente i giovani.

In questa ricerca ci siamo proposti di individuare e valutare i principali fattori che influenzano le esperienze di studio e lavoro post-diploma, ponendo speciale attenzione ai fattori di contesto legati alla scuola frequentata e alla zona di residenza, nel tentativo di pervenire ad una valutazione dell'efficacia degli istituti scolastici.

Come discusso nel paragrafo 1.4, il raggiungimento degli obiettivi sopra menzionati ha richiesto l'utilizzo di appropriati modelli multilivello per variabili di risposta categoriche, i cui aspetti teorici e computazionali sono stati approfonditi nel secondo capitolo.

La disponibilità di un dataset particolarmente ricco di informazioni, qual è quello derivato dall'indagine Istat sui Percorsi di Studio e Lavoro dei Diplomati, ha consentito di costruire una serie di modelli piuttosto articolati, in grado di includere gran parte dei fattori che la teoria e i precedenti lavori empirici giudicavano rilevanti. A questo proposito, va menzionato l'inserimento delle variabili relative alla situazione del mercato del lavoro, che in molte applicazioni è reso impossibile dalle carenze informative dei dati. Particolarmente importante si è rivelato l'utilizzo del tasso di disoccupazione giovanile regionale nei modelli per gli sbocchi occupazionali (capitoli 4 e 5), sebbene il livello regionale non sia certo quello ottimale (riteniamo infatti che, qualora fosse possibile, l'impiego dei tassi a livello provinciale comporterebbe un sensibile miglioramento in termini di adattamento del modello). Peraltro, la concomitante presenza in tali modelli delle variabili relative alle circoscrizioni

geografiche (Nord, Centro e Sud) lascia intendere che il tasso di disoccupazione utilizzato non è un indicatore esaustivo delle difficoltà incontrate dai giovani nell'inserimento lavorativo.

In generale, i modelli multilivello che abbiamo utilizzato si sono rivelati uno strumento particolarmente adatto allo studio delle esperienze post-diploma, consentendo di analizzare i vari aspetti del fenomeno sia a livello di individuo che a livello di scuola.

Relativamente alla probabilità di occupazione (capitolo 4), appaiono molto importanti tutte le variabili che si riferiscono all'esperienza di studio o di lavoro immediatamente successive al diploma. Invece l'unica variabile del background familiare che si è mostrata rilevante è la professione dei genitori, mentre il curriculum degli studi sembra esercitare la propria influenza solo attraverso il voto di maturità e il tipo di diploma ottenuto.

L'analisi dei tempi di ingresso al lavoro (capitolo 5) ha permesso di mettere in luce alcuni interessanti aspetti dell'inserimento lavorativo dei diplomati, come la ripresa della probabilità di occupazione nel terzo anno successivo al conseguimento del diploma. Questo fenomeno, parzialmente inatteso, ha sollecitato ulteriori approfondimenti, che hanno richiesto la costruzione, effettuata dall'Isfol su nostra richiesta, di un'adeguata serie storica di ingressi al lavoro dipendente al fine di valutare il ruolo svolto dalla domanda di lavoro. L'impiego di questa serie storica nelle analisi descrittive e nei modelli di sopravvivenza ha permesso di concludere che l'andamento degli ingressi al lavoro dei diplomati dell'indagine PSLD non dipende tanto da fluttuazioni cicliche della domanda di lavoro, quanto da dinamiche relative alla coorte dei diplomati. Ad esempio, la ripresa della probabilità di occupazione nel terzo anno successivo al conseguimento del diploma sembra in gran parte attribuibile al fatto che un buon numero di diplomati non è immediatamente disponibile a lavorare, perché impegnato nel servizio militare o in corsi di vario tipo.

Lo studio della probabilità di immatricolazione all'università (capitolo 6) ha invece evidenziato che in questo caso il ruolo chiave è svolto dal background familiare (professione e titolo dei genitori, numero di fratelli) e dal curriculum degli studi (giudizio di scuola media, ripetenze durante la scuola secondaria superiore, voto di maturità, tipo di diploma, tipo di scuola - pubblica o privata). Riguardo alle differenze fra i sessi, il modello segnala una minore probabilità di immatricolazione per le femmine, sebbene la probabilità marginale sia maggiore (le analisi effettuate hanno messo in luce che questa differenza è dovuta alla migliore performance scolastica delle femmine).

La variabilità residua attribuibile alle scuole è risultata consistente in tutte le analisi effettuate. In particolare, nello studio della probabilità di occupazione (capitolo 4) la presenza di tale varianza residua suggerisce l'e-

sistenza di importanti differenze nell'efficacia delle scuole (con l'avvertenza, però, che tali differenze sono in parte attribuibili alla variabilità infraregionale delle condizioni del mercato del lavoro). Nello studio della probabilità di immatricolazione all'università (capitolo 6) la varianza residua attribuibile alle scuole è risultata maggiore, presumibilmente a causa del ruolo esercitato dagli insegnanti e dai compagni di classe sulla scelta del giovane; inoltre le stime hanno evidenziato un'interessante differenziazione della variabilità fra i sessi, con un valore nettamente superiore per le femmine.

La presenza di un'importante variabilità residua attribuibile alle scuole, relativamente allo status occupazionale dei diplomati, ha motivato l'uso dei residui come stime di efficacia, al fine di valutare la qualità del servizio offerto dalle scuole. Tuttavia, nella presente applicazione tale valutazione è risultata di fatto impossibile, a causa del notevole margine di incertezza associato ai residui, attribuibile al limitato numero di diplomati per scuola (cfr. par. 4.3). La scarsa affidabilità dei residui come misure di efficacia è stata evidenziata anche dalle simulazioni riportate nell'appendice B. Tali simulazioni hanno inoltre mostrato che, in un contesto analogo a quello del capitolo 4, il livello di confidenza effettivo degli intervalli relativi ai residui è sensibilmente inferiore a quello nominale, lasciando intendere che i confronti di efficacia sono ancora più difficili di quanto emerga dalle procedure standard. Questi risultati rappresentano comunque delle utili indicazioni per la progettazione di eventuali studi mirati alla valutazione dell'efficacia delle scuole, soprattutto in termini di definizione delle dimensioni campionarie.

Da un punto di vista metodologico abbiamo affrontato una serie di questioni relative alla specificazione e stima dei modelli multilivello per dati categorici.

Innanzitutto, nel paragrafo 2.2 abbiamo presentato una trattazione unitaria dei modelli lineari generalizzati multilivello che, ricorrendo anche ai concetti di variabile latente e soglia, evidenzia gli aspetti che accomunano modelli apparentemente lontani, come quelli per i dati dicotomici e quelli per i tempi di sopravvivenza. Le problematiche legate alla specificazione e stima dei modelli di sopravvivenza in tempo discreto ad effetti casuali sono state poi approfondite nel capitolo 5, nel quale abbiamo confrontato le estensioni a rischi non proporzionali di due versioni in tempo discreto del modello di Cox, dette *grouped continuous* (McCullagh, 1980) e *continuation ratio* (Prentice & Gloeckler, 1978). Le analisi effettuate sui dati dell'indagine PSLD ci hanno fatto preferire il secondo di questi modelli, più complesso negli aspetti di specificazione e stima, ma in grado di descrivere adeguatamente, con un numero limitato di parametri, l'andamento del rischio per vari sottoinsiemi della popolazione; a ciò va aggiunta la possibilità di includere covariate tempo-dipendenti, che nell'applicazione specifica sono rappresentate

dalla serie storica di ingressi al lavoro dipendente sopra menzionata.

Relativamente ai metodi di stima, in questo lavoro abbiamo confrontato la Quasi-Verosimiglianza Penalizzata del secondo ordine con la Massima Verosimiglianza con integrazione numerica: in tutte le applicazioni le differenze nelle stime sono risultate minime, com'era prevedibile data l'elevata dimensione campionaria. Nei modelli con un solo effetto casuale l'efficienza computazionale dei due metodi è risultata simile; tuttavia, con l'aggiunta di un secondo effetto casuale (nel modello per la probabilità di immatricolazione all'univerisità presentato nel capitolo 6), il metodo con integrazione numerica ha mostrato tempi di calcolo nettamente maggiori.

Un altro aspetto metodologico considerato riguarda il possibile effetto del piano di campionamento sulle stime, che ha motivato la sperimentazione di alcuni metodi di stima pesata per il modello logit multilivello (paragrafo 4.4). In particolare abbiamo esteso la procedura sviluppata da Pfeffermann *et al.* (1998) per il modello lineare, che consente di inserire agevolmente i pesi relativi ad entrambi i livelli della gerarchia. I risultati sono stati convalidati tramite la comparazione con quelli ottenuti applicando in modo diretto, ma limitatamente ai pesi di livello 2, il classico metodo della Pseudo-Verosimiglianza. I metodi di stima pesata che abbiamo sperimentato sono sembrati affidabili nel presente contesto, anche se la carenza di risultati teorici ed empirici suggerisce una certa cautela nella loro applicazione generalizzata. L'analisi pesata ha mostrato che, nel nostro caso, i pesi rilevanti per le stime sono quelli di livello 2, cioè quelli relativi alle probabilità di inclusione delle scuole. Tuttavia, l'analisi pesata, pur facendo registrare qualche differenza apprezzabile rispetto all'analisi standard, non ha mutato il quadro d'insieme, confermando la validità dei risultati ottenuti nelle altre parti di questa ricerca.

Pur avendo a nostro avviso sviluppato diverse interessanti analisi e affrontato problemi rilevanti e spesso trascurati dalla letteratura corrente, rimangono numerose questioni che meritano di essere approfondite nelle future ricerche in questo campo. In primo luogo riteniamo che sarebbe particolarmente importante approfondire la questione dell'analisi dei residui, studiando in modo puntuale l'influenza della dimensione campionaria sugli intervalli di confidenza e quindi sulle comparazioni fra gruppi, anche al fine di fornire precise indicazioni per eventuali indagini campionarie mirate alla valutazione dell'efficacia degli istituti scolastici. Sempre sul tema della valutazione dell'efficacia, un altro filone di ricerca potrebbe riguardare la definizione di metodi alternativi per il calcolo dei residui, basati su criteri diversi da quello bayesiano empirico comunemente usato. Infine, un ulteriore argomento che necessita di approfondimenti teorici ed empirici è quello delle procedure di stima pesata per i modelli multilivello non lineari, che in questo lavoro hanno trovato una prima applicazione.

Appendice A

Sistema formativo e mercato del lavoro giovanile in Italia

Nella presente appendice, basandoci sulle statistiche ufficiali della scuola, dell'università e del lavoro, presentiamo un quadro della realtà italiana relativamente al sistema scolastico ed alle esperienze post-diploma dei maturi della scuola secondaria superiore. Il materiale di questa appendice costituisce un utile complemento per una migliore comprensione delle analisi dei capp. 4-6, tutte basate sui dati dell'indagine Istat sui Percorsi di Studio e Lavoro dei Diplomati (PSLD), relativa ai diplomati delle scuole secondarie superiori italiane dell'anno 1995.

A.1 Cenni sull'ordinamento scolastico e le scuole secondarie superiori

Iniziamo con una breve descrizione dell'ordinamento scolastico italiano, con particolare riferimento alle scuole secondarie superiori, *relativamente al 1995*, anno in cui i diplomati oggetto dell'indagine PSLD hanno concluso gli studi. Le informazioni sono tratte da Istat (1997).

L'ordinamento scolastico italiano suddivide l'istruzione in due grandi settori:

a) *Istruzione scolastica*, impartita in quegli istituti che perseguono il fine di educare ed istruire le nuove generazioni. Si suddivide nei seguenti livelli:

1. educazione prescolastica (scuole materne);
2. istruzione primaria (scuole elementari);

3. istruzione secondaria di primo grado (scuole medie);
4. istruzione secondaria di secondo grado (scuole secondarie superiori);
5. istruzione post-secondaria non universitaria;
6. istruzione universitaria.

b) *Istruzione extra-scolastica*, che comprende quell'insieme di iniziative finalizzate all'apprendimento diretto di arti o mestieri (formazione professionale) o all'avanzamento culturale degli adulti.

Le scuole secondarie superiori, che interessano in questo lavoro, hanno corsi di durata generalmente quinquennale. Il passaggio da un anno di corso all'altro avviene sulla base dei voti ottenuti durante l'anno e al termine dell'ultimo anno è previsto un esame per il conseguimento del diploma di maturità. Tale diploma, se conseguito al termine di un corso di 5 anni, consente l'iscrizione a qualsiasi facoltà universitaria. Gli istituti magistrali e i licei artistici hanno corsi di durata quadriennale e il relativo diploma permette l'accesso solo ad alcune facoltà[1].

Le scuole secondarie superiori si distinguono in:

- *Licei* - I licei sono stati istituiti al fine di fornire una preparazione di base per gli studi universitari. Le due tipologie principali sono il liceo classico (studi umanistici) e il liceo scientifico (studi scientifici). Esistono poi licei con indirizzo linguistico, pedagogico-sociale ed artistico (quest'ultimo con corsi di durata quadriennale).

- *Istituto e scuola magistrale* - L'istituto magistrale, il cui corso ha durata quadriennale, è nato con lo scopo di preparare gli insegnanti di scuola elementare; invece la scuola magistrale, il cui corso ha durata triennale, forma gli insegnanti della scuola materna.

- *Istituti tecnici* - Gli istituti tecnici preparano all'esercizio di alcune professioni o allo svolgimento di funzioni tecniche o amministrative nel campo dell'agricoltura, dell'industria e del commercio. Si distinguono vari indirizzi: agrario, industriale, commerciale, per geometri, aereonautico, nautico, per il turismo, per periti aziendali, femminili.

[1] Il diploma del liceo artistico consente l'iscrizione all'Accademia di Belle Arti o alla facoltà di Architettura a seconda dell'indirizzo seguito. Invece il diploma magistrale consente l'iscrizione alle facoltà di Magistero e Lingue. In entrambi i casi, gli studenti che vogliono accedere ad altre facoltà devono frequentare un corso integrativo della durata di un anno.

- *Istituti professionali* - Gli istituti professionali hanno finalità analoghe a quelle degli istituti tecnici, ma l'insegnamento ha un carattere più pratico e orientato all'immediato inserimento nel mondo del lavoro. Accanto al corso di 5 anni è presente anche un ciclo di studi di durata inferiore (2, 3 o 4 anni) al termine del quale viene rilasciato un diploma di qualifica professionale[2].

- *Istituti d'arte* - Gli istituti d'arte forniscono una formazione artistica di tipo applicato. Prevedono un triennio, che consente di ottenere il diploma di maestro d'arte, e un successivo biennio per il conseguimento del diploma di maturità d'arte applicata.

A seconda della gestione, le scuole secondarie superiori possono essere classificate nel seguente modo:

$$\text{scuole sec. sup.} \begin{cases} \text{pubbliche} \begin{cases} \text{statali} \\ \text{non statali} \end{cases} \\ \text{private} \end{cases}$$

Le scuole pubbliche statali dipendono direttamente dal Ministero della Pubblica Istruzione, mentre quelle non statali sono gestite da Enti locali territoriali (Comuni, Provincie e Regioni) o da altri Enti pubblici.

A.2 Percorsi formativi e diplomi di maturità

Il sistema scolastico italiano degli anni novanta è caratterizzato da una diminuzione del numero complessivo di studenti, attribuibile al calo demografico, e da un contemporaneo aumento dei tassi di scolarità. In particolare, il tasso di scolarità della scuola secondaria superiore è passato dal 68.3% del 1990/91 all'86.8% del 1997/98[3].

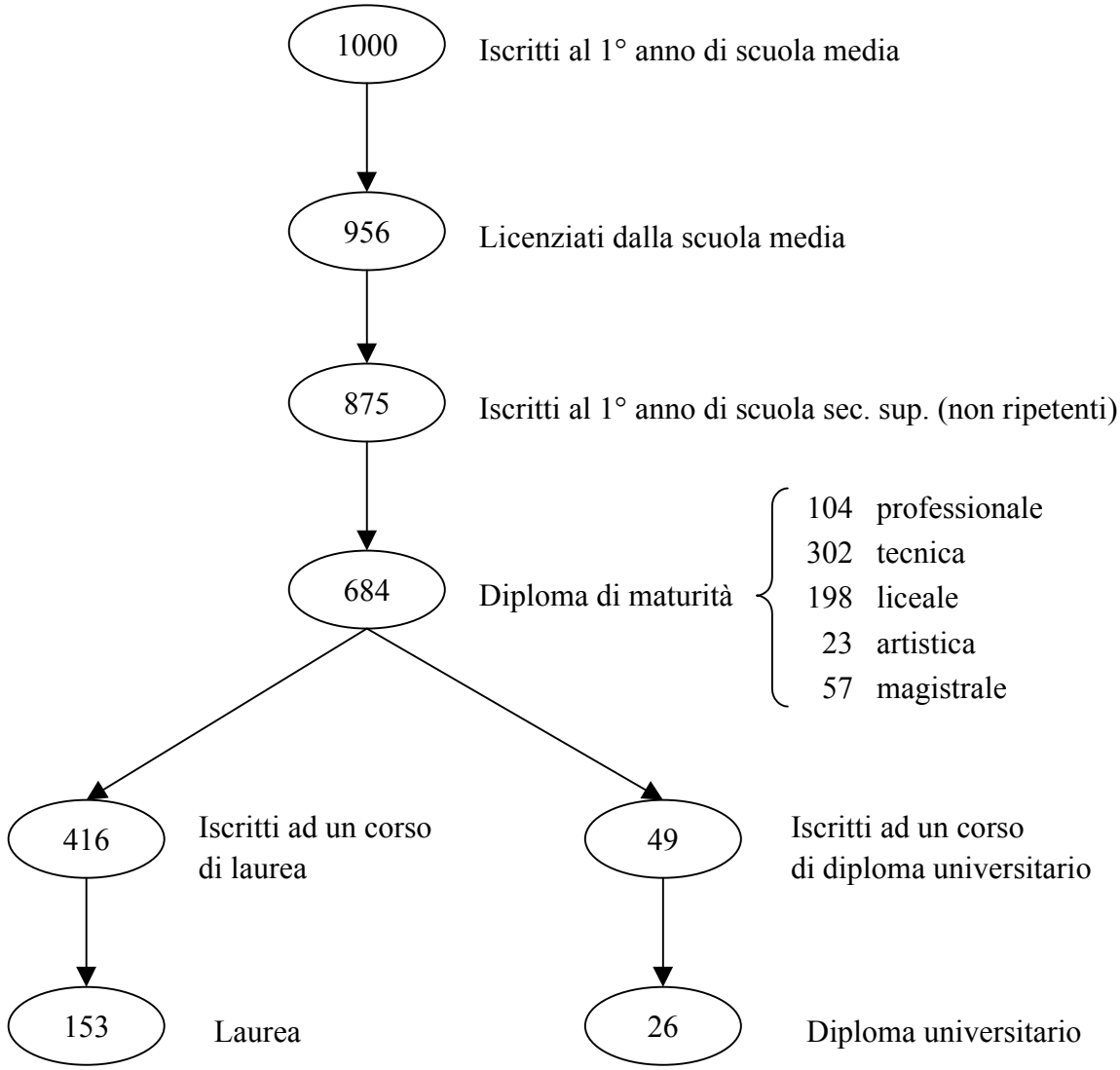
I flussi del sistema scolastico sono evidenziati nella fig. A.1, i cui valori sono stati calcolati con la metodologia *per contemporanei* utilizzando l'ultimo anno disponibile, cioè il 1996. A fronte di una elevata permanenza nel sistema scolastico (quasi la metà dei giovani si iscrive all'università), si registrano considerevoli tassi di abbandono nella scuola secondaria superiore e, soprattutto, nell'università (riesce a laurearsi circa una matricola su tre).

Esaminiamo adesso più in dettaglio la situazione relativa ai diplomati (maturi) dell'a.s. 1994/95, che costituiscono l'universo dell'indagine PSLD.

[2]Gli studenti che terminano gli studi con il diploma di qualifica professionale non rientrano nella popolazione di riferimento dell'indagine PSLD.

[3]Cfr. ISFOL (1998). Il tasso di scolarità è dato dal rapporto percentuale tra i frequentanti e i giovani di età compresa fra i 14 e i 18 anni.

Fig. A.1 - I percorsi nel sistema formativo italiano: il modello tendenziale



Fonte: ISFOL (1998). Elaborazioni ISFOL su dati ISTAT, ISCO e ISFOL con il metodo per contemporanei (tassi del 1996).

Tab. A.1 - Maturi e immatricolati all'università (corsi di laurea e di diploma) per tipo di scuola negli anni 1994-1996

Tipo di scuola	Anno solare 1994		Anno solare 1995		Anno solare 1996	
	Maturi	Immatric. università	Maturi	Immatric. università	Maturi	Immatric. università
Istituti Professionali	70745	21774	72439	19535	74349	20068
	14.3%	6.4%	14.8%	5.8%	15.2%	6.3%
Istituti Tecnici	229288	118217	223320	110756	215991	103101
	46.2%	34.8%	45.5%	33.0%	44.1%	32.2%
Licei Scientifici	86847	95204	87353	95151	91047	97890
	17.5%	28.0%	17.8%	28.4%	18.6%	30.5%
Licei Classici	42717	47690	42886	48674	43824	47896
	8.6%	14.0%	8.7%	14.5%	9.0%	14.9%
Istituti Magistrali	37931	23218	37497	22466	38502	20589
	7.7%	6.8%	7.6%	6.7%	7.9%	6.4%
Altre scuole	28245	33766	26853	38766	25664	30916
	5.7%	9.9%	5.5%	11.6%	5.2%	9.6%
Totale	495773	339869	490348	335348	489377	320460
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Fonte: Nostra elaborazione su dati ISTAT (Statistiche delle scuole secondarie superiori, Statistiche dell'istruzione universitaria).

Note: per ogni anno i maturi sono quelli dell'a.s. che si conclude in tale anno e gli immatricolati sono quelli dell'a.a. che inizia in tale anno.

La tab. A.1 mostra la composizione dei maturi per tipo di scuola, da cui emerge il ruolo dominante svolto dagli Istituti tecnici (45.5% dei diplomati). Dalle statistiche sulle scuole secondarie superiori (Istat, 1997) si desume che tale composizione in linea di massima vale sia per il Nord che per il Sud del Paese, seppur con qualche eccezione (ad esempio, i diplomati dell'Istituto magistrale rappresentano il 5.5% dei diplomati nel Centro-Nord e il 10.8% nel Sud). La composizione dei maturi per sesso vede una leggera prevalenza delle femmine (54.2%), mentre le scuole di provenienza sono per l'88.2% pubbliche (di cui oltre il 99% statali) e per l'11.8% private[4].

Il tasso di passaggio all'università per l'a.a. 1995/96 è del 68.4%[5]. La tab. A.1 presenta inoltre la composizione per scuola di provenienza degli immatricolati a corsi di laurea o di diploma universitario nell'a.a. 1995/96. Ciò fornisce alcune indicazioni di massima sui tassi di passaggio per tipo di scuola frequentata: l'iscrizione all'università riguarda la quasi totalità dei diplomati liceali, poco meno della metà dei diplomati tecnici e una minoranza dei diplomati professionali. Le immatricolazioni a corsi di diploma universitario sono appena il 7.1% del totale.

A.3 Neodiplomati e mercato del lavoro

Il periodo di osservazione dell'indagine PSLD va dal 1995 al 1998, anni in cui il mercato del lavoro è rimasto sostanzialmente immutato. Il fatto più rilevante è l'aumento di 187000 unità (pari allo 0.9%) del totale degli occupati (cfr. tab. A.3, discussa più avanti).

La tab. A.2 mostra che, in quegli anni, il tasso di disoccupazione è stabile a livello nazionale, ma tende a diminuire nel Centro-Nord e a crescere nel Sud. Non ci sono invece cambiamenti di rilievo nel rapporto fra i sessi, che vede le femmine nettamente svantaggiate.

Alcune indicazioni sulla situazione specifica dei neodiplomati emergono dall'analisi per classe di età e titolo di studio dell'Indagine sulle Forze di Lavoro dell'Istat. La tab. A.3 mostra che il tasso di disoccupazione della classe 20-24 anni è ben superiore a quello generale della popolazione con 15 anni e oltre (32.7% contro 12.0% nel 1995). Entrambi i tassi rimangono sostanzialmente immutati nel periodo 1995-1998. Se restringiamo poi l'attenzione

[4]Le scuole private sono caratterizzate, rispetto a quelle pubbliche, da una maggiore presenza di Licei e da un numero molto esiguo di Istituti professionali.

[5]Cfr. ISFOL (1998). La serie dei tassi di passaggio è la seguente: 64% (a.a. 1994/95), 68.4% (a.a. 1995/96), 67.9% (a.a. 1996/97) e 65.9% (a.a. 1997/98). Il calo degli ultimi due anni è probabilmente attribuibile da un lato al crescente peso delle tasse universitarie e dall'altro alle difficili prospettive occupazionali che interessano anche i laureati.

Tab. A.2 - Tassi di disoccupazione (%) degli anni 1995-1998 per ripartizione geografica e sesso (popolazione di 15 anni e oltre)

	Nord				Centro				Mezzogiorno				Italia			
	1995	1996	1997	1998	1995	1996	1997	1998	1995	1996	1997	1998	1995	1996	1997	1998
Maschi	4.4	4.2	4.3	4.0	7.4	7.5	7.2	7.2	16.8	17.5	17.9	18.2	9.2	9.4	9.5	9.5
Femmine	10.3	10.2	10.0	9.8	15.0	14.6	14.8	14.1	29.7	30.2	31.0	31.8	16.7	16.6	16.8	16.8
Maschi e Femmine	6.8	6.6	6.6	6.4	10.3	10.3	10.2	10.0	21.0	21.7	22.2	22.8	12.0	12.1	12.3	12.3

Fonte: Indagine ISTAT sulle Forze di Lavoro.

Tab. A.3 - Condizione occupazionale e tassi di attività e di disoccupazione negli anni 1995-1998 per alcuni sottoinsiemi della popolazione (valori assoluti in migliaia di unità)

	Classe di età 20-24				Classe di età 20-24 con diploma [(a)]				Popolazione totale (di 15 anni e oltre)			
	1995	1996	1997	1998	1995	1996	1997	1998	1995	1996	1997	1998
Occupati	1611	1568	1585	1587	537	552	589	640	20010	20088	20086	20197
In cerca di occupaz.	781	775	771	755	374	377	389	383	2725	2763	2805	2837
Non forze lavoro	2079	2107	2103	2070	1394	1447	1471	1505	25229	25196	25253	25306
Tasso di attività (%)	53.5	52.7	52.8	53.1	39.5	39.1	39.9	40.5	47.4	47.6	47.5	47.6
Tasso disoccupaz. (%)	32.7	33.1	32.7	32.2	41.1	40.6	39.8	37.4	12.0	12.1	12.3	12.3

Nota: (a) diploma che consente l'accesso all'Università.

Fonte: Nostre elaborazioni su dati ISTAT (Indagine sulle Forze di Lavoro).

alla classe dei giovani con 20-24 anni *e* possesso di diploma di maturità, osserviamo che il tasso di attività è inferiore (39.5% nel 1995, a fronte del 53.5% relativo all'intera classe 20-24 anni), mentre il tasso di disoccupazione è maggiore (41.1% nel 1995, a fronte del 32.7% relativo all'intera classe 20-24 anni)[6]. Tuttavia, a differenza degli altri tassi riportati in tab. A.3, il tasso di disoccupazione dei giovani diplomati diminuisce in modo lento, ma costante dal 1995 al 1998 (41.1%, 40.6%, 39.8%, 37.4%). La conclusione che si può trarre dai dati della tab. A.3 è che i giovani diplomati incontrano notevoli difficoltà nell'inserimento professionale, sebbene gli ultimi anni mostrino una tendenza al miglioramento.

Alcune situazioni di vantaggio o svantaggio rispetto alle possibilità di occupazione a breve termine possono essere individuate grazie alla tab. A.4, che riporta le percentuali di occupazione ad aprile 1998 delle persone che nell'aprile dell'anno precedente stavano cercando lavoro:

- le donne continuano ad avere più difficoltà dei maschi nella ricerca del lavoro (le percentuali di occupazione sono, rispettivamente, del 24.4% e del 31.6%);

- l'area geografica di residenza svolge un ruolo fondamentale, poiché la percentuale di occupazione va dal 42.4% del Nord Italia al 21.2% del Sud (il Centro ha un valore intermedio, 30.5%);

- per quanto riguarda il titolo di studio, osserviamo che i diplomati, con una percentuale del 27.9%, non godono di alcun vantaggio rispetto alle persone con titolo inferiore (mentre i laureati si trovano in una situazione relativamente privilegiata, con il 41.4%);

- un altro aspetto rilevante del mercato del lavoro è la difficoltà del primo inserimento, testimoniata dal fatto che le possibilità di trovare un impiego per coloro che sono in cerca di prima occupazione sono quasi la metà rispetto a coloro che avevano già lavorato (19.6% contro 37.7%);

- infine, le opportunità di lavoro diminuiscono sensibilmente all'aumentare della lunghezza del periodo di disoccupazione: ad esempio, la percentuale di occupazione è del 41.5% per chi è nella condizione di disoc-

[6]La riduzione del tasso di attività è imputabile al fatto che molti giovani diplomati continuano a studiare in modo esclusivo (cioè restando fuori dalle forze lavoro). Invece l'aumento del tasso di disoccupazione può ascriversi al fatto che i giovani diplomati che fanno parte delle forze lavoro hanno iniziato la ricerca del lavoro da poco tempo e quindi hanno avuto scarse opportunità (considerando anche che, come discuteremo tra breve, il possesso del diploma non è, di per sé, una credenziale sufficiente per un rapido inserimento professionale).

Tab. A.4 - Percentuali di occupazione ad aprile 1998
per le persone in cerca di lavoro ad aprile 1997

Sesso	
Maschi	31.6
Femmine	24.4
Area geografica	
Nord	42.4
Centro	30.5
Sud	21.2
Condizione lavorativa	
In cerca di nuova occupazione	37.7
In cerca di prima occupazione	19.6
Titolo di studio	
Laurea o dipl. univ.	41.4
Diploma	27.9
Media	26.1
Elementare o nessuno	28.2
Mesi di ricerca un anno prima	
Fino a 6 mesi	41.5
Da 7 a 12 mesi	32.1
Da 13 a 24 mesi	22.9
Oltre 24 mesi	19.1

Fonte: ISFOL (1998). Elaborazione ISFOL su dati ISTAT, Fdl, 1997,1998.

Tab. A.5 - Profilo dei nuovi ingressi al lavoro (ad aprile 1998) per coloro che erano in cerca di prima occupazione (ad aprile 1997) e confronto con gli occupati totali (valori percentuali)

Caratteristiche dell'occupazione	Nuovi ingressi (in cerca di prima occup. ad aprile 1997)	Totale occupati 1998
Carattere occupazione		
Permanente	63.9	92.1
Temporaneo	36.1	7.9
Totale	100.0	100.0
Tempo pieno / part-time		
Tempo pieno	83.0	92.7
Part-time	17.0	7.3
Totale	100.0	100.0
Cerca un'altra occupazione		
Sì	27.7	6.0
No	72.3	94.0
Totale	100.0	100.0
Tipo occupazione		
Indipendente	16.0	27.9
Dipendente	84.0	72.1
Totale	100.0	100.0
Settore attività		
Agricoltura	4.8	5.8
Industria	25.0	31.5
Terziario	70.2	62.7
Totale	100.0	100.0

Fonte: ISFOL (1998). Elaborazione ISFOL su dati ISTAT.

cupato da meno di 6 mesi e del 19.1% per chi si trova in tale condizione da più di due anni.

Una lettura congiunta di questi dati sembra suggerire che, ai fini dell'ottenimento di un impiego, le esperienze di lavoro e di disoccupazione sono molto più importanti del possesso di un titolo di studio.

Concludiamo questo paragrafo con un cenno al profilo dei nuovi ingressi al lavoro. A questo proposito, la tab. A.5 riporta le caratteristiche dell'occupazione, rilevate ad aprile 1998, per i nuovi ingressi che nell'aprile dell'anno precedente erano in cerca di *prima* occupazione. L'aspetto più evidente riguarda la maggiore flessibilità dei nuovi entranti rispetto agli occupati storici: infatti, il 36.1% dei nuovi entranti ha un lavoro temporaneo e il 17.0% un part-time, mentre queste percentuali scendono, rispettivamente, al 7.9% e 7.3% se riferite al totale degli occupati del 1998. Questi dati sembrano evidenziare la tendenza delle imprese ad offrire contratti più flessibili, anche se bisogna tener presente che il precariato e la sotto-occupazione rappresentano modalità tipiche dei primi approcci con il lavoro.

Un altro aspetto interessante riguarda il tipo di occupazione dei nuovi ingressi: nel 16.0% dei casi si tratta di un lavoro indipendente, contro il 27.9% del totale degli occupati. Ciò si può ascrivere al fatto che, in molti casi, per mettersi in proprio occorre aver maturato un'esperienza in qualità di lavoratore dipendente.

Infine, a proposito del settore di attività, i nuovi ingressi sono impegnati per il 4.8% nell'agricoltura, per il 25.0% nell'industria e per il 70.2% nel terziario. Dal confronto con la composizione degli occupati totali emerge il ruolo crescente del terziario, che sottrae occupazione agli altri due settori, in particolare all'industria.

Appendice B

Alcune simulazioni relative alla componente di varianza e ai residui

Come discusso nel par. 2.2.6, la stima dei parametri di un modello multilivello non lineare è un'operazione alquanto complessa, il cui risultato non sempre è soddisfacente. Gli studi di simulazione sinora svolti hanno riguardato principalmente le procedure di stima dei modelli lineari, ponendo particolare attenzione al ruolo svolto dalla numerosità campionaria (numero di gruppi e numerosità all'interno dei gruppi)[1]. Più rari sono gli studi di simulazione relativi ai modelli non lineari[2]. In ogni caso, la valutazione dei residui è un aspetto generalmente trascurato, sebbene i residui vengano spesso usati per il controllo delle assunzioni del modello e, soprattutto, per confrontare fra di loro i gruppi (ad esempio costruendo delle graduatorie). Poiché nella presente ricerca i residui sono stati utilizzati per confrontare l'efficacia delle scuole (par. 4.3), abbiamo ritenuto interessante condurre delle simulazioni per valutare alcune proprietà dei residui (oltre che della componente di varianza) forniti dal programma MLwiN (Goldstein *et al.*, 1998) nel caso di un modello logit ad intercetta casuale. A tal fine abbiamo scritto delle apposite macro-istruzioni con i comandi di MLwiN.

I dati impiegati nelle simulazioni sono quelli contenuti nel foglio di lavoro Bes83.ws distribuito con MLwiN. Si tratta di dati costituiti da una variabile di risposta binaria e da quattro covariate continue. Le 800 unità di primo

[1] Una rassegna di questi studi di simulazione è contenuta in Kreft & DeLeeuw (1998), par. 5.4.2.

[2] Per i modelli a risposta binaria ricordiamo Breslow & Clayton (1993), Rodriguez & Goldman (1995) e Goldstein & Rasbash (1996). Per i modelli a risposta politomica segnaliamo il lavoro di Yang (1997).

livello sono raccolte in 110 gruppi, la cui numerosità varia da 1 a 16, con un valore mediano di 7. Si noti che, sebbene il numero di gruppi sia molto minore, la numerosità dei gruppi è analoga a quella del sottocampione utilizzato nelle analisi presentate nei capp. 4 e 5, rendendo plausibile l'estensione a quel contesto dei risultati relativi ai residui.

Il modello usato per le simulazioni è il seguente (cfr. par. 2.2.2):

$$\begin{aligned} & y_{ij} \,|\, u_j \sim Bin(1, \pi_{ij}), \\ & \text{logit}(\pi_{ij}) = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + u_j, \\ & u_j \overset{iid}{\sim} N(0, \sigma_u^2), \end{aligned}$$

dove $j = 1, \ldots, J$ denota i gruppi, $i = 1, \ldots, n_j$ denota le unità elementari, y_{ij} è la variabile di risposta binaria, $\mathbf{x}_{ij}$ è il vettore delle covariate, $\boldsymbol{\beta}$ è il vettore dei parametri fissi, u_j è l'effetto casuale per il j-mo gruppo, σ_u^2 è la componente di varianza (parametro casuale). La simulazione della variabile di risposta y_{ij} a partire da questo modello si compone dei seguenti passi:

1. Scelta dei valori di $\boldsymbol{\beta}$ e di σ_u^2.
2. Campionamento casuale degli effetti casuali $\{u_j : j = 1, \ldots, J\}$ in modo indipendente dalla distribuzione $N(0, \sigma_u^2)$.
3. Calcolo dei valori di π_{ij}.
4. Campionamento casuale dei dati binari $\{y_{ij} : j = 1, \ldots, J;\ i = 1, \ldots, n_j\}$ in modo indipendente da distribuzioni $Bin(1, \pi_{ij})$.

Ogni esercizio di simulazione è costituito da una serie di repliche che indicheremo con $r = 1, \ldots, R$, secondo due varianti:

a) Il passo 1 viene eseguito all'inizio e ogni replica consiste nella ripetizione dei passi da 2 a 4.

b) I passi da 1 a 3 vengono eseguiti all'inizio e ogni replica consiste nella ripetizione del passo 4.

Nel corso di ogni replica, la variabile di risposta generata secondo il procedimento appena descritto viene impiegata per la stima dei parametri fissi e casuali e per il calcolo dei residui. Le quantità stimate che prenderemo in considerazione sono, per ogni replica $r = 1, \ldots, R$, le seguenti:

$$\hat{\sigma}^2_{u,r},\ e\hat{s}(\hat{\sigma}^2_{u,r}),\ \hat{u}_{j,r},\ e\hat{s}c(\hat{u}_{j,r}),$$

dove $e\hat{s}(\hat{\sigma}^2_{u,r})$ è la stima dell'errore standard di $\hat{\sigma}^2_{u,r}$, mentre $e\hat{s}c(\hat{u}_{j,r})$ è la stima dell'errore standard *comparativo* di $\hat{u}_{j,r}$ (cfr. par. 2.1.8).

B.1 Simulazione n. 1

La prima simulazione che abbiamo condotto si basa sulla variante a), che consiste nel campionamento ripetuto sia degli effetti casuali che della variabile di risposta. In questo modo gli effetti casuali variano da replica a replica e ciò permette di fare valutazioni non condizionate a particolari realizzazioni degli effetti casuali. In particolare sono state valutate: 1) l'affidabilità degli intervalli di confidenza dei residui basati sugli errori standard *comparativi*; 2) l'attendibilità delle graduatorie dei gruppi costruite sulla base dei residui. In entrambi i casi si può ipotizzare che l'affidabilità generale aumenti all'aumentare della componente di varianza, σ_u^2, e che l'affidabilità relativa ad uno specifico gruppo sia tanto maggiore quanto più quel gruppo è numeroso. L'ipotesi sul ruolo della componente di varianza è stata verificata ripetendo le simulazioni per diversi valori di σ_u^2, mentre la dipendenza dalla numerosità del gruppo è stata studiata considerando, oltre alla graduatoria generale, anche la graduatoria ristretta al sottoinsieme di gruppi con numerosità superiore al valore mediano.

Allo stesso tempo abbiamo verificato alcuni aspetti relativi alla stima della componente di varianza, che notoriamente rappresenta il punto debole delle procedure di quasi-verosimiglianza (Goldstein & Rasbash, 1996).

I risultati della simulazione sono riportati nella tab. B.1. I valori di $\boldsymbol{\beta}$ usati in tutte le simulazioni sono quelli stimati con la procedura PQL2 (cfr. par. 2.2.6) sui dati originari, mentre i valori di σ_u^2 sono stati scelti di volta in volta (il valore stimato sui dati originari è 0.157). Per ogni valore di σ_u^2 sono state effettuate $R = 100$ repliche, stimando i parametri con il metodo PQL2 (ottenendo sempre la convergenza dell'algoritmo). I valori della tab. B.1 sono stati calcolati come segue:

(i) Componente di varianza stimata media:

$$\frac{1}{R}\sum_{r=1}^{R}\hat{\sigma}_{u,r}^2.$$

(ii) Percentuale di repliche con varianza stimata nulla[3]:

$$\frac{100}{R}\sum_{r=1}^{R} I\{\hat{\sigma}_{u,r}^2 = 0\}.$$

[3] $I\{.\}$ è la funzione indicatrice che vale 1 se e solo se l'evento in parentesi è vero.

(iii) Percentuale di repliche in cui l'ipotesi $\sigma_u^2 = 0$ non viene rifiutata (basandosi sulla distribuzione normale asintotica di $\hat{\sigma}_u^2$):

$$\frac{100}{R}\sum_{r=1}^{R} I\{\hat{\sigma}_{u,r}^2 - 1.96 \cdot e\hat{s}(\hat{\sigma}_{u,r}^2) \leq 0\}.$$

(iv) Percentuale di repliche in cui l'intervallo di confidenza al 95% per la componente di varianza include σ_u^2 (basandosi sulla distribuzione normale asintotica di $\hat{\sigma}_u^2$):

$$\frac{100}{R}\sum_{r=1}^{R} I\{\sigma_u^2 \in (\hat{\sigma}_{u,r}^2 - 1.96 \cdot e\hat{s}(\hat{\sigma}_{u,r}^2),\ \hat{\sigma}_{u,r}^2 + 1.96 \cdot e\hat{s}(\hat{\sigma}_{u,r}^2)\}.$$

(v) Copertura media percentuale dell'intervallo di confidenza al 95% per gli effetti casuali (basandosi sulla distribuzione normale asintotica di $\hat{u}_j$):

$$\frac{1}{J}\sum_{j=1}^{J} c_j,$$

dove c_j è la percentuale di repliche in cui l'intervallo di confidenza al 95% include u_j:

$$\frac{100}{R-k} \sum_{\substack{r=1 \\ \{r:\hat{\sigma}_{u,r}^2 > 0\}}}^{R} I\{u_j \in (\hat{u}_{j,r} - 1.96 \cdot e\hat{s}c(\hat{u}_{j,r}),\ \hat{u}_{j,r} + 1.96 \cdot e\hat{s}c(\hat{u}_{j,r}))\},$$

con k pari al numero di repliche in cui $\hat{\sigma}_{u,r}^2 = 0$ (in tali repliche i residui non possono essere calcolati).

(vi) Indice di cograduazione di Spearman medio fra i ranghi dei residui e degli effetti casuali:

$$\frac{100}{R-k} \sum_{\substack{r=1 \\ \{r:\hat{\sigma}_{u,r}^2 > 0\}}}^{R} s_r,$$

dove s_r è l'indice di Spearman fra i ranghi dei residui della replica r, $\rho_{\hat{u}_{j,r}}$, e i ranghi degli effetti casuali, ρ_{u_j} (comuni a tutte le repliche):

$$1 - \frac{6}{J(J^2-1)}\sum_{j=1}^{J}(\rho_{\hat{u}_{j,r}} - \rho_{u_j})^2.$$

Tab. B.1 – Risultati della simulazione n. 1

Valore della componente di varianza σ^2_u	**0.15**	**0.30**	**0.60**	**1.20**	**2.40**
Componente di varianza stimata media	0.133	0.294	0.578	1.192	2.170
% repliche con componente di varianza stimata nulla	12	2	0	0	0
% repliche in cui l'ipotesi $\sigma^2_u=0$ non viene rifiutata	85	51	7	0	0
% repliche in cui l'intervallo di confidenza al 95% per la componente di varianza include σ^2_u	87	91	88	91	81
Copertura media percentuale dell'intervallo di confidenza al 95% per gli effetti casuali	85.4	91.0	93.0	94.2	93.7
Indice di cograduazione di Spearman medio fra i ranghi dei residui e degli effetti casuali (campo di variazione fra parentesi)	0.38 (0.22-0.56)	0.50 (0.29-0.65)	0.62 (0.45-0.73)	0.73 (0.61-0.87)	0.82 (0.72-0.90)
Indice di cograduazione di Spearman medio fra i ranghi dei residui e degli effetti casuali, relativamente ai gruppi con più di 7 unità (campo di variazione fra parentesi)	0.44 (0.08-0.70)	0.58 (0.23-0.80)	0.70 (0.50-0.85)	0.80 (0.58-0.89)	0.86 (0.73-0.95)

Nota: Numero di repliche R=100.

(vii) Indice di cograduazione di Spearman medio fra i ranghi dei residui e degli effetti casuali, relativamente ai gruppi con più di 7 unità: si procede come al punto precedente, limitando però la sommatoria all'insieme $\{j : n_j > 7\}$.

I risultati della tab. B.1 suggeriscono alcune riflessioni:

- Come ci si attendeva, lo stimatore della componente di varianza è distorto verso il basso. La distorsione più rilevante (11.3%) si ha nel caso $\sigma_u^2 = 0.15$, a causa dell'elevato numero di repliche in cui la stima è nulla. L'altra distorsione rilevante (9.6%) riguarda $\sigma_u^2 = 2.40$ e ciò è coerente con il fatto che altri studi di simulazione hanno individuato una relazione positiva fra entità della distorsione e valore della componente di varianza.

- Quando la componente di varianza è piccola occorre molta cautela, poiché è possibile ottenere stime nulle ed inoltre il test di Wald per l'ipotesi $\sigma_u^2 = 0$ è del tutto fuorviante.

- Il livello di copertura dell'intervallo di confidenza per la componente di varianza è inferiore a quello nominale.

- Il livello di copertura dell'intervallo di confidenza per gli effetti casuali è inferiore a quello nominale, ma di poco. La differenza sembra attenuarsi al crescere della componente di varianza.

- I valori dell'indice di Spearman dimostrano che l'attendibilità della graduatoria stilata sulla base dei residui dipende in modo cruciale dall'entità della componente di varianza: quanto maggiore è σ_u^2, tanto più affidabile è la graduatoria. Per valori di σ_u^2 inferiori a 0.30 l'indice di Spearman non raggiunge lo 0.50.

- I valori dell'indice di Spearman calcolato rispetto ai gruppi con più di 7 unità (valore mediano) indicano che le graduatorie riguardanti gruppi numerosi tendono ad essere più affidabili. Tuttavia, il miglioramento rispetto alla graduatoria generale è abbastanza modesto e nella singola applicazione si può avere addirittura un peggioramento, come si deduce dal confronto dei campi di variazione.

B.2 Simulazione n. 2

La variante b) della procedura di simulazione consente di studiare la distorsione dei residui visti come stimatori degli effetti casuali.

Nel caso dei modelli multilivello lineari, il residuo del gruppo j, condizionatamente a u_j, è esprimibile come (cfr. par. 2.1.9)

$$\hat{u}_j = s(n_j, \tau) \cdot u_j,$$

dove $s(n_j, \tau) = \left(1 + \frac{1}{n_j \tau}\right)^{-1}$ è lo *shrinkage factor* e $\tau = \frac{\sigma_u^2}{\sigma_e^2}$ è il rapporto fra le componenti di varianza. Pertanto la distorsione condizionata di $\hat{u}_j$ è

$$-(1 - s(n_j, \tau)) \cdot u_j,$$

e quindi la distorsione *relativa* condizionata di $\hat{u}_j$ è, in valore assoluto,

$$1 - s(n_j, \tau). \tag{B.1}$$

Tale distorsione è funzione decrescente di τ e di n_j.

Nel caso dei modelli multilivello non lineari non esistono formule che mettono in relazione la distorsione con la componente di varianza e la numerosità del gruppo. Tuttavia è lecito supporre che σ_u^2 e n_j abbiano un effetto analogo a quello che hanno nella (B.1).

Al fine di verificare la presenza di una relazione analoga alla (B.1) abbiamo effettuato delle simulazioni con la variante b), calcolando le seguenti quantità:

(i) Componente di varianza stimata media:

$$\frac{1}{R} \sum_{r=1}^{R} \hat{\sigma}_{u,r}^2.$$

(ii) Percentuale di repliche con varianza stimata nulla[4]:

$$\frac{100}{R} \sum_{r=1}^{R} I\{\hat{\sigma}_{u,r}^2 = 0\}.$$

(iii) Distorsione relativa percentuale media, in valore assoluto, dei residui

$$\frac{1}{J} \sum_{j=1}^{J} d_j,$$

con

$$d_j = 100 \cdot \left| \frac{\bar{\hat{u}}_j - u_j}{u_j} \right|,$$

[4] $I\{.\}$ è la funzione indicatrice che vale 1 se e solo se l'evento in parentesi è vero.

dove

$$\bar{\hat{u}}_j = \frac{1}{R-k} \sum_{\substack{r=1 \\ \{r:\hat{\sigma}^2_{u,r}>0\}}}^{R} \hat{u}_{j,r} \ .$$

Come in precedenza, k è il numero di repliche in cui $\hat{\sigma}^2_{u,r} = 0$ (in tali repliche i residui non possono essere calcolati).

(iv) Percentuale di gruppi per i quali la distorsione dei residui è positiva:

$$\frac{100}{J} \sum_{j=1}^{J} I\{\bar{\hat{u}}_j - u_j > 0\}.$$

(v) Coefficiente della regressione lineare semplice della distorsione relativa percentuale (in valore assoluto) sulla numerosità dei gruppi. Si tratta del coefficiente β della seguente regressione:

$$d_j = \alpha + \beta n_j + \varepsilon_j.$$

Per ogni valore di σ^2_u sono state effettuate 5 simulazioni con $R = 100$ repliche per ciascuna. Le 5 simulazioni differiscono per gli effetti casuali e ciò consente di valutare, in prima approssimazione, quanto la distorsione è legata alla particolare realizzazione degli effetti casuali.

La tab. B.2 riporta i risultati delle simulazioni. I parametri sono stati stimati con l'algoritmo PQL2, che è sempre giunto a convergenza. I risultati suggeriscono le seguenti osservazioni:

- La distorsione relativa in valore assoluto dei residui diminuisce all'aumentare della componente di varianza, confermando l'esistenza di una relazione analoga alla (B.1).

- La relazione fra la distorsione relativa in valore assoluto dei residui e la numerosità dei gruppi è invece meno chiara. Quando la componente di varianza è piccola il coefficiente di regressione, pur essendo negativo, presenta valori modesti e comunque spesso non significativi. Invece con valori elevati della componente di varianza la relazione diviene manifesta, anche se rimane una forte dipendenza dalla particolare realizzazione degli effetti casuali: emblematico è il caso del valore 0.76 ottenuto nella seconda simulazione con $\sigma^2_u = 2.40$.

- La distorsione dei residui tende ad essere negativa, indipendentemente dal valore della componente di varianza. Ciò costituisce un risultato inatteso che necessita, però, di ulteriori approfondimenti.

Tab. B.2 – Risultati della simulazione n. 2

Valore della componente di varianza σ^2_u	**0.15**	**0.30**	**0.60**	**1.20**	**2.40**
Componente di varianza stimata media	0.143 0.126 0.123 0.142 0.088 0.124	0.233 0.203 0.257 0.340 0.347 0.276	0.519 0.419 0.570 0.557 0.597 0.532	1.075 1.366 1.015 0.800 1.138 1.079	2.224 2.263 1.943 1.997 2.063 2.098
% repliche con componente di varianza stimata nulla	11 13 12 16 23 15	5 2 2 1 1 2.2	0 0 0 0 0 0	0 0 0 0 0 0	0 0 0 0 0 0
Distorsione relativa percentuale media in valore assoluto dei residui	87 94 106 86 99 94.4	83 85 94 78 73 82.6	71 68 64 76 59 67.6	51 51 53 53 51 51.8	48 50 41 59 42 48
% gruppi in cui la distorsione dei residui è positiva	48 43 42 45 54 46.4	42 53 40 54 48 47.4	35 51 45 37 47 43	47 57 46 45 45 48	30 71 55 40 45 48.2
Coefficiente di regressione della distorsione relativa percentuale in valore assoluto sulla numerosità dei gruppi	-0.53 -0.94 0.31 -1.36 -2.36 -0.98	-2.39 -2.53* -0.75 -2.27 -3.13* -2.21	-3.73* -3.80* -1.64 -0.52 -0.96 -2.13	-3.56* -4.60* -3.47* -4.44* -2.99 -3.81	-4.39* 0.76 -5.92* -6.46 -3.85* -3.97

Note: Numero di repliche R=100. Ogni cella contiene i risultati di 5 simulazioni e la relativa media.
L'asterisco indica significatività al 95%.

Appendice C

Il questionario dell'indagine Istat sui Percorsi di Studio e Lavoro dei Diplomati

Riteniamo utile riportare integralmente il questionario dell'indagine da cui abbiamo tratto gran parte dei dati elaborati in questo lavoro, poiché alla data in cui scriviamo la relativa pubblicazione dell'Istat non è ancora disponibile. D'altra parte, il questionario è necessario per comprendere le caratteristiche e i limiti dell'insieme informativo utilizzato, che hanno motivato molte delle scelte discusse nei capp. 3-6.

ISTAT

Istituto Nazionale di Statistica

Servizio Istruzione Formazione Lavoro

INDAGINE SUI PERCORSI DI STUDIO E DI LAVORO DEI DIPLOMATI DEL 1995[1]

[1] *Durante le interviste, sia per le domande che per le risposte, non sono state lette le parti in carattere corsivo, ma solo quelle in tondo. Le parti in tondo poste dentro le parentesi dopo le domande sono state lette solo nel caso in cui l'intervistato richiedeva chiarimenti.*

SEZIONE 1: CURRICULUM SCOLASTICO

1. La scuola in cui ti sei diplomato è dello **stesso tipo** di quella a cui ti eri iscritto dopo la Licenza Media?
(Non considerare i cambiamenti che sono avvenuti nell'ambito dello stesso tipo di scuola; per esempio, dall'Istituto tecnico commerciale "Cristoforo" all'Istituto tecnico commerciale "Arduino".)

NO..1 ☐
SI....2 ☐ *(passare al quesito 3)*

2. Dopo la Licenza Media a che tipo di scuola ti eri iscritto?

3. La scuola a cui ti sei iscritto dopo la licenza media era pubblica o privata?

- *pubblica* ... 1 ☐
- *privata* ... 2 ☐

4. Quanti anni avevi quando ti sei iscritto per la prima volta alla scuola secondaria superiore?

Anni ...

5. Durante le scuole superiori hai frequentato scuole per il recupero di anni scolastici?
(Non considerare eventuali ripetizioni private.)

NO..... 1 ☐
SI..... 2 ☐

6. Sempre durante le superiori, sei mai stato respinto?

NO..... 1 ☐
SI..... 2 ☐

7. In quale classe?
*(**Attenzione**: sono possibili più risposte.)*

- *in prima* (in 4ª ginnasio)..... 1 ☐
- *in seconda* (in 5ª ginnasio)..... 2 ☐
- *in terza* (in 1ª liceo classico)... 3 ☐
- *in quarta* (in 2ª liceo classico)... 4 ☐
- *in quinta* (in 3ª liceo classico)... 5 ☐

8. Hai frequentato l'anno integrativo?
(Rispondi a questa domanda se la scuola secondaria superiore in cui hai conseguito la maturità durava quattro anni; altrimenti passa al quesito 9).

NO..... 1 ☐
SI, lo sto frequentando..... 2 ☐
SI, l'ho frequentato... 3 ☐

9. Hai sostenuto l'esame di maturità come privatista?

NO..... 1 ☐
SI..... 2 ☐

SEZIONE 2: CORSI DI FORMAZIONE POST SECONDARIA

10. Dopo il conseguimento del diploma ti sei iscritto a un Corso di Formazione professionale organizzato dalla **Regione**? (Non considerare eventuali corsi privati a pagamento anche se riconosciuti dalla Regione.)

NO.... 1 ☐ *(passare al quesito 15)*
SI.... 2 ☐

11. Lo stai frequentando, lo hai concluso o lo hai interrotto?
(Se hai frequentato più di un corso, fai riferimento a quello che consideri più importante.)

- *si è iscritto, ma non l'ha ancora iniziato*... 1 ☐
- *lo sta frequentando/lo ha appena iniziato*..... 2 ☐
- *lo ha concluso*... 3 ☐
- *lo ha interrotto*... 4 ☐

12. Qual è la durata complessiva dell'intero corso?

- fino a 600 ore *(fino a 6 mesi circa)*..... 1 ☐
- da 601 a 1200 " *(da 6 a 12 mesi ")*..... 2 ☐
- da 1201 a 2400 " *(da 12 a 24 mesi ")*..... 3 ☐
- da 2401 a 3600 " *(da 24 a 36 mesi ")*..... 4 ☐

13. Quando ti eri iscritto a questo corso stavi attivamente cercando un lavoro continuativo?

NO, non lo stavo cercando perché già lavoravo..... 1 ☐ *(passare al quesito 15)*
NO, stavo cercando un lavoro occasionale... 2 ☐
NO, non stavo cercando un lavoro continuativo... 3 ☐
SI, lo stavo cercando... 4 ☐

14. Se avessi potuto scegliere, avresti preferito:

- *frequentare comunque il corso*..... 1 ☐
- *iniziare stabilmente un lavoro*..... 2 ☐
- *frequentare il corso e lavorare insieme*... 3 ☐

15. Dal conseguimento del diploma a oggi hai frequentato o stai frequentando
corsi di formazione privati a pagamento?
Per esempio: di lingue, di informatica, di dattilografia? *(Escludi eventuali corsi di attività artistica o ricreativa.)*
Se hai frequentato o stai frequentando più di un corso, fai riferimento a quello che consideri più importante.)

SI
- di informatica?..... 1 ☐
- di dattilografia/stenografia?..... 2 ☐
- di lingue?..... 3 ☐
- per brevetti sportivi?..... 4 ☐
- *altro*... 5 ☐

NO, nessuno... 6 ☐

SEZIONE 3: STUDI UNIVERSITARI

16.1. Hai conseguito un Diploma Universitario o di Scuola Diretta a Fini Speciali?

NO ..... 1 ☐
SI, un Diploma universitario ..... 2 ☐ *(passare al quesito 18)*
SI, di Scuola Diretta a Fini Speciali ..... 3 ☐ *(passare al quesito 18)*

16.2. Fai uno dei seguenti corsi di studio?
*(Considera anche i corsi per i quali ti stai preparando all'iscrizione. **Attenzione**: sono possibili più risposte.)*

SI →

- L'Accademia di belle arti?	01 ☐	
- L'Isef?	02 ☐	→ *(passare al quesito 17)*
- Il Conservatorio?	03 ☐	
- Una delle Scuole di FF.AA./Polizia *(esclusi i corsi di addestramento per militari di leva)*	04 ☐	
- Un Corso di Laurea?	05 ☐	
- Un Corso di Diploma universitario? *(escluso Isef)*	06 ☐	→ *(passare al quesito 18)*
- Una Scuola Diretta a Fini Speciali?	07 ☐	
NO, nessuno di questi	08 ☐	*(passare al quesito 17)*

17. Qual è il motivo principale per cui non sei iscritto all'Università?
*(**Attenzione**: è possibile una sola risposta.)*

- *mi sono iscritto/mi iscriverò nell'a.a. 1998/99*	01 ☐	
- *non mi piace studiare*	02 ☐	
- *mi ero iscritto, ma successivamente ho rinunciato*	03 ☐	
- *preferisco frequentare altri corsi*	04 ☐	→ *(passare al quesito 18)*
- *per il costo troppo elevato*	05 ☐	
- *preferisco lavorare*	06 ☐	
- *per motivi personali (salute, matrimonio, assistenza figli, ...)*	07 ☐	
- *per obblighi di leva*	08 ☐	
- *altro*	09 ☐	

18. In che regione si trova l'Ateneo in cui sei iscritto?
(Si riferimento all'effettiva sede in cui ero iscritto e non all'eventuale sede madre.
***Attenzione**: se studi all'estero passa al quesito 33)*

19. In che Università in particolare?
(Scrivi il nome dell'Ateneo)

20. A quale Corso di Laurea/Diploma universitario/Scuola Diretta a Fini Speciali sei iscritto/hai conseguito?
(Oppure se alla domanda 16.1 hai risposto di avere un Diploma Universitario o di una Scuola Diretta a Fini Speciali (modalità 2 o 3), proponi la domanda: "Quale Diploma Universitario/Scuola Diretta a Fini Speciali hai conseguito?" e passa alla domanda 24)

21. Hai dato esami in quel corso? Escludi quelli eventualmente sostenuti dopo l'estate.
(Escludi gli esami eventualmente sostenuti in un altro corso. Le integrazioni valgono come un esame sostenuto nel corso attuale. Le eventuali prove di accertamento o di idoneità, come ad es. lingue e informatica, non vanno considerate.)

NO ..... 1 ☐ *(passare al quesito 24)*
SI ..... 2 ☐

22. Quanti ne avevi superati complessivamente fino al 31 luglio?
(Escludi gli esami eventualmente sostenuti in un altro corso. Le integrazioni valgono come un esame sostenuto nel corso attuale. Le eventuali prove di accertamento o di idoneità, come ad es. lingue e informatica, non vanno considerate.
Se l'intervistato ha superato solo alcune prove (o moduli) di uno stesso esame, non conteggiare l'esame tra quelli superati.
Gli esami organizzati in più prove (o moduli) vanno conteggiati come se fossero un unico esame.)

N° esami superati | |

23. In che mese e anno hai superato l'ultimo esame? Ti ricordo ancora di escludere quelli eventualmente sostenuti dopo l'estate.

Mese | |
Anno 19 | |

24. Nell'anno accademico 1997/98, durante il periodo delle lezioni, quante volte a settimana le hai frequentate?

- *mai/quasi mai* ..... 1 ☐
- *uno due giorni a settimana* ..... 2 ☐
- *tre o più giorni a settimana* ..... 3 ☐
- *non ero ancora iscritto* ..... 4 ☐

25. Da quando ti sei immatricolato ad oggi, hai/avevi frequentato le lezioni di almeno un insegnamento fondamentale?

NO ..... 1 ☐ *(passare al quesito 28)*
SI, meno di dieci lezioni ..... 2 ☐ *(passare al quesito 28)*
SI, dieci o più lezioni ..... 3 ☐

26. Pensa all'ultimo insegnamento fondamentale che hai frequentato. Puoi dirmi in che misura sei/eri soddisfatto del docente titolare del corso, rispetto a:
(Se hai frequentato più esami fondamentali contemporaneamente, fai riferimento al primo di cui hai sostenuto o sosterrai l'esame.)

	Molto	*Abbastanza*	*Poco*	*Per niente*	*Non ho avuto contatti*
DOCENTE TITOLARE					
- la sua competenza sulla materia?	01 ☐	02 ☐	03 ☐	04 ☐	05 ☐
- la chiarezza espositiva?	06 ☐	07 ☐	08 ☐	09 ☐	
- la presenza/puntualità agli orari di ricevimento?	10 ☐	11 ☐	12 ☐	13 ☐	
- la presenza/puntualità alle lezioni?	14 ☐	15 ☐	16 ☐	17 ☐	
- la disponibilità al rapporto con gli studenti?	18 ☐	19 ☐	20 ☐	21 ☐	

(Se il codice della risposta è 05, passa immediatamente alla domanda 27 senza rispondere alle ulteriori domande.)

27. Adesso ti leggerò una breve lista di attrezzature didattiche della Facoltà. Puoi dirmi in che misura sei soddisfatto rispetto a:

	Molto	*Abbastanza*	*Poco*	*Per niente*	*Non esistono*
DOCENTE TITOLARE					
- le aule di lezione ?	01 ☐	02 ☐	03 ☐	04 ☐	05 ☐
- gli ambienti di studio/sale di lettura?	06 ☐	07 ☐	08 ☐	09 ☐	10 ☐
- le biblioteche?	11 ☐	12 ☐	13 ☐	14 ☐	15 ☐
- i laboratori linguistici?	16 ☐	17 ☐	18 ☐	19 ☐	20 ☐
- le aule/laboratori informatici ?	21 ☐	22 ☐	23 ☐	24 ☐	25 ☐
- i laboratori di altro tipo?	26 ☐	27 ☐	28 ☐	29 ☐	30 ☐

28. Nell'Anno Accademico 1997/98 hai presentato domanda di borsa di studio presso l'Università o presso enti universitari per il diritto allo studio?

NO ..... 1 ☐ *(passare al quesito 30)*
SI ..... 2 ☐
- non ero ancora iscritto .. 3 ☐ *(passare al quesito 32)*

29. Sei risultato idoneo?

NO ..... 1 ☐ *(passare al quesito 31)*
SI ..... 2 ☐
Non so ..... 3 ☐ ↓

30. L'hai ottenuta?

NO ..... 1 ☐
SI ..... 2 ☐
Solo in parte ..... 3 ☐

31. Nell'anno accademico 1997/98 sei stato esonerato in tutto o in parte dal pagamento delle tasse?

NO ..... 1 ☐
SI, parzialmente ..... 2 ☐
SI, del tutto ..... 3 ☐
Non so ..... 4 ☐

*(**Attenzione**: se alla domanda 16.1 hai risposto di avere un Diploma Universitario o di Scuola Diretta a Fini Speciali (modalità 2 o 3) passa alla domanda 33.)*

32. Ti sei immatricolato all'Università subito dopo il conseguimento del diploma?
(Si riferimento alla tua prima iscrizione, anche se successivamente hai cambiato corso o sede.)

NO ..... 1 ☐
SI ..... 2 ☐ *(passare al quesito 34)* ↓

33. Quanti anni dopo?

- *dopo un anno*	1 ☐
- *dopo due anni*	2 ☐ *(passare al quesito 35)*
- *dopo tre anni si è appena iscritto*	3 ☐ *(passare al quesito 41)*

34. Al 31 luglio eri iscritto in corso, ripetente o fuori corso?

- *regolarmente in corso* ..... 1 ☐
- *ripetente* ..... 2 ☐
- *fuori corso* ..... 3 ☐

35. Il corso in cui sei iscritto è lo stesso in cui ti sei immatricolato per la prima volta?
(Non considerare i cambiamenti di sede che non abbiano comportato un cambiamento di corso di studi.
Se ha cambiato corso dopo l'estate rispondere SI.)

NO ..... 1 ☐
SI ..... 2 ☐ *(passare al quesito 41)* ↓

36. Ti eri immatricolato a a un Corso di Laurea, a un corso di Diploma universitario, o a una Scuola Diretta a Fini Speciali?

- *Corso di Laurea* ..... 1 ☐
- *Diploma universitario* ..... 2 ☐
- *Scuola Diretta a Fini Speciali* ..... 3 ☐

37. Qual era la denominazione del corso?

38. Avevi dato esami in quel corso?
(Le eventuali prove di accertamento o di idoneità, come ad es. lingue e informatica, non vanno considerate.)

NO ..... 1 ☐ *(passare al quesito 41)*
SI ..... 2 ☐

39. Quanti ne avevi superati?
(Le eventuali prove di accertamento o di idoneità, come ad es. lingue e informatica, non vanno considerate.
Se aveva superato solo alcune prove (o moduli) di uno stesso esame, non conteggiare l'esame tra quelli superati.
Gli esami organizzati in più prove (o moduli) vanno conteggiati come se fossero un unico esame.)

N° esami superati | |

40. Quanti te ne hanno riconosciuti nel corso in cui sei attualmente iscritto?
(Le eventuali prove di accertamento o di idoneità, come ad es. lingue e informatica, non vanno considerate.)

N° esami riconosciuti | | |

41. Ti sei immatricolato per la prima volta nello stesso Ateneo in cui sei iscritto?

NO ..... 1 ☐
SI ..... 2 ☐ *(passare al quesito 44)*

42. In che regione si trovava l'Ateneo in cui eri iscritto?
(Si riferimento all'effettiva sede in cui eri iscritto/non all'eventuale sede madre.
***Attenzione**: se studi all'estero passa al quesito 53)*

43. In che Università in particolare?
(Scrivi il nome dell'ateneo)

44. Quando ti sei iscritto all'Università, stavi attivamente cercando un lavoro continuativo?

NO, non lo stavo cercando perché già lavoravo ..... 1 ☐ *(passare al quesito 53)*
NO, stavo cercando un lavoro occasionale ..... 2 ☐
NO, non stavo cercando un lavoro continuativo ..... 3 ☐
SI, lo stavo cercando ..... 4 ☐

45. Se avessi potuto scegliere, avresti preferito:

- *proseguire gli studi*..... 1 ☐
- *trovare stabilmente un lavoro*.... 2 ☐
- *studiare e lavorare insieme*..... 3 ☐

Passare al quesito 53

SEZIONE 4: INTERRUZIONE DEGLI STUDI

46. Dopo il conseguimento del diploma hai interrotto:
*(**Attenzione**: una possibile risposta)*

- L'Accademia di belle arti?..... 01 ☐
- L'Isef?..... 02 ☐
- Il Conservatorio?..... 03 ☐
- Una delle Scuole di FF.AA./Polizia (escluso i corsi di addestramento per militari di leva)?..... 04 ☐
- Un Corso di Laurea?..... 05 ☐
- Un Corso di Diploma universitario?(escluso l'Isef)... 06 ☐
- Una Scuola Diretta a Fini Speciali?..... 07 ☐

- *NO, nessuno di questi* ..... 08 ☐ *(passare al quesito 55)*

47. Qual è il motivo principale per cui hai interrotto gli studi ?
*(**Attenzione**: è possibile una sola risposta)*

- *gli studi erano troppo difficili non mi piacevano*..... 01 ☐
- *per frequentare altri corsi* ..... 02 ☐
- *il corso intrapreso offriva poche sbocchi per il futuro* ..... 03 ☐
- *gli studi erano troppo costosi* ..... 04 ☐
- *per lavoro* ..... 05 ☐
- *per motivi personali (salute, matrimonio, assistenza figli o altri famigliari, ecc.)* ..... 06 ☐
- *per obblighi di leva* ..... 07 ☐
- *altro* ..... 08 ☐

48. Ti eri immatricolato all'Università subito dopo il conseguimento del diploma?

NO..... 1 ☐
SI..... 2 ☐ *(passare al quesito 50)*

49. Quanti anni dopo?

- *dopo un anno*..... 1 ☐
- *dopo due anni*..... 2 ☐

50. Avevi dato qualche esame nel corso che hai interrotto?
*(**Attenzione**: le cosiddette prove di accertamento o di idoneità, come ad es. lingue e informatica, non vanno considerate)*

NO.... 1 ☐ *(passare al quesito 52)*
SI..... 2 ☐

51. Quanti ne avevi superati?
*(**Attenzione**: le cosiddette prove di accertamento o di idoneità, come ad es. lingue e informatica, non vanno considerate*
Se avevi superato solo alcune prove (o moduli) di uno stesso esame, non conteggiare l'esame tra quelli superati
Gli esami organizzati in più prove (o moduli) vanno conteggiati come se fossero un unico esame.)

N° esami superati : :

52. Avevi frequentato le lezioni di almeno un insegnamento fondamentale?

NO..... 1 ☐ *(passare al quesito 55)*
SI, meno di dieci lezioni.. 2 ☐ *(passare al quesito 55)*
SI, dieci o più lezioni..... 3 ☐

53. Pensa all'ultimo insegnamento fondamentale che hai frequentato. Puoi dirmi in che misura eri soddisfatto del docente titolare del corso, rispetto a:
(se hai frequentato più esami fondamentali contemporaneamente,fai riferimento al primo di cui hai sostenuto o avresti voluto sostenere l'esame.)

	Molto	*Abbastanza*	*Poco*	*Per niente*	*Non ho avuto contatto*
DOCENTE TITOLARE					
- la sua competenza sulla materia,?....	01 ☐	02 ☐	03 ☐	04 ☐	05 ☐
- la chiarezza espositiva?.....	06 ☐	07 ☐	08 ☐	09 ☐	
- la presenza/puntualità agli orari di ricevimento?.....	10 ☐	11 ☐	12 ☐	13 ☐	
- la presenza/puntualità alle lezioni?...	14 ☐	15 ☐	16 ☐	17 ☐	
- la disponibilità al rapporto con gli studenti?.....	18 ☐	19 ☐	20 ☐	21 ☐	

(Se il codice della risposta è 05 passa immediatamente alla domanda 54 senza rispondere alle ulteriori domande.)

54. Adesso ti leggerò una breve lista di attrezzature didattiche della Facoltà. Puoi dirmi in che misura sei soddisfatto rispetto a:

	Molto	*Abbastanza*	*Poco*	*Per niente*	*Non esistono*
DOCENTE TITOLARE					
- le aule di lezione ?.....	01 ☐	02 ☐	03 ☐	04 ☐	05 ☐
- gli ambienti di studio/sale di lettura?	06 ☐	07 ☐	08 ☐	09 ☐	10 ☐
- le biblioteche?.....	11 ☐	12 ☐	13 ☐	14 ☐	15 ☐
- i laboratori linguistici?.....	16 ☐	17 ☐	18 ☐	19 ☐	20 ☐
- le aule/laboratori informatici ?.....	21 ☐	22 ☐	23 ☐	24 ☐	25 ☐
- i laboratori di altro tipo?	26 ☐	27 ☐	28 ☐	29 ☐	30 ☐

SEZIONE 5: LAVORO

55. Dal momento in cui hai conseguito il diploma di scuola secondaria superiore ad oggi hai avuto qualche opportunità di lavoro continuativo che hai rifiutato?
*(**Attenzione**: fai riferimento a eventuali lavori continuativi anche se non regolarizzati, a quelli a tempo indeterminato, a quelli con contratto a termine, con contratto di formazione e lavoro e con contratto di apprendistato.)*

NO..... 1 ☐ *(passare al quesito 57)*
SI..... 2 ☐

56. Qual è il motivo principale per cui hai rifiutato?
*(**Attenzione**: è possibile una sola risposta)*

- *avevo già un lavoro ero in attesa di un altro lavoro*..... 01 ☐
- *non mi piaceva quel tipo di lavoro*..... 02 ☐
- *non mi garantiva stabilità e/o sicurezza*..... 03 ☐
- *non ero soddisfatto del trattamento economico*..... 04 ☐
- *volevo studiare*..... 05 ☐
- *avevo impegni familiari o personali che non mi consentivano di accettarla (servizio militare, accudimento figli o parenti, ecc.)* ..... 06 ☐
- *Altro*..... 07 ☐

57. Dopo il conseguimento del diploma hai iniziato uno o più lavori che successivamente hai interrotto?

NO..... 1 ☐ *(passare al quesito 60)*
SI..... 2 ☐

58. Di che tipo di lavoro si trattava, continuativo, stagionale od occasionale?
(Se hai interrotto più di un lavoro, fai riferimento al primo.)

Continuativo..... 1 ☐
Stagionale..... 2 ☐
Occasionale..... 3 ☐

59. Quando lo hai iniziato?

Mese : :
Anno 19 : :

60. **Attualmente** svolgi un'attività lavorativa retribuita? Ti segnalo che tirocinii, stage o borse non vengono considerati lavori, anche se retribuiti.

- *NO*..... 1 ☐
- *SI*..... 2 ☐ *(passare al quesito 63)*

61. Anche se attualmente non sei occupato, la settimana scorsa hai effettuato qualche ora di lavoro retribuito?

NO..... 1 ❑ *(passare al questo 63)*
SI..... 2 ❑

62. Quante?

N°ore : :

Passare al questo 65.

63. Cerchi lavoro?

NO..... 1 ❑
SI..... 2 ❑ *passare al questo 92*

64. Qual è il motivo principale per cui non cerchi lavoro?

- *Voglio proseguire gli studi*..... 01 ❑
- *Non trovo lavori che mi interessano*..... 02 ❑
- *Sono in attesa di un concorso*..... 03 ❑
- *Sto facendo uno stage presso una Impresa (Ente pubblico,)*..... 04 ❑
- *Collaboro ad un'attività familiare*..... 05 ❑
- *per motivi personali (salute, matrimonio, assistenza altri familiari,)*..... 06 ❑
- *per obblighi di leva*..... 07 ❑
- *altro*..... 08 ❑

Passare al quesito 99

65. E' un lavoro occasionale, stagionale o continuativo?

*(**Attenzione**: il lavoro continuativo include anche i contratti a termine e a progetto (contratti di formazione e lavoro).)*

- *occasionale (di tanto in tanto)*..... 1 ❑ *(passare al questo 91)*
- *stagionale*..... 2 ❑ *(passare al questo 91)*
- *continuativo (compresa libera professione e il lavoro autonomo)*..... 3 ❑

66. In che modo hai trovato il tuo attuale lavoro?
*(**Attenzione**: è possibile una sola risposta.*
Indica una sola risposta, facendo riferimento al modo che ritieni più importante)

- *su segnalazione o dietro di lavoro da parte di familiari, amici o conoscenti*..... 01 ❑
- *per conoscenza diretta del datore di lavoro*..... 02 ❑
- *per chiamata diretta dell'azienda*..... 03 ❑
- *con inserzioni sui giornali / rispondendo a offerte di lavoro pubblicate sui giornali/ su Internet*..... 04 ❑
- *inviando domande ai datori di lavoro (presentandomi di persona, telefonando,)*..... 05 ❑
- *per pubblico concorso*..... 06 ❑
- *iniziando un'attività autonoma (da solo o con altri)*..... 07 ❑
- *collaborando ad un attività familiare*..... 08 ❑
- *attraverso l'iscrizione presso un ufficio pubblico di collocamento*..... 09 ❑
- *attraverso agenzie di collocamento specializzate*..... 10 ❑
- *altro*..... 11 ❑

67. Il lavoro che stai svolgendo è iniziato prima o dopo il conseguimento del diploma di scuola secondaria superiore?

Prima..... 1 ❑ *(passare al questo 69)*
Dopo..... 2 ❑

68. Quando lo hai iniziato?

Mese : :

Anno 19 : :

69. Per accedere al tuo attuale lavoro possedevi un diploma di scuola secondaria superiore era requisito necessario?

NO..... 1 ❑ *(passare al questo 71)*
SI..... 2 ❑

70. Era necessario un diploma qualsiasi o di uno specifico indirizzo?

- *un diploma qualsiasi*..... 1 ❑
- *di uno specifico indirizzo*..... 2 ❑

71. Per svolgere il tuo attuale lavoro hai dovuto cambiare città?
*(**Attenzione**: i pendolari devono barrare no.)*

NO..... 1 ❑
SI..... 2 ❑

72. Lavori a tempo pieno o part-time *(a tempo parziale)*?

- *a tempo pieno*..... 1 ❑ *(passare al questo 74)*
- *part-time*..... 2 ❑

73. Lavori part-time per mancanza di altre opportunità o per scelta?

- *per mancanza di altre opportunità*..... 1 ❑
- *per scelta*..... 2 ❑

74. Quante sono le ore di lavoro che svolgi abitualmente in una settimana?
(Includi, se sono abituali, eventuali ore di straordinario.)

N. : :

75. Qual è il tuo guadagno mensile lordo e quello netto per questo lavoro? Ti ricordi che le risposte sono coperte dal segreto statistico
(Considera la media degli ultimi tre mesi.
Se non ti ricordi esattamente, indica quello di cui sei più sicuro.)

- *guadagno mensile lordo* LIRE 000
- *guadagno mensile netto* LIRE 000

76. Descrivi quali sono i compiti principali che svolgi nel tuo lavoro, senza usare termini generici come impiegato o operaio. *(Attenzione: è essenziale che indichi il luogo dove esercita la tua attività, ad es. **cassiere in banca**, **infermiera a domicilio**, **ragioniere contabile in azienda**, ecc.)*

_____
_____
_____

77. Sei un:

- Lavoratore dipendente?..... 1 ❑ *(passare al questo 80)*
- Consulente collaboratore?..... 2 ❑ *(passare al questo 79)*
- Lavoratore indipendente?..... 3 ❑ *(passare al questo 78)*

78. Quale tra le seguenti voci descrive meglio il tuo lavoro?
*(**Attenzione**: è possibile una sola risposta)*

- Commerciante..... 01❑
- Artigiano..... 02❑
- Coltivatore diretto..... 03❑
- Collaboratore ad un'attività autonoma famigliare..... 04❑
- Lavoratore autonomo senza qualificazione *(ad es. baby sitter, tradizionatore)*..... 05❑
- Socio di una cooperativa..... 06❑
- Imprenditore..... 07❑
- Libero professionista: *(ad es. infermiere a domicilio,)*..... 08❑
- *Altro*..... 09❑

79. Hai una partita IVA?

NO..... 1 ❑
SI..... 2 ❑

Passare al quesito 87

80. Lavori con contratto di formazione e lavoro?

NO..... 1 ❑
SI..... 2 ❑ *(passare al questo 83)*

81. Lavori con contratto a termine?

NO..... 1 ❑
SI..... 2 ❑

82. Il tuo datore di lavoro versa regolarmente i contributi per la tua pensione?

NO..... 1 ❑
SI..... 2 ❑

83. Da quando hai iniziato questa attività hai partecipato a corsi di formazione professionale organizzati dall'azienda o comunque dal tuo datore di lavoro?

- *NO*..... 1 ❑ *(passare al questo 85)*
- *SI*..... 2 ❑

84. Quanti giorni sono durati in tutto?
(Fai riferimento a tutti i corsi organizzati dall'azienda o dal datore di lavoro dall'inizio dell'attuale lavoro.)

- *Meno di 1 settimana/1 settimana*..... 1 ❑
- *2 settimane*..... 2 ❑
- *3 settimane*..... 3 ❑
- *4 settimane*..... 4 ❑
- *più di 4 settimane*..... 5 ❑

85. Tra le seguenti voci quale descrive meglio il tuo lavoro?
*(**Attenzione**: è possibile una sola risposta)*

- Quadro, funzionario. *(inclusi direttivi e ufficiali)*..... 01❑
- Insegnante..... 02❑
- Tecnico o impiegato ad alta/media qualificazione. *(es. analisti di dati, geometri e periti tecnici, impiegati amministrativi, capi segreteria, capi commessi, infermieri professionali, ecc.)*..... 03❑
- Impiegato esecutivo *(es. segretarie, addetti agli sportelli, telefonisti, ecc.)*..... 04❑
- Militare di carriera. *(delle FF.AA. o di Polizia, inclusi i sottufficiali)*..... 05❑
- Capo operaio o operaio qualificato..... 06❑
- Collaboratore ad un'attività autonoma famigliare:
  con contratto..... 07❑
  o senza contratto..... 08❑
- Lavoratore senza qualificazione nell'agricoltura e nell'industria *(ex manovali edili, braccianti ecc.)*..... 09❑
- Lavoratore senza qualificazione in altre attività. *(ex pony express, commessi, shampisti, camerieri, tecnici, custodi, bidelli, pulitori, benzinai, puntatori, ecc.)*..... 10❑
- Lavoratore nel proprio domicilio per conto di imprese, un apprendista..... 11❑
- *Altro*..... 12❑

86. Svolgi il tuo lavoro nel settore privato o pubblico?

- *privato*..... 1 ❑
- *pubblico*..... 2 ❑

87. *Attenzione se sei un lavoratore indipendente* In quale settore eserciti la tua professione?
se sei un lavoratore dipendente In quale settore opera l'impresa o l'amministrazione presso cui lavori?

- Agricoltura, Caccia e Pesca ..... 1 ☐ *(passare al quesito 90)*
- Industria ..... 2 ☐ *(passare al quesito 90)*
- Altre Attività ..... 3 ☐

88. Più in particolare:
(Attenzione è possibile una sola risposta)

- Nel Commercio, Alberghi e pubblici esercizi ..... 01 ☐
- Nei trasporti, viaggi e comunicazioni ..... 02 ☐
- Nel Credito, Assicurazioni, ecc ..... 03 ☐
- Nelle Attività immobiliari, nel Noleggio ..... 04 ☐
- Nella pubblicità, pubbliche relazioni, giornalismo ..... 05 ☐
- Nella Pubblica amministrazione/Difesa ..... 06 ☐ *(passare al quesito 90)*
- Nell'Informatica e attività connesse ..... 07 ☐
- Nell'Istruzione e formazione ..... 08 ☐
- Nella Sanità e assistenza sociale ..... 09 ☐
- Nei Servizi ricreativi, sportivi e culturali ..... 10 ☐
- In Altri servizi ..... 11 ☐

89. Quante persone, oltre te, lavorano abitualmente nell'impresa, ente o studio nel quale svolgi la tua attività?
(Il riferimento è a TUTTE le persone che vi lavorano, sia della sede madre, sia di eventuali altre sedi)

- nessuno oltre me ..... 01 ☐
- da 1 a 5 persone ..... 02 ☐
- da 6 a 14 ..... 03 ☐
- da 15 a 49 ..... 04 ☐
- da 50 a 99 ..... 05 ☐
- 100 e oltre ..... 06 ☐

90. Rispetto alle mansioni che svolgi, ritieni che avere un diploma di scuola secondaria superiore sia:

- eccessivo ..... 1 ☐
- sufficiente ..... 2 ☐
- insufficiente ..... 3 ☐

91. Attualmente cerchi un nuovo lavoro?

NO ..... 1 ☐ *(passare al quesito 99)*
SI ..... 2 ☐

SEZIONE 6: RICERCA DI LAVORO

92. Quanti mesi fa hai preso l'ultima iniziativa concreta per cercare lavoro? Per esempio, rispondendo a inserzioni sui giornali, prendendo contatti con datori di lavoro, partecipando a un concorso, iscrivendoti presso l'Ufficio di collocamento?

- *non ho ancora preso nessuna iniziativa* ..... 1 ☐
- *negli ultimi trenta giorni* ..... 2 ☐
- *da più di un mese a sei mesi fa* ..... 3 ☐
- *oltre sei mesi fa* ..... 4 ☐

93. In questo momento preferiresti un lavoro occasionale, stagionale o continuativo?

- *occasionale* ..... 1 ☐
- *stagionale* ..... 2 ☐
- *continuativo* ..... 3 ☐

94.1. Preferibilmente, adesso vorresti lavorare a tempo pieno, part-time, oppure con qualsiasi orario?

- *a tempo pieno* ..... 1 ☐
- *part-time* ..... 2 ☐
- *con qualsiasi orario* ..... 3 ☐

94.2. Vorresti lavorare come dipendente o in modo autonomo?

- *dipendente* ..... 1 ☐
- *autonomo* ..... 2 ☐
- *non ho preferenze* ..... 3 ☐

95. Adesso, dove saresti disposto a lavorare?

- Ovunque, sia in Italia che all'estero ..... 1 ☐ *(passare al quesito 97)*
- Solo in Italia? ..... 2 ☐

96. Per lavorare saresti disposto a cambiare città?

NO ..... 1 ☐
SI ..... 2 ☐

97. Qual è la cifra **minima** che saresti disposto ad accettare mensilmente **al netto** per un lavoro come quello che hai appena descritto nelle 5 domande precedenti?

Lire mensili nette : : : : 000

98. Se trovassi un lavoro con le caratteristiche appena indicate potresti iniziarlo entro le prossime due settimane o ci sono dei motivi per cui dovresti rinviare?

NO, ci sono dei motivi per cui dovrei rinviare ..... 1 ☐
SI, potrei iniziarlo entro le prossime due settimane ..... 2 ☐

SEZIONE 7: NOTIZIE SULLA FAMIGLIA

99. Hai fratelli o sorelle?

NO ..... 1 ☐ *(passare al quesito 101)*
SI ..... 2 ☐

100. Quanti?

- *uno* ..... 1 ☐
- *due* ..... 2 ☐
- *tre* ..... 3 ☐
- *quattro o oltre* ..... 4 ☐

101. Vivi abitualmente in casa dei tuoi genitori (o di chi ne fa le veci)?
(Con riferimento alla situazione prevalente durante l'anno
Per "abitualmente" si intende 4 giorni a settimana)

NO ..... 1 ☐
SI ..... 2 ☐ *(passare al quesito 106)*

102. Vivi nella stessa città dei tuoi genitori *(o di chi ne fa le veci)*, in altra città o all'estero?

- *nella stessa città* ..... 1 ☐
- *in altra città* ..... 2 ☐
- *all'estero* ..... 3 ☐

103. Qual è il principale motivo per cui non vivi abitualmente in casa dei tuoi genitori *(o di chi ne fa le veci)*?

- *per lavoro* ..... 1 ☐ *(passare al quesito 106)*
- *per motivi di studio* ..... 2 ☐
- *sono sposato/a - convivo* ..... 3 ☐ *(passare al quesito 106)*
- *altro* ..... 4 ☐

104. Vivi in una Casa dello Studente?

NO ..... 1 ☐
SI ..... 2 ☐

105. Nel posto in cui vivi paghi un affitto?

NO ..... 1 ☐
SI ..... 2 ☐

106. I tuoi genitori (o chi ne fa le veci) provvedono al tuo mantenimento:

- totalmente o quasi totalmente ..... 1 ☐
- più o meno per i tre quarti ..... 2 ☐
- per circa la metà ..... 3 ☐
- per meno della metà ..... 4 ☐
- per niente o quasi per niente ..... 5 ☐

107. Quando avevi 14 anni, qual era il titolo di studio di tuo padre?
(Indicalo anche se tuo padre è/era già deceduto)

- *analfabeta/senza titolo* ..... 01 ☐
- *licenza elementare* ..... 02 ☐
- *licenza media/avviamento professionale* ..... 03 ☐
- *qualifica professionale (2-3 anni)* ..... 04 ☐
- *diploma di scuola media superiore (4-5 anni)* ..... 05 ☐
- *diploma universitario o ex Scuole parauniversit.* ..... 06 ☐
- *laurea o dottorato di ricerca* ..... 07 ☐

108. E quello di tua madre?
(Indicalo anche se tua madre è/era già deceduta)

- *analfabeta/senza titolo* ..... 01 ☐
- *licenza elementare* ..... 02 ☐
- *licenza media/avviamento professionale* ..... 03 ☐
- *qualifica professionale (2-3 anni)* ..... 04 ☐
- *diploma di scuola media superiore (4-5 anni)* ..... 05 ☐
- *diploma universitario o ex Scuole parauniversit.* ..... 06 ☐
- *laurea o dottorato di ricerca* ..... 07 ☐

109. Uno dei tuoi nonni ha o aveva un diploma di scuola secondaria superiore o una laurea?

- *NO* ..... 1 ☐
- *SI* ..... 2 ☐

110. Sempre quando avevi 14 anni, tuo padre era:
(Attenzione indicare il codice 4 se all'epoca deceduto)

- occupato ..... 1 ☐
- in cerca di occupazione ..... 2 ☐
- pensionato ..... 3 ☐
- *altra condizione* ..... 4 ☐

111. Descrivi quali erano i compiti principali che tuo padre svolgeva nel suo lavoro, senza usare termini generici come impiegato o operaio.
*(Attenzione è essenziale che indichi il luogo dove il padre svolgeva la sua attività, ad es. scrivere **in banca**, informare **a domicilio**, ragioniere contabile **in azienda**, ecc)*
Se non ricordi il lavoro che tuo padre svolgeva quando aveva 14 anni, indicare comunque quello attuale
Se deceduto, pensionato o disoccupato, fai riferimento all'ultima situazione lavorativa)

112. Tuo padre è/era un:

- Lavoratore dipendente? ..... 1 ☐ *(passare al quesito 114)*
- Coniuente collaboratore? ..... 2 ☐ *(passare al quesito 115)*
- Lavoratore indipendente? ..... 3 ☐ *(passare al quesito 113)*

113. Quale tra le seguenti voci descrive meglio il lavoro di tuo padre?
(Attenzione è possibile una sola risposta)

- Commerciante ..... 01❑
- Artigiano ..... 02❑
- Coltivatore diretto ..... 03❑
- Collaboratore ad un'attività autonoma famigliare ..... 04❑
- Lavoratore autonomo senza qualificazione (*ad es. baby sitter, traslocatore*) ..... 05❑
- Socio parte di una cooperativa ..... 06❑
- Imprenditore ..... 07❑
- Libero professionista ..... 08❑
- *Altro* ..... 09❑

Passare al quesito 115

114. Quale tra le seguenti voci descrive meglio il lavoro di tuo padre?
(Attenzione è possibile una sola risposta)

- Dirigente *(inclusi i medici)* ..... 01❑
- Docente universitario *(I e II fascia)* ..... 02❑
- Quadro, funzionario *(inclusi direttivi e ufficiali FF.AA.)* ..... 03❑
- Ricercatore ..... 04❑
- Insegnante:
  - di scuola media inferiore o superiore ..... 05❑
  - o di scuola elementare/materna ..... 06❑
- Tecnico o impiegato ad alta media qualificazione (*es. analisti di dati, geometri e periti tecnici, insegnanti, impiegati amministrativi, capi segreteria, capi commessi, infermieri professionali, ecc.*) ..... 07❑
- Impiegato esecutivo (*a es. segretari, addetti agli sportelli, telefonisti, ecc.*) ..... 08❑
- Graduato o militare di carriera (*delle FF.AA. o di Polizia, inclusi i sottufficiali*) ..... 09❑
- Capo operaio o operaio qualificato ..... 10❑
- Collaboratore ad un'attività autonoma famigliare ..... 11❑
- Lavoratore senza qualificazione nell'agricoltura e nell'industria (*es. manovali edili, braccianti ecc.*) ..... 12❑
- Lavoratore senza qualificazione in altre attività (*a es commessi, shampisti, camerieri, uscieri, custodi, bidelli, pulitori, benzinai, portantini, ecc.*) ..... 13❑
- Lavorante nel proprio domicilio per conto di imprese, un apprendista ..... 14❑
- *Altro* ..... 15❑

115. Sempre quando avevi 14 anni, tua madre era:
(Attenzione indicare il codice 4 se all'epoca deceduta)

- occupata ..... 1❑
- in cerca di occupazione ..... 2❑
- casalinga ..... 3❑ *(passare al quesito 120)*
- pensionata ..... 4❑
- *altra condizione* ..... 5❑

116. Descrivi quali erano i compiti principali che tua madre svolgeva nel suo lavoro senza usare termini generici come impiegato o operaio.
*(Attenzione è sconsigliato che indichi il luogo dove tua madre esercitava la sua attività, ad es. cassiere **in banca**, infermiere **a domicilio**, bagnante custode **in azienda**, ecc)*
Se non ricordi il lavoro che tua madre svolgeva quando aveva 14 anni, indica comunque quello attuale
Se deceduta, pensionata o disoccupata, fai riferimento all'ultima situazione lavorativa.
Se tua madre è era deceduta quando aveva 14 anni e precedentemente era casalinga passa al quesito 120)

_____

_____

- *casalinga* .. 1❑ *(passare al quesito 120)*

117. Tua madre è era una:

- Lavoratrice dipendente? ..... 1❑ *(passare al quesito 119)*
- Consulente/collaboratore? ... 2❑ *(passare al quesito 120)*
- Lavoratrice indipendente? ... 3❑ *(passare al quesito 118)*

118. Quale tra le seguenti voci descrive meglio il lavoro di tua madre?
(Attenzione è possibile una sola risposta)

- Commerciante ..... 01❑
- Artigiano ..... 02❑
- Coltivatrice diretta ..... 03❑
- Collaboratore ad un'attività autonoma famigliare ..... 04❑
- Lavoratore autonomo senza qualificazione (*ad es. baby sitter, traslocatore*) ..... 05❑
- Socio di una cooperativa ..... 06❑
- Imprenditore ..... 07❑
- Libero professionista ..... 08❑
- *Altro* ..... 09❑

Passare al quesito 120

119. Quale tra le seguenti voci descrive meglio la posizione lavorativa di tua madre?
(Attenzione è possibile una sola risposta)

- Dirigente *(inclusi i medici)* ..... 01❑
- Docente universitario *(I e II fascia)* ..... 02❑
- Quadro, funzionario *(inclusi direttivi e ufficiali FF.AA.)* ..... 03❑
- Ricercatore ..... 04❑
- Insegnante:
  - di scuola media inferiore o superiore ..... 05❑
  - o di scuola elementare/materna ..... 06❑
- Tecnico o impiegata ad alta/media qualificazione (*es. analisti di dati, geometri e periti tecnici, insegnanti, impiegati amministrativi, capi segreteria, capi commessi, infermiera professionale, ecc.*) ..... 07❑
- Impiegata esecutiva (*es. segretarie, addetti agli sportelli, telefoniste, ecc.*) ..... 08❑
- Militare di carriera (*delle FF.AA. o di Polizia, inclusi i sottufficiali*) ..... 09❑
- Capo operaio o operaio qualificato ..... 10❑
- Collaboratore ad un'attività autonoma famigliare ..... 11❑
- Lavoratrice senza qualificazione nell'agricoltura e nell'industria (*es. manovali edili, braccianti ecc.*) ..... 12❑
- Lavoratrice senza qualificazione in altre attività (*es. commessi, shampisti, camerieri, uscieri, custodi, bidelli, pulitori, benzinai, portantini, ecc.*) ..... 13❑
- Lavorante nel proprio domicilio per conto di imprese, un apprendista ..... 14❑
- *Altra* ..... 15❑

SEZIONE 6: NOTIZIE ANAGRAFICHE

120. Hai la cittadinanza italiana?
(Attenzione nel caso di doppia cittadinanza seleziona SI)

NO ..... 1❑
SI ..... 2❑ *(passare al quesito 122)*

121. Qual è la tua cittadinanza?

- *Paesi Unione Europea (Danimarca, Irlanda, Regno Unito, Paesi Bassi, Belgio, Lussemburgo, Germania, Grecia, Francia, Portogallo, Spagna, Austria, Finlandia, Svezia)* ..... 1❑
- *altri paesi europei* ..... 2❑

EXTRAEUROPEA
- *Africa* ..... 3❑
- *America* ..... 4❑
- *Asia* ..... 5❑
- *Oceania* ..... 6❑

122. In che provincia hai la residenza?

- *provincia* : : : :

123. E' la stessa in cui vivi abitualmente?
(Attenzione per "abitualmente" si intende 4 5 giorni a settimana)

NO ... 1❑
SI ... 2❑ *(passare al quesito 125)*

124. Qual è quella in cui vivi abitualmente?

- *provincia* : : : :

125.1. *Sesso*

M ..... 1❑
F ..... 2❑ *(passare al quesito 126)*

125.2. Qual è la tua posizione nei confronti degli obblighi militari?

- *li ha già assolti* ..... 1❑
- *li sta assolvendo* ..... 2❑
- *li deve ancora assolvere* ... 3❑
- *è stato esonerato* ..... 4❑

126. In che anno sei nato?

Anno 19: :

127. Qual è il tuo Stato civile?

- celibe/nubile ..... 1 ☐
- coniugato-a convivente ..... 2 ☐
- separato-a/divorziato-a ..... 3 ☐
- vedovo-a ..... 4 ☐

Ti ringraziamo per la preziosa collaborazione

Istituto Nazionale di Statistica
SCHEDA INFORMATIVA SUI MATURI DEL 1995

I dati raccolti nell'ambito della presente indagine sono tutelati dal segreto statistico. E' fatto obbligo alle amministrazioni, enti ed organismi pubblici, nonché ai soggetti privati, per le rilevazioni indicate nel D.P.R. 4 Dicembre 1996, di fornire tutti i dati e le notizie richieste nel modello di rilevazione.

SEZIONE I: DATI RIGUARDANTI IL MATURO N. | | | *(inserire un numero progressivo)*

La famiglia *(scrivere in stampatello)*

Cognome e nome del padre *(separare con uno spazio)*

Cognome e nome della madre *(separare con uno spazio)*

Residenza: Via/Piazza e n. civico

CAP Comune

Provincia | | *(sigla)* Telefono | | | | | | | | | |
prefisso numero telefonico

Dati anagrafici *(scrivere in stampatello)*

Cognome e nome *(separare con uno spazio)*

Sesso | | *(M = 1, F = 2)*

Data di nascita: mese | | *(da 01 a 12)* anno | | | | *(ultime due cifre)*

SEZIONE II: CURRICULUM DEL MATURO

Cambi di indirizzo scolastico (a)	Privatista all'esame di maturità		Passaggi da pubblica a privata e/o viceversa (b)	Ripetenze durante gli studi superiori	Voto di maturità	Voto di licenza media inferiore (c)	Anno scolastico di conseguimento della licenza media
N. |__| *(indicare 0 se nessuno)*	SI	1	N. |__| *(indicare 0 se nessuno)*	N. |__| *(indicare 0 se nessuna)*	|__|__| /60°	|__|	19|__|__|/|__|__|
	NO	2					

Note

(a) Indicare il numero degli eventuali passaggi da un tipo scuola ad un altro (es. passaggio da un liceo scientifico a un istituto tecnico commerciale; passaggio da un istituto professionale industriale ad un istituto professionale alberghiero).

(b) Considerare esclusivamente i passaggi da scuole legalmente riconosciute o parificate a scuole pubbliche, e viceversa. Non considerare quindi le iscrizioni di coloro che nel corso degli studi hanno superato esami come candidati esterni. Escludere anche quanti hanno superato gli esami di maturità da "privatisti".

(c) Indicare con 1=sufficiente, 2=buono, 3=distinto, 4=ottimo.

Bibliografia

[1] Aitkin, M., Anderson, D. & Hinde, J. (1981). Statistical modelling of data on teaching styles (con discussione). *Journal of the Royal Statistical Society*, A., 144, pp. 148-61.

[2] Aitkin, M. & Longford, N. (1986). Statistical modelling in school effectiveness studies (con discussione). *Journal of the Royal Statistical Society*, A, 149, pp. 1-43.

[3] Anderson, D.A. & Aitkin, M. (1985). Variance components models with binary response: interviewer variability. *Journal of the Royal Statistical Society*, B, 47, pp. 203-210.

[4] Agresti, A. (1984). *Analysis of ordinal categorical data*. Wiley, New York.

[5] Agresti, A. (1990). *Categorical data analysis*. Wiley, New York.

[6] Allison, P.D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, S. Leindhardt Editor, San Francisco, pp. 61-98.

[7] Allmendinger, J. (1989). Educational Systems and Labor Market Outcomes. *European Sociological Review*, 5, pp. 231-250.

[8] Battistoni, L. & Ruberto, A. (1988). *I percorsi giovanili di studio e lavoro. Indagine sulla entrata nella vita attiva*. Franco Angeli, Milano.

[9] Bennett, K.D. (1995). Contextual Factors Surrounding Transitions from Education to the Labour Market. In *Education and Employment*, OECD - Centre for Educational Research and Innovation, CERI, Parigi.

[10] Bernardi, L. & Trivellato, U. (1986). *Problemi e orientamenti nelle indagini sulla transizione dalla scuola alla vita attiva: una nota metodologica*. Rapporto 90/1, COSES, Venezia.

[11] Biggeri, L., Bini, M. & Grilli, L. (1999). *The transition from university to work: a multilevel approach to the analysis of the time to get the first job.* Working Papers n. 79, Dipartimento di Statistica, Firenze.

[12] Bini, M (1999). *Valutazione dell'efficacia dell'istruzione universitaria rispetto al mercato del lavoro.* MURST - Osservatorio per la Valutazione del Sistema Universitario, Roma.

[13] Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, 88, pp. 9-25.

[14] Brown, C.C. (1975). On the use of indicator variables for studying the time dependence of parameters in a response-time model. *Biometrics*, 31, pp. 863-872.

[15] Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical Linear Models.* Sage, Newbury Park.

[16] Burnstein, L., Kim, K.S. & Delandshere, G. (1989). Multilevel investigations of systematically varying slopes: issues, alternatives and consequences. In Bock, R.D. (a cura di), *Multilevel Analysis of educational data*, Academic Press, New York.

[17] Calzolari, G., Mealli, F. & Rampichini, C. (1999). *Indirect estimation of logit multilevel models.* Articolo presentato alla Seconda Conferenza Internazionale di Amsterdam sull'Analisi Multilivello. Amsterdam, 30-31 marzo 1999.

[18] Cariani, G. (1998). Nuovi orientamenti della statistica ufficiale sulle aree istruzione, formazione e lavoro. Contributo presentato al convegno *L'educazione come processo permanente istruzione-formazione-riqualificazione*, SIEDS, Milano, 30 maggio 1998.

[19] Coleman, J.S. (1984). The transition from School to Work. *Research in Social Stratification and Mobility*, 3, pp. 27-59.

[20] Conaway, M.R. (1990). A random effects model for binary data. *Biometrics*, 46, pp. 317-328.

[21] Cox, D.R. (1972). Regression models and life tables (con discussione). *Journal of the Royal Statistical Society,* B, 34, pp. 187-220.

[22] D'Agostino, A. (1998). *Specificazione e stima di un modello a stati ed episodi multipli per lo studio della transizione scuola-lavoro*. Tesi di dottorato in Statistica Applicata. Dipartimento di Statistica dell'Università di Firenze.

[23] Dex, S. (1989). Gender and the Labour Market. In *Employment in Britain* (ed. D. Gallie), Blackwell, Oxford.

[24] Diggle, P.J., Liang, K.-Y. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*.Clarendon Press, Oxford.

[25] Dolton, P. J., Makepeace, G. H. and Treble, J. G. (1993). *The youth training scheme and the school to work transition*. Research Paper 93/22. University of Wales, Bangor.

[26] Duncan, O. D. (1966). Path Analysis: Sociological Examples. *American Journal of Sociology*, 72, pp. 1-16.

[27] Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, 83, pp. 414-425.

[28] Fabbri, D., Fazioli, R. & Filippini, M. (1996). *L'intervento pubblico e l'efficienza possibile*. Il Mulino, Bologna.

[29] Fahrmeir L. & Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*. Springer-Verlag, New York.

[30] Fiorito, J. (1981). The School-to-Work Transition of College Graduates. *Industrial and Labor Relations Review*, 35, pp. 103-114.

[31] Fondazione Giacomo Brodolini (1992). *Giovani, interventi formativi e mercato del lavoro nel Mezzogiorno*, Fase 1, Rapporto di avanzamento, mimeo, Roma.

[32] Ghellini, G. (1992). Progetto LEVA. Successi e insuccessi scolastici dei giovani in Lombardia nel triennio post-obbligo. *Notiziario Statistico Regionale*, 87, Regione Lombardia.

[33] Gibbons, R.D. & Hedeker, D. (1997). Random Effects Probit and Logistic Regression Models for Three-Level Data. *Biometrics*, 53, pp. 1527-1535.

[34] Gibbons, R.D., Hedeker, D., Charles, S.C. & Frisch, P. (1994). A Random-Effects Probit Model for Predicting Medical Malpractice Claims. *Journal of the American Statistical Association*. 89. pp. 760-767.

[35] Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, Londra.

[36] Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, 73, pp. 43-56.

[37] Goldstein, H. (1987). Multilevel covariance component models. *Biometrika*, 74, pp. 430-31.

[38] Goldstein, H. (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika*, 76, pp. 622-23.

[39] Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika*, 78, pp. 45-51.

[40] Goldstein, H. (1995). *Multilevel Statistical Models*. Edward Arnold, Londra.

[41] Goldstein, H., & Healy, M.J.R. (1994). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society*, A, 158, pp. 175-7.

[42] Goldstein, H. & Rasbash, J. (1992). Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalised least squares. *Computational Statistics and Data Analysis*, 13, pp. 63-71.

[43] Goldstein, H. & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, A. 159, pp. 505-13.

[44] Goldstein, H. & Spiegelhalter, D.J. (1996). Leage tables and their limitations: statistical issues in comparisons of institutional performances (con discussione). *Journal of the Royal Statistical Society*, A. 159, pp. 385-443.

[45] Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. and Healy, M.J.R. (1998). *A User's Guide to MLwin*. Institute of Education, Londra.

[46] Gori, E. (1992). La valutazione dell'efficienza e efficacia dell'istruzione. *Atti della XXXVI Riunione Scientifica della Società Italiana di Statistica*, CISU, Pescara, pp. 219-230.

[47] Green, P.J. (1987). Penalized Likelihood for General Semi-Parametric Regression Models. *International Statistical Review*, 55, pp. 245-259.

[48] Hartog, J., Pfann, G. & Ridder, G. (1989). (Non-) Graduation and the Earnings Function. *European Economic Review*, 33(7), pp. 1373-1395.

[49] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, pp. 320-340.

[50] Hedeker, D. (1998). *MIXNO: a computer program for mixed-effects nominal logistic regression*. University of Illinois, Chicago.

[51] Hedeker, D. & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50, pp. 933-944.

[52] Hedeker, D. & Gibbons, R. D. (1996). *MIXOR: a computer program for mixed-effects ordinal regression analysis*. University of Illinois, Chicago.

[53] Hedeker, D., Siddiqui, O. & Hu, F.B. (1999). Random-Effects Regression Analysis of Correlated Gruoped-Time Survival Data. *Articolo in corso di pubblicazione.*

[54] Holford, T.R. (1976). Life table with concomitant information. *Biometrics*, 32, pp. 587-597.

[55] Isfol (1989). *Percorsi giovanili di studio e lavoro*. Franco Angeli, Milano.

[56] Isfol (1998). *Formazione e occupazione in Italia e in Europa. Rapporto 1998.* Franco Angeli, Milano.

[57] Istat (1996a). *Indagine sulle Forze di Lavoro. Media 1995.* Istat, Roma.

[58] Istat (1996b). *Inserimento professionale dei laureati. Indagine 1995.* Informazioni n. 10, Istat, Roma.

[59] Istat (1997). *Statistiche sulle scuole secondarie superiori. Anno scolastico 1995-96*. Annuari. Istat, Roma.

[60] Istat (1998). *Conti economici delle imprese. Anno 1995.* Informazioni n. 102, Istat, Roma.

[61] Istat (1999a). *Indagine sulle Forze di Lavoro. Media 1998.* Istat, Roma.

[62] Istat (1999b). *Percorsi di studio e di lavoro dei diplomati. Indagine 1998.* In corso di pubblicazione.

[63] Kalbfleish, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

[64] Kreft, I.G., de Leeuw, J. (1998). *Introducing Multilevel Modelling.* Sage, Londra.

[65] Läärä, E. & Matthews, J.N.S. (1985). The equivalence of two models for ordinal data. *Biometrika*, 72, pp. 206-7.

[66] Langford, I. & Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society*, A., 161, pp. 121-160.

[67] Layder, D., Ashton, D. and Sung, J. (1991) The Empirical Correlates of Action and Structure: the Transition from School to Work. *Sociology*, 25, pp. 447-464.

[68] Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models (con discussione). *Journal of the Royal Statistical Society,* B, 57, pp. 619-678.

[69] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika*, 73, pp. 45-51.

[70] Lin, X. (1997). Variance Component Testing in Generalised Linear Models with Random Effects. *Biometrika*, 84, pp. 309-326.

[71] Little, R. J. (1998). Missing Data. In *Encyclopedia of Biostatistics*, pp. 2622-35, Wiley.

[72] Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74, pp. 817-27.

[73] Longford, N.T. (1993). *Random Coefficient Models.* Clarendon Press, Oxford.

[74] Longford, N.T. (1996). Discussione dell'articolo di Lee & Nelder (1986). *Journal of the Royal Statistical Society,* B, 57, pp. 619-678.

[75] Lynch, L.M. (1987). Individual Differences in the Youth Labour Market: A Cross-section Analysis of London Youths. In *From School to Unemployment? The Labour Market for Young People* (P. N. Junakar ed.). MacMillan Press, Londra, pp. 185-211.

[76] Lynch, L.M. (1985). State Dependence in Youth Unemployment: A Lost Generation?. *Journal of Econometrics*, 28, pp.71-84.

[77] Mansky, C.F. & Garfinkel, I. (1992) (ed.) *Evaluating Welfare and Training Programs*. Harvard University Press, Cambridge (MA).

[78] Mantel, N. & Hankey, B. (1978). A logistic regression analysis of response-time data where the hazard function is time dependent. *Communications in Statistics - Theory and Methods*, A7, pp. 333-347.

[79] Mare, R.D. (1980). Social background and school continuation decisions. *Journal of the American Statistical Association*, 75, pp. 295-305.

[80] McCullagh, P. (1980). Regression models for ordinal data (con discussione). *Journal of the Royal Statistical Society*, B, 42, pp. 109-142.

[81] McCullagh, P. & Nelder, J.A. (1989). *Generalised Linear Models* (2nd edition). Chapman and Hall, Londra.

[82] McDonald, R.P. (1994). Two random effects models for multivariate binary data. *Biometrics*, 50, pp. 164-72.

[83] Micali, A. (1993a). Le fonti statistiche sulla transizione famiglia-formazione-lavoro. Parte I: le tecniche di indagine. *Economia & Lavoro*, anno XXVII, n. 2, pp. 31-41.

[84] Micali, A. (1993b). Le fonti statistiche sulla transizione famiglia-formazione-lavoro. Parte II: le fonti nazionali. *Economia & Lavoro*, anno XXVII, n. 3, pp. 63-82.

[85] Micali, A. & Ungaro, P. (1998). Il sistema di indagini Istat sulla transizione scuola-lavoro. Contributo presentato al convegno *L'educazione come processo permanente istruzione-formazione-riqualificazione*, SIEDS, Milano, 30 maggio 1998.

[86] Micklewright, J. (1988). *Schooling Choice, Educational Maintainance Allowances and Panel Attrition*. QMC, Dep. of Economics, Working Paper n. 185, Londra.

[87] Micklewright, J. (1989). Choice at sixteen. *Economica*, 56, pp. 25-39.

[88] Montagni, M. (1997). *I modelli multilivello. Un'applicazione all'analisi di efficacia dei corsi di laurea*. Tesi di dottorato in Statistica Applicata. Dipartimento di Statistica dell'Università di Firenze.

[89] Muthén, B.O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22, pp. 376-398.

[90] Myers, M., Hankey, B.F. & Mantel, N. (1973). A logistic-exponential model for use with response-time data involving regressor variables. *Biometrics*, 29, pp. 257-269.

[91] Neuhaus, J.M., Kalbfleisch, J.D. & Hauck, W.W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, 59(1), pp. 25-35.

[92] OECD (1997). *Education at a Glance: OECD Indicators*. CERI, Parigi.

[93] O'Higgins, N. (1992). School-leaving at sixteen: an empirical analysis using individual data. *Economia & Lavoro*, anno XXVI, n. 1, pp. 3-22.

[94] Ordine, P. (1992). *Labour Market Transitions of Youth and Prime Age Italian Unemployed*. DYNAMIS - Quaderno 3/92, IDSE, Milano.

[95] Patterson, H.D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, pp. 545-554.

[96] Pedenzini, C. & Zaccarin, S. (1992). Dalla scuola verso dove. Indagine sui percorsi formativi e professionali dei diplomati nel quadro delle iniziative della Regione Veneto e del COSES sulla transizione scuola/lavoro nel Veneto. *COSES Informazione*, numero unico (marzo), Venezia.

[97] Pfeffermann, D., Skinner, C. J., Holmes, D., Goldstein, H. & Rasbash, J. (1997). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, B, 60, pp. 23-40.

[98] Prentice, R.L. and Gloeckler, L.A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics,* 34, pp. 57-67.

[99] Rampichini, C. (1994). SISCPVAR - Un programma per la stima di modelli a componenti di varianza, in *Software sperimentale per la statistica*. SIS ed. Centro Duplicazione Offset, Firenze.

[100] Rampichini, C. & Mealli, F. (1999). Estimating binary multilevel models through indirect inference. *Computational Statistics & Data Analysis*, 29, pp. 313-324.

[101] Rasbash, J., Yang, M., Woodhouse, G., & Goldstein, H. (1995). *Mln command reference*. Institute of Education, Londra.

[102] Raudenbush, S.W. & Willms, J.D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20.*

[103] Rees, A. & Gray, W. (1982). Family Effects in the Youth Unemployment. In R.B. Freeman & D.A. Wise (eds), *The Youth Labour Market Problem: its Nature, Casues and Consequences*, University of Chicago, Chicago.

[104] Rinaldelli, C. (1997). *Indagine sugli sbocchi professionali dei maturi del 1995. Principali aspetti metodologici del disegno di indagine e di campionamento*. Manoscritto non pubblicato, Istat, Roma.

[105] Robinson, W.S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review*, 15, pp. 351-57.

[106] Rodriguez, G. & Goldman, L. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society*, A, 159, pp. 73-89.

[107] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

[108] Santoro, M. & Pisati, M. (1996). *Dopo la laurea. Status, sfide e strategie*. Il Mulino, Bologna.

[109] Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*. Wiley, New York.

[110] Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys* (ed. C.J. Skinner, D. Holt e T.M.F. Smith). Wiley, Chichester, pp. 59-87.

[111] Snjiders, T. & Bosker, R. (1999). *An Introduction to Basic and Advanced Multilevel Modeling*. Sage, London.

[112] Torelli, N. & Trivellato, U. (1988). Modelling Job-Search Duration from the Italian Labuor Force Data. *Labour*, 2, pp. 91-111.

[113] Trivellato, U. & Bernardi, L. (1994). Scolarità e formazione professionale nel Mezzogiorno: nuove evidenze da un'analisi dei flussi. *Economia & Lavoro*, anno XXVIII, n. 3-4, pp. 1-34.

[114] Tronti, L. & Mariani, P. (1994). La transizione Università-Lavoro in Italia. Un'esplorazione delle evidenze dell'indagine Istat sugli sbocchi professionali dei laureati. *Economia & Lavoro*, anno XXV, n. 3, pp. 3-26.

[115] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, 61, pp. 439-47.

[116] Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, 31, pp. 144-8.

[117] Willis, R.J. & Rosen, S. (1979). Education and Self-selection. *Journal of Political Economy*, 87, n. 5, pp. s7-s36.

[118] Willms, J.D. (1992) *Monitoring School Performance: A Guide for Educators*. Falemer, Londra.

[119] Winship, C. & Mare, R.D. (1983). Structural equations and path analysis for discrete data. *American Journal of Sociology,* 89, pp. 54-110.

[120] Yang, M. (1997). Multilevel models for multiple category responses - a simulation. *Multilevel Modelling Newsletter*, vol. 9, n.1, pp. 10-16.

[121] Zaccarin, S. (1994). Indagini longitudinali sulla transizione scuola-lavoro: alcune riflessioni di metodo. *Economia & Lavoro*, anno XXVIII, n. 2, pp. 27-46.

[122] Zeger, S.L., Liang, K-Y., & Albert, P.S. (1988). Models for longitudinal data: a generalised estimating equation approach. *Biometrics*, 44, pp. 1049-60.

[123] Zeger, S.L. & Karim, M.R. (1991). Generalised linear models with random effects: a Gibbs Sampling approach. *Journal of the American Statistical Society*, 86, pp. 79-102.